# Methoden en Technieken van Onderzoek in de Psychologie

**Eric Maris**

16 april 2012

# Inhoudsopgave


**1 Dataverzamelingsmethoden** **15**
1.1 Variabelen . . . 17
1.1.1 Wat zijn Variabelen? . . . 17
1.1.2 Persoons- en Situatievariabelen . . . 18
1.2 Effectonderzoek . . . 20
1.2.1 Experimentele en Natuurlijke Variabelen . . . 20
1.2.2 Het Gevaar bij het Gebruik van Natuurlijke Variabelen: Storende Persoonsvariabelen . . . 22
1.2.3 Het Gevaar bij een Gebrekkige Experimentele Controle van de Condities: Storende Situatievariabelen . . . 24
1.2.4 Het Binnen-Subjectenexperiment . . . 27
1.2.5 Wanneer is een Onderzoek een Experiment? . . . 30
1.3 Standaardisering . . . 33
1.4 Observatiemethoden . . . 33
1.4.1 Vrije (niet-gestandaardiseerde) Observatie . . . 34
1.4.2 Participerende Observatie . . . 35
1.4.3 Gestandaardiseerde Observatie . . . 36
1.4.4 Registratie . . . 40
1.4.5 Enkele Problemen bij het Gebruik van Observatiemethoden . . . 43
1.5 Interview, Vragenlijst en Test . . . 50
1.5.1 Het Vrije (niet-gestandaardiseerde) Interview . . . 50
1.5.2 Het Gestandaardiseerde Interview en de Vragenlijst . . . 50
1.5.3 De Psychologische Test . . . 57

**2 Meten** **61**
2.1 De Plaats van het Meten binnen de Psychologie . . . 62
2.1.1 Voorbeelden . . . 63
2.1.2 Psychologische Constructen . . . 65
2.1.3 Operationele Definities . . . 66
2.1.4 Het Meten van Persoonskenmerken . . . 68
2.2 Betrouwbaarheid . . . 70
2.2.1 Betrouwbaarheid=Herhaalbaarheid . . . 70
2.2.2 De Basisformule van de Klassieke Testtheorie . . . 74
2.2.3 De Definitie van Betrouwbaarheid . . . 79

2.2.4 Parallelle Tests . . . 80
2.2.5 Testverlenging en Betrouwbaarheid . . . 82
2.2.6 Praktische Methoden voor het Bepalen van de Betrouwbaarheid Gebaseerd op Parallelliteit . . . 84
2.2.7 Cronbachs Alpha . . . 86
2.2.8 Itemanalyse . . . 93
2.2.9 Het Restriction-of-Range Fenomeen . . . 95
2.2.10 De Standaardfout van Meting . . . 98
2.2.11 Correctie voor Attenuatie . . . 98
2.2.12 Berekenen van de Betrouwbaarheid met de SPSS-procedure Reliability . . . 100
2.3 Validiteit . . . 104
2.3.1 Inhoudsvaliditeit . . . 104
2.3.2 Constructvaliditeit . . . 105
2.3.3 Criteriumvaliditeit . . . 109
2.3.4 Criteriumvaliditeit en Beslissingen . . . 113
2.4 Appendix bij Hoofdstuk 2 . . . 122

**3 Variantie- en Covariantieanalyse 129**
3.1 Inleiding . . . 130
3.1.1 Voorbeeld: Mobiliteit en Sociale Activiteit . . . 130
3.1.2 Experimenteel en Niet-Experimenteel Onderzoek . . . 131
3.1.3 De Toewijzing aan de Condities . . . 132
3.1.4 Experimentele Controle van de Condities . . . 135
3.2 Variantieanalyse met Eén Factor . . . 137
3.2.1 Voorbeeld: Mobiliteit en Sociale Activiteit . . . 137
3.2.2 De Notatie voor de Gegevens . . . 139
3.2.3 Het Rekenwerk . . . 140
3.2.4 De F-Toets . . . 143
3.2.5 Het Variantieanalysemodel . . . 145
3.2.6 De Effectkwadratensom . . . 148
3.2.7 De Errorkwadratensom, Totale Kwadratensom, en de Kwadratensomontbinding . . . 152
3.2.8 Berekening en Interpretatie van de F-Statistiek . . . 154
3.2.9 De Proportie Verklaarde Variantie . . . 156
3.3 Variantieanalyse met Twee Factoren . . . 160
3.3.1 Voorbeeld 1: Mobiliteit en Sociale Activiteit . . . 160
3.3.2 Wat zijn Ongebalanceerde Designs en Wanneer Krijgt men Ermee te Maken? . . . 161
3.3.3 Voorbeeld 2: Pesten in de Klas . . . 164
3.3.4 De Notatie voor de Gegevens . . . 169
3.3.5 De Statistische Toetsing . . . 174
3.3.6 Het Rekenwerk in een Gebalanceerd Design . . . 178
3.3.7 Het Rekenwerk in een Ongebalanceerd Design . . . 184
3.3.8 F-statistieken en Kwadratensommen . . . 190
3.3.9 Interactie tussen Factoren . . . 208

3.3.10 Het Rekenwerk in een Gebalanceerd Design (Met Interactie) . . . . 211
3.3.11 Het Rekenwerk in een Ongebalanceerd Design (Met Interactie) . . 217
3.3.12 F-statistieken en Kwadratensommen (Met Interactie) . . . . . 221
3.3.13 Het Nut van het Toevoegen van Factoren . . . . . 237
3.3.14 Een Relativering van de Kracht van Statistische Uitzuivering . . . 238
3.3.15 Designs met Lege Cellen . . . . . 239
3.3.16 De Variantieanalyse met Meer dan Twee Factoren . . . . . 240
3.3.17 Uitvoeren van een Variantieanalyse met de SPSS-procedure GLM-Univariate . . . . . 240
3.4 Uitzuivering voor een Kwantitatieve Variabele met Covariantieanalyse . . 252
3.4.1 Voorbeeld: Pesten in de Klas . . . . . 252
3.4.2 Uitgezuiverde Verschillen en Gemiddelden . . . . . 258
3.4.3 De Nulhypothese voor het Uitgezuiverde Populatieverschil . . . . . 260
3.4.4 De F-statistieken . . . . . 262
3.4.5 De Effectkwadratensommen . . . . . 262
3.4.6 De Errorkwadratensom en de Kwadratensom-ontbinding . . . . . 267
3.4.7 Nogmaals De F-Statistieken . . . . . 267
3.4.8 Overzichtstabellen . . . . . 269
3.4.9 Uitvoeren van een Covariantieanalyse met de SPSS-procedure GLM-Univariate . . . . . 271
3.4.10 Twee Redenen om Uit te Zuiveren voor een Covariaat . . . . . 272
3.4.11 Wanneer Stel je de Covariaatscores Vast? . . . . . 274
3.4.12 Een Relativering van de Kracht van Statistische Uitzuivering . . . 275
3.5 Designs met een Voor- en een Nameting . . . . . 285
3.5.1 Het Design met Eén Groep . . . . . 285
3.5.2 Het Design met Twee Groepen . . . . . 287
3.6 Appendix bij Hoofdstuk 3 . . . . . 290

**4 Regressieanalyse 295**
4.1 Inleiding . . . . . 297
4.1.1 Multivariate Analyse . . . . . 297
4.1.2 Regressieanalyse . . . . . 298
4.1.3 Voorbeeld . . . . . 301
4.2 Voorspellen Zonder Predictoren . . . . . 302
4.2.1 Kleinste-kwadratenschatting . . . . . 302
4.2.2 De Gemiddelde Kwadratische Fout . . . . . 303
4.2.3 Voorbeeld . . . . . 305
4.3 Voorspellen met één Predictor: Enkelvoudige Regressieanalyse . . . . . 306
4.3.1 Enkelvoudige Regressieanalyse voor Gestandaardiseerde Variabelen 306
4.3.2 Enkelvoudige Regressieanalyse voor Ongestandaardiseerde Variabelen . . . . . 308
4.3.3 Verklaarde en Onverklaarde Variantie . . . . . 309
4.3.4 De Proporties Verklaarde en Onverklaarde Variantie . . . . . 310
4.3.5 De Standaardschattingsfout van de Regressie . . . . . 311
4.3.6 Een Zuivere Schatter van de Gemiddelde Kwadratische Fout in de Populatie . . . . . 311

4.3.7 Voorbeeld . . . 312
4.4 Voorspellen met Twee Predictoren: Meervoudige Regressieanalyse . . . 318
4.4.1 Meervoudige Regressieanalyse voor Gestandaardiseerde Variabelen 319
4.4.2 Meervoudige Regressieanalyse voor Ongestandaardiseerde Variabelen . . . 321
4.4.3 De Meervoudige Correlatiecoëfficiënt . . . 322
4.4.4 Voorbeeld . . . 325
4.5 Voorspellen met Willekeurig Veel Predictoren . . . 328
4.5.1 Theorie . . . 328
4.5.2 Voorbeeld . . . 331
4.6 De Semi-Partiële of Partcorrelatiecoëfficiënt . . . 333
4.6.1 Theorie . . . 333
4.6.2 Voorbeeld . . . 336
4.7 Enige Problemen bij de Interpretatie van Regressiegewichten . . . 341
4.7.1 Onechte Correlatie . . . 341
4.7.2 Suppressie . . . 344
4.7.3 Collineariteit . . . 348
4.8 De Statistische Significantie van Regressiegewichten . . . 351
4.9 Het Selecteren van Predictoren . . . 362
4.10 Uitvoeren van een Regressieanalyse met de SPSS-procedure Regression . . 365
4.11 Hiërarchische Meervoudige Regressieanalyse . . . 371
4.11.1 HMRA binnen het Onderzoek naar de Rol van Storende Variabelen372
4.11.2 HMRA binnen het Onderzoek naar de Rol van Mediërende Variabelen . . . 376
4.12 Regressieanalyse met Categorische Predictoren . . . 378
4.12.1 Voorbeeld . . . 378
4.12.2 De Regressievergelijking . . . 379
4.12.3 Statistische Significantie . . . 381
4.12.4 Het Algemene Lineaire Model . . . 387
4.12.5 Groepsspecifieke Regressiegewichten . . . 389
4.12.6 Een Overzicht van de Verschillende Types F-toetsen binnen het Algemene Lineaire Model . . . 393
4.12.7 Dummycodering . . . 398
4.12.8 Uitvoeren van een Regressieanalyse met de SPSS-procedure GLM-Univariate . . . 403
4.13 Regressieanalyse met Interactie Tussen Twee Kwantitatieve Predictoren . 411
4.13.1 Voorbeeld . . . 411
4.13.2 De Regressievergelijking met een Interactievariabele . . . 415
4.13.3 Inhoudelijke Interpretatie van de Binnen-score Regressielijnen . . . 419
4.13.4 De Regressievergelijking met Niet-gecentreerde Predictoren . . . 420
4.13.5 Het Belang van de Regressieanalyse met Interactie Tussen Twee Kwantitatieve Predictoren . . . 421
4.14 Logistische Regressieanalyse . . . 423
4.14.1 Kansen in plaats van Scores . . . 424
4.14.2 Het Berekenen van de Logistische Regressiecoëfficiënten . . . 427

4.14.3 Eigenschappen van de Logistische Regressiecoëfficiënten . . . . . 429
4.14.4 De Kwaliteit van de Voorspelling op basis van een Logistische Regressieanalyse . . . . . 429
4.14.5 Het Evalueren van de Effectgrootte . . . . . 433
4.14.6 Statistische Toetsen . . . . . 438
4.14.7 Uitvoeren van een Logistische Regressieanalyse met de SPSS-procedure Logistic Regression . . . . . 440
4.15 Appendix bij Hoofdstuk 4 . . . . . 452

**5 Factoranalyse 461**
5.1 Het Model met Eén Gemeenschappelijke Factor . . . . . 463
5.1.1 Het Model en zijn Implicaties voor de Correlaties . . . . . 463
5.1.2 Schattingen van de Factorladingen . . . . . 467
5.2 Het Model met Twee Gemeenschappelijke Factoren . . . . . 470
5.2.1 Het Model en zijn Implicaties voor de Correlaties . . . . . 470
5.2.2 Schattingen van de Factorladingen . . . . . 472
5.3 Ladingen, Communaliteit en Eigenwaarde . . . . . 474
5.3.1 Ladingen als Correlaties . . . . . 474
5.3.2 De Communaliteit van een Variabele . . . . . 475
5.3.3 De Eigenwaarde van een Factor . . . . . 476
5.3.4 Grafische Weergave van de Factoroplossing . . . . . 477
5.4 Hoofdcomponentenanalyse . . . . . 484
5.4.1 Het Model en zijn Implicaties voor de Correlaties . . . . . 485
5.4.2 Schattingen van de Factorladingen . . . . . 486
5.4.3 Een Vergelijking met het Factormodel met Ongecorreleerde Residuen 488
5.5 De Onbepaaldheid van een Factoroplossing . . . . . 491
5.5.1 Illustratie en Omschrijving . . . . . 491
5.5.2 Het Belang van de Onbepaaldheid . . . . . 494
5.6 Roteren naar Interpreteerbare Factoren . . . . . 496
5.6.1 Enkele Richtlijnen bij de Interpretatie . . . . . 496
5.6.2 Oblieke Rotatie . . . . . 498
5.7 Het Model met Meer dan Twee Gemeenschappelijke Factoren . . . . . 506
5.7.1 Veralgemening van de Formules . . . . . 506
5.7.2 Grafische Weergave van de Factoroplossing . . . . . 507
5.7.3 Roteren met Meer dan Twee Factoren . . . . . 507
5.8 De Keuze van het Aantal Gemeenschappelijke Factoren . . . . . 508
5.8.1 Inspectie van de Residuele Correlaties. . . . . 509
5.8.2 Inspectie van de Eigenwaarden . . . . . 510
5.8.3 De Interpreteerbaarheid van de Gevonden factoroplossing. . . . . 510
5.8.4 De Uitslag van een Statistische Toets . . . . . 511
5.9 Het Bepalen van de Factorscores . . . . . 512
5.10 Uitvoeren van een Factoranalyse met de SPSS-procedure Factor Analysis 513
5.11 Appendix bij Hoofdstuk 5 . . . . . 522
Literatuur . . . . . 524

**Index 526**

# Voorwoord

## Over het Boek

Dit is een boek over methoden en technieken van onderzoek in de psychologie. Het is geschikt als cursusboek voor het academisch en hoger beroepsonderwijs (in België, het niet-universitair hoger onderwijs). Het is ook geschikt voor zelfstudie en als naslagwerk. Het boek is in eerste instantie bedoeld voor psychologen, maar het behandelt onderwerpen die ook van belang zijn voor andere gedrags- en sociale wetenschappers (pedagogen, onderwijskundigen, sociologen, communicatiewetenschappers, bedrijfskundigen, beleidswetenschappers, ...).

Dit boek bouwt verder op begrippen uit de beschrijvende en de inferentiële (toetsende) statistiek. Wat betreft de beschrijvende statistiek, zijn vooral de begrippen gemiddelde, variantie, covariantie, correlatie, en regressie van belang. Wat betreft de inferentiële statistiek, komen onder andere de begrippen nulhypothese, toetsstatistiek en p-waarde aan de orde. Deze begrippen worden telkens kort uitgelegd, maar voor een lezer die niet reeds een inleiding heeft gehad in de beschrijvende en de inferentiële statistiek, is deze uitleg mogelijk onvoldoende.

De technieken die in dit boek behandeld worden (onderzoeken van de betrouwbaarheid en de validiteit van een test, variantie- en covariantieanalyse, regressieanalyse, factoranalyse) volstaan voor het merendeel van het niet-experimenteel onderzoek in de psychologie. Bij de behandeling van de variantieanalyse heb ik me beperkt tot tussen-subjecten designs. De variantieanalyse voor binnen-subjecten designs of herhaalde metingen (*mixed model* en MANOVA), die veel gebruikt wordt binnen experimenteel onderzoek (o.a., in de functieleer, de biologische psychologie en de experimentele sociale psychologie) is de enige courante techniek die niet in dit boek behandeld wordt.

Bij het schrijven van dit boek heb ik geprobeerd om zoveel mogelijk uit te leggen wat de bedoeling is van bepaalde berekeningen. Hierbij moet men bijvoorbeeld denken aan de verschillende toetsstatistieken waarmee statistische hypothesen getoetst worden. De keuze voor het uitleggen van het waarom van bepaalde berekeningen heeft ervoor gezorgd dat er ook aandacht besteed wordt aan de precieze structuur van een aantal formules. Van de lezer wordt dus verwacht dat hij of zij ook de formules bestudeert. De ervaring heeft geleerd dat dit geen problemen oplevert voor tweedejaars psychologiestudenten. Sommige studenten moeten wel leren hun normale leestempo te vertragen om de symbolen in de formules te kunnen vertalen naar woorden. Voor een aantal belangrijke formules wordt een afleiding gegeven in een appendix. De tekst kan echter

ook gelezen worden zonder de afleidingen te bestuderen. De afleidingen hebben een didactische functie: ze laten zien waar bepaalde formules vandaan komen en dat is vaak verhelderend.

Een belangrijk onderdeel van het boek zijn de opgaven. Met deze opgaven kan de lezer toetsen welke delen van de stof hij of zij beheerst en welke delen nog niet. Het merendeel van de opgaven zijn gestileerde onderzoeksvragen; het gestileerde zit in het feit dat één of enkele methodologische punten centraal staan en dat complicaties omwille van andere redenen uit de weg gegaan worden. Een groot deel van de opgaven (ongeveer 80 procent) zijn bewerkte oude tentamenvragen. De ervaring heeft geleerd dat het maken van de opgaven essentieel is voor een goed begrip van de stof.

Het boek bevat tien computeropgaven. Deze opgaven moet men maken met SPSS. De databestanden die bij deze opgaven horen zijn verkrijgbaar via de webpagina van uitgeverij ACCO (www.uitgeverij.acco.be/didactisch materiaal). Alle bestanden zijn SPSS-bestanden en hebben de extensie ".sav". Als SPSS op je computer geïnstalleerd is, dan kun je met dubbelklikken op de bestandsnaam tegelijkertijd SPSS starten en het bestand in de data-editor laden. Eenmaal dit gebeurd is, zijn de aanwijzingen in het boek voldoende om de analyse uit te voeren en de relevante informatie in de uitdraai te selecteren.

Eén van de opgaven bij de paragraaf over logistische regressieanalyse is een zelfstudieopdracht. De functie van deze zelfstudieopdracht is niet het toetsen van je kennis, maar het toepassen van deze kennis in een bijna echte onderzoekssituatie. De kern van deze zelfstudieopdracht is het heranalyseren van de data uit het boek "The bell curve" van Herrnstein en Murray (1994).

De vijf hoofdstukken van het boek zijn in belangrijke mate onafhankelijk van elkaar te lezen. Er zijn slechts drie duidelijke gevallen van afhankelijkheid:

1. Het centrale thema in hoofdstuk 3 (*Variantie- en Covariantieanalyse*) is uitzuivering voor storende variabelen en de technieken waarmee men dit kan doen. Dit thema wordt voorbereid in paragraaf 2 van hoofdstuk 1 (*Dataverzamelingsmethoden*).

2. In de deelparagrafen 12.4 en 12.6 van hoofdstuk 4 (*Regressieanalyse*) wordt expliciet verwezen naar begrippen en formules die in hoofdstuk 3 (*Variantie- en Covariantieanalyse*) zijn behandeld.

3. Op verschillende plaatsen in hoofdstuk 5 (*Factoranalyse*) wordt de factoranalyse voorgesteld als regressieanalyse met latente predictoren. Dit veronderstelt dat de lezer eerst hoofdstuk 4 (*Regressieanalyse*) heeft bestudeerd.

## Aan de Student

Het leren van de inhoud van dit boek vereist een inspanning van de student. Geen enkele didactische aanpak of docent kan ervoor zorgen dat men methoden en technieken van onderzoek zonder inspanning kan leren. Sommige begrippen kunnen moeilijk zijn bij een eerste lezing. In dat geval, herlees de tekst, indien nodig, meerdere malen; denk erover

na; formuleer vragen voor je medestudenten en je docent. Stel het niet uit, want jouw begrip van vandaag is noodzakelijk voor je begrip van morgen.

In het cursusboek staan twee soorten formules: formules met een kader en formules zonder kader. De formules met kader zijn belangrijker dan de formules zonder kader. De formules met kader zijn, ofwel belangrijke definities, ofwel belangrijke resultaten. De formules zonder kader kunnen tussenresultaten zijn, bijzondere gevallen van andere formules, minder belangrijke definities en resultaten, of toepassingen op een getallenvoorbeeld.

In de appendices bij elk van de hoofdstukken staan de afleidingen van een aantal belangrijke formules. Ook al kan de tekst gelezen worden zonder de afleidingen te bestuderen, het is aan te raden deze afleidingen toch te bestuderen. De afleidingen hebben immers een didactische functie: ze laten zien waar bepaalde formules vandaan komen en dat is vaak verhelderend.

Het maken van opgaven is de beste manier om je begrip te evalueren en te verdiepen. Van sommige opgaven weet je het antwoord onmiddellijk en voor anderen moet je even gaan zitten. De belangrijkste opgaven zijn die waarbij je het geleerde moet toepassen in nieuwe situaties en die waarbij je een redenering moet opzetten. Wees niet ontmoedigd als je niet alle opgaven kunt maken, want sommige opgaven zijn vrij moeilijk. De overgrote meerderheid van de opgaven zou je echter moeten kunnen maken, al zullen sommige een behoorlijke inspanning kosten. Als je een bepaalde opgave erg moeilijk vindt, doe dan het volgende:

1. Lees en herlees de opgave totdat je precies begrijpt welke informatie in de opgave gegeven wordt en wat van jou verwacht wordt (iets beargumenteren, iets uitrekenen, ...).

2. Soms is het verhelderend als je een schema tekent met de relaties tussen de variabelen waar het in de opgave om draait.

3. Maak gebruik van alles wat je al weet. Vertrek van de uitgewerkte voorbeelden in de tekst; zoek naar relaties tussen deze opgave en de opgaven die je reeds gemaakt hebt; in het algemeen, probeer patronen te herkennen.

4. Soms kan een opgave vereenvoudigd worden als je een extra veronderstelling maakt. Als je een vereenvoudigde versie van de opgave kunt maken, dan zal dit jou helpen om ook de originele opgave te maken.

5. Als je een antwoord gevonden hebt, vraag je dan af of het antwoord ook redelijk is. Als het dat niet is, probeer dan de stap in de redenering te vinden waar je een fout gemaakt hebt.

## Aan de Docent

Dit boek is bij uitstek geschikt voor gebruik binnen het onderwijs. Voor dit doel heb ik in Tabel 1 het boek opgedeeld in 19 vrij homogene blokken. Voor het grondig bestuderen van één blok en het maken van een vijftal opgaven uit dit blok, heeft een student tussen

| Blok | Paragrafen |
|---|---|
| 1 | 1.1, 1.2 |
| 2 | 1.3 t/m 1.5 |
| 3 | 2.1, 2.2.1 t/m 2.2.4 |
| 4 | 2.2.5 t/m 2.2.12 |
| 5 | 2.3 |
| 6 | 3.1, 3.2 |
| 7 | 3.3.1 t/m 3.3.8 |
| 8 | 3.3.9 t/m 3.3.17 |
| 9 | 3.4, 3.5 |
| 10 | 4.1 t/m 4.3 |
| 11 | 4.4 t/m 4.6 |
| 12 | 4.7, 4.8 |
| 13 | 4.9 t/m 4.11 |
| 14 | 4.12 |
| 15 | 4.13 |
| 16 | 4.14 |
| 17 | 5.1 t/m 5.3 |
| 18 | 5.4 t/m 5.6 |
| 19 | 5.7 t/m 5.10 |

Tabel 1: Verdeling van de inhoud van het boek over 19 blokken.

de drie en de acht uur nodig. Voor een onderwijsvorm met wekelijks terugkerende studie-activiteiten, raad ik aan om elke week één blok te behandelen.

Wat betreft de onderwijsvorm waarin het boek gebruikt kan worden, heb ik goede ervaringen met een combinatie van zelfstudie, werkgroepen. Ik behandel één blok per week. In de eerste helft van de week is er gelegenheid tot zelfstudie en het maken van vijf á zes opgaven uit het blok van die week. In de tweede helft van de week zijn er werkgroepen. Elke werkgroep bestaat uit ongeveer vijf groepjes van vier à acht studenten. Binnen elk groepje vergelijken en bespreken de studenten hun uitwerkingen van de opgaven. Voor elke werkgroep is er één begeleider (meestal een studentassistent) aan wie de studenten uitleg kunnen vragen.

Van de opgaven die ik zelf in de werkgroepen behandel heb ik de uitwerkingen opgeschreven. Docenten die dit boek voor hun onderwijs gebruiken kunnen bij mij (maris@nici.ru.nl) een set van deze uitwerkingen bestellen.

## Met Dank Aan ...

De inhoud van het boek is gebaseerd op de kennis van vele anderen. Het is onmogelijk om hier alle personen, boeken en tijdschriftartikelen te noemen die eraan hebben bijgedragen. Echter, er zijn drie boeken die het verdienen om genoemd te worden. Ten eerste, bijna alles wat in het boek staat over variantie-, covariantie- en regressieanalyse is terug te vinden of af te leiden uit *The Analysis of Variance* van Harold Scheffé

(1959). Jammer genoeg, is het een boek voor statistici en volstrekt ontoegankelijk voor psychologiestudenten. Ten tweede, voor het hoofdstuk over dataverzamelingsmethoden heb ik verschillende voorbeelden ontleend aan Shaughnessy, Zechmeister en Zechmeister (2000), een mooi inleidend methodenboek voor eerstejaars psychologiestudenten. En ten derde, ik heb enkele opgaven van Maxwell en Delaney (1990) bewerkt en in het boek opgenomen. Dit wordt telkens vermeld bij het begin van deze bewerkte opgaven.

Van de personen die hebben bijgedragen aan dit boek moet in de eerste plaats mijn collega Mathieu Koppen genoemd worden. Hij heeft de eerste versies geschreven van de hoofdstukken over regressie- en factoranalyse en in de huidige versies staan nog steeds passages van zijn hand. Ook de werkgroepbegeleiders met wie ik de afgelopen jaren heb samengewerktmogen niet onvernoemd blijven. Het zijn er teveel om ze allemaal individueel te noemen. Door hun scherpe opmerkingen hebben zij mij gedwongen hele stukken te herschrijven. Ook hebben zij voorstellen gedaan voor tentamenvragen, die later als opgaven in het boek terechtkwamen. Sommige voorstellen waren zo goed dat ik ze vrijwel ongewijzigd kon overnemen. Mijn vriendin heeft het boek twee keer nagekeken op kromme zinnen, spellingsfouten en inconsistent taalgebruik. Alle fouten en onduidelijkheden die er nu nog in staan, zijn natuurlijk voor mijn rekening.

Eric Maris

# Hoofdstuk 1

# Dataverzamelingsmethoden

## Inhoudsopgave


**1.1 Variabelen . . . . . 17**
1.1.1 Wat zijn Variabelen? . . . . . 17
1.1.2 Persoons- en Situatievariabelen . . . . . 18
**1.2 Effectonderzoek . . . . . 20**
1.2.1 Experimentele en Natuurlijke Variabelen . . . . . 20
1.2.2 Het Gevaar bij het Gebruik van Natuurlijke Variabelen: Storende Persoonsvariabelen . . . . . 22
1.2.3 Het Gevaar bij een Gebrekkige Experimentele Controle van de Condities: Storende Situatievariabelen . . . . . 24
1.2.4 Het Binnen-Subjectenexperiment . . . . . 27
1.2.5 Wanneer is een Onderzoek een Experiment? . . . . . 30
**1.3 Standaardisering . . . . . 33**
**1.4 Observatiemethoden . . . . . 33**
1.4.1 Vrije (niet-gestandaardiseerde) Observatie . . . . . 34
1.4.2 Participerende Observatie . . . . . 35
1.4.3 Gestandaardiseerde Observatie . . . . . 36
1.4.4 Registratie . . . . . 40
1.4.5 Enkele Problemen bij het Gebruik van Observatiemethoden . 43
**1.5 Interview, Vragenlijst en Test . . . . . 50**
1.5.1 Het Vrije (niet-gestandaardiseerde) Interview . . . . . 50
1.5.2 Het Gestandaardiseerde Interview en de Vragenlijst . . . . . 50
1.5.3 De Psychologische Test . . . . . 57


Dit boek gaat over methoden en technieken van empirisch onderzoek. Empirisch onderzoek is onderzoek waarin gebruik gemaakt wordt van waarnemingen van gedrag. Deze waarnemingen noemen we ook *gegevens* of *data*.

Er is ook onderzoek dat geen gebruik maakt van data. Dit is zo bij het literatuuronderzoek. Hier baseert men zich op boeken en tijdschriftartikelen. Deze boeken en tijdschriftartikelen kunnen natuurlijk gebaseerd zijn op data, en daarom kan het resultaat van een literatuuronderzoek best bepaald zijn door data, zij het op een onrechtstreekse manier. Een literatuuronderzoek is echter geen empirisch onderzoek omdat het niet rechtstreeks gebaseerd is op data.

In dit hoofdstuk bespreken we een aantal methoden om data te verzamelen. Men kan de verschillende dataverzamelingsmethoden op meerdere manieren beschrijven en in dit hoofdstuk doen we dit op drie manieren, telkens vertrekkend van een ander uitgangspunt:

1. Wat is het *doel* van het onderzoek?

2. *Op welke wijze* worden de waarnemingen gedaan?

3. *Wat* wordt er waargenomen?

Elk van deze drie uitgangspunten leidt tot een andere indeling van de verschillende dataverzamelingsmethoden. Dit betekent niet dat deze drie indelingen volledig los staan van elkaar. Bijvoorbeeld, hoe en wat men waarneemt hangt mede af van het doel van het onderzoek.

Wat betreft het doel van het onderzoek, kunnen we meerdere typen van onderzoek onderscheiden. Eén daarvan is het *beschrijvend* onderzoek. Een voorbeeld van beschrijvend onderzoek is een onderzoek naar het gemiddeld aantal romans dat gelezen wordt door jongeren tussen de 16 en de 18 jaar. Een ander voorbeeld is een onderzoek naar het verband tussen het aantal gelezen romans en de omvang van de vriendenkring, zonder dat men geïnteresseerd is in de richting van dit verband. Indien de richting van het verband wel van belang zou zijn (bijvoorbeeld, "Heeft het aantal romans dat men leest een effect op de omvang van zijn of haar vriendenkring?"), dan zou het gaan om een *effectonderzoek*. Effectonderzoek probeert de vraag te beantwoorden of een bepaalde variabele een effect heeft op een andere variabele. In dit hoofdstuk, beperken we ons tot effectonderzoek; zeer veel onderzoek is van dit type en er is veel interessants over te vertellen.

Wat betreft de wijze waarop de waarnemingen gedaan worden, beperken we ons tot slechts één aspect: *standaardisering*. Standaardisering betekent dat men op voorhand bepaalt naar welke gedragsaspecten men gaat kijken en op welke manier men dat gaat doen. Standaardisering van de waarnemingen is van belang omdat het een voorwaarde is voor de herhaalbaarheid van het onderzoek.

Wat betreft datgene wát waargenomen wordt, maken we een onderscheid tussen observeren en het stellen van vragen. Binnen deze twee categorieën bespreken we een aantal dataverzamelingsmethoden die van elkaar verschillen wat betreft de mate van standaardisering.

Voordat we beginnen met de bespreking van de verschillende dataverzamelingsmethoden, staan we even stil bij het begrip *variabele*. Dit is een centraal begrip in de terminologie van de methoden van psychologisch onderzoek en van de dataverzamelingsmethoden in het bijzonder.

# 1.1 Variabelen

## 1.1.1 Wat zijn Variabelen?

Essentieel voor een variabele is dat zij verschillende waarden kan aannemen. Deze idee kan men uitdrukken door een variabele te beschouwen als een bepaald soort functie. Karakteristiek voor deze functie is dat zij een *index* als argument heeft.

We kunnen dit illustreren met een paar voorbeelden. We beginnen met de scores van 100 personen op een intelligentietest. Deze scores duiden we aan met $X$ en de personen duiden we aan met de index $p$ ($p = 1, \ldots, 100$). De score van de p-de persoon duiden we dan aan met $X_p$. Deze $X_p$ kunnen we beschouwen als de functiewaarde van een functie $X$ met argument $p$. Dit duiden we als volgt aan: $X(p)$. Door andere waarden te kiezen voor $p$, krijgen we telkens andere scores als functiewaarden. Dus, als we spreken van de variabele INTELLIGENTIETESTSCORE, dan bedoelen we de functie $X$. Hier en in de rest van dit boek wordt een variabele meestal aangeduid door haar naam in hoofdletters te schrijven. Een variabele wordt meestal voorgesteld in een tabel. Zo'n tabel wordt getoond in Tabel 1.1. In de eerste kolom van deze tabel staat de index $p$ en in de tweede kolom de variabele INTELLIGENTIETESTSCORE.

| Index (persoonsindex) | INTELLIGENTIE-TESTSCORE | MENING OVER HET BROEIKASEFFECT |
|---|---|---|
| 1 | 112 | 4 |
| 2 | 93 | 1 |
| 3 | 125 | 5 |
| 4 | 103 | 5 |
| 5 | 89 | 5 |
| ⋮ | ⋮ | ⋮ |
| 98 | 109 | 3 |
| 99 | 131 | 2 |
| 100 | 91 | 2 |

Tabel 1.1: De variabelen INTELLIGENTIETESTSCORE en MENING OVER HET BROEIKASEFFECT, voorgesteld in de vorm van een tabel.

Het is niet altijd zo dat een andere waarde voor $p$ (i.e., een andere persoon) ook resulteert in een andere functiewaarde, want verschillende personen hebben vaak dezelfde score op een variabele. Dit is bijvoorbeeld zo voor het antwoord op de stelling "Het broeikaseffect is een ernstige bedreiging voor het milieu" gegeven op een vijfpuntenschaal (1=helemaal oneens, 2=eerder oneens, 3=onbeslist, 4=eerder eens, 5=helemaal eens). Omdat er slechts vijf antwoordcategorieën zijn, zullen veel personen dezelfde score hebben op de variabele MENING OVER HET BROEIKASEFFECT. Deze variabele wordt getoond in de derde kolom van Tabel 1.1.

De bovenstaande twee voorbeelden waren variabelen die als functie van een *persoonsindex* uitgedrukt worden. In het volgende voorbeeld gaat het om een variabele die als

functie van een *situatie-index* uitgedrukt wordt. Darley en Latané (1968) toonden aan dat naarmate het aantal omstanders toeneemt, de kans afneemt dat iemand hulp verleent in een noodsituatie. Stel dat dit onderzoek uit vier condities bestaat met, respectievelijk, 0, 1, 3 en 7 omstanders. Deze vier situaties (condities) duiden we aan met de situatie-index $s$ ($s = 1, \ldots, 4$). De waarde van de s-de situatie duiden we aan met $X_s$. Deze $X_s$ is de functiewaarde van een functie $X$ met argument $s$, hetgeen aangeduid wordt als $X(s)$. Dus, de situatievariabele $X$, AANTAL OMSTANDERS, heeft de volgende waarden: $X_1 = X(1) = 0$, $X_2 = X(2) = 1$, $X_3 = X(3) = 3$ en $X_4 = X(4) = 7$. Deze variabele wordt voorgesteld in Tabel 1.2.

| Index (situatie-index) | AANTAL OMSTANDERS |
|---|---|
| 1 | 0 |
| 2 | 1 |
| 3 | 3 |
| 4 | 7 |

Tabel 1.2: De variabele AANTAL OMSTANDERS, voorgesteld in de vorm van een tabel.

Er bestaan ook variabelen die alleen maar dienen om aan te geven dát personen of situaties van elkaar verschillen, maar verder niet aangeven op welke manier zij dat doen. Bijvoorbeeld, stel dat men wil onderzoeken of er een verband is tussen iemands sociale stijl en de mate waarin hij of zij sympathiek gevonden wordt. Met andere woorden, men wil onderzoeken of personen met verschillende sociale stijlen ook verschillen wat betreft de mate waarin ze sympathiek gevonden worden. We zijn in de volgende vier sociale stijlen geïnteresseerd: kameraadschappelijk, koel, schuchter en verleidend. We kunnen dan de variabele SOCIALE STIJL definiëren, die de waarden 1, 2, 3 en 4 kan aannemen, elk corresponderend met één van de vier sociale stijlen. Waar het hier om gaat, is dat deze variabele slechts dient om aan te geven dat de personen verschillend zijn wat betreft sociale stijl, en verder niets. Een dergelijke variabele wordt met meerdere benamingen aangeduid. Soms spreekt men van een *nominale* variabele, maar ook de benamingen *categorische* en *kwalitatieve* variabele worden gebruikt.

Tot nu toe werden alle variabelen geschreven als functie van slechts één index. Variabelen kunnen echter ook meerdere indices als argument hebben. Bijvoorbeeld, in het kader van een onderzoek naar de rekenvaardigheid van basisschoolleerlingen neemt men bij een groep leerlingen een toets af bestaande uit 25 items. Voor elke leerling $p$, hebben we 25 itemscores die we aanduiden met $Y_{pi}$ ($i = 1, \ldots, 25$). Deze $Y_{pi}$ heeft als mogelijke waarden 1 (voor een correct antwoord) en 0 (voor een incorrect antwoord). In functievorm wordt $Y_{pi}$ geschreven als $Y(p, i)$, waarin $Y$ de variabele ITEMSCORE is. Deze variabele wordt voorgesteld in Tabel 1.3.

### 1.1.2 Persoons- en Situatievariabelen

Voor het onderscheid tussen persoons- en situatievariabelen is het goed om te vertrekken van Kurt Lewins (1951) definitie van psychologie. Lewin stelt dat psychologie de weten-

| | Itemindex | | | | | |
|---|---|---|---|---|---|---|
| Persoonsindex | 1 | 2 | 3 | ... | 24 | 25 |
| 1 | 1 | 1 | 0 | ... | 0 | 1 |
| 2 | 0 | 1 | 0 | ... | 0 | 0 |
| 3 | 1 | 1 | 1 | ... | 0 | 1 |
| 4 | 1 | 0 | 1 | ... | 0 | 1 |
| 5 | 0 | 1 | 1 | ... | 1 | 1 |
| ⋮ | ⋮ | ⋮ | ⋮ | | ⋮ | ⋮ |
| 98 | 0 | 1 | 1 | ... | 0 | 1 |
| 99 | 1 | 1 | 1 | ... | 1 | 1 |
| 100 | 1 | 0 | 0 | ... | 1 | 0 |

Tabel 1.3: De variabele ITEMSCORE, voorgesteld in de vorm van een tabel.

schap is die relaties bestudeert tussen, enerzijds, gedrag, en anderzijds, persoons- en situatiekenmerken. Dit kan men als volgt schrijven:

$$\boxed{G = f(P, S)} \tag{1.1}$$

In woorden, staat er in (1.1) dat gedrag, aangeduid met $G$, uitgedrukt wordt als functie van persoonskenmerken, aangeduid met $P$, en situatiekenmerken, aangeduid met $S$. De functie zelf wordt aangeduid met $f$. Deze functie kan verschillende vormen aannemen, afhankelijk van de wijze waarop een bepaald gedrag (het al-dan-niet verlenen van hulp, het behalen van bepaalde studieresultaten, het kiezen voor het ene of het andere beroep, ... ) bepaald wordt door persoons- en/of situatiekenmerken (het aantal omstanders bij hulpverlenend gedrag, intelligentie bij studieresultaten, sociale bewogenheid bij beroepskeuze, ... ).

Een persoonsvariabele is een variabele die een persoonskenmerk weergeeft en een situatievariabele is een variabele die een situatiekenmerk weergeeft. Meestal geldt voor een persoonsvariabele dat, geschreven in de vorm van een functie, deze variabele een persoonsindex als argument heeft. En voor een situatievariabele geldt meestal dat, geschreven in de vorm van een functie, deze variabele een situatie-index als argument heeft. Een voorbeeld van een persoonsvariabele is de score op een intelligentietest, die de persoonsindex $p$ als argument heeft ($X_p = X(p)$). Een voorbeeld van een situatievariabele is het aantal omstanders in het onderzoek van Darley en Latané (1968), die de situatie-index $i$ als argument heeft ($X_i = X(i)$).

Het feit dat de index van een variabele een persoons- of situatie-index is, is niet altijd een adequate manier om te bepalen of deze variabele een persoons- of een situatievariabele is. Dit is bijvoorbeeld het geval in het onderzoek van Darley en Latané (1968). Hier is het in principe mogelijk om het aantal omstanders niet alleen als een situatievariabele maar ook als een soort persoonsvariabele te beschouwen. Elke proefpersoon werd immers aan één van de vier condities toegewezen en het aantal omstanders

in deze conditie (0, 1, 3 of 7) kan beschouwd worden als de waarde van deze persoon op een persoonsvariabele. Omdat het de onderzoeker was die de proefpersonen aan de verschillende condities heeft toegewezen, zegt deze variabele natuurlijk niets over de proefpersonen zelf. Door de toekenning van proefpersonen aan condities hebben we van een situatievariabele een *pseudo*-persoonsvariabele gemaakt.

## 1.2 Effectonderzoek

In een effectonderzoek probeert men de vraag te beantwoorden of een bepaalde variabele een effect heeft op een andere variabele, en zo ja, in welke mate. Een andere manier om dit uit te drukken is zeggen dat men wil onderzoeken of er een *causaal verband* is tussen deze twee variabelen en hoe sterk dat verband is. We noemen een verband causaal als het *in isolatie* variëren van de ene variabele (dit wil zeggen, zonder dat enige andere variabele mee varieert) voldoende is voor het optreden van variatie in de andere variabele. De variabele waarvan men het effect onderzoekt is het feit of de proefpersoon tot de ene of de andere conditie behoort. Dit wil zeggen, het gaat bij effectonderzoek altijd over het effect van een situatievariabele en nooit van een persoonsvariabele. Een situatievariabele heeft een causaal verband met een andere variabele als het behoren tot de ene of de andere conditie voldoende is voor het optreden van variatie in die andere variabele. De eerste variabele wordt verder de *onafhankelijke variabele* genoemd en de tweede de *afhankelijke variabele*.

Effectonderzoek gebeurt meestal met behulp van een *experiment*. Men spreekt van een experiment als er sprake is van *manipulatie*. Een handig hulpmiddel om dit te onthouden is via de slogan "Experimenteren is manipuleren". Er zijn drie manieren waarop men iets kan manipuleren: (1) manipulatie door elke proefpersoon aan één van de condities toe te wijzen, (2) manipulatie van de condities zelf en (3) manipulatie door elke proefpersoon aan alle condities toe te wijzen. De eerste soort manipulatie (toewijzing aan één van de condities) is de basis van het onderscheid tussen *experimentele* en *natuurlijke* variabelen. Dit onderscheid is van belang als we het hebben over de controle van *storende persoonsvariabelen*. De tweede soort manipulatie (manipulatie van de condities zelf) speelt een rol bij de controle van *storende situatievariabelen*. De derde soort manipulatie (toewijzing aan alle condities) resulteert in een zogenaamd *binnen-subjectenexperiment*. Het toewijzen van de proefpersonen aan één van de condities gebeurt in een *tussen*-subjectenexperiment en het toewijzen van de proefpersonen aan álle condities gebeurt in een *binnen*-subjectenexperiment. Het probleem van storende persoonsvariabelen doet zich alleen voor in een tussen-subjectenexperiment. Het binnen-subjectenexperiment kent zijn eigen probleem: *volgorde-effecten*. Op al deze punten wordt in de volgende paragrafen ingegaan.

### 1.2.1 Experimentele en Natuurlijke Variabelen

Het onderscheid tussen experimentele en natuurlijke variabelen heeft te maken met de wijze waarop personen terechtkomen in een conditie. Bij een experimentele variabele is het de onderzoeker die bepaalt in welke conditie een persoon geobserveerd wordt. Bij

een natuurlijke variabele bepaalt de onderzoeker dit niet, maar registreert hij gewoon in welke conditie een persoon zich bevindt.

In een onderzoek naar het effect van sociale vaardigheidstraining op alcoholgebruik bij alcoholverslaafden, is de variabele TRAINING een experimentele variabele indien de onderzoeker bepaalt in welke van twee groepen de alcoholverslaafde terechtkomt. In dit onderzoek zijn dit de volgende twee groepen: (1) een groep die sociale vaardigheidstraining krijgt en (2) een groep die op dezelfde tijdstippen samenkomt voor ontspannende activiteiten. Ook in een onderzoek naar het effect van verschillende verpakkingen op de voorkeur voor dit waspoeder is het zo dat de variabele VERPAKKINGSVORM een experimentele variabele is indien de onderzoeker bepaalt of hij het waspoeder in de ene of de andere verpakking aanbiedt. Merk op dat, in het waspoederonderzoek, het erg voor de hand ligt dat de onderzoeker de verpakking bepaalt. In het sociale vaardigheidsonderzoek, daarentegen, ligt het niet zo voor de hand dat de onderzoeker bepaalt of een verslaafde in de groep met of de groep zonder training terechtkomt. In het sociale vaardigheidsonderzoek is men eerder geneigd om het aan de proefpersoon (i.c., de alcoholverslaafde) over te laten of hij de training wel of niet wil volgen. Dit brengt ons bij de natuurlijke variabelen.

Een voorbeeld van een natuurlijke variabele vinden we in een onderzoek naar het effect van verhuizing op geestelijk gezondheid. De variabele VERHUIZING werd vastgesteld als antwoord op de vraag of men nog steeds in dezelfde gemeente woont als die waar men het grootste gedeelte van zijn of haar jeugd heeft doorgebracht. De mate van geestelijke gezondheid werd vastgesteld met een vragenlijst waarin gepeild werd naar het voorkomen van gedragingen en belevingen die beschouwd worden als manifestaties van een psychische stoornis (bijvoorbeeld, irrationele angsten, periodes van lusteloosheid, achtervolgingsideeën, ...). Het belangrijke punt hier is dat de personen zelf beslist hadden om te verhuizen of niet; de onderzoeker had hier geen invloed. De variabele VERHUIZING noemen we daarom een natuurlijke variabele.

Een tweede voorbeeld van een natuurlijke variabele vinden we in een onderzoek naar het effect van twee vormen van verdoving op de ervaren pijn bij een tandheelkundige ingreep. De twee vormen van verdoving (de onafhankelijke variabele) zijn de volgende: (1) de klassieke injectie met een verdovend middel en (2) hypnose. En de afhankelijke variabele is de beoordeling van de hoeveelheid ervaren pijn op een schaal van 0 (helemaal geen pijn) tot 10 (ondraaglijke pijn). In dit onderzoek konden de patiënten zelf kiezen welke vorm van verdoving ze wensten. De variabele TYPE VERDOVING noemen we daarom een natuurlijke variabele.

Het moet benadrukt worden dat het onderscheid tussen een experimentele en een natuurlijke variabele niets te maken heeft met de condities waar men in terecht kan komen. Het is de wijze waarop de personen in de condities terechtkomen die bepalend is voor dit onderscheid.

Een onderzoek waarin men het effect wil onderzoeken van een natuurlijke variabele wordt een *observationeel onderzoek* genoemd. In een observationeel onderzoek kiezen de proefpersonen zelf de conditie waar ze in terechtkomen. Twee voorbeelden van een observationeel onderzoek zijn (1) het onderzoek naar het effect van verhuizing op geestelijke gezondheid en (2) het onderzoek naar het effect van verschillende vormen van verdoving op de ervaren pijn bij een tandheelkundige ingreep. Zoals verder uitgelegd zal worden,

is het grote gevaar bij een observationeel onderzoek dat storende persoonsvariabelen niet onder controle gehouden worden. Hier ligt het grote voordeel van experimentele variabelen. Omdat de onderzoeker zelf de proefpersonen aan de condities toewijst, kan hij dat op zo'n manier doen dat storende persoonsvariabelen onder controle gehouden worden. Zoals verder uitgelegd zal worden, is dit mogelijk door de proefpersonen op *zuiver toevallige (gerandomiseerde) wijze* aan de condities toe te wijzen.

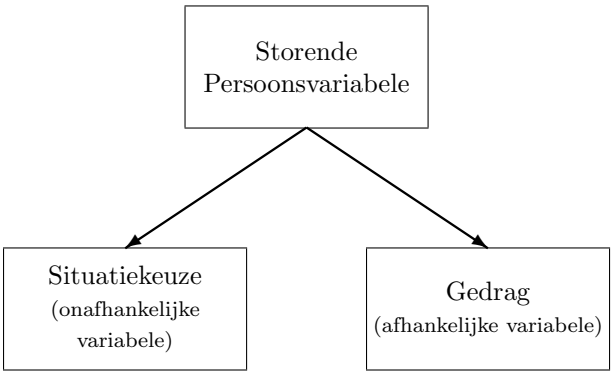
Volgens sommige boeken is randomisering zo belangrijk dat men alleen van een experiment spreekt als de toewijzing aan de condities gerandomiseerd is. In dit boek zullen we dit een *gerandomiseerd experiment* noemen. We spreken van een experiment zondermeer als de onderzoeker de proefpersonen toewijst aan de condities. Hij bepaalt zelf op welke manier hij dat doet. Het spreekt echter voor zich dat de onderzoeker bij de toewijzing aan de condities niet zomaar wat doet. Meestal zal hij de toewijzing zo uitvoeren dat de storende persoonsvariabelen onder controle gehouden worden. De beste manier om storende persoonsvariabelen onder controle te houden is randomisering.

### 1.2.2 Het Gevaar bij het Gebruik van Natuurlijke Variabelen: Storende Persoonsvariabelen

Indien de onafhankelijke variabele een natuurlijke variabele is, dan is het gevaar groot dat een effectonderzoek geen antwoord kan geven op de vraag of het verband tussen de onafhankelijke en de afhankelijke variabele causaal is. De reden hiervoor is dat deze natuurlijke variabele kan samenhangen met een storende persoonsvariabele die ook samenhangt met de afhankelijke variabele. Hieruit volgt dat het geobserveerde verband tussen de natuurlijke en de afhankelijke variabele niet noodzakelijk toe te schrijven is aan een causaal verband tussen beide. En hieruit volgt ook dat de áfwezigheid van een geobserveerd verband tussen de natuurlijke en de afhankelijke variabele niet noodzakelijk toe te schrijven is aan de áfwezigheid van een causaal verband tussen beide.

De reden waarom een natuurlijke variabele kan samenhangen met een storende persoonsvariabele is dat het bij een natuurlijke variabele zo kan zijn dat de personen zelf de conditie kiezen waar ze in terechtkomen. Het is goed mogelijk dat de keuze voor een bepaalde conditie (bijvoorbeeld, injectie/hypnose, wel/niet verhuizen) samenhangt met persoonsvariabelen die eveneens samenhangen met het gedrag dat men onderzoekt, de afhankelijke variabele (i.c., de hoeveelheid ervaren pijn en de mate van geestelijke gezondheid). In het meest extreme geval is het zelfs zo dat de samenhang met deze persoonsvariabelen volledig verantwoordelijk is voor een eventueel verschil tussen de condities op de afhankelijke variabele; er is dan geen extra invloed van de variabele waarvan men het effect onderzoekt. Dit extreme geval wordt grafisch voorgesteld in Figuur 1.1. In het pijn-onderzoek kan de samenhang met de persoonsvariabele ZELFVERTROUWEN er zowel voor zorgen dat patiënten eerder voor een onbekende verdoving (i.c., hypnose) kiezen als voor het feit dat de patiënten minder pijngevoelig zijn. En in het verhuisonderzoek kan de samenhang met de persoonsvariabele CONTACTVAARDIGHEID er zowel voor zorgen dat men eerder geneigd is om niet te verhuizen (omwille van de vele vriendschapsrelaties die men zou moeten verbreken), als voor het feit dat men minder vatbaar is voor psychische stoornissen.

Het kiezen van een bepaalde situatie moet hier ruim opgevat worden. Zo is het ook




Figuur 1.1: Het gevaar bij natuurlijke variabelen: een persoonsvariabele die samenhangt met zowel situatiekeuze als gedrag.

mogelijk dat een persoon in een bepaalde situatie geplaatst wordt door anderen die deze beslissing nemen op basis van de kenmerken van de persoon. Bijvoorbeeld, het zou zo kunnen zijn dat intelligente kinderen minder populair zijn en daardoor minder op feestjes uitgenodigd worden, zodat er een verband ontstaat tussen het aantal feestjes dat iemand bezoekt en zijn studieresultaten, een verband dat echter volledig toe te schrijven is aan de samenhang met de persoonsvariabele INTELLIGENTIE.

Het is verleidelijk om de pijltjes in Figuur 1.1 te interpreteren als causale verbanden. Deze causale interpretatie houdt in dat het in isolatie variëren van de storende variabele voldoende is voor het optreden van variatie in de twee andere variabelen. Hoewel deze interpretatie niet fout is, is het niet de enig mogelijke. Men kan de pijltjes net zo goed opvatten als verbanden zonder richting, hetgeen neerkomt op het wegdenken van de punten van de pijltjes. Immers, ook zonder richting in het verband tussen de variabelen, is het zo dat het verband tussen de afhankelijke en de onafhankelijke variabele (verschillen tussen de condities op de afhankelijke variabele) toe te schrijven is aan de storende persoonsvariabele, en wel in de volgende zin: indien de storende persoonsvariabele constant gehouden wordt (bijvoorbeeld, in het pijn-onderzoek, door alleen te kijken naar personen met dezelfde mate van zelfvertrouwen) dan verdwijnt het verband tussen de onafhankelijke en de afhankelijke variabele (i.c., wel-of-geen injectie en de hoeveelheid ervaren pijn). Op deze manier is de storende persoonsvariabele verantwoordelijk voor het verband tussen de onafhankelijke en de afhankelijke variabele zonder dat er sprake is van het in isolatie variëren van de storende persoonsvariabele.

Indien een onafhankelijke variabele niet natuurlijk is maar experimenteel, dan is het mogelijk om storende persoonsvariabelen onder controle te houden. Dit kan op twee manieren: met *matching* of met *randomisering*. Bij matching gaat men ervan uit dat men alle relevante kenmerken van de proefpersonen kent, dit wil zeggen, alle kenmerken die samenhangen met de afhankelijke variabele (zelfvertrouwen, contactvaardigheid, intelligentie, ...). Men zorgt er dan voor dat de groepen gelijk samengesteld (in het Engels, *matched*) zijn wat betreft deze kenmerken. Jammer genoeg, kan men er nooit zeker van zijn dat men alle relevante kenmerken kent. Daarom kan matching ook niet garanderen dat er geen systematische verschillen bestaan tussen de groepen; men kan

immers altijd relevante kenmerken over het hoofd gezien hebben.

Het grote voordeel van randomisering (zuiver toevallige toewijzing) is dat men helemaal niets hoeft te weten over de persoonskenmerken die samenhangen met de afhankelijke variabele. We kunnen randomisering uitleggen aan de hand van het onderzoek naar het effect van sociale vaardigheidstraining op alcoholgebruik bij een groep verslaafden. De onderzoeker kan deze verslaafden op zuiver toevallige wijze (in het Engels, *at random*) toewijzen aan één van de twee condities (training of geen training) door een muntstuk op te gooien. Bij kruis krijgt de verslaafde sociale vaardigheidstraining en bij munt krijgt hij in plaats daarvan ontspannende activiteiten. De kracht van deze toewijzingsprocedure is erin gelegen dat het resultaat van deze muntworp zuiver toevallig is en daarom met geen enkele persoonsvariabele kan samenhangen. Met andere woorden, het zuiver toevallig zijn van de toewijzing maakt storende persoonsvariabelen onmogelijk.

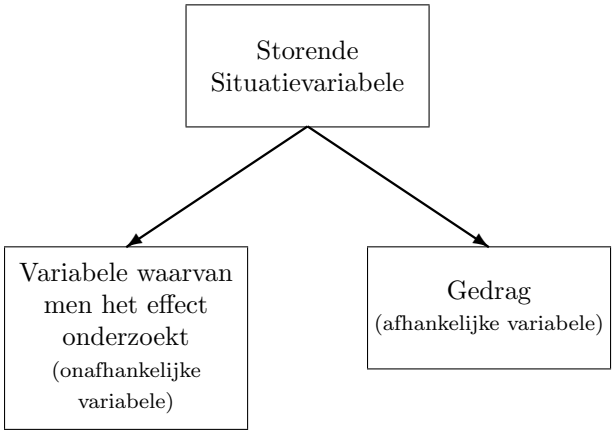
Randomisering garandeert niet dat er geen *toevallige* samenhang kan zijn tussen het resultaat van de toewijzingsprocedure en een storende persoonsvariabele. Bijvoorbeeld, het kan toevallig zo uitpakken dat de ernstig verslaafden in de conditie zonder training terechtkomen en de minder ernstig verslaafden in de conditie met training. In dit geval zal men ook zonder een effect van de training een statistisch significant verschil observeren. De *kans* op een dergelijke "ongelukkige" toewijzing is echter klein. Men kan langs wiskundige weg aantonen dat deze kans vrijwel gelijk is aan het alfaniveau van de statistische toets (0.05 of 0.01).

### 1.2.3 Het Gevaar bij een Gebrekkige Experimentele Controle van de Condities: Storende Situatievariabelen

Bij effectonderzoek is men vaak geïnteresseerd in één welbepaald aspect van de condities die men met elkaar vergelijkt. Als voorbeeld nemen we een onderzoek bij een groep werkende mannen en vrouwen naar het effect van het behoren tot een een- of een tweeverdienersgezin (de variabele EEN- VERSUS TWEEVERDIENER) op hun tevredenheid met de relatie met hun partner (de variabele TEVREDENHEID MET DE RELATIE). In dit onderzoek is men geïnteresseerd in het effect van EEN- VERSUS TWEEVERDIENER op zich en niet van een andere variabele die hiermee samenhangt. Het is echter zeer goed mogelijk dat de condities (in ons voorbeeld, eenverdienersgezin en tweeverdienersgezin) op meer aspecten van elkaar verschillen dan degene waar men in geïnteresseerd is. Het is dan mogelijk dat het geobserveerde verband tussen de onafhankelijke en de afhankelijke variabele (blijkend uit een verschil tussen de condities) niet toe te schrijven is aan het aspect waar men in geïnteresseerd is, maar wel aan die andere aspecten. Die andere aspecten worden *storende situatievariabelen* genoemd.

We geven nu een voorbeeld. Uit het genoemde onderzoek naar het effect van EEN- VERSUS TWEEVERDIENER op TEVREDENHEID MET DE RELATIE zou kunnen blijken dat de partners in een tweeverdienersgezin gemiddeld meer tevreden zijn dan de partners in een eenverdienersgezin. Dit betekent echter niet dat het verband tussen deze variabelen causaal is, in de zin dat het toe te schrijven is aan de variabele EEN- VERSUS TWEEVERDIENER op zich. Het kan zijn dat de storende situatievariabele OPLEIDINGSNIVEAU PARTNER samenhangt met de variabele TEVREDENHEID MET DE RELATIE, maar dat door de samenhang met de variabele EEN- VERSUS TWEEVER-

DIENER ook deze laatste variabele een samenhang vertoont met TEVREDENHEID MET DE RELATIE. Anders uitgedrukt, het is mogelijk dat men gemiddeld meer tevreden is met de relatie met zijn of haar partner naarmate het opleidingsniveau van die partner hoger is, en dat het opleidingsniveau van deze partner sterk samenhangt met het feit of men tot een- of een tweeverdienersgezin behoort. Het is immers aannemelijk dat, naarmate de partner een hoger opleidingsniveau heeft, hij of zij ook mee voor het inkomen zorgt. Dit patroon van verbanden wordt grafisch voorgesteld in Figuur 1.2.




Figuur 1.2: Het gevaar bij een gebrekkige experimentele controle: een storende situatievariabele die zowel samenhangt met de variabele waarvan men het effect wil onderzoeken als met het gedrag.

Door randomisering wordt een systematische invloed van storende *persoons*variabelen uitgesloten, maar het is geen oplossing voor het probleem van storende *situatie*variabelen. Immers, stel je voor dat, in een denkbeeldig onderzoek, een groep werkenden op zuiver toevallige wijze wordt toegewezen aan een partner die zelf ook een betaalde baan heeft of een partner die dat niet heeft. Een eventueel verschil in tevredenheid tussen deze twee condities kan dan nog steeds verklaard worden door de storende situatievariabele OPLEIDINGSNIVEAU PARTNER.

Wat wel een oplossing is voor het probleem van storende situatievariabelen, is experimentele controle van de condities. Experimentele controle van de condities houdt in dat men de verschillende condities vergelijkbaar maakt wat betreft de mogelijke storende situatievariabelen. Voor het onderzoek naar de tevredenheid met de relatie zou dit inhouden dat men ervoor zorgt dat de partners in de eenverdienersconditie gemiddeld hetzelfde opleidingsniveau hebben als de partners in de tweeverdienersconditie. Hoe men dit concreet moet doen, is niet onmiddellijk duidelijk, en het toont aan waar het probleem ligt bij experimentele controle: experimentele controle van de condities is slechts mogelijk als de onderzoeker zelf de aard van de condities kan bepalen. In veel onderzoek buiten het laboratorium, is dit niet mogelijk; daar moet men de condities nemen zoals ze zijn.

We geven nu een voorbeeld van een laboratoriumonderzoek waarin experimentele controle van de condities wel mogelijk is. In dit onderzoek wil men nagaan of er een

effect is van de frequentie van voorkomen van een woord op de tijd die nodig is om dit woord te lezen en daarna uit te spreken. Een voorbeeld van een vaak voorkomend woord is "goed" en een voorbeeld van een weinig voorkomend woord is "gard". De twee condities in dit onderzoek bestaan uit een groep vaak voorkomende en een groep weinig voorkomende woorden. Het is de taak van de onderzoeker om aannemelijk te maken dat een eventueel verschil tussen deze twee condities alleen toegeschreven kan worden aan de frequentie van voorkomen van de woorden. Hij dient er bijvoorbeeld voor te zorgen dat de woorden in deze twee groepen uit evenveel letters bestaan en opgebouwd zijn uit lettercombinaties met dezelfde uitspreekbaarheid (vergelijk "kla" met "wra"). Omdat de onderzoeker zelf kan bepalen hoe de condities eruit zullen zien (i.c., welke woorden hij zal aanbieden), is experimentele controle van de condities hier mogelijk.

De kracht van experimentele controle van de condities als onderdrukker van storende situatievariabelen is meestal niet zo groot als de kracht van randomisering als onderdrukker van storende persoonsvariabelen. De reden hiervoor is dat randomisering de systematische invloed van alle mogelijke storende persoonsvariabelen onmogelijk maakt, terwijl experimentele controle van de condities meestal alleen die storende situatievariabelen onderdrukt waarvan men de mogelijke invloed onderkent. Niets garandeert dat de situatievariabelen die men controleert ook de enige zijn die een invloed hebben.

Er is één vorm van experimentele controle van de condities die wel garandeert dat alle mogelijke storende situatievariabelen onder controle gehouden worden. Jammer genoeg, is deze vorm van experimentele controle bijna nooit praktisch uitvoerbaar. Hij komt erop neer dat de situaties op zuiver toevallige wijze toegewezen worden aan het aspect van de condities waarin men geïnteresseerd is. Dus, net zoals men storende persoonsvariabelen onder controle kan houden door de proefpersonen op zuiver toevallige wijze aan de condities toe te wijzen, zo kan men storende situatievariabelen onder controle houden door de situaties op zuiver toevallige wijze toe te wijzen aan de verschillende niveaus van het aspect van de condities waarin men geïnteresseerd is. Bijvoorbeeld, in het onderzoek naar het effect van EEN- VERSUS TWEEVERDIENER op TEVREDENHEID MET DE RELATIE zou men in principe alle partners (de situaties) op zuiver toevallige wijze kunnen toewijzen aan de eenverdienersconditie (als de partner een betaalde baan heeft, dan moet hij ontslag nemen) of de tweeverdienersconditie (als de partner nog geen betaalde baan heeft, dan moet hij er een nemen). De kracht van zuiver toevallige toewijzing is erin gelegen dat het resultaat ervan met geen enkele situatievariabele kan samenhangen (bijvoorbeeld, OPLEIDINGSNIVEAU PARTNER). Zelfs in laboratoriumonderzoek is het vaak onmogelijk om deze vorm van experimentele controle uit te voeren. Bijvoorbeeld, in het onderzoek naar het effect van de frequentie van voorkomen van een woord op de tijd die nodig is om dit woord te lezen en daarna uit te spreken, is het onmogelijk om een woord (de situatie) op zuiver toevallige wijze toe te wijzen aan de hoog of de laag frequente conditie. Voor de duidelijkheid, experimentele controle van de condities zoals hierboven beschreven (i.e., door middel van zuiver toevallige toewijzing van de situaties) wordt verder gewoon *experimentele controle van de condities* genoemd en zuiver toevallige toewijzing van de proefpersonen aan de condities wordt verder *randomisering* genoemd.

Randomisering en experimentele controle van de condities zijn beide vormen van experimentele controle: met randomisering worden storende persoonsvariabelen onder

**Statistische Uitzuivering**
Het is vaak onmogelijk om storende variabelen te controleren door middel van randomisering of experimentele controle van de condities. Dit is bijvoorbeeld het geval in alle onderzoeken waarbij de onafhankelijke variabele natuurlijk is in plaats van experimenteel (bijv., als de proefpersoon zelf zijn conditie kiest). Wil men hier het effect onderzoeken van de onafhankelijke variabele, dan moet men zich behelpen met andere methoden om de invloed van storende variabelen uit te sluiten. De meest bekende van deze methoden is de statistische uitzuivering van het effect van storende variabelen. De covariantieanalyse en de variantieanalyse voor ongebalanceerde designs zijn twee technieken waarmee men een statistische uitzuivering kan doen. Jammer genoeg zijn deze technieken niet zo krachtig als de zuiver toevallige toewijzing. Zij kunnen namelijk niet garanderen dat na uitzuivering alleen het zuivere effect van de onafhankelijke variabele overblijft.

controle gehouden en met experimentele controle van de condities worden storende situatievariabelen onder controle gehouden. Deze twee vormen van experimentele controle kunnen los van elkaar voorkomen. Bijvoorbeeld, in een *gerandomiseerd* onderzoek naar het effect van EEN- VERSUS TWEEVERDIENER op TEVREDENHEID MET DE RELATIE (de werkende proefpersonen worden at random toegewezen aan een partner die zelf ook een betaalde baan heeft of een partner die dat niet heeft) worden de storende persoonsvariabelen onder controle gehouden maar niet de storende situatievariabelen. En in een onderzoek naar het uitspreken van woorden waarbij de proefpersonen zelf hun conditie mogen kiezen worden storende situatievariabelen onder controle gehouden (tenminste, als de onderzoeker het goed doet) maar geen storende persoonsvariabelen (misschien kiezen zwakke lezers de gemakkelijke conditie en sterke lezers de moeilijke).

### 1.2.4 Het Binnen-Subjectenexperiment

In een binnen-subjectenexperiment bestaat de manipulatie erin dat elke proefpersoon aan álle condities toegewezen wordt. Een voorbeeld van een binnen-subjectenexperiment is het volgende onderzoek naar de cognitieve verwerking van actieve en passieve zinnen: een proefpersoon beantwoordt een reeks vragen over plaatjes, gesteld in de actieve vorm (bijvoorbeeld, "Bijt de hond de man?"), en een reeks volledig vergelijkbare vragen, gesteld in de passieve vorm (bijvoorbeeld, "Wordt de man door de hond gebeten?"). Men is hier geïnteresseerd in het verschil in gemiddelde antwoordtijd op actieve en passieve vragen. Dit geeft ons een aanwijzing over de wijze waarop deze vragen cognitief verwerkt worden. Het belangrijke punt is dat elke proefpersoon zowel actieve als passieve vragen beantwoordt; dit is wat het onderzoek tot een binnen-subjectenexperiment maakt.

In een binnen-subjectenexperiment observeert men dezelfde proefpersonen in alle condities. Er is dus geen reden om ons zorgen te maken over storende persoonsvariabelen. Het feit dat men elke proefpersoon in alle condities observeert wordt ook uitgedrukt door te zeggen dat elke proefpersoon *als zijn eigen controle fungeert*. Storende situatievariabelen kunnen echter wel een invloed hebben. In een binnen-subjectenexperiment worden storende situatievariabelen op dezelfde manier gecontroleerd als in een tussen-subjectenexperiment, namelijk door middel van experimentele controle van de condities.

Het binnen-subjectenexperiment heeft zijn eigen probleem: volgorde-effecten. Om-

dat men de ene conditie voor of na de andere conditie observeert, zijn de observaties in deze condities niet noodzakelijk vergelijkbaar. Een probleem dat zich hier kan voordoen is *transfer* van de eerste naar de volgende conditie(s). We spreken van transfer als de ervaring uit de eerste conditie(s) het gedrag in de volgende condities beïnvloedt. We kunnen dit illustreren aan de hand van een onderzoek van Sharma en Moskowitz (1972) naar het effect van marihuana op de waarneming. Sharma en Moskowitz lieten hun proefpersonen op vier verschillende dagen naar hun laboratorium komen. Daar rookten zij twee sigaretten die één van vier mogelijke doses marihuana bevatten (0, 50, 100 of 200 eenheden tetrahydrocannabinol, het actieve bestanddeel in marihuana). Onmiddellijk daarna werd getest in welke mate zij gevoelig waren voor een bepaalde autokinetische illusie. Concreet, zij werden in een verduisterde kamer geplaatst en moesten zich concentreren op een lichtpuntje. De illusie die proefpersonen vaak rapporteren is dat het lichtpuntje lijkt te bewegen, ook al blijft het in werkelijkheid op dezelfde plaats. Sharma en Moskowitz stelden vast dat de mate van autokinetische illusie (i.e., de grootte van de schijnbare beweging) toenam met de dosis marihuana. Voor de interpretatie van deze vaststelling is het van belang te weten wat de volgorde van de condities (0, 50, 100 of 200 eenheden) was. Immers, het is niet uitgesloten dat de mate van autokinetische illusie afhangt van ervaring (i.e., het aantal keer dat men deze perceptuele taak reeds gedaan heeft) en daarom kan de volgorde van de condities bepalend geweest zijn voor de vaststelling dat de mate van autokinetische illusie toenam met de dosis marihuana.

De mate waarin men zich zorgen maakt over de invloed van transfer hangt af van de aard van de taak. Als Sharma en Moskowitz, in plaats van de mate van autokinetische illusie, de prestatie op een psychomotorische taak (bijvoorbeeld, kleine schroefjes indraaien) als afhankelijke variabele hadden gekozen, dan zou de interpretatie van hun resultaten veel meer aan kritiek onderhevig zijn. Immers, van dergelijke taken is geweten dat men er beter op presteert naarmate men ze vaker gedaan heeft. Dat dit ook het geval is bij de autokinetische illusie is niet aangetoond, maar het is ook niet uitgesloten.

Om te voorkomen dat transfer de resultaten beïnvloedt, wordt er vaak gebruik gemaakt van *contrabalancering*. Contrabalancering is het op een slimme manier bepalen van de volgorde van de condities met de bedoeling de invloed van transfer op te heffen. Bijvoorbeeld, Sharma en Moskowitz zouden bij één helft van de proefpersonen de volgorde $(0, 50, 100, 200)$ kunnen hanteren en bij de andere helft de volgorde $(200, 100, 50, 0)$. (In werkelijkheid, gebruikten Sharma en Moskowitz een meer gesofisticeerde vorm van contrabalancering, het zogenaamde *Latijns vierkant*. Deze vorm van contrabalancering wordt in detaïl beschreven door Maxwell en Delaney (1990).)

Contrabalancering werkt niet in alle situaties. Het werkt bijvoorbeeld niet als er sprake is van *differentiële transfer*. Differentiële transfer betekent dat de mate van transfer van conditie A naar conditie B anders is dan die van conditie B naar conditie A. Bijvoorbeeld, stel dat men geïnteresseerd is in een vergelijking van twee didactische methoden voor het leesonderwijs: methode A en methode B. De afhankelijke variabele in dit onderzoek is een test voor leesvaardigheid. Omdat het effect van een didactische methode duurzaam is (de verworven leesvaardigheid verliest men niet), is er hier sprake van transfer. En indien een van de twee methoden volstrekt inefficiënt is (een kind leert er helemaal niets van), dan is er ook sprake van differentiële transfer. Er is dan immers

transfer van de efficiënte naar de inefficiënte methode, maar niet omgekeerd.

Of er sprake is van transfer, en van welke vorm deze transfer is (differentieel of niet-differentieel), kan men niet bepalen op basis van de gegevens van het binnen-subjectenexperiment zelf. Bepalen of er sprake kan zijn van transfer gebeurt voordat men de gegevens verzamelt en meestal doet men dit op basis van gezond verstand. In het onderzoek naar de effectiviteit van de twee didactische methoden voor het leesonderwijs is het zeer aannemelijk dat er sprake is van transfer. Maar het is zeker niet altijd zo duidelijk of er nu sprake is van transfer of niet.

In de praktijk maakt men zich minder zorgen over gewone transfer dan over differentiële transfer. Door gebruik te maken van contrabalancering kan men immers voorkomen dat gewone transfer de resultaten beïnvloedt. Dit geldt niet voor differentiële transfer. Differentiële transfer is nooit met zekerheid uit te sluiten op basis van gezond verstand alleen. Het is wel zo dat differentiële transfer in sommige onderzoeken meer of minder aannemelijk is dan in andere. Stel dat we een onderzoek doen naar een situatievariabele met een momentaan in plaats van een duurzaam effect. Een voorbeeld hiervan is een onderzoek naar het effect van verschillende verpakkingen op de voorkeur voor een bepaald waspoeder. Elke conditie binnen dit onderzoek correspondeert met een bepaalde verpakking (een grote rode doos, een klein wit trommeltje met een tekening, ... ). De voorkeur wordt operationeel gedefinieerd als het bedrag dat men bereid is te betalen voor het waspoeder (in een bepaalde verpakking). In vergelijking met het onderzoek naar de twee didactische methoden voor het leesonderwijs, is er hier sprake van een situatievariabele waarvan het effect veel minder of zelfs helemaal *niet* duurzaam is. We verwachten immers niet dat de persoon verandert door het geven van zijn voorkeur voor een bepaalde verpakking. Dit drukken we uit door te zeggen dat het effect van de situatievariabele VERPAKKINGSVORM momentaan is.

Als het effect van de situatievariabele VERPAKKINGSVORM momentaan is, dan kan men deze variabele probleemloos onderzoeken in een binnen-subjectenexperiment, en is contrabalancering zelfs niet nodig. Immers, als een effect momentaan is, dan kan men ervan uitgaan dat de voorkeur voor een bepaalde verpakkingsvorm niet beïnvloed wordt door de verpakkingsvormen waarvoor men *eerder* een voorkeursoordeel gegeven heeft.

Het probleem met deze redenering is dat hij vertrekt van de veronderstelling dat het effect van de situatievariabele VERPAKKINGSVORM momentaan is. Deze veronderstelling hoeft niet noodzakelijk op te gaan en dan is er sprake van transfer, en is het niet uitgesloten dat deze transfer ook differentieel is. De Britse psycholoog Poulton (1973, 1975, 1982) heeft in verschillende artikelen laten zien dat het gevaar van differentiële transfer vaak ten onrechte onderschat wordt. Poulton is dan ook een van de grote pleitbezorgers van het gerandomiseerde tussen-subjectenexperiment. Omdat elke proefpersoon slechts in één conditie terechtkomt, en de proefpersonen in de verschillende condities vergelijkbaar zijn (door de randomisering), zijn de resultaten van zo'n experiment eenduidig.

Tot slot, vermelden we nog een speciaal soort binnen-subjectenexperiment dat regelmatig ten onrechte gebruikt wordt om een causaal verband aan te tonen. Het gaat hier over het onderzoeksdesign met een voor- en een nameting. Stel dat men wil onderzoeken of meditatie een positief effect heeft op de gemoedsgesteldheid. De onderzoeksgroep is een groep patiënten van een huisartsenpraktijk die zich aangemeld had met als voor-

naamste klacht lusteloosheid. Bij aanvang van het onderzoek vullen zij een klachtenvragenlijst in (de voormeting), daarna volgen zij een cursus mediteren (de behandeling) en tenslotte vullen zij de klachtenvragenlijst opnieuw in (de nameting). Nu blijkt dat bij de nameting de lusteloostheid duidelijk lager is dan bij de voormeting. Mag men hieruit besluiten dat mediteren een positief effect heeft op gevoelens van lusteloosheid? Dit is duidelijk niet het geval, want het kan best zijn dat de lusteloosheid ook afgenomen zou zijn zonder de cursus mediteren. Anders uitgedrukt, de voor- en de nameting verschillen niet alleen van elkaar wat betreft het gevolgd hebben van de cursus mediteren, maar ook wat betreft het tijdstip waarop men registreert.

### 1.2.5 Wanneer is een Onderzoek een Experiment?

De meeste studenten willen weten of een bepaald onderzoek nu wel of geen experiment is. Bijvoorbeeld, men wil weten of een laboratoriumonderzoek met sterke experimentele controle van de condities maar zonder randomisering, toch een experiment genoemd mag worden. Het antwoord op deze vraag hangt af van hoe het begrip "experiment" gedefinieerd wordt. Op zijn beurt hangt deze definitie af van zijn ingrediënten en of deze ingrediënten noodzakelijk en/of voldoende zijn. De mogelijke ingrediënten zijn de volgende: randomisering, experimentele controle van de condities, toewijzing aan meerdere condities (het binnen-subjectenexperiment), en contrabalancering.

We geven hier geen definitie van het begrip "experiment" omdat het niet van belang is voor de praktijk. Het enige wat van belang is, is het antwoord op de vraag of er een causaal effect is van de onafhankelijke variabele en hoe groot dit effect is. Voor het antwoord op deze vraag, zijn de volgende drie deelvragen van belang:

1. Is er sprake van storende persoonsvariabelen?

2. Is er sprake van storende situatievariabelen?

3. Is er sprake van transfer?

Het antwoord op deze drie deelvragen hangt onder andere af van het feit of er sprake was van randomisering, experimentele controle van de condities, en contrabalancering. Hoe het begrip "experiment" gedefinieerd is, helpt ons niet bij de beantwoording van deze drie deelvragen.

---

*Opgave 1*
Hoe heet een variabele die alleen maar dient om aan te geven dát personen of situaties van elkaar verschillen, maar verder niet aangeeft op welke manier zij dat doen?

*Opgave 2*
Wanneer is een verband causaal?

*Opgave 3*
Aan welke voorwaarde moet een onderzoek voldoen om van een experiment te kunnen spreken? Geef minstens twee voorbeelden van een onderzoek dat een experiment genoemd kan worden.

*Opgave 4*
Wat is een gerandomiseerd experiment?

*Opgave 5*
Wat is het verschil tussen een experimentele en een natuurlijke onafhankelijke variabele?

*Opgave 6*
Wat is het gevaar bij natuurlijke variabelen? Leg uit hoe randomisering een oplossing is voor dit probleem.

*Opgave 7*
Wat is het gevaar bij een gebrekkige experimentele controle van de condities? Leg dit uit aan de hand van een voorbeeld.

*Opgave 8*
Een onderzoekster was geïnteresseerd in het effect van de woonsituatie op vereenzaming bij bejaarden. Zij onderzocht twee groepen bejaarden: een groep die zelfstandig woonde en een groep die in zogenaamde serviceflats woonde. (Serviceflats zijn appartementen met speciale voorzieningen voor de bejaarden, zoals de verstrekking van warme maaltijden, onmiddellijke medische bijstand, e.d.). De zelfstandig-wonende groep was een zuiver toevallige trekking uit de populatie van zelfstandig-wonende bejaarden in Nederland. En de groep bejaarden die in een serviceflat wonen was een zuiver toevallige trekking uit de populatie van alle bejaarden in Nederland die in een serviceflat wonen. De onderzoekster stelde vast dat de bejaarden die zelfstandig woonden minder sociale contacten hadden dan de bejaarden in serviceflats. (Hierbij werden de contacten met het verzorgend personeel niet meegeteld.) Mag men hieruit besluiten dat men de sociale contacten van bejaarden kan doen toenemen door serviceflats te promoten? Leg uit waarom wel of waarom niet.

*Opgave 9*
Een onderzoeker was geïnteresseerd in het effect van arbeidsomstandigheden op de arbeidstevredenheid bij seizoenarbeiders (aspergestekers, fruitplukkers, ... ). Meer bepaald, hij wilde weten of er een effect was van het feit of de baas zelf op het veld meewerkte of niet. De onderzoeker kreeg de toestemming van enkele arbeidsbemiddelingsbureaus om de kandidaatwerknemers op zuiver toevallige wijze toe te wijzen aan de bedrijven waar de baas meewerkte en bedrijven waar dit niet gebeurde. Hij stelde vast dat de arbeidstevredenheid veel groter was in de bedrijven waar de baas zelf meewerkte. Mag men hieruit besluiten dat men de arbeidstevredenheid kan verhogen door de baas mee te laten werken op het land? Op welke manier wordt er wel rekening gehouden met factoren die dit besluit in de weg kunnen staan en op welke manier wordt er geen rekening mee gehouden?

*Opgave 10*
Een psycholoog van het Ministerie van Justitie was geïnteresseerd in het effect van alternatieve straffen voor jongeren. Hij kreeg de indruk dat jongeren veel gemakkelijker hun straf volbrengen als zij binnen de hun toegewezen organisatie (gemeentedienst, verzorgingstehuis, ... ) taken moesten uitvoeren die tot het vaste takenpakket van de organisatie behoorden, dan als zij klusjes moesten doen die anders toch maar zouden blijven liggen. Om hierover duidelijkheid te krijgen, verdeelde hij de organisaties op zuiver toevallige wijze in twee groepen: één groep kreeg de instructie om de jongeren alleen taken te laten uitvoeren die tot het vaste takenpakket van de organisatie behoren, en de andere groep om de jongeren alleen klusjes te laten doen die anders toch zouden blijven liggen. Afhankelijk van de voorgeschiedenis van de jonge delinquent besliste de rechter of hij/zij aan het ene of het andere type organisatie werd toegewezen (met een taak uit het vaste takenpakket of een klusje). Het aantal jongeren dat de straf probleemloos volbracht was groter in de groep met een taak uit het vaste takenpakket dan in de groep met

een klusje. Mag men hieruit besluiten dat het type taak (vast of klusje) een effect heeft op het voorkomen van problemen bij het volbrengen van de straf? Op welke manier wordt er wel rekening gehouden met factoren die dit besluit in de weg kunnen staan en op welke manier wordt er geen rekening mee gehouden?

*Opgave 11*
Een onderzoekster wil het effect evalueren van een nieuwe therapie voor dyslexie. Daartoe vergelijkt zij deze nieuwe therapie (aangeduid met A) met de oude therapie (aangeduid met B). Zij krijgt de medewerking van een school waar in totaal 60 dyslectische leerlingen verblijven.

1. Stel dat de onderzoekster de medewerking krijgt van een therapeut die beide therapievormen beheerst. Op welke manier kan men gebruik maken van zuiver toevallige toewijzing om storende patiëntvariabelen onder controle te houden? Leg tevens uit hoe deze storende variabelen wel een rol kunnen spelen als men geen gebruik maakt van zuiver toevallige toewijzing.
2. Stel nu dat de onderzoekster de medewerking krijgt van in totaal 20 therapeuten. Het zijn allen afgestudeerden van een gespecialiseerde opleiding in de behandeling van leesproblemen. In deze opleiding kregen zij een grondige training in ofwel therapie A, ofwel therapie B. De keuze voor de therapievorm maakten zij zelf. Is het nu mogelijk dat een storende situatievariabele van invloed is op het resultaat van de vergelijking van therapie A met therapie B? Indien ja, leg uit hoe dit kan.
3. Stel nu dat de 20 therapeuten een grondige training kregen in zowel therapie A als therapie B. Op welke manier kan men gebruik maken van zuiver toevallige toewijzing om *zowel* storende patiëntvariabelen als storende therapeutvariabelen onder controle te houden? Welke storende therapeutvariabelen kunnen wel een rol spelen als men geen gebruik maakt van zuiver toevallige toewijzing?

*Opgave 12*
Wat is het belangrijkste probleem bij de interpretatie van de resultaten van een binnen-subjectenexperiment? Leg dit uit aan de hand van een voorbeeld.

*Opgave 13*
Geef van de onderstaande stellingen aan of ze waar of onwaar zijn. Geef een motivatie voor elk van de onware stellingen.

1. Door proefpersonen op zuiver toevallige wijze toe te wijzen aan de condities, worden alle storende variabelen onder controle gehouden.
2. Contrabalancering zorgt ervoor dat de resultaten van een binnen-subjectenexperiment niet beïnvloed worden door differentiële transfer.
3. Men kan een uitspraak doen over het effect van een sociale vaardigheidscursus door bij de deelnemers aan deze cursus een voor- en een nameting te verrichten en deze twee metingen met elkaar te vergelijken.

*Opgave 14*
Een onderwijspsychologe wil een vergelijking maken van twee didactische methoden voor het oplossen van rekenvraagstukken. Zij wil dit doen met behulp van een binnen-subjectenexperiment.

1. Leg uit hoe differentiële transfer de resultaten van dit experiment kan beïnvloeden.
2. Kan contrabalancering gebruikt worden om te voorkomen dat differentiële transfer de resultaten beïnvloedt? Motiveer je antwoord.

## 1.3 Standaardisering

We nemen nu als uitgangspunt de *wijze waarop* de data verzameld worden. Het belangrijkste kenmerk van dataverzamelingsmethoden voor wat betreft dit aspect, is de mate waarin zij *gestandaardiseerd* zijn. Er zijn twee aspecten van de dataverzameling die men kan standaardiseren: (1) de registratie zelf en (2) de situatie(s) waarin registreerd wordt. Onder standaardisering van de registratie zelf wordt verstaan dat op voorhand bepaald wordt naar welke gedragsaspecten men gaat kijken/vragen en op welke manier men dat gaat doen. Onder standaardisering van de situatie(s) wordt verstaan dat men het gedrag observeert in op voorhand welomschreven situaties. Bijvoorbeeld, indien een psychotherapeut zijn opvatting over depressie baseert op de gesprekken met zijn cliënten, dan is dat een weinig gestandaardiseerde dataverzameling. Daarentegen, indien een neuropsycholoog zijn opvatting over het effect van valium baseert op de gemiddelde aandachttestscores van een groep valiumgebruikers en een placebogroep, dan is dat een sterk gestandaardiseerde dataverzameling.

Het voordeel van een gestandaardiseerde dataverzameling is dat het onderzoek op precies dezelfde manier herhaald kan worden door een andere onderzoeker. Als dezelfde waarneming gedaan kan worden door een andere onderzoeker, dan noemen we deze waarneming *objectief*. Bij een gestandaardiseerde dataverzameling moet men op voorhand bepalen naar welk gedragsaspect men gaat kijken.

Door deze gerichte waarneming loopt men het risico dat belangrijke fenomenen niet opgemerkt worden. Het is altijd belangrijk om goed je ogen open te houden, ook als je je data op een gestandaardiseerde manier verzamelt. Dit kunnen we toelichten aan de hand van een hypothetisch onderzoek naar het effect van een nieuwe therapie voor dyslexie. In dit onderzoek werd één groep dyslectici behandeld met de oude therapie en een andere met de nieuwe. Door op een gestandaardiseerde manier naar het leesgedrag te kijken (bijvoorbeeld, met behulp van een leestest) werd op een objectieve manier vastgesteld dat met de nieuwe therapie meer woorden foutloos gelezen werden dan met de oude. De onderzoeker beperkte zich echter niet tot het berekenen van de score op de leestest, maar observeerde de kinderen ook tijdens de testafname. Dit gebeurde op een niet-gestandaardiseerde manier omdat hij niet op voorhand kon zeggen naar welk gedragsaspect hij zou kijken. Het viel hem op dat de kinderen die de nieuwe therapie volgden bij het lezen veel vaker de woorden opsplitsten in lettergrepen. Deze waarneming kan de aanzet zijn voor een meer precieze omschrijving van de werkzame factoren in deze nieuwe therapie.

## 1.4 Observatiemethoden

We bespreken nu de verschillende dataverzamelingsmethoden met als uitgangspunt datgene wát waargenomen wordt. We maken een onderscheid tussen *observeren*, hetgeen in deze paragraaf behandeld wordt, en *het stellen van vragen*, hetgeen in de volgende paragraaf behandeld wordt. Wat betreft het observeren, wordt er meestal een onderscheid gemaakt tussen het observeren van het gedrag zelf en het observeren van sporen en producten van gedrag. Bij dit laatste moet men denken aan krantenartikelen, kunstwerken, tv-programma's, archieven, en dergelijke. Deze categorie dataverzamelingsmethoden

wordt in dit boek niet besproken. Wat betreft de eerste categorie, het observeren van het gedrag zelf, maken we een onderscheid tussen drie dataverzamelingsmethoden op basis van de mate van standaardisering: (1) vrije (niet-gestandaardiseerde) observatie, (2) participerende observatie, en (3) gestandaardiseerde observatie. De gestandaardiseerde observatie wordt dan nog verder onderverdeeld in gestructureerde observatie, het veldexperiment en het laboratoriumexperiment.

### 1.4.1 Vrije (niet-gestandaardiseerde) Observatie

Een onderzoeker die deze methode gebruikt stelt zich op als een passieve waarnemer van wat er gebeurt. Hij wijst de proefpersonen niet toe aan condities en hij richt de omgeving niet speciaal in voor het doel van het onderzoek. Er is dus geen controle van storende persoons- en situatievariabelen. De situatie waarin het gedrag geobserveerd wordt is dus niet gestandaardiseerd. Men spreekt dan ook vaak van een *natuurlijke omgeving*. De gedrags*registratie* kan zowel gestandaardiseerd zijn als niet-gestandaardiseerd.

Het doel van vrije observatie is het beschrijven van het gedrag in de natuurlijke omgeving en het onderzoeken van relaties tussen variabelen die zich voordoen. Een voorbeeld van vrije observatie is het onderzoek van Hartup (1974) naar de frequentie en het type van agressie bij kleuters. Hartup maakte een onderscheid tussen vijandige agressie (gericht tegen een persoon) en instrumentele agressie (gericht op het herwinnen van een voorwerp, territorium of privilege). Het bleek dat jongens agressiever waren dan meisjes, maar dat er geen verschil was tussen jongens en meisjes wat betreft het type agressie: de verhouding tussen het aantal vijandig agressieve en het aantal instrumenteel agressieve uitingen was hetzelfde bij jongens en bij meisjes.

De natuurlijke omgeving wordt vaak geplaatst tegenover het psychologisch laboratorium. Men hoort wel eens de kritiek dat bevindingen op basis van laboratoriumonderzoek niet te veralgemenen zijn naar de natuurlijke omgeving. Gedragsobservatie in de natuurlijke omgeving wordt dan vaak voorgesteld om de veralgemeenbaarheid van deze bevindingen na te gaan. De observatie wordt dan wel wat minder vrij, omdat men nu heeft vastgelegd naar welke gedragsaspecten men gaat kijken. Met andere woorden, de gedrags*registratie* is nu gestandaardiseerd. De *situatie waarin* men het gedrag observeert blijft echter ongestandaardiseerd.

Vrije observatie is soms nuttig in de beginfase van een onderzoek. Soms weet men nog niet hoe de onderzoeksvraag precies geformuleerd moet worden of welke data relevant zijn voor een bepaalde onderzoeksvraag. De vrije observatie kan nuttig zijn bij het aanscherpen van de onderzoeksvraag en het bepalen van de relevante data.

Psychologen zijn niet de enige onderzoekers die het gedrag onderzoeken in de natuurlijke omgeving. Observatie in de natuurlijke omgeving is de belangrijkste dataverzamelingsmethode van de ethologen. Ethologen onderzoeken het gedrag van organismen in relatie met de natuurlijke leefomgeving van deze organismen. Zij bekijken het gedrag vooral vanuit het perspectief van de natuurlijke selectie. Meer bepaald, zij vragen zich af hoe de natuurlijke selectie ervoor gezorgd kan hebben dat een bepaald gedragspatroon zich ontwikkeld heeft. Ethologen proberen het gedrag van hogere diersoorten te verklaren op basis van aangeboren gedragspatronen van diersoorten lager op de evolutionaire ladder. Zo proberen zij ook menselijk gedrag te verklaren. Bijvoorbeeld, Barash (1977) deed een onderzoek naar de manier waarop mannen en vrouwen een gevaarlijke weg over-

steken (standaardisering van de registratie maar niet van de situatie). Als een volwassen man en vrouw samen waren, dan waren de mannen veel oplettender bij het oversteken dan de vrouwen (blijkend uit het feit dat zij vaker naar links en naar rechts kijken). Dit was zowel het geval als er kinderen bij waren als wanneer dit niet zo was. De verklaring van de etholoog was gebaseerd op de evolutietheorie: volwassen mannetjesapen hebben in de vrije natuur de functie van waarschuwer bij dreigend gevaar.

### 1.4.2 Participerende Observatie

Observatie door iemand die een actieve rol speelt in de situatie waarin hij observeert, wordt *participerende observatie* genoemd. In *onvermomde* participerende observatie, weten degenen die geobserveerd worden dat de observator aanwezig is om hun gedrag te observeren. Deze methode wordt vaak gebruikt door antropologen; met als doel het gedrag en de cultuur van een groep te leren begrijpen, delen zij het dagelijks leven van deze groep. We spreken van *vermomde* participerende observatie als degenen die geobserveerd worden niet weten dat de observator aanwezig is om hun gedrag te observeren. De reden om te kiezen voor vermomde participerende observatie is de verwachting dat degenen die geobserveerd worden zich anders zullen gedragen als zij weten dat zij geobserveerd worden.

Rosenhan (1973) gebruikte vermomde participerende observatie om te onderzoeken op basis waarvan men een diagnose stelt in een psychiatrische instelling. Acht personen (een aantal psychologen, een kinderarts en een huisvrouw) vervalsten hun naam en beroep en meldden zich aan bij twaalf verschillende psychiatrische instellingen. Elk van hen klaagde over hetzelfde symptoom: dat hij of zij stemmen hoorde. Rosenhan beschreef dit symptoom en de verantwoording voor de keuze ervan als volgt:

> "Op de vraag wat de stemmen zeiden, antwoordde hij (de pseudopatiënt) dat ze vaak onduidelijk waren, maar dat voor zover hij het wist ze "leeg", "hol" en "dof" zeiden. De stemmen waren onbekend en van hetzelfde geslacht als het geslacht van de pseudopatiënt. De keuze voor deze symptomen was ingegeven door hun schijnbare overeenkomst met existentiële symptomen. Dergelijke symptomen zouden voortvloeien uit pijnlijke zorgen over de ervaren zinloosheid van iemands leven. Het is alsof de symptomen zeggen "Mijn leven is leeg en hol.". De keuze voor deze symptomen was ook bepaald door de afwezigheid van ook maar één gerapporteerd geval van existentiële psychose in de literatuur. Behalve het melden van de symptomen en het vervalsen van naam, beroep en opleiding, werden er geen andere wijzigingen van persoon, geschiedenis of omstandigheden aangebracht." (Rosenhan, 1973, p. 251)

Onmiddellijk nadat de onderzoekers op de psychiatrische afdeling opgenomen werden, stopten zij met klagen over enig symptoom en gedroegen zich normaal. Tijdens hun opname observeerden zij de interacties tussen patiënten en personeel. Zij waren met name geïnteresseerd in de interacties tussen het personeel en henzelf. De onderzoekers-pseudopatiënten maakten aantekeningen van hun observaties en deden geen moeite om dit te verbergen voor patiënten en personeel. Zoals later bleek, werd dit gedrag (het maken van aantekeningen) door het personeel geïnterpreteerd als symptoom van hun

psychopathologie (i.c., wantrouwen, paranoia). Sommige patiënten, daarentegen, hadden een beter oog voor de ware toedracht van het gedrag van de onderzoekers: zij suggereerden dat de onderzoekers er waren om de instelling te controleren. De opnameduur varieerde van 7 tot 52 dagen. De onderzoekers werden nooit genezen verklaard, en bij ontslag werd gezegd dat hun schizofrenie *in remissie* was (i.e., een afzwakking van de ziekteverschijnselen). Eenmaal dat de pseudopatiënten het label "schizofreen" hadden gekregen, bleven ze dat label houden, wat ook de aard was van hun daaropvolgend gedrag.

Met participerende observatie krijgt men toegang tot situaties die meestal niet open zijn voor een meer gestandaardiseerde observatie. Bovendien is de participerende observator vaak in een positie om hetzelfde mee te maken als de personen die hij onderzoekt. Zo konden de pseudopatiënten in het onderzoek van Rosenhan (1973) ervaren hoe het was om als schizofreen gediagnosticeerd te zijn en niet te weten hoe lang het zou duren voor ze weer naar huis konden gaan. Een nog duidelijker voorbeeld hiervan vinden we in het boek "Black Like Me" van John Howard Griffin (1960) die met behulp van participerende observatie onderzocht hoe het is om als neger slachtoffer te zijn van rassendiscriminatie. De blanke Griffin verdonkerde zijn huid door het gebruik van medicatie en kleurstof om als zwarte aanzien te worden. Gedurende meer dan een maand reisde hij (lopend, liftend en met de bus) door de zuidelijke staten van de Verenigde Staten die, zeker in de beginjaren zestig, bekend stonden om hun rassendiscriminatie.

### 1.4.3 Gestandaardiseerde Observatie

Gestandaardiseerde observatie verschilt van vrije observatie doordat zowel de registratie als de situatie waarin men observeert meer gestandaardiseerd zijn. Het is meestal zo dat, naarmate de situatie meer gestandaardiseerd is, dit ook geldt voor de registratie. Er zijn meerdere vormen van gestandaardiseerde observatie. Ze verschillen van elkaar wat betreft de mate van standaardisering. We maken hier een onderscheid tussen *gestructureerde observatie*, *veldexperimenten* en *laboratoriumexperimenten*.

#### Gestructureerde Observatie

Men kan gestructureerde observatie beschouwen als bijna-gestandaardiseerde observatie. Soms richt een observator een situatie zo in dat een bepaald gedrag de kans krijgt om zich te vertonen, maar dat er geen sprake is van manipulatie. En soms is er wel sprake van manipulatie, maar het doel van deze manipulatie is dan niet om twee condities met elkaar te vergelijken. Het doel van deze manipulatie is wel om een bepaald gedrag uit te lokken. Dit wordt het beste duidelijk aan de hand van enkele voorbeelden.

Klinisch psychologen maken regelmatig gebruik van gedragsobservaties in gestandaardiseerde situaties voor het stellen van een diagnose. Dit gebeurt bijvoorbeeld om de interacties tussen ouders en kinderen in kaart te brengen. Een hulpvragende ouder en zijn of haar kind worden uitgenodigd om samen een bepaalde activiteit te beginnen (iets bouwen, een spelletje spelen, ...). Tijdens deze activiteit worden zij geobserveerd, en meestal gebeurt dit vanachter een eenrichtingsspiegel. Men registreert zowel verschillende specifieke gedragingen als het meer algemene karakter van de ouder-kind interacties. Mede op basis van deze waarnemingen maakt de therapeut een behandelplan

waarmee hij de gedragsproblemen zal aanpakken.

In de ontwikkelingspsychologie is de gestructureerde observatie populair geworden door het werk van Jean Piaget. In veel van zijn onderzoeken gaf hij een kind eerst een probleem om op te lossen, en daarna gaf hij het verschillende variaties op dit probleem om de grenzen te toetsen van het begrip van het kind. Hierbij stelde Piaget ook vragen met de bedoeling om het redeneerproces van het kind te achterhalen. De bevindingen uit dit type onderzoek vormde de basis voor Piagets theorie over de cognitieve ontwikkeling. Meer bepaald, deze bevindingen hebben geleid tot de theorie dat de cognitieve ontwikkeling verloopt in een vast aantal stadia (met daartussen cognitieve sprongen) die in een vaste volgorde doorlopen worden. Het onderzoek van Piaget was niet vrij van kritiek: hij volgde niet altijd dezelfde procedure van de ene observatie naar de andere (gebrekkige standaardisering) en zijn observaties waren vaak afkomstig van slechts enkele kinderen (vaak Piagets eigen kinderen!).

Ethologen maken gebruik van gestructureerde observatie om de grenzen te onderzoeken van de situatie waaronder een bepaald gedrag gesteld wordt. Het onderzoek van Konrad Lorenz naar het fenomeen *imprinting* is van dit type. Imprinting is het fenomeen dat er bij bepaalde diersoorten een kritische periode bestaat waarin zij zich hechten aan een voorwerp uit hun omgeving. Bijvoorbeeld, zodra een jonge gans uit het ei kruipt zal het zich hechten aan het eerste bewegend voorwerp in zijn omgeving. Onder normale omstandigheden, hecht de jonge vogel zich aan zijn ouder. Deze aangeboren hechtingsrespons is dan ook functioneel voor het overleven van het jonge dier. Om imprinting beter te begrijpen hebben ethologen kunstouders gemaakt en hiermee de hechtingsrespons van het jonge dier onderzocht. Het gedrag van de jonge dieren werd onder andere geobserveerd bij een bewegende kunstouder, eentje die geluid maakte en eentje waarbij de hechtingsrespons bestraft werd.

### Veldexperimenten

Het veldexperiment is een experiment dat in een natuurlijke omgeving wordt uitgevoerd. Net zoals ieder ander experiment, wordt ook het veldexperiment gekenmerkt door manipulatie. In een veldexperiment bestaat deze manipulatie uit het toewijzen – meestal op zuiver toevallige wijze – van de proefpersonen aan de condities. Dus, in een veldexperiment hebben we te maken met een experimentele variabele en niet met een natuurlijke, want de proefpersoon kiest niet zelf de conditie waar hij in terechtkomt. (Het natuurlijke in een veldexperiment is de omgeving waarin men het gedrag observeert, en niet de wijze waarop de proefpersoon in een conditie terechtkomt.) In een veldonderzoek is de gedrags*registratie* gestandaardiseerd. Immers, de bedoeling van een veldexperiment, net zoals van ieder ander experiment, is het onderzoeken van het effect van een onafhankelijke op een welbepaalde afhankelijke variabele. Het is dus op voorhand duidelijk waar de onderzoeker naar gaat kijken. De situaties (condities) waarin men observeert, zijn tot op zekere hoogte gestandaardiseerd. Omdat men in een natuurlijke omgeving observeert, is het soms niet mogelijk om deze situaties precies in te richten zoals men dat wil (met het doel storende situatievariabelen onder controle te houden). Dit is het verschil met het laboratoriumexperiment.

Het veldexperiment wordt vaak gebruikt in de sociale psychologie. Meestal zijn de proefpersonen er zich niet van bewust dat zij deelnemen aan een experiment. Vaak

wordt er ook gebruik gemaakt van zogenaamde *pseudo-proefpersonen* (in het Engels, *confederates*). Een pseudo-proefpersoon is een medewerker van de onderzoeker die zich op een bepaalde manier gedraagt om zo een bepaalde experimentele conditie tot stand te brengen. In onderzoek van Darley en Latané (1968) werden pseudo-proefpersonen opgedragen om iemand te beroven, en werd geregistreerd hoe lang het duurde voordat een omstander reageerde. En in onderzoek van Milgram, Liberty, Toledo en Wackenhut (1986) werden pseudo-proefpersonen opgedragen om voor te kruipen in een wachtrij, en werd geregistreerd hoe de anderen reageerden.

Als voorbeeld, nemen we een onderzoek van Crusco en Wetzel (1984) naar het effect van aanraking op de omvang van de fooi in restaurants. De diensters (pseudo-proefpersonen) raakten ofwel de hand, ofwel de schouder van de klant. De onderzoekers verwachtten dat een aanraking van de hand een positief gevoel naar de dienster zou opwekken, maar dat een aanraking van de schouder een negatief gevoel zou opwekken. Deze laatste verwachting werd uitgelegd door de veronderstelling dat een aanraking van de schouder als teken van dominantie gezien kon worden en dat dit, zeker door mannen, als negatief ervaren zou worden. De proefpersonen waren 114 gasten van twee restaurants. Zij werden op zuiver toevallige wijze toegewezen aan één van de drie condities: (1) handaanraking, waarbij de dienster twee keer de hand van de klant aanraakte gedurende een halve seconde bij het teruggeven van wisselgeld; (2) schouderaanraking, waarbij de dienster haar hand gedurende één à anderhalve seconde op de schouder van de klant legde bij het teruggeven van wisselgeld; en (3) geen aanraking. De belangrijkste afhankelijke variabele was de omvang van de fooi. Het bleek dat (1) mannen een grotere fooi gaven dan vrouwen, (2) zowel mannen als vrouwen een grotere fooi gaven als zij aangeraakt waren, en (3) dat het er niet toe deed of men aan de hand of aan de schouder aangeraakt werd.

### Laboratoriumexperimenten

Het verschil tussen een veld- en een laboratoriumexperiment is de omgeving waarin het experiment uitgevoerd wordt: een veldexperiment wordt uitgevoerd in een natuurlijke omgeving, een laboratoriumexperiment wordt uitgevoerd in een laboratorium. Dit laatste moet men niet te letterlijk nemen: we gebruiken de term laboratorium voor een omgeving die volledig te manipuleren is. Het voordeel van een laboratoriumexperiment boven een veldexperiment is immers dat de condities waarin men observeert volledig te manipuleren zijn. Hierdoor kan men in een laboratoriumexperiment de storende situatievariabelen veel makkelijker onder controle houden dan in een veldexperiment. Immers, men heeft de aard van de condities in eigen hand, zodat men ervoor kan zorgen dat deze condities alleen van elkaar verschillen wat betreft het aspect waarin men geïnteresseerd is. Met andere woorden, in een laboratoriumexperiment zijn de condities in sterke mate gestandaardiseerd.

We kunnen dit illustreren aan de hand van het veldexperiment van Crusco en Wetzel (1984) naar het effect van aanraking op de omvang van de fooi in restaurants. Een kritiek die men op dit experiment kan geven, is dat de condities met en zonder aanraking op meerdere punten van elkaar verschillen dan het al-dan-niet aangeraakt worden. Concreet, het is bijvoorbeeld mogelijk dat, door de aanraking, er vaker oogcontact gemaakt wordt in de conditie met dan in de conditie zonder aanraking. Het zou dan kunnen zijn

dat het verschil in fooi niet toe te schrijven is aan het al-dan-niet aangeraakt worden, maar aan de hoeveelheid oogcontact. Met andere woorden, de hoeveelheid oogcontact kan een storende situatievariabele zijn.

Indien men in een laboratoriumexperiment met een gelijkaardig probleem geconfronteerd zou worden als in het onderzoek van Crusco en Wetzel (1984), dan is het in principe mogelijk om de storende situatievariabele(n) onder controle te houden. Dit kunnen we illustreren aan de hand van een onderzoek naar *sociale luiheid* (in het Engels, *social loafing*). Sociale luiheid is de neiging van personen om zich minder hard in te spannen wanneer zij zich samen met anderen ergens voor inzetten dan wanneer zij alleen verantwoordelijk zijn. Men zou dit kunnen onderzoeken door personen in te delen in groepjes van, bijvoorbeeld, een, twee en zes personen, en de groep (mogelijk bestaande uit één persoon) een bepaalde taak te laten uitvoeren. Deze taak zou erin kunnen bestaan om de plaatselijke volleybalploeg aan te moedigen. Een mogelijke afhankelijke variabele is de hoeveelheid geluid die geproduceerd wordt door een groep. Als een groep van twee of zes personen minder geluid produceert dan te verwachten is op basis van de hoeveelheid geluid die iemand alleen produceert, dan zou men kunnen concluderen dat er sprake is van sociale luiheid. De resultaten van een dergelijk veldexperiment zijn echter op meerdere manieren te verklaren. De condities met een of meerdere personen per groep verschillen niet alleen van elkaar wat betreft het besef dat men er alleen of samen voor staat, maar ook wat betreft de fysieke aanwezigheid van anderen. Een eenvoudige alternatieve verklaring is dat de proefpersonen minder luid schreeuwen in groepen omdat ze geen zuivere feedback krijgen van hun eigen bijdrage: van het geschreeuw dat ze horen (afkomstig van zichzelf en eventueel van de anderen) schrijven ze een onevenredig groot deel aan zichzelf toe.

Latané, Williams, en Harkin (1979) deden een laboratoriumexperiment naar de determinanten van sociale luiheid waarbij zij de mogelijke storende invloed van de situatievariabele FYSIEKE AANWEZIGHEID VAN ANDEREN onder controle hielden. Alle proefpersonen maakten deel uit van een groep van zes personen. De proefpersonen namen plaats in een halve cirkel, werden geblinddoekt en kregen een koptelefoon op. De taak bestond erin om zo luid mogelijk te schreeuwen. Tijdens de schreeuw kreeg elke proefpersoon hetzelfde, vooraf opgenomen, geschreeuw te horen via zijn koptelefoon. De proefpersoon hoorde dus noch zichzelf, noch de anderen. De onafhankelijke variabele was het aantal personen dat tegelijkertijd moest schreeuwen. Dit aantal was een, twee of zes. Uit de resultaten bleek duidelijk dat een groep van twee of zes personen minder geluid produceert dan te verwachten is op basis van de hoeveelheid geluid die iemand alleen produceert.

Uit een vervolgonderzoek van Williams, Harkin en Latané (1981) werd duidelijk wat de cruciale variabele is die verantwoordelijk is voor het fenomeen. In plaats van het geluid op te nemen met een groepsmicrofoon, zoals in het eerdere onderzoek, kreeg elke proefpersoon nu zijn eigen microfoon, en werd hem gezegd dat hiermee het volume van zijn eigen schreeuw vastgesteld kon worden (niet vermengd met het geschreeuw van de anderen). In deze conditie verdween het effect van het aantal personen dat samen schreeuwt: als het volume van zijn eigen schreeuw gemeten wordt, dan doet het er niet toe of hij alleen of samen met anderen schreeuwt. Dit toont ons het grote voordeel van het laboratoriumexperiment: het opsporen van de cruciale situatievariabele

die verantwoordelijk is voor een bepaald fenomeen.

### 1.4.4 Registratie

**Het Observatieverslag**

Met een observatieverslag poogt men zo getrouw mogelijk het gedrag weer te geven zoals het heeft plaatsgevonden. Men moet de term "observatieverslag" ruim interpreteren: het is niet noodzakelijk een schriftelijk verslag, maar het kan ook een band- of een video-opname zijn. Indien het observatieverslag schriftelijk is, dan is het van belang dat de observator zo weinig mogelijk interpreteert en zo kort mogelijk bij het observeerbare gedrag blijft. Het bestuderen, ordenen, classificeren en interpreteren van het observatieverslag gebeurt nadat alle observaties gedaan zijn. Dit is een belangrijk verschil met andere manieren van registreren, zoals de beoordelingsschaal (in het Engels, *rating scale*) en de checklist, waarbij de beoordeling of classificatie van het gedrag gebeurt op het moment van de observatie.

We vinden een voorbeeld van een observatieverslag in een boek van Piaget (1965) over de ontwikkeling van het begrip "hoeveelheid" bij het kind. Piaget doet verslag van de wijze waarop een vierjarig meisje een probleem probeert op te lossen dat hij haar heeft voorgelegd. Voorafgaand aan het observatieverslag geeft hij de volgende beschrijving van de proef en van de conclusie die hij getrokken heeft op basis van zijn observaties:

> "Het kind krijgt twee glazen $A$ en $L$ van dezelfde hoogte, waarvan $A$ breed is en $L$ smal. $A$ wordt gevuld tot een bepaalde hoogte (een vierde of een vijfde) en het kind wordt gevraagd om dezelfde hoeveelheid in $L$ te schenken. De dimensies van de glazen zijn zo dat, met dezelfde hoeveelheid vloeistof, het niveau van die vloeistof in $L$ vier keer zo hoog is als die in $A$. Ondanks het opvallende verschil in de dimensies van de glazen, is het kind in dit stadium van zijn ontwikkeling niet in staat om in te zien dat de kleinere diameter van $L$ zorgt voor een hoger vloeistofniveau. De kinderen die zich in dit stadium bevinden zijn tevreden dat er "dezelfde hoeveelheid te drinken" is in $L$ wanneer zij dit glas tot hetzelfde niveau gevuld hebben als $A$." (Piaget, 1941/1952, p. 12)

Piaget illustreert dan zijn conclusie met het volgende observatieverslag:

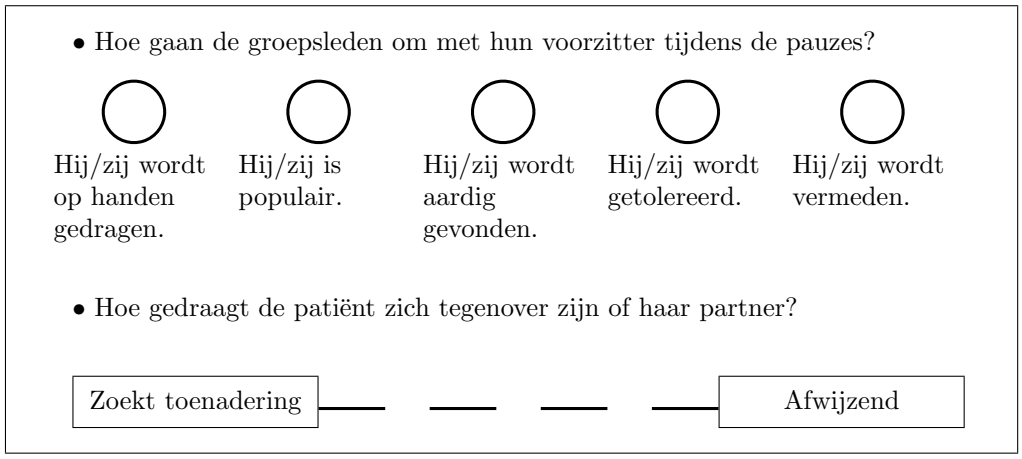
> "Kijk, jouw mamma heeft een glas limonade voor zichzelf ingeschonken ($A$) en zij geeft jou dit glas ($L$). Wij willen dat jij in jouw glas net zoveel limonade schenkt als jouw mamma heeft in haar glas.— (Zij schenkt nogal snel in en gaat voorbij aan het niveau in $A$ dat zij probeerde te bereiken.)—Hebben jullie nu evenveel?—*Nee.*— Wie heeft er meer?—*Ik.*—Laat me zien tot waar je moet inschenken zodat jullie allebei evenveel hebben.—(Zij schenkt in tot aan hetzelfde niveau.)—Hebben jij en mamma nu evenveel te drinken?—*Ja.*—Ben je daar zeker van?—*Ja.*" (Piaget, 1941/1952, p. 12)

Dit observatieverslag bevat geen interpretaties van de onderzoeker en het lijkt erop dat Piaget de gedragingen beschrijft in dezelfde volgorde als waarin hij ze geobserveerd heeft.

Observatieverslagen zijn niet altijd zo gedetaïlleerd als die van Piaget. Vaak zijn observatieverslagen een neerslag van de gebeurtenissen, omstandigheden, personen en gedragingen die de onderzoeker belangrijk vindt. Dergelijke verslagen worden vaak *veldnota's* genoemd. Veldnota's worden gebruikt door, onder andere, journalisten, antropologen, sociale werkers en ethologen. Veldnota's zijn niet altijd een getrouwe weergave van de gebeurtenissen omdat (1) alleen gebeurtenissen en gedragingen geregistreerd worden die de onderzoeker belangrijk vindt en (2) deze gebeurtenissen en gedragingen meestal geïnterpreteerd worden in termen van de expertise en de opvattingen van de onderzoeker.

### De Beoordelingsschaal

Soms is het op voorhand duidelijk welk(e) gedragsaspect(en) men gaat observeren. In dit geval kan men gebruik maken van zogenaamde beoordelingsschalen. Op een beoordelingsschaal geeft men aan wat de positie is van een persoon of een situatie op een bepaalde psychologische dimensie. Meestal wordt de schaal aangeduid door twee omschrijvingen aan de uiteinden (bijvoorbeeld, blij versus triest). De schaal is meestal ingedeeld in vijf of zeven categorieën, maar het kan ook een doorlopende lijn zijn. De observator moet een teken zetten bij dat vakje of op die positie die hij het meest kenmerkend vindt voor de beoordeelde persoon of situatie. In Figuur 1.3 worden twee voorbeelden gegeven van een beoordelingsschaal. De categorieën van de eerste beoordelingsschaal hebben elk hun eigen omschrijving maar de interpretatie van de verschillende posities op de lijn van de tweede beoordelingsschaal wordt aan de beoordelaar overgelaten.




Figuur 1.3: Twee voorbeelden van een beoordelingsschaal.

### Codering

Observatieverslagen (schriftelijk, video-opnamen, bandopname) worden meestal gecodeerd. Coderen is het beschrijven van gedragingen door ze onder te brengen in een aantal categorieën. Bales (1950, 1952) ontwierp zo'n categorieënsysteem voor het beschrijven van

de interactie in kleine taakgerichte groepen (discussiegroepen, vergaderingen, therapiegroepen,...). Bales maakt een onderscheid tussen 12 gedragscategorieën:

- Sociaal-emotioneel positief gedrag
  1. tekenen van solidariteit, hulp, beloning
  2. tekenen van ontspanning, grapjes, gelach
  3. mee eens zijn, tekenen van acceptatie
- Taakgericht gedrag (antwoorden)
  4. doen van suggesties, geven van richtlijnen
  5. geven van opinies, waarderingen, analyses
  6. geven van informatie
- Taakgericht gedrag (vragen)
  7. vragen om informatie
  8. vragen om opinies, waarderingen
  9. vragen om suggesties, richtlijnen
- Sociaal-emotioneel negatief gedrag
  10. niet mee eens zijn, afwijzing
  11. tekenen van spanning
  12. tekenen van antagonisme, vijandigheid

Naast deze 12 categorieën is er ook nog *voor zich uit praten* dat aangeduid wordt met X. Ieder lid van de groep wordt aangeduid met een nummer of een letter en de groep als geheel wordt aangeduid met 0. Iedere verbale uiting van een groepslid wordt opgesplitst in een of meerdere *acts* en iedere act wordt gescoord wat betreft (1) de gedragscategorie waarin hij valt en (2) zijn richting (i.c., tegen wie wordt er wat gezegd). Bijvoorbeeld, als groepslid 3 tegen groepslid 5 zegt "Wat vind jij er nu van dat jouw collega's op het werk zo vaak hun afspraken niet nakomen? Ik heb de indruk dat wij er ons meer zorgen over maken dan jij." en tegen de groep "Vinden jullie dat ook niet?", dan worden deze drie acts gescoord als 3–5 (8), 3–5 (5) en 3–0 (8).

Na codering van de interacties met behulp van het systeem van Bales, kan men nagaan of de aard van de groepsinteractie (1) evolueert in de tijd of (2) afhangt van bepaalde condities. Bijvoorbeeld, het zou kunnen blijken dat beginnende therapiegroepen vooral taakgericht gedrag vertonen en dat het sociaal-emotionele gedrag in frequentie toeneemt naarmate de groep vaker samen geweest is. Het zou ook kunnen blijken dat, bij vergaderingen, de stijl van leidinggeven bepalend is voor de mate waarin er sociaal-emotioneel positief gedrag vertoond wordt.

### 1.4.5 Enkele Problemen bij het Gebruik van Observatiemethoden

#### Overeenstemming tussen de Observatoren

Als twee observatoren naar hetzelfde gedrag kijken en toch iets anders registreren, dan zeggen we dat de overeenstemming tussen de observatoren laag is. Een lage overeenstemming tussen de observatoren impliceert dat de conclusies die men uit het onderzoek trekt sterk afhangen van wie de observaties gedaan heeft. Omdat we willen dat wetenschappelijke kennis niet afhangt van de persoon die de kennis verzamelt, is een lage overeenstemming tussen de observatoren niet aanvaardbaar.

Een lage overeenstemming tussen de observatoren die te wijten is aan de observatoren zelf, is niet zo erg. Bijvoorbeeld, het kan zijn dat de observatoren weinig ervaring hebben, zodat ze nog niet goed weten wat de relevante gedragsaspecten zijn, of dat ze weinig gemotiveerd zijn. In deze gevallen kan men de overeenstemming tussen de observatoren vergroten door hen te trainen en te motiveren. Een voor de hand liggende vorm van training, die ook effectief blijkt te zijn, is het geven van feedback over de discrepanties tussen de eigen registraties en die van de andere observatoren.

Het is erger als een lage overeenstemming tussen de observatoren te wijten is aan het feit dat de omschrijving van een gedrag onvoldoende precies is. Bijvoorbeeld, stel dat in het onderzoek van Hartup (1974) aan de observatoren gevraagd werd om het agressieve gedrag van kinderen te registreren zonder dat zij een precieze omschrijving kregen van agressief gedrag. Het zou kunnen dat sommige observatoren alleen spreken van agressief gedrag als er sprake is van een fysieke aanval, terwijl andere observatoren ook beledigingen als een vorm van agressie beschouwen. En wellicht zijn ze het ook niet met elkaar eens als ze een speels duwtje moeten onderscheiden van een boze zet. Zonder een precieze omschrijving van wat men moet registreren, kan men geen redelijke overeenstemming tussen de observatoren verwachten. Behalve een precieze omschrijving, zorgen ook concrete voorbeelden voor een grotere overeenstemming tussen de observatoren.

Een grote overeenstemming tussen de observatoren betekent niet dat men ook datgene registreert wat men wil registreren. Men wil bijvoorbeeld niet dat de registraties mede bepaald worden door de verwachtingen van de observator (over de uitkomsten van het onderzoek, over hoe de mens in elkaar zit, ...). Op dit punt wordt in een van de volgende paragrafen dieper ingegaan.

Als men werkt met beoordelingsschalen of een coderingssysteem, dan is het mogelijk om de mate van overeenstemming tussen de observatoren in een getal weer te geven. In dit geval spreekt men meestal van *interbeoordelaarsovereenstemming* omdat er, zowel bij het gebruik van een beoordelingsschaal als van een coderingssysteem, sprake is van een beoordeling van het geobserveerde gedrag. Als men gebruik maakt van een beoordelingsschaal dan wordt de interbeoordelaarsovereenstemming uitgedrukt in een correlatiecoëfficiënt. Concreet, voor een paar van observatoren bekijkt men hoe elk van hen een reeks van gedragingen beoordeelt en berekent daar de correlatie tussen. Als men gebruik maakt van een coderingssysteem, dan wordt de interbeoordelaarsovereenstemming uitgedrukt als het percentage gedragingen dat door beide observatoren op dezelfde manier gecodeerd wordt (in dezelfde categorie gescoord). Er bestaat geen harde minimumcorrelatie (voor beoordelingsschalen) of een hard minimumpercentage

(voor een coderingssysteem) aan de hand waarvan men kan bepalen of de interbeoordelaarsovereenstemming voldoende is. Kijkt men naar de literatuur, dan blijkt dat de interbeoordelaarsovereenstemming meestal groter is dan 0.85 (voor de correlatie), respectievelijk, 85 (voor het percentage). Dit kan als richtlijn fungeren.

### Reactiviteit van de Proefpersoon

Onder *reactiviteit* verstaan we het feit dat proefpersonen zich soms anders gedragen wanneer ze geobserveerd worden dan wanneer ze niet geobserveerd worden. Anders gezegd, proefpersonen reageren soms op de aanwezigheid van een observator. Als er sprake is van reactiviteit, dan is het geobserveerde gedrag niet representatief voor het gedrag dat men zou observeren als de observator afwezig zou zijn. Dit is een probleem omdat men juist in dit laatste gedrag geïnteresseerd is.

We maken een onderscheid tussen twee vormen van reactiviteit: (1) reactiviteit in situaties waarin de proefpersonen zich bewust zijn van het feit dat zij geobserveerd worden en (2) reactiviteit in situaties waarin zij zich er niet van bewust zijn. Het meest bekende voorbeeld van reactiviteit van het eerste type is het zogenaamde *Hawthorne effect* (Roethlisberger, 1977). Deze naam verwijst naar een grootschalig onderzoek in het Hawthorne-complex van de Western Electric Company in de buurt van Chicago. In dit onderzoek was men geïnteresseerd in de relaties tussen arbeidsomstandigheden, productiviteit, tevredenheid en motivatie. Een welbepaalde relatie waarin men geïnteresseerd was, was die tussen de mate van verlichting en de productiviteit bij het controleren van kleine onderdelen en het assembleren van printplaten en spoelen. In een experiment werden de arbeiders verdeeld in twee groepen. In de ene groep werkten de arbeiders bij een constante lichtsterkte en in de andere groep werkten zij bij een lichtsterkte die in de loop van het experiment geleidelijk toenam. Het resultaat was dat de productiviteit in beíde groepen toenam, in ongeveer dezelfde mate. Andere experimenten van hetzelfde type gaven gelijkaardige vreemde resultaten. Het leek erop dat de arbeiders eerder reageerden op het feit dat de onderzoekers begaan waren met hun arbeidsomstandigheden dan op de fysische veranderingen in lichtsterkte.

Sociaal psychologen zijn zich zeer goed bewust van het feit dat gedrag beïnvloed wordt door de aanwezigheid van anderen. (Als men sociale psychologie letterlijk opvat, is dat zelfs de kern van hun onderzoeksgebied.) Uit onderzoek blijkt dat de aanwezigheid van een andere persoon de prestatie op eenvoudige taken verhoogt (sociale facilitatie) en de prestatie op complexe taken verlaagt (sociale inhibitie). Deze effecten blijken het duidelijkst wanneer de rol van de andere persoon erin bestaat om iemands prestatie te evalueren, zoals vaak het geval is als die andere persoon de onderzoeker is. Sociale facilitatie en sociale inhibitie worden meestal verklaard door de opwinding (in het Engels, *arousal*) die ontstaat door het bewustzijn van het feit dat men geëvalueerd wordt. Men veronderstelt hierbij dat opwinding ervoor zorgt dat goed geleerde gedragspatronen (nodig bij eenvoudige taken) meer beschikbaar komen ten koste van de minder goed geleerde gedragspatronen (nodig bij complexe taken).

Het tweede type reactiviteit (in situaties waarin de proefpersonen zich niet bewust zijn van het feit dat zij geobserveerd worden) kan zich voordoen bij vermomde participerende observatie. Daar speelt de observator immers een actieve rol in de situatie waarin hij observeert. Dit kunnen we illustreren aan de hand van een onderzoek van Fes-

tinger, Riecken en Schachter (1956) naar de wijze waarop een groep reageert op het feit dat een gezamenlijke verwachting wordt tegengesproken. Zij waren vooral geïnteresseerd in de eventuele veranderingen in groepscohesie. De onderzoekers infiltreerden in een groep die beweerde contacten te onderhouden met wezens van een andere planeet. De leider van de groep beweerde dat hij een boodschap had ontvangen van deze wezens waarin voorspeld werd dat er zich een grote overstroming zou voordoen op een bepaalde datum. Deze overstroming zou zich uitstrekken van de noordpoolcirkel tot aan de Golf van Mexico. Omdat de groep nogal afwijzend stond ten overstaan van "ongelovigen", moesten de onderzoekers bizarre verhalen verzinnen om toegelaten te worden. Deze tactiek werkte echter té goed want niet alleen werden de vermomde onderzoekers toegelaten tot de groep, maar hun verhalen werden gezien als tekens van de buitenaardse wezens. Eén van de onderzoekers werd zelfs gezien als een ruimtevaarder met een boodschap. Onbedoeld hadden de onderzoekers het geloof van de groep versterkt en op die manier hadden zij een invloed op het verdere verloop van de gebeurtenissen. Tussen haakjes, de grote overstroming vond nooit plaats, maar toch waren er een aantal groepsleden die deze disconfirmatie aangrepen om hun geloof te versterken. Ze begonnen zelfs nieuwe leden te werven met het argument dat hun geloof de voorspelde overstroming had voorkomen.

### Verwachtingen van de Proefpersoon

De verwachtingen van de proefpersoon ten aanzien van het onderzoek zijn een bijzondere vorm van reactiviteit. Deze verwachtingen worden uitgelokt door de wetenschap dat men aan een onderzoek gaat deelnemen. Er zijn twee soorten verwachtingen die van belang zijn: (1) verwachtingen ten aanzien van het gedrag dat de onderzoeker van hen verwacht (verlangt) en (2) verwachtingen ten aanzien van het effect van de behandeling die men ondergaat (de conditie waartoe men behoort). Orne (1962) heeft de aandacht gevestigd op het eerste soort verwachtingen. Hij spreekt in dit verband van de *vraagkenmerken* (in het Engels, *demand characteristics*) van de onderzoekssetting. Orne gaat ervan uit dat proefpersonen zich afvragen wat er van hen verwacht wordt. Ze kunnen zich hier een idee van vormen door aandachtig te zijn voor bepaalde signalen in de onderzoekssituatie. Eén van de belangrijkste elementen van de onderzoekssituatie is de onderzoeker zelf. Op basis van zijn gedrag kan de proefpersoon een verwachting vormen van hoe hij zich hoort te gedragen. Als het gedrag van de onderzoeker sterk bepalend is voor het gedrag van de proefpersoon, dan is het geregistreerde gedrag niet representatief voor het gedrag dat men zou observeren buiten de onderzoekssetting.

Het tweede soort verwachtingen zijn de verwachtingen ten aanzien van het effect van de behandeling. Deze verwachtingen zijn verantwoordelijk voor het zogenaamde *placebo-effect*. Het placebo-effect is vooral bekend uit de geneeskunde. Daar verstaat men eronder dat patiënten die een bepaalde behandeling krijgen het beter doen dan patiënten die niets krijgen, ook al heeft het veronderstelde werkzame bestanddeel van die behandeling geen effect. Het is bijvoorbeeld bekend dat patiënten met vage klachten er baat bij hebben als zij een spuitje met water toegediend krijgen; zij hebben daarna minder last van die klachten dan patiënten die niets krijgen.

In de psychologie verstaat men onder het placebo-effect het fenomeen dat proefpersonen die weten dat ze zich in een bepaalde conditie bevinden zich anders gedragen, ook

al heeft het veronderstelde werkzame aspect van die conditie geen effect. Bijvoorbeeld, stel dat men wil onderzoeken of een bepaalde training de geheugencapaciteit doet toenemen. Aan één groep proefpersonen geeft men deze training samen met de boodschap dat men verwacht dat hierdoor de geheugencapaciteit zal toenemen. Daarna neemt men een geheugentest af. Aan een andere groep geeft men geen training, maar neemt men alleen de geheugentest af. (De proefpersonen werden op zuiver toevallige wijze aan één van deze twee condities toegewezen.) Er is sprake van een placebo-effect als de groep met training een hogere gemiddelde testscore behaalt dan de groep zonder training, ook al heeft de training op zich geen effect.

De voor de hand liggende vraag is hoe men ervoor kan zorgen dat de resultaten van een onderzoek niet toe te schrijven zijn aan een mogelijk placebo-effect. Deze vraag is snel beantwoord als men ziet dat het placebo-effect het resultaat is van een storende situatievariabele. De condities die met elkaar vergeleken worden verschillen niet alleen van elkaar wat betreft het aspect waar men in geïnteresseerd is (het spuitje, de training) maar ook wat betreft het al-dan-niet weten dat men een behandeling gekregen heeft. De oplossing voor het probleem van storende situatievariabelen is experimentele controle van de condities. Concreet, men zorgt ervoor dat de condities alleen van elkaar verschillen wat betreft het aspect waarin men geïnteresseerd is door in beíde condities een behandeling te geven, maar slechts in één van beide een behandeling waar het werkzame aspect in zit. De andere conditie wordt de *placeboconditie* genoemd. Bijvoorbeeld, Pishkin en Shurley (1983) maakten gebruik van een placeboconditie om aan te tonen dat twee medicijnen (dioxepin en hydroxyzine) de opwinding verlaagde bij psychiatrische patiënten die faalden bij de oplossing van een cognitieve taak. De proefpersonen in de placeboconditie kregen dezelfde capsules als de proefpersonen in de andere twee condities (dioxepin en hydroxyzine), maar ditmaal bevatten de capsules alleen lactose (lactose is een suiker die men in melk aantreft). Omdat de proefpersonen in de placeboconditie niet wisten dat hun capsule geen medicijn bevatte, waren de drie condities gelijk wat betreft het bewustzijn van het feit dat zij een medicijn hadden ingenomen. Bijgevolg kan men een eventueel verschil tussen de condities met een groter vertrouwen toeschrijven aan de werkzaamheid van het medicijn op zich.

Placebocondities zijn van groot belang in het onderzoek naar de effectiviteit van psychotherapie. Placebocondities worden hier gebruikt om de zogenaamde *niet-specifieke factoren* die werkzaam zijn in psychotherapie onder controle te houden. Voorbeelden van niet-specifieke factoren zijn het feit dat men wekelijks tijd vrijmaakt om samen met een ander te werken aan zijn eigen problemen, het feit dat men die ander betaalt, en de verwachtingen die men heeft ten aanzien van het effect van de therapie. Men noemt deze factoren *niet-specifiek* omdat ze gemeenschappelijk zijn aan alle therapievormen. Niet-specifieke factoren zijn storende situatievariabelen: binnen het onderzoek naar de effectiviteit van psychotherapie is men er niet in geïnteresseerd, maar als men de condities *met psychotherapie* en *zonder psychotherapie* met elkaar vergelijkt, dan verschillen zij ook op deze variabelen. Placebocondities in psychotherapieonderzoek zijn behandelingen die theoretisch inert zijn, in de zin dat ze geen aspect bevatten van de veronderstelde werkzame factoren van de psychotherapie waarmee men vergelijkt. Een mooie placeboconditie is een behandeling door een persoon die geen psychotherapieopleiding gevolgd heeft, maar die in de ogen van de cliënt wel een volwaardige psychotherapeut is. Het

blijkt vaak dat cliënten in een dergelijke placeboconditie het ongeveer even goed doen als patiënten die een echte psychotherapie volgen.

### Verwachtingen van de Onderzoeker

Men ziet vaak wat men verwacht te zien, en dit geldt niet alleen voor de leek, maar ook voor de onderzoeker. In het merendeel van het wetenschappelijk onderzoek heeft de onderzoeker een verwachting over het gedrag dat in een bepaalde situatie zal optreden. Deze verwachting kan gebaseerd zijn op de resultaten van vroeger onderzoek of op de hypothese die men heeft over het mechanisme achter het te observeren gedrag. Deze verwachtingen kunnen leiden tot systematische observatiefouten. Dit kunnen we illustreren aan de hand van een onderzoek van Rosenthal en Fode (1963). Zij vroegen aan studenten om ratten te testen in een doolhof. Aan de studenten werd verteld dat de ratten die zij moesten testen van een speciale soort waren. Sommige studenten kregen te horen dat het slimme ratten waren, en andere studenten dat het domme ratten waren. In werkelijkheid was er geen verschil tussen de ratten. Uit de resultaten bleek dat de "slimme" ratten duidelijk minder fouten maakten in de doolhof dan de "domme" ratten. De studenten werden geobserveerd tijdens het testen van de ratten, en uit niets bleek dat zij de resultaten vervalsten. Dus, het verschil tussen de twee groepen moet het gevolg zijn van vrij subtiele verschillen in de manier waarop de studenten met de ratten omgingen.

Niet iedereen is het er mee eens dat de verwachtingen van de onderzoeker een sterke invloed hebben op de resultaten. Barber en Silver (1968a, 1968b) analyseerden het bestaande onderzoek over de invloed van de verwachtingen van de onderzoeker (onderzoek zoals dat van Rosenthal en Fode, 1963) en concludeerden dat er nog onvoldoende evidentie is om te kunnen besluiten dat deze effecten zich altijd en overal voordoen.

Als men vermoedt dat de verwachtingen van de onderzoeker een invloed kunnen hebben op de resultaten, dan is de beste oplossing om te werken met een of meerdere proefleiders die niet weten in welke conditie zij observeren. Anders uitgedrukt, de proefleiders zijn *blind* voor de condities. Bijvoorbeeld, in het onderzoek van Rosenthal en Fode (1963) zouden de studenten-proefleiders blind zijn voor de condities als hen niet verteld werd dat zij slimme of domme ratten moesten testen.

Het blind-zijn van de proefleiders voor de condities (het middel om de invloed van de verwachtingen van de proefleiders uit te schakelen) is de tegenhanger van het gebruik van een placeboconditie (het middel om de invloed van de verwachtingen van de proefpersonen uit te schakelen). Soms worden beide controlemiddelen samen gebruikt en dan spreekt men van een *dubbel-blind experiment*. Dubbel-blind experimenten worden vaak gebruikt om het effect van geneesmiddelen te onderzoeken. Het onderzoek van Pishkin en Shurley (1983) zou als volgt als een dubbel-blind experiment opgezet kunnen worden: (1) de onderzoeker vult de capsules met de medicijnen (doxepin en hydroxyzine) en het middel voor de placeboconditie (lactose) en geeft de capsules een code, en (2) een andere persoon (hier, de *proefleider* genoemd) deelt de capsules uit en noteert bij elke proefpersoon de code van de capsule die hij of zij gekregen heeft. Op deze manier weet de onderzoeker in welke conditie elke proefpersoon zit, maar de proefleider en de proefpersonen zelf niet.

Ook al hebben de verwachtingen van de onderzoeker een invloed op de resultaten,

toch mag men de impact van dit fenomeen niet overschatten. Wetenschappelijke kennis is immers niet gebaseerd op de resultaten van afzonderlijke experimenten. Als een onderzoeker interessante resultaten presenteert (mogelijk beïnvloed door zijn eigen verwachtingen) dan zullen andere onderzoekers (met andere verwachtingen) deze resultaten proberen te repliceren. Als zij hetzelfde vinden, dan kan men erop vertrouwen dat deze resultaten niet het product zijn van de verwachtingen van de onderzoeker. Als zij niet hetzelfde vinden, dan is een mogelijke verklaring dat de verwachtingen van de onderzoeker een belangrijke rol gespeeld hebben. Maar wat ook de verklaring is, een resultaat dat niet door andere onderzoekers gerepliceerd kan worden zal een stille dood sterven.

---

*Opgave 15*
Een onderzoekster wil weten of er een effect is van een tv-spotje over de nadelige gevolgen van overmatig alcoholgebruik. Zij vraagt aan een aantal voorbijgangers op straat wat zij vinden van een feest waarop er veel gedronken wordt. De voorbijgangers moeten hun mening geven op een beoordelingsschaal die loopt van *volledige afkeuring* tot *volledige goedkeuring*. Daarna vraagt de onderzoekster of zij het tv-spotje gezien hebben. Zij vergelijkt vervolgens de gemiddelde beoordeling van de mensen die het tv-spotje gezien hebben met de gemiddelde beoordeling van hen die het niet gezien hebben.

1. Welke vorm van dataverzameling wordt hier gebruikt: vrije observatie, participerende observatie, gestructureerde observatie, een veldexperiment of een laboratoriumexperiment? Motiveer je antwoord.
2. Stel dat de onderzoekster een significant verschil vaststelt tussen de gemiddelde beoordeling van de mensen die het tv-spotje gezien hebben en zij die het niet gezien hebben (degenen die het gezien hebben geven een meer negatieve beoordeling). Vind jij deze vaststelling overtuigende evidentie voor het feit dat er een effect is van het tv-spotje? Zo nee, leg uit waarom niet en zeg hoe jij het onderzoek zou opzetten om meer overtuigende evidentie te verzamelen.

*Opgave 16*
Een onderzoekster is geïnteresseerd in de invloed van de kleding die iemand draagt op zijn/haar humeur. Om deze invloed te onderzoeken, registreert ze van 100 treinreizigers de kleur van hun kleding (licht/donker). Bovendien laat ze deze 100 reizigers op een beoordelingsschaal aangeven hoe goed ze zich voelen die dag. Ze vergelijkt vervolgens het humeur van de mensen die lichte kleding dragen met het humeur van de mensen die donkere kleding dragen.

1. Welke vorm van dataverzameling wordt hier gebruikt: vrije observatie, participerende observatie, gestructureerde observatie, een veldexperiment of een laboratoriumexperiment? Motiveer je antwoord.

Een collega-onderzoeker gebruikt een andere methode. Hij verdeelt 100 kledingverkopers op zuiver toevallige wijze over twee groepen. De ene groep instrueert hij om lichte kleding te dragen en de andere groep om donkere kleding te dragen. Beide instructies worden gegeven onder het mom van het promoten van een nieuwe collectie. In de middagpauze moeten de verkopers de beoordelingsschaal invullen, waarop ze aangeven hoe goed ze zich voelen die dag. Ook deze onderzoeker vergelijkt vervolgens het humeur van de mensen die lichte kleding dragen met het humeur van de mensen die donkere kleding dragen.

2. Welk type onderzoek is dit: vrije observatie, participerende observatie, gestructureerde observatie, een veldexperiment of een laboratoriumexperiment? Motiveer je antwoord.

3. Als je één van de bij 1. en 2. genoemde dataverzamelingsmethoden zou moeten gebruiken om de onderzoeksvraag te beantwoorden, welke methode zou je dan kiezen? Wat is het voordeel van deze methode in vergelijking met de andere?

*Opgave 17*
Een aankomend uitgever wil graag weten op basis waarvan kinderen een leesboek uitzoeken. Om hierover een idee te krijgen, bezoekt hij gedurende drie weken elke middag een basisschool om te kijken welke boeken de kinderen uit groep 7 uit de boekenkast halen voor hun vrij-lezen uurtje. De uitgever beschikt over een lijst met de titels van alle boeken. Op deze lijst turft hij het aantal keren dat elk boek uit de boekenkast gehaald werd.

1. Welke vorm van dataverzameling wordt hier gebruikt? Motiveer je antwoord.
2. Is er in dit onderzoek sprake van standaardisering? Motiveer je antwoord.

De uitgever denkt dat kinderen meer zullen lezen als een verhaal gecombineerd wordt met een puzzel. Om zijn hypothese te toetsen, maakt hij twee versies van een verhaal. De eerste versie bestaat uit alleen het verhaal, en bij de tweede versie wordt na elk hoofdstuk een vraag gesteld (bijvoorbeeld, "Door wie werd de hond losgelaten in de tuin?"). Met de antwoorden op deze vragen kan het kind een zin vormen die de clou van het verhaal vormt. Op zuiver toevallige wijze wordt aan elk van de kinderen uit groep 7 een van de twee versies gegeven. Na een week kijkt de uitgever hoeveel bladzijden de kinderen gelezen hebben.

3. Welke vorm van dataverzameling wordt hier gebruikt? Motiveer je antwoord.
4. Sommige kinderen wilden niet meedoen met het onderzoek en weigerden het boek. (De kinderen wisten niet van tevoren welke versie van het boek ze aangeboden werd.) Denk je dat dit de resultaten verstoord heeft? Motiveer je antwoord. (In de laatste paragraaf van dit hoofdstuk wordt uitgelegd wat *respons* is, en na het lezen van deze paragraaf zal je wellicht een ander antwoord geven.)

*Opgave 18*
Geef aan, voor elk van de onderstaande vragen, of deze te beantwoorden zijn met (1) vrije observatie, (2) participerende observatie, (3) gestructureerde observatie, (4) een veldexperiment en/of (5) een laboratoriumexperiment. Sommige vragen zijn op meerdere manieren te beantwoorden. Geef argumenten waarom de door jou voorgestelde dataverzamelingsmethode te verkiezen is.

1. Uit welk gedrag blijkt dat iemand de leider is van een jeugdbende en welke factoren spelen een mogelijke rol bij het ontstaan van dit leiderschap?
2. Welke kenmerken van een instructie (woordkeuze, intonatie, het uiterlijk van degene die de instructie geeft, ...) hebben een invloed op de kans dat de instructie ook opgevolgd wordt?
3. Zorgt het vertonen van een schokkende scène (dood, lijden, agressie, ...) op een affiche ervoor dat de tekst op de affiche beter onthouden wordt?

*Opgave 19*
Geef van de onderstaande stellingen aan of ze waar of onwaar zijn. Geef een motivatie voor elk van de onware stellingen.

1. In een experiment zijn de condities waarin men observeert volledig gestandaardiseerd.
2. Als proefpersonen at random aan de condities toegewezen worden, kan er zich geen placebo-effect voordoen.

3. In een vergelijkend onderzoek naar het effect van een bepaalde psychotherapie, kan men de niet-specifieke factoren controleren door de patiënten op zuiver toevallige wijze aan de psychotherapie toe te wijzen.

---

## 1.5 Interview, Vragenlijst en Test

Het interview, de vragenlijst en de psychologische test zijn alledrie vormen van *vragen stellen*. In dit opzicht verschillen ze van de *observatie*methoden. Een belangrijk aspect van het interview, de vragenlijst en de psychologische test is de mate waarin ze gestandaardiseerd zijn. Er zijn twee aspecten van de dataverzameling die men kan standaardiseren: (1) de vragen die men stelt en (2) de antwoordmogelijkheden van de proefpersoon. Standaardisering van de vragen betekent dat op voorhand bepaald wordt welke vragen men gaat stellen. Bij het vrije interview (zie verder) zijn de vragen niet gestandaardiseerd; bij het gestandaardiseerde interview, de vragenlijst en de psychologische test zijn ze wel gestandaardiseerd. Standaardisering van de antwoordmogelijkheden betekent dat de proefpersoon op een vooraf omschreven manier moet antwoorden (door een optie te kiezen, door een kruisje te zetten op een lijn, door een cijfer te geven, ...). Bij het vrije interview zijn de antwoordmogelijkheden niet gestandaardiseerd; bij het gestandaardiseerde interview, de vragenlijst en de psychologische test zijn ze soms gestandaardiseerd en soms niet.

### 1.5.1 Het Vrije (niet-gestandaardiseerde) Interview

Een vrij interview is een gesprek waarin van tevoren niets is vastgelegd behalve wellicht het onderwerp waarover gesproken wordt. Het vrije interview is vergelijkbaar met de vrije observatie en heeft dezelfde voor- en nadelen. Zo is het vrije interview soms nuttig in de beginfase van een onderzoek wanneer men nog niet weet hoe de onderzoeksvraag precies geformuleerd moet worden of welke data relevant zijn voor een bepaalde onderzoeksvraag.

Een bijzondere vorm van het vrije interview is het *diepte-interview* . Dit is een type interview waarin men net zolang op een bepaald onderwerp ingaat tot men het gevoel heeft dat alle aspecten van dit onderwerp belicht zijn, ook-en-vooral die aspecten waar de geïnterviewde zich niet goed bewust van is. Het diepte-interview is een vermenging van dataverzameling, hypothesevorming en hypothesetoetsing. Als het onderwerp van het gesprek het persoonlijke leven van de geïnterviewde is (wat niet noodzakelijk zo is), dan is een diepte-interview moeilijk te onderscheiden van psychodynamische gesprekstherapie.

### 1.5.2 Het Gestandaardiseerde Interview en de Vragenlijst

Kenmerkend voor het gestandaardiseerde interview is dat de vragen op voorhand vastliggen. Als men deze vragen opschrijft en ze voorlegt aan een proefpersoon die er schriftelijk op moet antwoorden, dan heeft men een vragenlijst. Er is dus geen wezenlijk verschil

tussen een gestandaardiseerd interview en een vragenlijst. We behandelen ze hier dan ook samen en duiden ze verder samen aan als *vragenlijst* (een gestandaardiseerd interview is ook een "lijst van vragen").

### De Vragenlijst en zijn Afname

Het doel van een vragenlijst is het bevragen van feiten en meningen. Voorbeelden van feiten die men kan bevragen met een vragenlijst zijn de volgende:

1. Lectuur. Wat leest men, hoe vaak, afhankelijk van welke omstandigheden, ...?
2. Vakantie. Waar gaat men naartoe, met wie, hoe vaak, ...?
3. Verplaatsing. Met welk vervoermiddel verplaatst men zich, over welke afstand, met welke regelmaat, ...?
4. Voeding. Hoeveel maaltijden per dag, hoe zijn ze samengesteld, wat wordt nooit gegeten, ...?
5. Uitgaan. Hoe vaak, met een groep of alleen, hoeveel wordt er gepraat, ...?

En voorbeelden van meningen die men kan bevragen met een vragenlijst zijn de volgende:

1. Wat vindt men van zijn of haar ouders?
2. Hoe staat men tegenover buitenlanders?
3. Welke veranderingen verwacht men als men ouder wordt?
4. Wat vindt men van de relatie met zijn partner?
5. Hoe staat men tegenover euthanasie en abortus?

Niets belet een onderzoeker om in een vragenlijst zowel naar feiten als naar meningen te vragen. Zo kan het bijvoorbeeld interessant zijn om in dezelfde vragenlijst zowel te vragen naar de mening over zijn of haar ouders als naar de soort en de frequentie van activiteiten die men samen met zijn of haar ouders onderneemt.

In een vragenlijst zijn de vragen gestandaardiseerd. De antwoordmogelijkheden daarentegen, zijn soms gestandaardiseerd en soms niet. Voorbeelden van gestandaardiseerde antwoordmogelijkheden zijn de volgende: (1) Ja/Nee, (2) Eens/Oneens, (3) een cijfer tussen 1 en 7, en (4) een kruisje op een lijnstuk. Als de antwoordmogelijkheid volledig vrij (ongestandaardiseerd) is, dan noemen we de vraag een *open vraag* . Een voorbeeld van een open vraag is de volgende: "Beschrijf in termen van een aantal adjectieven hoe je vader met je omgaat."

Als men zelf een vragenlijst construeert, dan is het belangrijk om de vragen zorgvuldig te formuleren. Men vindt een aantal richtlijnen voor de formulering van vragen in een handig en beknopt Nederlandstalig boekje van Brinkman (1994).

Een vragenlijst kan mondeling, schriftelijk of telefonisch afgenomen worden. Deze drie vormen van afname hebben elk hun eigen voor- en nadelen, maar het is niet bijzonder interessant om ze allemaal de revue te laten passeren, en dat doen we hier ook niet.

### Populatie en Steekproef

Het is belangrijk om stil te staan bij de groep personen waarbij men de vragenlijst afneemt. Bij elk onderzoek (dus niet alleen bij vragenlijstonderzoek) moet men zich in principe afvragen over welke groep personen men een uitspraak wil doen op basis van het onderzoek. Deze groep noemen we de *populatie* . Voorbeelden van populaties zijn de volgende:

1. Alle Nederlanders boven de 18 jaar.
2. Alle Nederlanders boven de 65 jaar.
3. De bewoners van de Nederlandse bejaardentehuizen.
4. De Nederlandse universiteitsstudenten.
5. De Nijmeegse psychologiestudenten.

Het is niet zo dat de vraag naar wat de populatie is bij alle vormen van onderzoek even belangrijk is. In experimenteel onderzoek, bijvoorbeeld, stelt men zich die vraag zelden en soms is dat terecht. De reden waarom wij er hier bij stilstaan, is dat het bij vragenlijstonderzoek soms essentieel is om de populatie precies te omschrijven.

Als men in een bepaalde populatie geïnteresseerd is, dan zou men in het ideale onderzoek alle personen uit deze populatie in het onderzoek moeten betrekken. Omdat de meeste populaties zeer groot zijn, is deze werkwijze in de praktijk vrijwel nooit realiseerbaar. Daarom kiest men ervoor om een *representatieve steekproef* te trekken uit deze populatie en alleen deze personen in het onderzoek te betrekken. Een representatieve steekproef heeft alle kenmerken van de populatie waaruit zij getrokken is behalve de omvang. Anders uitgedrukt, een representatieve steekproef is de populatie in het klein. Bijvoorbeeld, als de populatie 20 procent academici bevat, dan geldt dat ook ongeveer voor een representatieve steekproef. Het woord "ongeveer" heeft hier een zeer precieze betekenis: de afwijking tussen steekproef en populatie kan door het toeval verklaard worden. Dit impliceert dat als men het verschil tussen steekproef en populatie statistisch zou toetsen, dit slechts zeer zelden tot een significant resultaat zou leiden.

Het belang van een representatieve steekproef kan geïllustreerd worden aan de hand van een voorbeeld uit de politicologie: het voorspellen van welke presidentskandidaat de verkiezingen zal winnen. Het voorbeeld heeft betrekking op de Amerikaanse presidentsverkiezingen van 1936. De populatie waarover men een uitspraak wil doen is de groep van alle stemgerechtigde Amerikanen. Onderzoekers trokken een steekproef van twee miljoen personen vertrekkende van een namenlijst die samengesteld was op basis van telefoongidsen, lijsten van tijdschriftabonnees, en ledenlijsten van automobielclubs. Met een dergelijke steekproef hadden de onderzoekers de uitslag van de vorige vier verkiezingen correct voorspeld. Deze keer voorspelden zij een overweldigende overwinning van Alfred M. Landon op Franklin D. Roosevelt. Toch was het Roosevelt die de verkiezingen won! De reden was dat de steekproef niet representatief was voor de populatie. In de jaren dertig was er een grote economische crisis waaronder vooral de armsten in de samenleving gebukt gingen. Deze sociale klasse trok in 1936 massaal naar de stembus, maar zij waren ondervertegenwoordigd op de namenlijst van de onderzoekers. Dat de steekproef verschrikkelijk groot was (twee miljoen!) kon hier niets aan verhelpen.

De verstoring van de resultaten door het gebruik van een niet-representatieve steekproef wordt *selectiebias* genoemd. Het woord *selectie* verwijst naar het feit dat de samenstelling van de steekproef selectief was: sommige categorieën waren oververtegenwoordigd ten koste van andere. Het woord *bias* komt uit het Engels en betekent *verstoring* (in dit geval, verstoring ten opzichte van de populatie).

Het gebruik van een niet-representatieve steekproef heeft soms ernstige gevolgen, zoals blijkt uit het volgende voorbeeld. Op het einde van de jaren vijftig werd in de Verenigde Staten een grootscheeps vragenlijstonderzoek opgezet om de behoeften van ouderen aan gezondheidszorg door de overheid in kaart te brengen. De steekproef werd zo getrokken dat de hogere sociale klassen oververtegenwoordigd waren. Bovendien werden de volgende categorieën uitgesloten: niet-blanke personen van boven de 65, ouderen die in verpleegtehuizen verbleven en ouderen die bijzondere financiële ondersteuning kregen. De steekproef was dus niet representatief voor de populatie van alle Amerikaanse ouderen. En omdat juist de ondervertegenwoordigde categorieën veel last hebben van de problemen van het ouder-worden, kan men verwachten dat de resultaten van het onderzoek beïnvloed zullen zijn door deze ondervertegenwoordiging. Het mag dan ook geen verbazing wekken dat uit de resultaten bleek dat de ouderen gezonder waren en meer tevreden met de bestaande gezondheidsvoorzieningen dan uit eerder onderzoek bleek. Jammer genoeg, vonden deze resultaten ook hun weg buiten de onderzoeksgemeenschap. Zij stonden onder andere aan de basis van een nationale campagne van de American Medical Association tegen overheidsprogramma's voor gezondheidszorg voor ouderen.

De manier waarop men een representatieve steekproef kan trekken is in principe zeer eenvoudig:

1. Men vertrekt van de namenlijst van alle leden van de populatie. Als men alleen geïnteresseerd is in personen van boven de 18, dan is de lijst van stemgerechtigden een uitstekend vertrekpunt.

2. Men trekt op zuiver toevallige wijze een steekproef uit deze lijst. Een steekproef wordt op zuiver toevallige wijze getrokken als iedere persoon op de lijst een even grote kans heeft om getrokken te worden. De handigste manier om dit te doen is door gebruik te maken van de randomtoets van een zakrekenmachine.

Het belangrijkste punt in deze procedure is het feit dat een zuiver toevallige trekking garandeert dat de steekproef representatief is, althans binnen de grenzen van toevallige schommelingen. De idee hierachter is dat, door het zuiver toevallig zijn van de trekking, geen enkele categorie personen over- of ondervertegenwoordigd zal zijn (opnieuw binnen de grenzen van toevallige schommelingen). De zuiver toevallige trekking is de belangrijkste, maar niet de enige, manier om aan een representatieve steekproef te komen. Alle andere manieren zijn echter variaties op de zuiver toevallige trekking.

Sociologen en politicologen maken zich meestal meer zorgen over de representativiteit van de steekproef dan psychologen. De reden hiervoor is dat de gevolgen van niet-representativiteit vooral duidelijk zijn als het gaat over de kenmerken van afzonderlijke variabelen. Anders dan sociologen en politicologen, zijn psychologen vooral geïnteresseerd in *relaties* tussen variabelen. Dit kunnen we het beste verduidelijken aan de hand van enkele voorbeelden. Het voorspellen van de verkiezingsuitslag (een zaak van sociologen en politicologen) komt in essentie neer op het berekenen van proporties:

de proportie personen in de steekproef die een voorkeur heeft voor elk van de kandidaten. We hebben hier te maken met slechts één (nominale, kwalitatieve) variabele: de kandidaat waarvoor men gaat kiezen. Een ander voorbeeld komt uit het Amerikaanse onderzoek naar de gezondheidszorgbehoeften van ouderen. Op basis van de antwoorden op de vraag "Hoeveel uur per maand krijgt u begeleiding van een fysiotherapeut?" kan men een schatting maken van de te verwachten behoefte aan fysiotherapie bij een bepaalde evolutie van de bevolkingssamenstelling (i.c., een bepaalde mate van vergrijzing). Opnieuw hebben we hier met slechts één variabele te maken: het aantal uur fysiotherapeutische begeleiding.

Een psycholoog zou eerder geïnteresseerd zijn in een vraag van het volgende type: is het zo dat adolescenten die hun ouders bij-de-tijd vinden ook meer activiteiten ondernemen met hun ouders? Dit zou men kunnen onderzoeken met een vragenlijst waarin men aan adolescenten zowel hun mening over hun ouders vraagt (bijvoorbeeld, "Zijn ze bij-de-tijd of ouderwets?") als naar feiten (bijvoorbeeld, "Hoe vaak hebben jullie de afgelopen maand samen wat gedaan?"). We hebben hier te maken met de relatie tussen twee variabelen (een mening en een feit) in plaats van met een kenmerk van slechts één variabele. Vaak is men ook geïnteresseerd in de relatie tussen meer dan twee variabelen. Men maakt dan gebruik van technieken zoals de *meervoudige regressieanalyse*, waarmee men onderzoekt of men één variabele kan voorspellen op basis van een aantal andere variabelen, en de *factoranalyse*, waarmee men de samenhang tussen een groot aantal variabelen probeert te beschrijven in termen van een klein aantal latente variabelen.

Wanneer we het hadden over slechts één variabele (presidentskandidaat, aantal uur fysiotherapeutische begeleiding) dan was het niet moeilijk om te verzinnen hoe een niet-representatieve steekproef verantwoordelijk kon zijn voor een verstoring van de resultaten. Wanneer we het hebben over de relatie tussen twee of meer variabelen, dan is dat een stuk moeilijker. Concreet, welk selectiemechanisme kan ervoor verantwoordelijk zijn dat bij de adolescenten in de steekproef, de relatie tussen hun mening over het bij-de-tijd zijn van hun ouders en de hoeveelheid gezamenlijke activiteiten met hun ouders, anders is dan in de populatie als geheel? Omdat het moeilijk is om een plausibel selectiemechanisme te verzinnen, maakt men zich meestal weinig zorgen over een mogelijke selectiebias wanneer men de relatie tussen twee of meer variabelen onderzoekt. Dat men zo'n selectiemechanisme niet kan verzinnen, betekent natuurlijk niet dat er geen selectiebias is!

Het probleem van een mogelijke selectiebias in het onderzoek naar de relatie tussen twee of meer variabelen is niet specifiek voor vragenlijstonderzoek; het probleem stelt zich ook in experimenteel onderzoek. Het verschil met vragenlijstonderzoek is dat men in een experiment het effect onderzoekt van een onafhankelijke op een afhankelijke variabele. Waar men in een vragenlijstonderzoek primair geïnteresseerd is in relaties tussen variabelen zondermeer, is men in experimenteel onderzoek uitsluitend geïnteresseerd in causale relaties. We noemen een relatie causaal als het in isolatie variëren van de onafhankelijke variabele (dit wil zeggen, zonder dat enige andere variabele mee varieert) voldoende is voor het optreden van variatie in de afhankelijke variabele. Maar een causale relatie is ook een relatie, en niets garandeert dat die relatie, vastgesteld in een steekproef, ook opgaat voor de populatie waar men in geïnteresseerd is. Dit is zeer relevant voor het experimenteel onderzoek in de psychologie, waar men zich vrijwel nooit zorgen maakt

over de representativiteit van zijn steekproef; men neemt gewoon de proefpersonen die beschikbaar zijn, en meestal zijn dat psychologiestudenten. Het is niet gegarandeerd dat het effect dat men observeert bij de groep psychologiestudenten even groot is als in de populatie waarin men geïnteresseerd is (voor een taalpsycholoog, bijvoorbeeld, de populatie van alle personen die kunnen lezen).

Men mag het belang van selectiebias in psychologisch onderzoek ook niet overschatten. Wetenschappelijke kennis is immers niet gebaseerd op de resultaten van afzonderlijke onderzoeken. Als een onderzoeker interessante resultaten presenteert (mogelijk specifiek voor de groep die hij onderzocht heeft) dan zullen andere onderzoekers deze resultaten proberen te repliceren. Eerst zal men het fenomeen proberen te repliceren met een vergelijkbare groep proefpersonen als in het oorspronkelijke onderzoek (meestal psychologiestudenten). Als deze replicatie succesvol is, zal op termijn ook de vraag gesteld worden in welke mate het fenomeen te generaliseren is naar andere groepen (niet-psychologiestudenten, kinderen, bejaarden, ...). Als de resultaten blijken te verschillen naargelang de groep die men onderzoekt, dan is dat niet noodzakelijk een slechte zaak. Soms kan het een aanwijzing bieden voor het mechanisme dat verantwoordelijk is voor het fenomeen (bijvoorbeeld, als de sterkte van het fenomeen afneemt met de leeftijd).

### De Respons

Niet elke persoon waarbij men de vragenlijst wil afnemen is ook bereid om mee te werken. De redenen hiervoor kunnen zeer divers zijn: men heeft er de tijd niet voor over, men wil niets prijsgeven over zichzelf, men voelt zich niet betrokken bij het onderwerp van de vragenlijst, .... Het aantal personen waarbij men effectief de vragenlijst heeft afgenomen noemt men de *respons*. Meestal wordt de respons uitgedrukt als een percentage: het percentage van de personen die men benaderd heeft waarbij men de vragenlijst ook afgenomen heeft. De respons hangt onder andere af van de wijze waarop de vragenlijst afgenomen wordt. Men krijgt de hoogste respons als men de vragenlijst uitdeelt tijdens een groepsbijeenkomst (bijvoorbeeld, een college). Als men de vragenlijst individueel afneemt, dan krijgt men de hoogste respons als dat gebeurt in een face-to-face interview (80 á 85 procent). Een vragenlijst die men met de post verstuurt, levert meestal slechts een respons van rond de 30 procent op. In meer gespecialiseerde boeken zoals dat van Babbie (1998) vindt men nuttige tips over hoe men de respons kan verhogen.

Het probleem van een lage respons is dat het goed mogelijk is dat de antwoorden die men effectief verkregen heeft niet representatief zijn voor de antwoorden die men verkregen zou hebben als iedereen meegewerkt had. De verstoring van de resultaten door een lage respons wordt *responsbias* genoemd. Een goed voorbeeld van een onderzoek waarin responsbias vrijwel zeker een rol gespeeld heeft, is het vragenlijstonderzoek waarop Shere Hite haar boek "Women and Love: A Cultural Revolution in Progress" (1987) gebaseerd heeft. Zij benaderde in totaal 100.000 vrouwen via verenigingen zoals feministische organisaties, kerkgemeenschappen en tuinierclubs. Haar vragenlijst bevatte 127 open vragen over onderwerpen die varieerden van vrijetijdsbezigheden tot de verstandhouding met hun echtgenoot. Slechts 4,5 procent van de aangeschreven vrouwen beantwoordde de vragenlijst.

Het eerste probleem met Shere Hites onderzoek is dat de 100.000 vrouwen die ze

aanschreef vrijwel zeker niet representatief zijn voor de populatie van Amerikaanse vrouwen; alle aangeschreven vrouwen waren immers lid van een of andere organisatie. Het tweede probleem, dat veel meer in het oog springt, is dat slechts 4,5 procent van de aangeschreven vrouwen de vragenlijst ook ingevuld heeft teruggestuurd. Met 100.000 verstuurde vragenlijsten, zijn dat er nog steeds 4.500, maar dat feit alleen maakt het nog geen goed onderzoek, ook al beweerde Shere Hite "It's 4.500 people. That's enough for me."

De resultaten van het onderzoek doen sterk vermoeden dat vooral de ontevreden vrouwen geneigd waren te antwoorden en daardoor oververtegenwoordigd waren in de steekproef. Zo bleek bijvoorbeeld dat 95 procent van de vrouwen zich emotioneel geteisterd voelde door hun mannen. Dit staat in schril contrast met de resultaten van een onderzoek van Peplau en Gordon (1985) dat in ongeveer dezelfde periode plaatsvond en dat gebaseerd was op een zuiver toevallige steekproef. Uit hun onderzoek bleek dat ongeveer de helft van de vrouwen volledig tevreden was met hun huwelijk en dat slechts 3 procent zei dat ze helemaal niet gelukkig was.

Omwille van dezelfde reden als bij de selectiebias, is ook de responsbias meestal een groter probleem voor sociologen en politicologen dan voor psychologen: (1) de gevolgen van een lage respons zijn vooral duidelijk als het gaat over de kenmerken van afzonderlijke variabelen (zoals in het onderzoek van Shere Hite) en (2) psychologen zijn vooral geïnteresseerd in relaties tussen variabelen. Vaak is het moeilijk om het mechanisme te achterhalen waardoor de lage respons de gevonden relaties verstoord kan hebben. Dit wil echter niet zeggen dat er geen verstoring is!

### Zijn de Antwoorden te Vertrouwen?

De vraag of de antwoorden op een vragenlijst wel te vertrouwen zijn, wordt het beste apart behandeld voor een vragenlijst die naar meningen peilt en een die naar feitelijk gedrag peilt. Vaak beschouwt men de antwoorden op een vragenlijst die naar meningen peilt als niet te vertrouwen als de proefpersonen niet hun "echte mening" geven. Bijvoorbeeld, misschien koesteren veel pubers een diepe haat tegen hun ouders, maar durven ze dat niet toe te geven (misschien zelfs niet tegen zichzelf!) omdat dit een sociaal ongewenst gevoel is, of omdat ze bang zijn voor de gevolgen als ze zich er tezeer bewust van worden. Dit probleem is geen echt probleem, gewoon omdat het niet vast te stellen is óf er überhaupt een probleem is: als de proefpersonen hun echte mening niet willen geven, is het ook niet mogelijk om vast te stellen of die mening afwijkt van de mening die ze wel gegeven hebben. Misschien is men niet zo vlug overtuigd door dit argument, en stelt men dat de echte mening wel achterhaald kan worden door middel van een bijzondere techniek, zoals het diepte-interview. Met deze tegenwerping wordt de bewijslast doorgeschoven naar die bijzondere techniek: van die techniek moet aangetoond worden dat hij de ongecensureerde meningen kan blootleggen. Het is echter niet duidelijk hoe men dit moet aantonen.

Bij vragenlijsten die naar feitelijk gedrag peilen moet men een onderscheid maken tussen vragen naar hoe men zich in het verleden effectief gedragen heeft, en vragen naar hoe men zich zóu gedragen in een bepaalde situatie. De antwoorden op het eerste soort vragen zijn niet te vertrouwen als zij niet overeenstemmen met hoe men zich feitelijk gedragen heeft. In een onderzoek van Parry en Crossley (1950) werden mensen

geïnterviewd over onderwerpen zoals giften aan liefdadigheidsorganisaties en deelname aan verkiezingen. Na de interviews werden de antwoorden gecontroleerd aan de hand van de archieven van overheids- en privé-organisaties. Het bleek dat 40 procent van de ondervraagden een fout antwoord gaf op een vraag naar hun bijdrage aan een bepaald liefdadigheidsfonds, en dat 25 procent een fout antwoord gaf op de vraag of ze gestemd hadden bij een recente verkiezing. We kunnen hieruit het volgende leren: als men geïnteresseerd is in hoe mensen zich gedragen, en we vertrouwen er niet op dat ze ons dat gewoon zeggen, dan gebruiken we beter geen vragenlijst. In dit geval is observatie (van het gedrag zelf of van sporen van het gedrag) veel beter geschikt.

Men kan in een vragenlijst ook vragen naar hoe men zich zou gedragen in een bepaalde situatie, zonder expliciet te refereren aan situaties die zich in het verleden hebben voorgedaan. Men beschouwt antwoorden op dergelijke vragen als niet te vertrouwen als mensen in die situatie een ander gedrag stellen dan wat ze beweren te zullen doen. Dit kunnen we illustreren aan de hand van het onderzoek van Darley en Latané (1968) naar het effect van het aantal omstanders op hulpverlenend gedrag in noodsituaties. Uit hun onderzoek bleek dat men eerder geneigd is om hulp te verlenen als er geen omstanders aanwezig zijn dan wanneer er wel omstanders aanwezig zijn. Wanneer zij echter aan een tweede groep proefpersonen vroegen of het aantal omstanders een invloed zou hebben op de kans dat ze het slachtoffer helpen, dan antwoordden ze allemaal met "Nee".

Bevindingen zoals deze wijzen erop dat wij ons vaak niet bewust zijn van de determinanten van ons gedrag. Als we dat wel zouden zijn, dan zou een groot deel van de wetenschappelijke psychologie overbodig zijn; in plaats van te experimenteren zouden wij gewoon kunnen vragen of een bepaalde variabele wel of geen rol speelt. Dat het irreëel is om te verwachten dat mensen zich bewust zijn van de determinanten van hun gedrag, wordt mooi geïllustreerd in een gedachtenexperiment van Kenneth Bowers (1987). In dit gedachtenexperiment veronderstelt Bowers dat de aard van de verlichting (hard of licht) een invloed heeft op het feit dat een eerste ontmoeting tussen twee personen in een vaste relatie uitmondt of niet. De achterliggende gedachte hierbij is dat een zachte verlichting een romantische sfeer creeërt waardoor mensen meer geneigd zijn om toenadering te zoeken tot elkaar. Bowers stelt zich voor dat hij na een aantal jaren iemand spreekt die destijds heeft deelgenomen aan dit onderzoek naar het effect van harde versus zachte verlichting op het beginnen van een vaste relatie. Hij vraagt de ex-proefpersoon hoe het toch kwam dat hij tijdens dat onderzoek een relatie begonnen is met zijn huidige vrouw, waarop de ex-proefpersoon antwoordt: "Dat kwam door die verdomde verlichting!"

Dat mensen zich niet bewust zijn van de determinanten van hun gevoelens en gedragingen is een normale zaak. Daarom mogen we van een vragenlijst ook niet verwachten dat deze ons deze determinanten openbaart. Ze kan ons wel informatie geven over wat mensen dénken dat de determinanten zijn van hun gedrag. Deze lekenpsychologie (Hoe verklaren mensen hun eigen gevoelens en gedrag?) is interessant als *voorwerp* van onderzoek maar ze voldoet niet als wetenschappelijke verklaring.

### 1.5.3 De Psychologische Test

De psychologische test is een instrument waarmee persoonskenmerken gemeten worden. Men moet hierbij denken aan het volgende:

1. Cognitieve vaardigheden zoals intelligentie, ruimtelijk voorstellingsvermogen, verbale vlotheid, creativiteit, ....

2. Persoonlijkheidskenmerken zoals extraversie, neuroticisme, vriendelijkheid, gewetensvolheid, ....

3. Attitudes zoals houding ten opzichte van buitenlanders, abortus, het milieu, drugsgebruik, ....

4. Tijdelijke gevoelens zoals boosheid, angst, opwinding, ontroering, ....

Het verschil met de vragenlijst is dat men met een psychologische test altijd een bepaald persoonskenmerk wil meten; bij een vragenlijst doet het er niet toe of de meningen en feiten waarnaar men vraagt betrekking hebben op een bepaald persoonskenmerk.

Het onderscheid tussen een psychologische test en een vragenlijst is niet altijd zo precies te maken als hier misschien gesuggereerd wordt. Dit komt omdat sommige psychologische tests vragenlijsten zíjn. Dit geldt bijvoorbeeld voor de bekende papier-en-pen persoonlijkheidstests. Deze persoonlijkheidstests (of persoonlijkheidsvragenlijsten) bestaan uit een reeks vragen over hoe men zich voelt, wat men denkt en hoe men zich gedraagt. Deze omschrijving voldoet aan wat we onder een vragenlijst verstaan. Maar het zijn ook psychologische tests, want op basis van hun antwoorden krijgen de proefpersonen een score op persoonlijkheidsdimensies zoals extraversie, neuroticisme, gewetensvolheid, en dergelijke.

Een psychologische test wordt soms ook een *schaal* genoemd. De term "schaal" wordt ook gebruikt als men met één test meerdere persoonskenmerken meet. Zo bestaat de WAIS (een intelligentietest) uit een verbaal- en een performantiegedeelte. Deze gedeelten worden ook aangeduid als de verbale- en performantieschaal. En de ABV (een persoonlijkheidsvragenlijst) bestaat uit vragen waarmee men neuroticisme meet en andere vragen waarmee men extraversie meet. Deze twee sets vragen worden aangeduid als de neuroticisme- en de extraversieschaal.

Aan een psychologische test worden twee eisen gesteld: *betrouwbaarheid* en *validiteit*. In het hoofdstuk over meten gaan we hier uitgebreid op in.

---

*Opgave 20*
Een vragenlijstonderzoeker had op zuiver toevallige wijze een steekproef geselecteerd. Hij nam de vragenlijst af als een gestandaardiseerd interview. Bij de afname van de vragenlijst kwam het regelmatig voor dat er niemand thuis was op het adres dat hij geselecteerd had. Hij had een eenvoudige oplossing voor dit probleem: hij interviewde de eerste buurman of buurvrouw die wel thuis was. Welke kritiek kun je geven op deze werkwijze?

*Opgave 21*
Een bedrijf brengt sinds enige tijd merk Q op de markt. De marketingafdeling van het bedrijf wilde evalueren hoe tevreden de gebruikers waren met het product. Daartoe stuurde zij per post een vragenlijst naar deze gebruikers. Uit de lijst van geregistreerde gebruikers werd op zuiver toevallige wijze een steekproef van 100 personen getrokken. Hiervan stuurden 30 personen de vragenlijst ingevuld terug en 28 van hen antwoordden dat zij zeer tevreden waren met het product. Daarna werd een reclamecampagne opgezet waarin vermeld werd dat 93 procent van de gebruikers zeer tevreden was met het product. Onder welke voorwaarde is deze

reclameboodschap verdedigbaar op basis van de resultaten van het onderzoek? Denk je dat het aannemelijk is dat aan deze voorwaarde voldaan is? Motiveer je antwoord.

*Opgave 22*
Een onderzoeker wil middels een korte vragenlijst nagaan welk percentage van de jongeren tussen de 14 en de 21 jaar het roken van marihuana op feestjes aanvaardbaar vindt. Welke vorm(en) van bias kan (kunnen) ervoor zorgen dat hij met dit vragenlijstonderzoek geen valide antwoord kan geven op zijn onderzoeksvraag? Leg uit.

*Opgave 23*
Geef van de onderstaande stellingen aan of ze waar of onwaar zijn. Geef een motivatie voor elk van de onware stellingen.

1. Door een vragenlijst individueel af te nemen (in een face-to-face interview) in plaats van hem met de post te versturen, wordt de responsbias verlaagd.
2. Het feit dat een onderzoekster voor haar onderzoek alleen vragenlijsten verstuurde naar vrouwen die lid waren van een of andere organisatie, duidt op responsbias.

# Hoofdstuk 2

# Meten

## Inhoudsopgave


**2.1 De Plaats van het Meten binnen de Psychologie . . . . . 62**
2.1.1 Voorbeelden . . . . . 63
2.1.2 Psychologische Constructen . . . . . 65
2.1.3 Operationele Definities . . . . . 66
2.1.4 Het Meten van Persoonskenmerken . . . . . 68
**2.2 Betrouwbaarheid . . . . . 70**
2.2.1 Betrouwbaarheid=Herhaalbaarheid . . . . . 70
2.2.2 De Basisformule van de Klassieke Testtheorie . . . . . 74
2.2.3 De Definitie van Betrouwbaarheid . . . . . 79
2.2.4 Parallelle Tests . . . . . 80
2.2.5 Testverlenging en Betrouwbaarheid . . . . . 82
2.2.6 Praktische Methoden voor het Bepalen van de Betrouwbaarheid Gebaseerd op Parallelliteit . . . . . 84
2.2.7 Cronbachs Alpha . . . . . 86
2.2.8 Itemanalyse . . . . . 93
2.2.9 Het Restriction-of-Range Fenomeen . . . . . 95
2.2.10 De Standaardfout van Meting . . . . . 98
2.2.11 Correctie voor Attenuatie . . . . . 98
2.2.12 Berekenen van de Betrouwbaarheid met de SPSS-procedure Reliability . . . . . 100
**2.3 Validiteit . . . . . 104**
2.3.1 Inhoudsvaliditeit . . . . . 104
2.3.2 Constructvaliditeit . . . . . 105
2.3.3 Criteriumvaliditeit . . . . . 109
2.3.4 Criteriumvaliditeit en Beslissingen . . . . . 113
**2.4 Appendix bij Hoofdstuk 2 . . . . . 122**

## 2.1 De Plaats van het Meten binnen de Psychologie

We vertrekken hier opnieuw van Kurt Lewins (1951) definitie van psychologie:

$$\boxed{G = f(P, S) \quad .}$$

In deze formule wordt gedrag ($G$) uitgedrukt als functie van persoons- ($P$) en situatiekenmerken ($S$). Deze functie kan verschillende vormen aannemen, afhankelijk van de wijze waarop een bepaald gedrag (het al dan niet verlenen van hulp, het behalen van bepaalde studieresultaten, het kiezen voor het ene of het andere beroep, ... ) bepaald wordt door persoons- en/of situatiekenmerken (het aantal omstanders bij hulpverlenend gedrag, intelligentie bij studieresultaten, sociale bewogenheid bij beroepskeuze, ... ).

Door de psychologie te definiëren met behulp van een formule zoals $G = f(P, S)$, lijkt het dat we met een erg exacte wetenschap te maken hebben. Onder "exact" verstaan we hier dat het mogelijk is om precies te bepalen welk gedrag ($G$) een bepaalde persoon (met kenmerken $P$) zal stellen in een bepaalde situatie (met kenmerken $S$). Zoals verder zal blijken, wordt de formule $G = f(P, S)$ meestal niet op zo een exacte manier gehanteerd, maar gebruikt men haar op een veel vrijblijvender manier, namelijk om op een gecondenseerde manier een hypothese weer te geven.

Een belangrijk thema in de methodologie van het psychologisch onderzoek is de manier waarop men omgaat met psychologische kenmerken (intelligentie, gespannenheid, taalvaardigheid, aandachtigheid, zelfstandigheid, ... ). Die psychologische kenmerken kunnen persoonskenmerken zijn, en dan corresponderen ze met de $P$ in $G = f(P, S)$ (bijvoorbeeld, intelligentie, taalvaardigheid, zelfstandigheid). Het kunnen ook psychologische kenmerken zijn die variëren omdat ze onder invloed staan van de situatie. Gespannenheid en aandachtigheid zijn twee voorbeelden van psychologische kenmerken die onder invloed staan van de situatie. Situatiekenmerken worden in $G = f(P, S)$ aangeduid met $S$. Psychologische kenmerken zijn niet rechtstreeks observeerbaar maar toch willen we er uitspraken over doen.

Meten is het getalsmatig uitdrukken van de niet-rechtstreeks observeerbare psychologische kenmerken. Er zijn meerdere benaderingen van het meetprobleem in de psychologie, maar in dit boek wordt er slechts één behandelend. De kerngedachte van deze benadering is dat een meting een *operationele definitie* is van een niet-rechtstreeks observeerbaar psychologisch kenmerk. We komen hier nog uitgebreid op terug.

Een alternatieve benadering, waar we hier niet in detaïl op ingaan, heeft als kerngedachte dat meten een toegift is van een *formele structuur* die bij de data past. Zo een formele structuur wordt ook vaak een *model* genoemd. Het nut van een formele structuur die bij de data past bestaat erin dat hij ons een inzicht verschaft in de wetmatigheden van een bepaald gedrag. Dit gebeurt op een gelijkaardige manier als die waarop Newtons mechanica ons een inzicht verschaft in de wetmatigheden van het gedrag (i.c., de bewegingen) van vallende en botsende voorwerpen. Volgens deze benadering kun je alleen maar meten als je een model hebt dat bij de data past. In de literatuur worden onder andere de volgende modellen gebruikt: het scalogram, het Rasch model, en het Mokken model.

In de volgende paragraaf bespreken we een aantal voorbeelden van hypotheses zoals die in de psychologie geformuleerd worden. Deze hypotheses zijn geformuleerd in termen

van niet-observeerbare kenmerken. Men kan deze hypotheses alleen toetsen als men operationele definities formuleert voor de niet-observeerbare kenmerken.

### 2.1.1 Voorbeelden

In wat volgt, gebruiken we de formule $G = f(P, S)$ als samenvattende beschrijving van een hypothese over een bepaald gedrag $G$. Hierbij is het zó dat deze hypothese geformuleerd is in termen van niet-observeerbare kenmerken en dat de bijdrage van de formule $G = f(P, S)$ tot een beter begrip van het gedrag in kwestie slechts gegeven is door het verbinden van deze termen met de symbolen $G$, $P$, $S$ en $f$. We geven nu vier voorbeelden.

1. *Studieresultaten en intelligentie.*
   Dit voorbeeld heeft betrekking op een hypothese over verschillen in studieresultaten. De studieresultaten corresponderen met $G$. De hypothese houdt in dat iemands studieresultaten hoger zijn naarmate zijn of haar intelligentie hoger is. De intelligentie correspondeert met $P$ en $f$ is een stijgende functie van $P$.

   In de hypothese wordt de invloed van de situatie niet genoemd en daarom kan $G = f(P, S)$ vereenvoudigd worden tot $G = f(P)$. Dit betekent níet dat situatiekenmerken geen invloed hebben op de studieresultaten; een motiverende houding van de ouders, bijvoorbeeld, kan best een invloed hebben. Het weglaten van de $S$ in de formule betekent alleen dat we te maken hebben met een hypothese waarin de invloed van situatiekenmerken niet genoemd wordt.

2. *Instemming en vertrouwdheid.*
   Tijdens een overleg, wordt een groot gedeelte van de tijd besteed aan het geven van en luisteren naar standpunten. Veel van deze standpunten zijn bekend, omdat men ze al vaker gehoord heeft. We beschouwen nu een hypothese die betrekking heeft op de mate waarin men het eens is met een standpunt. Dit wordt verder de *mate van instemming* genoemd. Deze hypothese houdt in dat de mate van instemming toeneemt naarmate een standpunt meer vertrouwd is (omdat men het reeds vaker gehoord heeft, er meer over nagedacht heeft, ...). De $G$ in $G = f(P, S)$ correspondeert hier met de mate van instemming met een standpunt en de $S$ met de mate waarin men vertrouwd is met dit standpunt.

   In deze hypothese wordt geen gewag gemaakt van een invloed van de personen van wie men het gedrag onderzoekt, en bijgevolg kan de formule $G = f(P, S)$ vereenvoudigd worden tot $G = f(S)$. Dit betekent niet dat persoonskenmerken geen invloed hebben op de mate waarin men het met een bepaald standpunt eens is. Het betekent alleen dat we het in de hypothese alleen over situatiekenmerken hebben.

3. *Aptitude-treatment interactie (ATI).*
   Net als in het eerste voorbeeld, hebben we hier te maken met een hypothese over verschillen in studieresultaten. Deze hypothese heeft betrekking op de wijze waarop studieresultaten bepaald worden door verschillende onderwijsstijlen. Er

wordt hier een onderscheid gemaakt tussen een democratische en een autoritaire onderwijsstijl. Deze twee onderwijsstijlen corresponderen met $S$.

Er wordt ook een onderscheid gemaakt tussen twee types van leerlingen, zelfstandige en volgzame. Deze twee types corresponderen met $P$. De hypothese is nu dat voor zelfstandige leerlingen de studieresultaten beter zijn bij een democratische dan bij een autoritaire onderwijsstijl, maar dat het omgekeerde geldt voor de volgzame leerlingen. De studieresultaten, corresponderend met $G$, worden dus beschouwd als een functie van een factor die varieert over de personen, namelijk zelfstandigheid, en als een functie van een factor die varieert over de situaties, namelijk democratisch/autoritair[1]. Dus, $G = f(P, S)$ kan niet vereenvoudigd worden door het weglaten van één van de argumenten, zoals in de vorige twee voorbeelden.

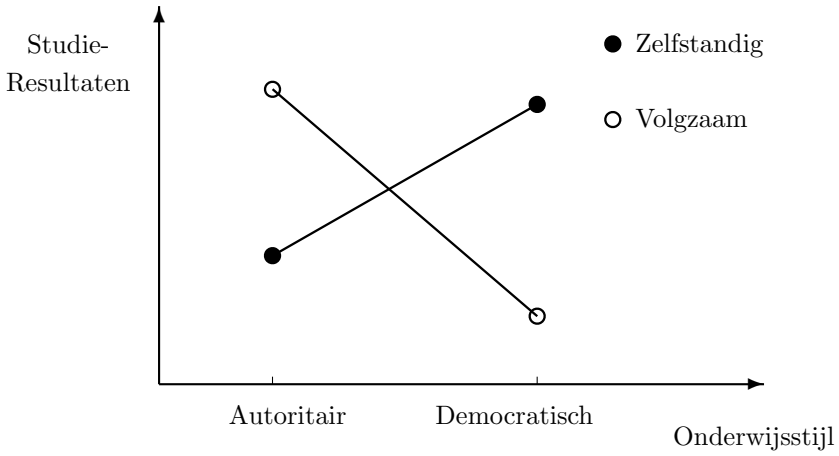
Een bijzonder kenmerk van de functie $f$ in dit voorbeeld, is haar vorm als functie van ofwel $P$, ofwel $S$, bij een vaste waarde van, respectievelijk, $S$ of $P$: bij de $P$-waarde zelfstandig is $f$ stijgend als $S$ verschuift van autoritair naar democratisch, maar bij de $P$-waarde volgzaam is zij dalend. Iets analoogs geldt voor $f$ als functie van $P$ bij vaste waarden van $S$. Dit fenomeen wordt *interactie* genoemd en in de specifieke context van het besproken voorbeeld spreekt men van *aptitude-treatment interaction* (ATI) (in het Nederlands, bekwaamheid-behandeling interactie). Het fenomeen ATI wordt uitgebreid behandeld door Cronbach en Snow (1977).

De functie $f$ wordt weergegeven in Figuur 2.1. De interactie blijkt uit het feit dat het lijnstuk voor de zelfstandige leerlingen het lijnstuk voor de volgzame leerlingen snijdt. Tussen haakjes, het snijden van de lijnstukken is geen noodzakelijk kenmerk van interactie; het feit dat ze niet parallel zijn is al voldoende om over interactie te kunnen spreken.

4. *Studieresultaten, intelligentie en zelfstandigheid.*
   In alle bovenstaande voorbeelden was er slechts één persoonskenmerk (intelligentie, zelfstandigheid) en één situatiekenmerk (mate van afhankelijkheid, democratisch/autoritair) waardoor het gedrag bepaald werd. Dit is niet altijd het geval. We kunnen bijvoorbeeld de hypothese formuleren dat studieresultaten bepaald worden door zowel intelligentie als zelfstandigheid, waarbij deze laatste factor interageert met het situatiekenmerk democratisch/autoritair. In $G = f(P, S)$ staat $P$ dan voor twee persoonskenmerken in plaats van één, zoals in het vorige voorbeeld.

Waar het in de bovenstaande voorbeelden op neerkomt is dat de formule $G = f(P, S)$ op een compacte wijze een bepaalde hypothese weergeeft. Het grote gevaar bij een dergelijk gebruik van $G = f(P, S)$, is dat het suggereert dat het mogelijk is om exact te bepalen welk gedrag ($G$) een persoon (met kenmerken $P$) zal stellen in een bepaalde situatie (met kenmerken $S$). Een dergelijke exacte bepaling is slechts mogelijk als aan de volgende voorwaarden voldaan is:

[1] Merk op dat de onderwijsstijl slechts dan een situatiekenmerk is, als het bekeken wordt vanuit het standpunt van de leerlingen (en het is hun gedrag dat we hier beschouwen), maar niet als het bekeken wordt vanuit het standpunt van de leerkrachten; vanuit het standpunt van de leerkrachten is onderwijsstijl een persoonskenmerk.




Figuur 2.1: Grafische voorstelling van het verband tussen studieresultaten, onderwijsstijl en zelfstandig/volgzaam, zoals beschreven in de tekst.

1. De persoonskenmerken $P$ en de situatiekenmerken $S$ kunnen met een getal weergegeven worden.
2. De functie $f$ kan exact bepaald worden.
3. Men kan aangeven met welke observeerbare gedragingen de verschillende functiewaarden van $f$ corresponderen.

Indien niet aan deze voorwaarden voldaan is, is er slechts sprake van schijnexactheid. Dit wordt mooi geïllustreerd aan de hand van een formule van de achttiende-eeuwse filosoof Davis Hartley (zie Miller, 1964, p. 9):

$$W = \frac{F^2}{L} \quad ,$$

waarin $W$ iemands liefde voor de wereld aanduidt en $F$ en $L$, respectievelijk, vrees voor God en liefde voor God aanduiden. Volgens Hartley zou $L$ naar oneindig gaan bij het ouder worden, waardoor $W$ naar nul zou gaan.

### 2.1.2 Psychologische Constructen

Eén van de grote problemen in de psychologie is dat veel theorieën geformuleerd zijn in termen van kenmerken waarvan het niet onmiddellijk duidelijk is hoe we ze kunnen observeren. Zo hadden we het in de voorbeelden onder andere over intelligentie, instemming, vertrouwdheid, afhankelijkheid, onderwijsstijl en zelfstandigheid.

Strikt genomen bestaan deze begrippen alleen in ons hoofd (dat van de onderzoeker, de psycholoog). Wij denken over menselijk gedrag in termen van deze begrippen. Het

is niet zo dat ons denken over gedrag louter beschrijving is van wat we observeren, want de begrippen die we hanteren verwijzen naar niet-rechtstreeks observeerbare kenmerken. In feite construeren we tot op zekere hoogte de werkelijkheid. Deze constructie wordt mede tot stand gebracht door het feit dat wij bepaalde observaties (van gedrag en van de situatie waarin het gesteld wordt) interpreteren in termen van deze begrippen. Bijvoorbeeld, sommige gedragingen worden beschouwd als manifestaties van instemming en andere niet. En zo worden ook sommige gedragingen gezien als manifestaties van intelligentie of zelfstandigheid en andere niet. We noemen deze begrippen *psychologische constructen* en geven hiermee aan dat wij zelf een construerende rol spelen als we gedrag beschrijven in deze termen.

De mate waarin wij zelf een construerende rol spelen bij het gebruik van psychologische constructen is niet hetzelfde voor alle constructen. Zo zijn er constructen die vrij dicht bij de waarneming staan en andere die er verder vanaf staan, en die daarom vaak als meer *theoretische* constructen aangeduid worden. Studieresultaten zijn een voorbeeld van een psychologisch construct dat kort bij de empirie staat: men kan zich vrij makkelijk een voorstelling maken van de wijze waarop studieresultaten vastgesteld kunnen worden. Velen zullen bij het construct studieresultaten denken aan het gemiddelde cijfer op de eindexamenvakken. Maar dit is zeker niet voor iedereen het geval. Bijvoorbeeld, voor sommigen is de mate waarin schoolse kennis toegepast kan worden een essentieel aspect van het construct studieresultaten. Dit illustreert dat, ook al zijn studieresultaten een construct dat kort bij de empirie staat, men van mening kan verschillen over hoe men het moet vaststellen.

Enkele voorbeelden van psychologische constructen die wat verder van de empirie af staan zijn verantwoordelijkheidsgevoel, zelfstandigheid en egosterkte. Bij deze constructen kan men zich veel moeilijker een voorstelling maken van de wijze waarop ze concreet vastgesteld kunnen worden. Zeer vaak zijn verschillende onderzoekers het met elkaar oneens over wat een goede indicatie is van dit construct en wat niet. Zo kan bijvoorbeeld de ene onderzoeker een score op Rorschach beschouwen als een goede indicatie van ego-sterkte terwijl een andere onderzoeker zweert bij een diepte-interview door een ervaren psychoanalyticus.

### 2.1.3 Operationele Definities

Een onderzoeker wil natuurlijk nagaan of zijn hypothese, die gesteld is in termen van relaties tussen psychologische constructen, ook overeenkomt met de werkelijkheid. Hij zal dus een verbinding moeten leggen tussen zijn gedachten en de observeerbare werkelijkheid. Hiervoor maakt hij gebruik van *operationele definities*:

> Een operationele definitie is een beschrijving van de wijze waarop men de hoedanigheid van een bepaald psychologisch construct kan vaststellen.

Zonder operationele definitie blijft een psychologisch construct iets dat uitsluitend in het hoofd van de onderzoeker bestaat.

We geven nu enkele voorbeelden van operationele definities. We doen dit voor de drie types psychologische constructen die elk corresponderen met één van de drie symbolen in de formule $G = f(P, S)$: $G$, $P$ en $S$. Ten eerste, zijn er de psychologische constructen die corresponderen met $G$. De studieresultaten en de mate van instemming met een

standpunt, die beschreven werden in de voorbeelden, zijn psychologische constructen van dit type. Elk van de constructen van dit type wordt operationeel gedefinieerd door observeerbare gedragingen. We veronderstellen dat deze gedragingen allemaal op dezelfde wijze afhangen van de persoons- en de situatiekenmerken. Zo is het mogelijk om het construct studieresultaten operationeel te definiëren door de score op (1) een CITO-toets, (2) een tentamen van de leerkracht of (3) een praktijkproef, en in elk van deze drie gevallen verwachten wij dezelfde functionele verbanden met de persoons- en situatiekenmerken (i.c., intelligentie, zelfstandigheid en democratisch/autoritair). Het construct instemming kan operationeel gedefinieerd worden door (1) de vraag hoe sterk men het eens is met het standpunt (te beantwoorden op een schaal van 0 tot 10), (2) het aantal keren dat men tijdens een overleg het standpunt steunt, en (3) door het aantal argumenten dat men kan geven voor dit standpunt. Ook hier verwachten wij in elk van deze drie gevallen hetzelfde functionele verband met het relevante situatiekenmerk (i.c., de mate van vertrouwdheid met het standpunt).

Ten tweede, zijn er de psychologische constructen die corresponderen met $P$, de persoonskenmerken. De kenmerken intelligentie en zelfstandigheid uit de voorbeelden zijn psychologische constructen van dit type. Elk van deze constructen wordt operationeel gedefinieerd door een *methode* waarmee verschillen met betrekking tot dit construct vastgesteld kunnen worden. Zo is het mogelijk om het construct intelligentie operationeel te definiëren door de score op (1) de WAIS, (2) de GIT, of (3) de Raven Progressive Matrices. In deze drie gevallen verwachten wij hetzelfde functionele verband met het gedrag waarin we geïnteresseerd zijn (i.c., de studieresultaten). Analoog kan men ook het construct zelfstandigheid op verschillende manieren operationeel definiëren: (1) door een vragenlijst waarin gepeild wordt naar iemands ervaringen met problemen die men alleen (zonder hulp van anderen) moest oplossen, (2) een vragenlijst waarin gepeild wordt naar iemands voorkeur voor het zelfstandig dan wel onder begeleiding uitvoeren van taken, of (3) een beoordeling door leerkrachten. Ook hier verwachten wij in elk van deze drie gevallen hetzelfde functionele verband met de studieresultaten.

Ten derde, zijn er de psychologische constructen die corresponderen met $S$, de situatiekenmerken. De kenmerken vertrouwdheid met het standpunt (in het voorbeeld over instemming) en onderwijsstijl (in het voorbeeld over ATI) zijn psychologische constructen van dit type. Elk van deze constructen wordt operationeel gedefinieerd door een set van *operaties* waarmee dit construct *gerealiseerd* kan worden. Zo kunnen verschillende graden van vertrouwdheid met een standpunt gecreëerd worden door (1) dit standpunt terloops te melden tijdens de rondvraag van de vorige vergadering, (2) dit standpunt kort uiteen te zetten in een nota die vóór de vergadering verspreid wordt, en (3) dit standpunt aan bod te laten komen in een bijdrage aan een lokale krant. In deze drie gevallen verwachten wij hetzelfde functionele verband met het gedrag waarin we geïnteresseerd zijn (de mate van instemming met het standpunt). En het construct onderwijsstijl kan operationeel gedefinieerd worden door leerkrachten in een democratische en een autoritaire groep in te delen op basis van (1) een zelfbeoordelingsvragenlijst waarin gepeild wordt naar iemands onderwijsstijl, (2) een beoordeling door ervaren beoordelaars van een lesuur gegeven door de betreffende leerkracht (opgenomen op video), of (3) een beoordeling door de leerlingen. Ook hier verwachten wij in elk van de drie gevallen hetzelfde functionele verband met het gedrag waarin we geïnteresseerd zijn (i.c.,

de studieresultaten).

Het gebruik van operationele definities heeft zowel een goede als een slechte kant. De goede kant is dat de psychologische constructen nu verankerd zijn in de empirie. Anders uitgedrukt, gegeven de operationele definities weten we precies waar we het over hebben als we spreken over een bepaald psychologisch construct. De slechte kant is dat er voor één psychologisch construct meerdere operationele definities gegeven kunnen worden die elk op andere manieren gerelateerd zijn met een operationele definitie van een ander psychologisch construct. Bijvoorbeeld, het kan zijn dat men eerder instemt met een standpunt waarmee men vertrouwd is dan met een standpunt waarmee men niet vertrouwd is, maar alléén als deze grotere vertrouwdheid terloops tot stand gekomen is (bijvoorbeeld, door het in de krant te lezen); als het standpunt expliciet verdedigd wordt tijdens een overleg, is er geen effect op de mate van instemming. Iets gelijkaardigs kan ook gebeuren in het ATI-onderzoek. In het ATI-onderzoek kan de voorspelde interactie gevonden worden als het construct zelfstandigheid operationeel gedefinieerd wordt op basis van beoordelingen door leerkrachten maar niet als het operationeel gedefinieerd wordt op basis van een zelfbeoordelingsvragenlijst.

### 2.1.4 Het Meten van Persoonskenmerken

Uit de vorige paragraaf onthouden we het volgende: datgene wát men wil meten is het psychologisch construct, en het instrument waarméé men dat wil doen is de operationele definitie. Wanneer we in de psychologie spreken over meten, dan bedoelen we bijna altijd het meten van persoonskenmerken. Men moet hier onder andere denken aan de volgende instrumenten: intelligentie- en vaardigheidstests, kennistoetsen, persoonlijkheidsvragenlijsten, reactietijden, leestijden, en beoordelingen door anderen (bijvoorbeeld, een beoordelaar geeft met een getal aan hoe sociaal vaardig hij iemand vindt). Al deze verschillende meetinstrumenten worden vaak aangeduid met de verzamelnaam *test*.

De aandacht voor het meten van persoonskenmerken betekent niet dat het meten van situatiekenmerken onbelangrijk zou zijn. Denk bijvoorbeeld aan het meten van de geluidshinder van verschillende types vliegtuigen in de buurt van een luchthaven. Het is een historisch feit dat meten binnen de psychologie vrijwel steeds geassocieerd wordt met persoonskenmerken. Wellicht is dit een gevolg van het feit dat een van de belangrijke toepassingen van de psychologie ligt in het selecteren en oriënteren van personen binnen onderwijs en bedrijfswereld.

We komen nu bij één van de grote problemen bij het meten van persoonskenmerken: het overgrote deel van de psychologisch constructen zijn niet op een eenduidige manier met een bepaalde operationele definitie (i.c., test) verbonden. Met andere woorden, het is niet duidelijk welke observeerbare gedragingen een manifestatie zijn van een bepaald psychologisch construct, en welke niet. De vraag is dan hoe we kunnen uitmaken of we met een goede of een slechte test te maken hebben. Hierop is geen echt goed antwoord te geven. Toch zijn de meeste psychologen tevreden met dit antwoord: een goede test moet *betrouwbaar* en *valide*. Onder *betrouwbaar* verstaan we het feit dat een test ongeveer dezelfde scores oplevert wanneer men hem afneemt onder vergelijkbare (in principe, identieke) condities. Het probleem van de betrouwbaarheid van een test vormt het onderwerp van de *Klassieke Testtheorie* die in de volgende paragraaf behandeld wordt.

*Valide* betekent *geldig*. In de literatuur onderscheidt men drie vormen van validiteit:

inhoudsvaliditeit, constructvaliditeit, en criteriumvaliditeit. Inhoudsvaliditeit betekent dat de test conform is met een definitie van het te meten psychologisch kenmerk. In zo'n definitie wordt gespecificeerd uit welk type items de test moet bestaan. Constructvaliditeit betekent dat een test voor een bepaald psychologisch kenmerk samenhangt met tests voor andere psychologische kenmerken op een wijze die conform is met een theorie over de samenhang tussen deze psychologische kenmerken. Met andere woorden, de validiteit van een test wordt hier opgehangen aan een theorie over het psychologisch kenmerk dat men wil meten. Criteriumvaliditeit, tenslotte, betekent dat je met de test een andere variabele kunt voorspellen. Meestal heeft de kennis van die voorspelde variabele een bepaalde nuttigheidswaarde (bijvoorbeeld, hoe goed zal iemand presteren in een bepaalde functie of opleiding?). Na de paragraaf over betrouwbaarheid, worden deze drie vormen van validiteit wat uitvoeriger besproken.

---

*Opgave 1*
Geef een omschrijving van

1. operationele definities van de $G$ in $G = f(P, S)$.
2. operationele definities van de $P$ in $G = f(P, S)$.
3. operationele definities van de $S$ in $G = f(P, S)$.

*Opgave 2*
Een onderzoeker was geïnteresseerd in de behandeling van depressie. Hij wilde nagaan wat het effect is van een nieuwe psychotherapievorm in vergelijking met een bestaande behandeling met pillen. Hij had de hypothese dat het effect van de behandelvorm afhangt van het feit of de patiënt gelooft in een psychogene dan wel een somatische oorsprong van zijn/haar problemen: bij patiënten die geloven in een psychogene oorsprong van hun problemen zal de psychotherapie een beter resultaat hebben dan de medicamenteuze behandeling, terwijl bij patiënten die geloven in een somatische oorsprong het omgekeerde het geval zal zijn. De patiënten werden op zuiver toevallige wijze toegewezen aan ofwel de behandeling met psychotherapie, ofwel de behandeling met pillen.

1. Beschrijf deze hypothese aan de hand van de formule $G = f(P, S)$ (waar staan $G$, $P$ en $S$ voor?) en beschrijf de functie $f$ aan de hand van een figuur.

Twee onderzoekers toetsten onafhankelijk van elkaar de bovenstaande hypothese. Ze maakten gebruik van verschillende maar volledig vergelijkbare patiëntengroepen. Beide onderzoekers namen een zuiver toevallige steekproef uit hetzelfde patiëntenbestand. In beide onderzoeken werd de ernst van de depressie vastgesteld met dezelfde vragenlijst, en werden de twee behandelvormen op dezelfde wijze toegepast (hetzelfde medicatievoorschrift, en twee groepen psychotherapeuten met dezelfde opleiding). Nu blijkt dat de gegevens van één onderzoeker overeenstemmen met de hypothese maar de gegevens van de andere onderzoeker niet.

2. Hoe kun je dit verschil verklaren aan de hand van datgene wat er in de tekst gezegd werd over psychologische constructen en operationele definities?

*Opgave 3*
Een therapeut die verbonden is aan de theateracademie heeft een behandeling ontwikkeld tegen plankenkoorts. Hij beweert dat zijn behandeling alleen effectief is bij mensen met een interne locus of control; bij mensen met een externe locus of control heeft de behandeling volgens hem geen effect. Om dit te onderzoeken, werden een aantal proefpersonen op zuiver toevallige wijze toegewezen aan de conditie mét of de conditie zónder behandeling.

1. Beschrijf deze hypothese aan de hand van de formule $G = f(P, S)$ (waar staan $G$, $P$ en $S$ voor?) en beschrijf de functie $f$ aan de hand van een figuur.

Twee onderzoekers toetsten onafhankelijk van elkaar de bovenstaande hypothese. Ze deden hun onderzoek met leerlingen van de theateracademie die last hadden van plankenkoorts. Van elke leerling werd eerst bepaald of hij/zij een interne of een externe locus of control had. Daarna werden zij op zuiver toevallige wijze toegewezen aan de conditie mét of de conditie zónder behandeling. Beide onderzoekers kregen de medewerking van dezelfde therapeut voor het geven van de behandeling. Nu bleek dat de gegevens van één onderzoeker overeenstemden met de hypothese, maar de gegevens van de andere onderzoeker niet.

2. Hoe kun je dit verschil verklaren aan de hand van datgene wat in de tekst gezegd wordt over psychologische constructen en operationele definities?

*Opgave 4*
Is de onderstaande stelling waar of onwaar? Indien ze onwaar is, motiveer dit.

In een onderzoek naar de invloed van het gedrag van een leraar op de leerprestaties van zijn leerlingen, is het gedrag van de leraar een persoonskenmerk.

---

## 2.2 Betrouwbaarheid

### 2.2.1 Betrouwbaarheid=Herhaalbaarheid

Veel meetresultaten worden uitgedrukt op een vrij grove schaal. Bijvoorbeeld, op de oude personenweegschalen (niet de digitale) kon je je gewicht aflezen met een nauwkeurigheid van ongeveer 200 gram. Zelfs indien deze weegschalen perfect zouden functioneren, dan nog zouden we op zijn best zo'n 100 gram van het correcte gewicht afwijken. Door nu gebruik te maken van een meetinstrument met een fijne schaalindeling (bijvoorbeeld, een digitale weegschaal) krijgt men de indruk dat de meetresultaten erg nauwkeurig zijn. Voor de personenweegschaal is dat vast ook wel zo, maar er zijn heel wat meetinstrumenten met een fijne schaalindeling die toch onnauwkeurige metingen produceren. Men krijgt pas echt een idee van de nauwkeurigheid als men de meetprocedure een aantal keer herhaalt bij dezelfde persoon. Bij metingen van iemands hartslag of bloeddruk krijgt men dan zeker behoorlijke schommelingen te zien. En met een fijnere schaalindeling krijgt men die schommelingen niet weg!

Het fenomeen van een bedrieglijke nauwkeurigheid van meetresultaten doet zich ook voor bij psychologische tests. Zeker wanneer de antwoorden op een objectieve manier gescoord worden (geen subjectieve interpretatie vanwege de beoordelaar), krijgt men makkelijk de indruk dat iemand die 30 items goed heeft wezenlijk anders is dan iemand die er 29 of 31 goed heeft. Het getal 30 is wezenlijk anders dan 29 en 31, maar als meting van iemands vaardigheid is het nog niet zo zeker dat een somscore van 30 wezenlijk anders is dan een somscore van 29 of 31. Zou een andere persoon met dezelfde vaardigheid ook een somscore van 30 halen? En als de persoon onder vergelijkbare omstandigheden opnieuw getest zou worden, zou hij dan opnieuw een somscore van 30 halen? Waarschijnlijk niet, en dit is de reden om onze aandacht te richten op *meetfouten*. We hebben het hier over meetfouten in de betekenis van toevallige schommelingen over herhaalde testafnamen.

De term "meetfouten" wordt nog in andere betekenissen gebruikt dan die van toevallige schommelingen over herhaalde testafnamen. Zo wordt ook wel eens van meetfouten gesproken als een persoon foute antwoorden geeft die het gevolg zijn van het op een onjuiste manier afnemen van de test (bijvoorbeeld, door verkeerde instructies). Dergelijke foute antwoorden worden géén meetfouten genoemd, omdat ze systematisch zijn; de foute antwoorden zouden opnieuw gegeven worden indien men de test nogmaals zou afnemen. (Hierbij gaan we er stilzwijgend vanuit dat geheugeneffecten geen rol spelen; we veronderstellen dat de persoon de eerder gegeven antwoorden vergeten is. In de praktijk kan men daar natuurlijk niet zeker van zijn. We komen hier later nog op terug.) Een ander voorbeeld van een situatie waarin van meetfouten gesproken wordt, maar niet in onze betekenis van toevallige schommelingen, is het geven van juiste antwoorden op meerkeuze items zonder dat men het juiste antwoord kent. Sommige personen slagen er immers in om het juiste antwoord te geven door de onjuiste antwoordalternatieven te elimineren. Men kiest dan het (juiste) overblijvende antwoordalternatief, ook al zou men dit antwoord niet kunnen geven indien het een open vraag was en geen meerkeuzevraag. Dergelijke antwoorden zijn geen meetfouten omdat ze systematisch zijn; ze zouden opnieuw gegeven worden indien men de test weer zou afnemen.

Onderzoekers in fysische laboratoria hebben ook vaak met niet-perfect-herhaalbare metingen te maken. Als het voor hen belangrijk is om betrouwbare gegevens te hebben, dan herhalen zij hun metingen een aantal keer en berekenen het gemiddelde. De gemiddelde meting is veel nauwkeuriger dan de afzonderlijke metingen, en bovendien krijg je informatie over de máte van herhaalbaarheid door te kijken naar de variabiliteit over de herhaalde metingen.

In de psychologie is het herhalen van metingen jammer genoeg niet zo makkelijk als in de natuurwetenschappen. Men moet bijvoorbeeld altijd rekening houden met geheugeneffecten (het zich herinneren van het vorige antwoord), vermoeidheid, en misschien zelfs het niet bereid zijn om nog verder mee te werken. Bij het toepassen van de idee "betrouwbaarheid=herhaalbaarheid" op psychologische tests, bekijken we hier de herhaalbaarheid van een testscore over *equivalente tests* in plaats van over dezelfde test. Op dit moment is het voldoende om te weten dat twee tests equivalent zijn als ze hetzelfde meten. Later zullen we zien dat equivalente tests een belangrijke rol spelen in de Klassieke Testtheorie. We zullen ze zelfs een formele definitie geven, en dan noemen we ze *parallelle tests*.

In Tabel 2.1 geven we de scores van 12 van in totaal 80 personen op twee equivalente tests voor hoofdrekenen, A en B. Deze personen werden geordend naar hun score op test A. Sommige personen behaalden ongeveer dezelfde score op de twee tests, maar voor een aantal anderen waren er toch behoorlijke verschillen. Bijvoorbeeld, persoon 1, 3 en 6 scoorden aanmerkelijk beter op test A dan op test B. En persoon 75, 76 en 80 scoorden aanmerkelijk beter op test B dan op test A. Gegeven dat de twee tests hetzelfde meten (want ze zijn equivalent), stellen we hier vast dat er behoorlijke meetfouten in het spel zijn.

Door onze aandacht te richten op de personen die sterk verschillend scoren op de twee tests, krijgen we misschien de indruk dat er helemáál geen verband is tussen de twee sets van scores. Dat dit niet het geval is, blijkt uit het feit dat de personen die hoog scoorden op test A gemiddeld hoger scoorden op test B dan de personen die laag scoorden op test

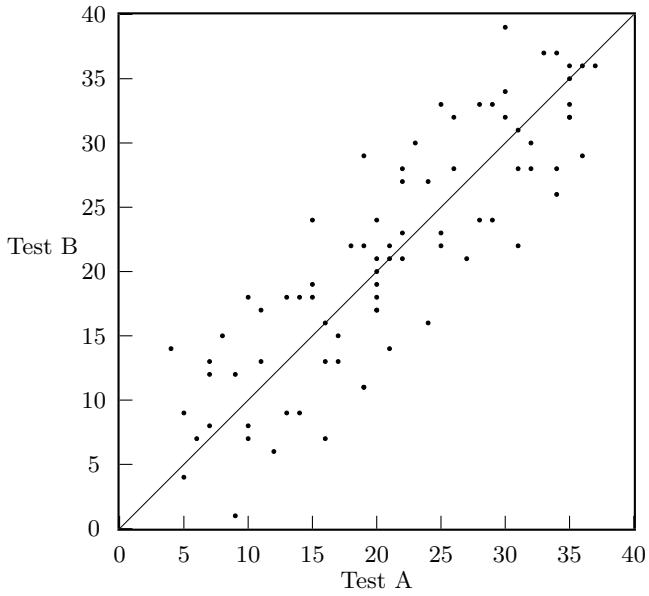
| Proef-persoon | Test A | Test B |
|---|---|---|
| 1 | 37 | 26 |
| 2 | 36 | 36 |
| 3 | 36 | 29 |
| 4 | 35 | 32 |
| 5 | 35 | 32 |
| 6 | 34 | 28 |
| ⋮ | ⋮ | ⋮ |
| 75 | 7 | 13 |
| 76 | 7 | 12 |
| 77 | 6 | 7 |
| 78 | 5 | 9 |
| 79 | 5 | 4 |
| 80 | 4 | 14 |

Tabel 2.1: Scores van 12 van de 80 personen op twee equivalente tests voor hoofdrekenen (A en B).

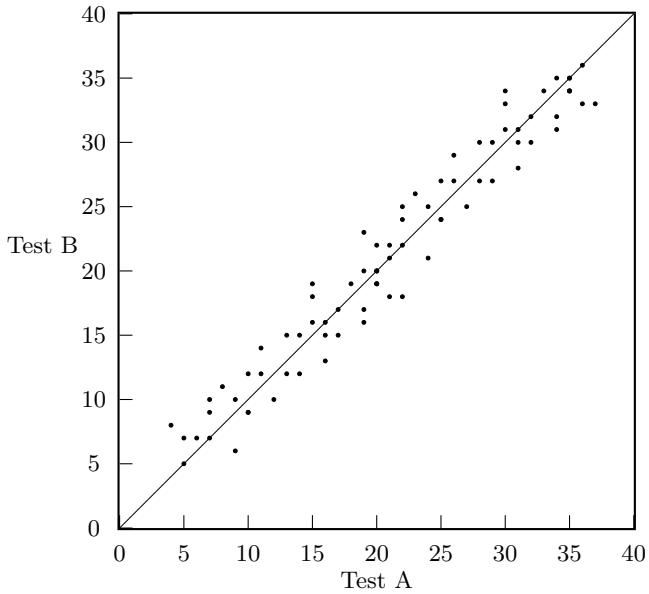
A. Het verband tussen de twee tests kunnen we nog beter zien als we de scores van alle 80 personen weergeven in een spreidingsdiagram. Dit wordt getoond in Figuur 2.2. Als elke persoon precies dezelfde score zou halen op de twee tests, dan zouden alle scores op de diagonale lijn liggen. Deze lijn definieert perfecte herhaalbaarheid van de metingen. Het feit dat de punten gespreid liggen rond deze lijn, geeft aan dat onze meting van hoofdrekenvaardigheid niet perfect betrouwbaar is. De mate waarin de punten gespreid liggen rond de diagonale lijn, is een indicatie van de graad van onbetrouwbaarheid. Naarmate de punten sterker afwijken van deze lijn, zal de rangorde van de personen op basis van hun score op test A sterker afwijken van de rangorde op basis van hun score op test B.

Om te verduidelijken dat we met verschillende graden van betrouwbaarheid te maken kunnen hebben, tonen we in Figuur 2.3 een spreidingsdiagram van de scores op twee equivalente woordenschattests. Hier zien we dat de mate van spreiding ten opzichte van de diagonale lijn veel kleiner is dan in Figuur 2.2. De woordenschattest is dus betrouwbaarder dan de hoofdrekentest.

We hebben hier gekeken naar de herhaalbaarheid (=betrouwbaarheid) van testscores door twee equivalente tests elk één keer af te nemen. Ook andere soorten herhalingen zijn mogelijk. Zo kunnen we bijvoorbeeld kijken naar een herhaalde afname van dezelfde test onder vergelijkbare condities. Omwille van de storende invloed van onder andere geheugeneffecten zal het in de praktijk vrijwel steeds onmogelijk zijn om de condities vergelijkbaar te houden. Dit belet ons echter niet om een gedachtenexperiment op te zetten waarin we een test een groot aantal keren afnemen onder hypothetisch vergelijkbare condities. Het nut van dit gedachtenexperiment is dat het ons een precieze definitie oplevert van betrouwbaarheid. De uitwerking van dit gedachtenexperiment is de kern van de Klassieke Testtheorie. Met deze theorie wordt een begin gemaakt in de volgende




Figuur 2.2: Spreidingsdiagram van de scores op de hoofdrekentests A en B voor 80 personen.




Figuur 2.3: Spreidingsdiagram van de scores op de woordenschattests A en B voor 80 personen.

paragraaf.

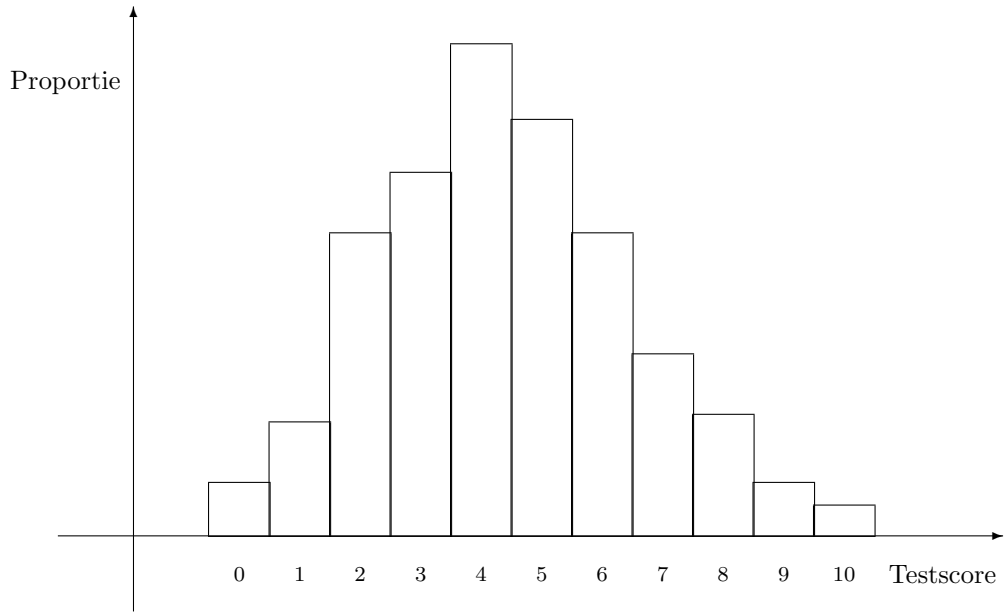
De Klassieke Testtheorie is nauw verwant aan de *Generaliseerbaarheidstheorie*. Het verschil tussen de twee theorieën vloeit voort uit een verschillend uitgangspunt: in de Klassieke Testtheorie vertrekt men van een hypothetisch experiment waarbij men een groot aantal keren dezelfde test afneemt, en in de Generaliseerbaarheidstheorie vertrekt men van een test die slechts één keer wordt afgenomen maar waarvan de items at random zijn getrokken uit een *universum* van mogelijke items. Sommige formules uit de Klassieke Testtheorie vind je ook in de Generaliseerbaarheidstheorie, maar hun interpretatie is daar anders. Een voordeel van de Generaliseerbaarheidstheorie is dat zij een onderscheid maakt tussen verschillende types van meetfouten, zoals meetfouten die toe te schrijven zijn aan (1) de samenstelling van de test, (2) de beoordelaars van de antwoorden, (3) het tijdstip van de afname, .... Jammer genoeg, is de Generaliseerbaarheidstheorie wat moeilijker dan de Klassieke Testtheorie. In dit boek beperken we ons tot de Klassieke Testtheorie.

### 2.2.2 De Basisformule van de Klassieke Testtheorie

Stel dat we bij één persoon een bepaalde test een oneindig aantal keer afnemen, en we doen dit onder (hypothetisch) vergelijkbare condities. Stel dat deze test een score oplevert tussen 0 en 10. We kunnen dan bijhouden hoe vaak elk van deze 11 verschillende scores geobserveerd werd bij deze persoon. Men kan dan de proportie berekenen van het totaal aantal observaties dat in elk van deze 11 scores resulteerde (door het aantal observaties per score te delen door het totaal aantal observaties). Deze proporties kan men dan in een histogram uitzetten. In Figuur 2.4 wordt zo'n histogram getoond.

Tot nu toe werd nog niet omschreven wat we bedoelen met "vergelijkbare condities". Voor we daar een omschrijving van geven, maken we een onderscheid tussen twee groepen factoren die verantwoordelijk kunnen zijn voor schommelingen in de testscores: de *endogene* en de *exogene* factoren. Onder endogene factoren verstaan we factoren ín de persoon, zoals wisselende aandacht en tijdelijke vermoeidheid. Endogene factoren zijn verantwoordelijk voor schommelingen in de testscores die optreden bij identieke fysieke condities. Onder exogene factoren verstaan we factoren buíten de persoon die de testprestatie beïnvloeden. Hierbij moet men denken aan geluiden in de omgeving die de aandacht afleiden, bekeken worden door omstanders, onderbrekingen van de testafname, en dergelijke. Een belangrijk verschil tussen endogene en exogene factoren, is dat de invloed van exogene factoren in principe uit te sluiten is, en dat dit niet mogelijk is bij endogene factoren. De invloed van exogene factoren is uit te sluiten door de fysieke condities constant te houden. Voor sommige exogene factoren, is het echter zeer moeilijk om hun invloed uit te sluiten. Dit geldt bijvoorbeeld voor geluiden in de omgeving. Voor de invloed van deze factoren doen we dan, bij wijze van spreken, een oogje dicht, en beschouwen het als meetfout. We zijn nu klaar om een precieze omschrijving te geven van "vergelijkbare condities": twee condities zijn vergelijkbaar als ze alleen verschillen wat betreft exogene factoren waarvan we de invloed niet kunnen controleren. De gezamenlijke invloed van (1) de niet-gecontroleerde exogene factoren en (2) de endogene factoren die wisselen over de testafnamen, beschouwen we dan als meetfout.

We introduceren nu een aantal symbolen. De testscores van persoon $p$ duiden we aan met $X_p$. Deze $X_p$ is een toevalsvariabele die 11 mogelijke waarden kan aannemen (van 0




Figuur 2.4: Histogram van de proporties observaties binnen elk van de 11 verschillende testscores.

tot 10). De toevalsverdeling van $X_p$ wordt gegeven door het histogram in Figuur 2.4: de hoogte van elk van de staafjes in het histogram is de kans om de bijbehorende testscore te observeren. Eén testscore noemen we een *trekking* uit deze toevalsverdeling. Als we de toevalsverdeling van de testscores van persoon $p$ met één getal willen karakteriseren, dan is het een goede keuze om hiervoor de *verwachte waarde* van deze verdeling te nemen. De verwachte waarde van een toevalsverdeling[2] is gelijk aan het gemiddelde van een oneindig aantal trekkingen uit deze verdeling. De verwachte waarde van de toevalsvariabele $X_p$ is dan het gemiddelde van een oneindig aantal testscores van persoon $p$. Deze verwachte waarde wordt aangeduid met $T_p$. We kunnen $T_p$ berekenen door elke mogelijke waarde van $X_p$ te vermenigvuldigen met zijn kans van voorkomen, en dan alles op te tellen. De kansen van de 11 mogelijke waarden van $X_p$ zijn de volgende (in volgorde van oplopende score): 0.02, 0.05, 0.13, 0.15, 0.20, 0.17, 0.12, 0.08, 0.05, 0.02 en 0.01. De verwachte waarde is dan de volgende:

$$\begin{aligned}\mathcal{E}(X_p) =& 0.02 \times 0 + 0.05 \times 1 + 0.13 \times 2 + 0.15 \times 3 + 0.20 \times 4 \\ &+ 0.17 \times 5 + 0.12 \times 6 + 0.08 \times 7 + 0.05 \times 8 + 0.02 \times 9 + 0.01 \times 10 \\ =& 4.37 \quad .\end{aligned}$$

We noemen $T_p$ de *ware-score* van persoon $p$ (de letter $T$ komt van het Engelse *true score*). De term "ware-score" wordt makkelijk fout geïnterpreteerd. Met ware-score bedoelen we niet wat iemand "echt waard is" op een test. Een ware-score is niets meer

[2]Men spreekt ook van de verwachte waarde van een toevals*variabele*.

of minder dan wat gesteld wordt in zijn definitie: het is het gemiddelde van iemands testscore over een oneindig aantal herhalingen. Stel dat iemand erg goed is in het beantwoorden van meerkeuzevragen, gewoon omdat hij via simpele regeltjes de onjuiste alternatieven kan elimineren. Zonder de antwoorden te kennen, kan hij op deze manier toch een redelijke score halen. Van deze persoon vinden we dus niet dat hij veel kennis heeft op het gebied dat in deze test getoetst wordt. Toch zal deze persoon een redelijke ware-score hebben. Deze overschatting van zijn kennis wordt hier niet gezien als het gevolg van meetfouten. We zien het wel als een onzuiverheid die het gevolg is van een speciale "vaardigheid" in het oplossen van meerkeuzevragen, die niets te maken heeft met de vaardigheid (kennis) die men met deze test wil meten. Onder meetfouten verstaan we alleen de spreiding rónd de ware-score, en niet de onzuiverheden in de ware-score zelf.

De meetfouten van persoon $p$ duiden we aan met $E_p$. Deze meetfouten zijn het verschil tussen de geobserveerde score $X_p$ en de ware-score $T_p$:

$$E_p = X_p - T_p \quad .$$

Net zoals $X_p$, is ook $E_p$ een toevalsvariabele. In het voorbeeld kan $E_p$ 11 mogelijke waarden aannemen. We vinden deze waarden door het getal 4.37 ($= T_p$) af te trekken van elk van de 11 mogelijke testscores (van 0 tot 10). De toevalsverdeling van $E_p$ wordt getoond in Figuur 2.5. Het is belangrijk om te zien dat de verwachte waarde van $E_p$ gelijk is aan 0. In de vorm van een formule:

$$\mathcal{E}(E_p) = 0 \quad ,$$

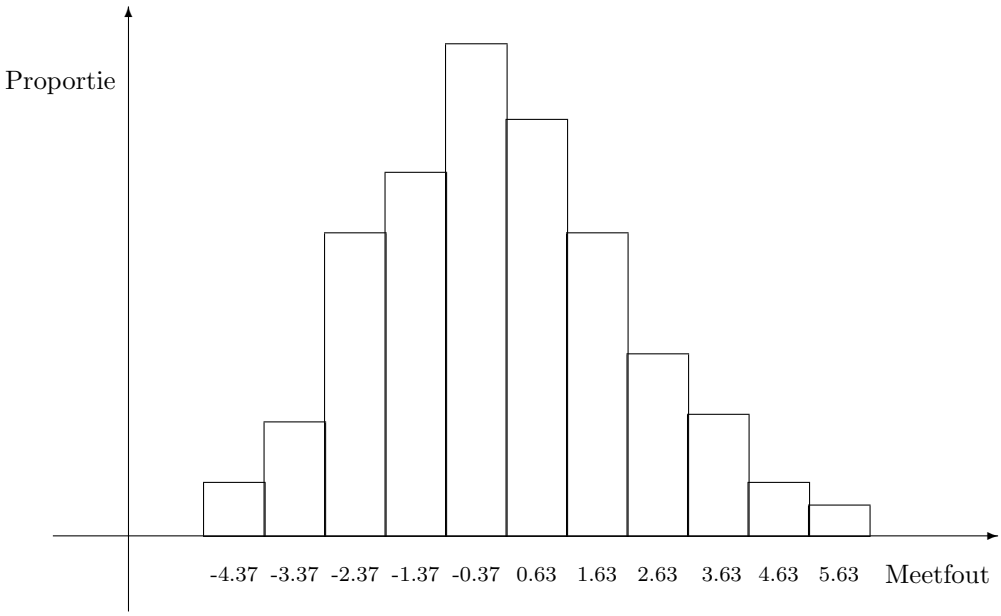
waarin $\mathcal{E}$ staat voor "expected value", het Engels voor "verwachte waarde". We krijgen dit resultaat als we de verwachte waarde nemen van $X_p - T_p$. De verwachte waarde van het verschil $X_p - T_p$ is gelijk aan het verschil tussen de verwachte waarden van $X_p$ en $T_p$:

$$\mathcal{E}(X_p - T_p) = \mathcal{E}(X_p) - \mathcal{E}(T_p) \quad .$$

Omdat de verwachte waarde van de constante $T_p$ triviaal gelijk is aan deze constante, zijn de verwachte waarden $\mathcal{E}(X_p)$ en $\mathcal{E}(T_p)$ beiden gelijk aan $T_p$. Hieruit volgt dat $\mathcal{E}(E_p)$ ($= \mathcal{E}(X_p - T_p)$) gelijk is aan 0.

De betrouwbaarheid van de testscores is rechtstreeks gerelateerd aan de variantie van de meetfouten. Deze variantie duiden we aan met $\text{VAR}(E_p)$. Naarmate de meetfouten meer gespreid liggen rond het punt 0, zijn de testscores minder betrouwbaar (want, minder herhaalbaar). Om een idee te krijgen van $\text{VAR}(E_p)$ zouden we de test een groot aantal keren moeten afnemen. Welnu, dat kan niet in de praktijk. Daarom is er een methode ontwikkeld waarmee we op basis van een groot aantal personen die de test slechts twéé keer afleggen een idee krijgen van de gemiddelde $\text{VAR}(E_p)$ van deze personen [3]. Dit leggen we nu uit, vertrekkende van de basisformule van de Klassieke Testtheorie.

[3] Dit is niet helemaal in overeenstemming met wat we later zullen zeggen. Daar zullen we het hebben over twee zogenaamde *parallelle tests* die elk één keer afgelegd worden. Voorlopig is het onderscheid tussen een herhaalde testafname en de afname van twee parallelle tests nog niet van belang.




Figuur 2.5: Toevalsverdeling van de meetfouten voor de verdeling van de testscores in Figuur 2.4.

De basisformule van de Klassieke Testtheorie krijgen we door de definitie van meetfouten als volgt te herschrijven:

$$\boxed{X_p = T_p + E_p \quad .}$$

In woorden, staat er in deze formule dat een testscore gelijk is aan een ware-score plus een meetfout. Deze formule geldt voor elke willekeurige persoon $p$. Daarom kunnen we de index $p$ ook weglaten. Met de variabele $X$ duiden we de testscores aan in een populatie van personen. De p-de persoon in deze populatie heeft testscore $X_p$. Op dezelfde manier duiden we met $T$ en $E$ de ware scores en de meetfouten aan in deze populatie. De basisformule voor de hele populatie is dan:

$$\boxed{X = T + E \quad .}$$

Aan de hand van Tabel 2.2 kunnen we het verschil tussen de twee bovenstaande formules toelichten. In deze tabel staan de hypothetische testscores van alle personen in de populatie op een oneindig aantal herhalingen van de testafname. We veronderstellen dat het aantal personen in de populatie oneindig ($\infty$) is. De formule $X_p = T_p + E_p$ heeft betrekking op één rij van deze tabel en de formule $X = T + E$ heeft betrekking op de gehele tabel. In elke rij, tussen de twee dubbele lijnen, staan de testscores van één persoon op een oneindig aantal herhalingen van de testafname. Deze testscores worden aangeduid met $X_p$. In de meest rechtse kolom staat de ware-score. Voor een vaste

persoon, is deze ware-score een constante, maar over de personen heen varieert hij. Als men, voor elke persoon (rij), de ware-score $T_p$ aftrekt van de geobserveerde scores $X_p$, dan houdt men de meetfouten $E_p$ over. De $X$ in de formule $X = T + E$ staat voor het binnenste deel van Tabel 2.2 (het deel dat begrensd wordt door de dubbele lijnen), de $T$ staat voor de meest rechtse kolom, en de $E$ staat voor het binnenste deel na aftrekking van de rechtse kolom.

| | | Herhalingen | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | $\cdots$ | $\infty$ | $T$ |
| Personen | 1 | 10 | 8 | $\cdots$ | 9 | 8.74 |
| | 2 | 4 | 2 | $\cdots$ | 5 | 3.15 |
| | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| | $p$ | 6 | 4 | $\cdots$ | 3 | 4.37 |
| | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| | $\infty$ | 4 | 7 | $\cdots$ | 5 | 6.61 |

Tabel 2.2: Hypothetische datamatrix met de testscores van een oneindig aantal personen op een oneindig aantal herhalingen.

De variabelen $E$ en $T$ blijken enkele handige eigenschappen te hebben. Ten eerste, de verwachte waarde van $E$ is gelijk aan 0. In de vorm van een formule:

$$\mathcal{E}(E) = 0 \quad .$$

Dit volgt onmiddellijk uit het feit dat de verwachte waarde van de meetfouten van één persoon gelijk is aan 0 ($\mathcal{E}(E_p) = 0$). Als dit geldt voor elk van de personen afzonderlijk, dan geldt het ook voor de hele populatie. Ten tweede, de correlatie tussen $T$ en $E$ is gelijk aan 0. In de vorm van een formule:

$$r_{TE} = 0 \quad .$$

In woorden, betekent dit dat er geen samenhang is tussen de ware-scores en de meetfouten. Concreet, het is niet zo dat grote ware-scores samengaan met positieve meetfouten en kleine ware-scores met negatieve, en het omgekeerde geldt ook niet. Deze nulcorrelatie volgt uit het feit dat de verwachte waarde van de meetfouten voor elke persoon gelijk zijn aan 0 ($\mathcal{E}(E_p) = 0$). Dus, er kan geen samenhang zijn tussen $T$ en $E$, want bij elke waarde van $T$ is de verwachte waarde van de bijbehorende $E$'s even groot (namelijk, 0).

### 2.2.3 De Definitie van Betrouwbaarheid

In het voorgaande hebben we betrouwbaarheid gelijkgesteld aan herhaalbaarheid. We zijn nu klaar om betrouwbaarheid wat preciezer te definiëren. We doen dit vertrekkende van een formule voor de variantie van $X$. Deze formule volgt uit de toepassing van de volgende algemene regel: de variantie van een som van twee ongecorreleerde variabelen is gelijk aan de som van hun beider varianties. Welnu, omdat $T$ en $E$ ongecorreleerd zijn, volgt uit de formule $X = T + E$ dat

$$\boxed{\mathrm{VAR}(X) = \mathrm{VAR}(T) + \mathrm{VAR}(E) \quad .}$$

Dus, de variantie van de testscores is gelijk aan de variantie van de ware-scores plus de variantie van de meetfouten. Men kan laten zien dat $\mathrm{VAR}(E)$ gelijk is aan het gemiddelde van de persoonsspecifieke $\mathrm{VAR}(E_p)$'s in de populatie. We noemen deze formule de *variantie-ontbinding* (omdat $\mathrm{VAR}(X)$ ontbonden wordt in $\mathrm{VAR}(T)$ plus $\mathrm{VAR}(E)$).

De afleiding van de bovenstaande formule wordt gegeven op pagina 123. In deze afleiding wordt gebruik gemaakt van de formule van de variantie van een som van twee variabelen (hier, $T$ en $E$). Deze formule wordt afgeleid op pagina 123. In deze twee afleidingen, en ook in afleidingen die later in dit hoofdstuk aan bod komen, wordt gebruik gemaakt van het begrip "covariantie" alsook van de relatie tussen de covariantie en de correlatie. Daarom wordt op pagina 122 uitgelegd wat de covariantie is en wat zijn relatie is met de correlatie.

We hebben te maken met betrouwbare testscores als $\mathrm{VAR}(T)$ groot is in vergelijking met $\mathrm{VAR}(E)$. In dat geval bestaat $\mathrm{VAR}(X)$ vooral uit ware-scorevariantie ($\mathrm{VAR}(T)$) en slechts voor een klein gedeelte uit meetfoutvariantie ($\mathrm{VAR}(E)$). Een equivalente manier om dit uit te drukken is zeggen dat $\mathrm{VAR}(T)$ kort bij $\mathrm{VAR}(X)$ ligt. We definiëren nu betrouwbaarheid als de verhouding van $\mathrm{VAR}(T)$ over $\mathrm{VAR}(X)$. In de vorm van een formule:

$$\boxed{\text{Betrouwbaarheid} = \frac{\mathrm{VAR}(T)}{\mathrm{VAR}(X)} \quad .}$$

We kunnen de betrouwbaarheid ook uitdrukken als functie van $\mathrm{VAR}(E)$, namelijk door $\mathrm{VAR}(T)$ te vervangen door $\mathrm{VAR}(X) - \mathrm{VAR}(E)$, hetgeen volgt uit de variantie-ontbinding. We krijgen dan de volgende uitdrukking:

$$\text{Betrouwbaarheid} = 1 - \frac{\mathrm{VAR}(E)}{\mathrm{VAR}(X)} \quad .$$

Dus, de betrouwbaarheid neemt toe als het aandeel van $\mathrm{VAR}(E)$ in $\mathrm{VAR}(X)$ kleiner wordt.

Uit haar definitie volgt dat de betrouwbaarheid een getal is tussen 0 en 1 dat groter is naarmate het aandeel van $\mathrm{VAR}(E)$ in $\mathrm{VAR}(X)$ kleiner is. Als de testscores alleen uit meetfouten bestaan, dan is de betrouwbaarheid gelijk aan 0, en als de testscores vrij zijn van meetfouten dan is de betrouwbaarheid gelijk aan 1.

### 2.2.4 Parallelle Tests

Op dit moment hebben we alleen nog maar een definitie van betrouwbaarheid. We weten nog niet hoe we die betrouwbaarheid kunnen berekenen in de praktijk. De parallelle test gaat ons hierbij helpen. We moeten hier onmiddellijk bij opmerken dat parallelle tests eigenlijk een theoretisch begrip is. Van twee concrete tests kan men nooit met zekerheid zeggen of ze parallel zijn. Men kan alleen zeggen dat parallelliteit zeer of niet zeer plausibel is. Ons vertrekpunt is de basisformule van de Klassieke Testtheorie die we toepassen op de scores van twee tests, aangeduid met $X$ en $X'$:

$$\begin{aligned} X_p &= T_p + E_p \\ X'_p &= T'_p + E'_p \quad . \end{aligned}$$

We zeggen nu dat de twee tests parallel zijn als:

1. iedere persoon dezelfde ware-score heeft op de twee tests ($T_p = T'_p$), en
2. beide tests dezelfde meetfoutvariantie hebben ($\text{VAR}(E) = \text{VAR}(E')$)
3. de meetfouten van de twee tests ongecorreleerd zijn ($r_{EE'} = 0$).

Kort uitgedrukt, twee tests zijn parallel als ze hetzelfde meten (=1), als ze dat even precies doen (=2), en als hun meetfouten ongecorreleerd zijn (=3).

Een situatie waarin voldaan kan zijn aan de voorwaarden van parallelle tests, is wanneer men van een test twee equivalente vormen maakt. Voor elk item in de eerste vorm heeft men een item in de tweede vorm dat er zoveel mogelijk op lijkt. "Zoveel mogelijk op lijken" kun je je voorstellen als het stellen van dezelfde vraag met andere woorden (bijvoorbeeld, met "bang" in plaats van "angstig") of het voorleggen van hetzelfde probleem met ander materiaal (bijvoorbeeld, "liters melk" in plaats van "kilo's appels" in een rekenvraagstuk). Een bekende test waarvoor twee equivalente vormen werden gemaakt, is een intelligentietest die meestal *de Terman* genoemd wordt (naar zijn maker, de Amerikaan Terman). Een andere situatie waarin voldaan kan zijn aan de voorwaarden van parallelle tests, is wanneer dezelfde test twee keer afgenomen wordt onder identieke condities (zonder geheugeneffecten, en dergelijke). Men beschouwt deze twee afnamen dan als twee verschillende tests. Dat deze twee tests uit dezelfde fysieke items bestaan is geen probleem, want parallelliteit is gedefinieerd op het niveau van ware-scores en meetfouten: wáár die precies vandaan komen doet er strikt genomen niet toe.

Volgens sommige onderzoekers correleren meetfouten met geen enkele andere variabele. Voor hen is het daarom overbodig om de derde eigenschap van parallelle tests (ongecorreleerde meetfouten) te noemen. Het ongecorreleerd-zijn van meetfouten is voor hen immers een onderdeel van de definitie van meetfout. Echter, niet iedereen defineert meetfouten op deze manier en daarom is het noemen van de derde eigenschap noodzakelijk.

Het rechtstreeks controleren van de drie eigenschappen van parallelle tests is onmogelijk, gewoon omdat de ware-scores en de meetfouten niet observeerbaar zijn. Een ónrechtstreekse controle is echter wel mogelijk. Men kan namelijk een aantal eigenschappen van parallelle tests afleiden, die men wel kan controleren. Twee parallelle tests hebben namelijk:

1. gelijke gemiddelden ($\mathcal{E}(X) = \mathcal{E}(X')$)
2. gelijke varianties ($\text{VAR}(X) = \text{VAR}(X')$), en
3. gelijke correlaties met andere tests ($r_{XY} = r_{X'Y}$, waarin $Y$ staat voor de score op een andere test).

Deze eigenschappen moeten gelden in de populatie. In een steekproef hoeven deze eigenschappen niet precies op te gaan, maar de afwijkingen moeten wel zo klein zijn dat ze te verklaren zijn als toevalsfluctuaties. Of het mogelijk is de afwijkingen te verklaren als steekproeffluctuaties, kan statistisch getoetst worden.

De nuttigheid van parallelle tests blijkt als we de correlatie berekenen tussen $X$ en $X'$ (aangeduid met $r_{XX'}$). Het volgende blijkt namelijk te gelden:

$$\boxed{r_{XX'} = \frac{\text{VAR}(T)}{\text{VAR}(X)} = \text{Betrouwbaarheid} \quad .}$$

(Deze formule wordt afgeleid op pagina 124.) Dus, de betrouwbaarheid is precies gelijk aan de correlatie tussen twee parallelle tests. Dus, als we over twee parallelle tests beschikken, dan kunnen we de betrouwbaarheid bepalen. Vanaf nu zullen we de betrouwbaarheid aanduiden met $r_{XX'}$.

---

*Opgave 5*
Welke van de onderstaande uitspraken zijn waar, en welke onwaar? Als een uitpraak onwaar is, leg uit waarom dat zo is.

1. Een groep leerlingen wordt getest met twee parallelle tests. Als de scores uitgezet worden in een spreidingsdiagram, dan blijkt dat alle punten op een lijn liggen die met een hoek van 45 graden door de oorsprong gaat. De betrouwbaarheid van deze tests is 1.
2. Een woordenschattest blijkt hoog te correleren met een hoofdrekentest. Dit is voldoende reden om te twijfelen aan het feit dat deze woordenschattest een ware-score heeft.

*Opgave 6*
Geef voor elk van de volgende situaties aan of er sprake is van meetfouten.

1. Een leerling die een wiskundetoets aflegt wordt gestoord door lawaai in de gang, en maakt daardoor enkele fouten.
2. Peter is zo angstig als hij een toets moet afleggen dat hij verschillende vragen onbeantwoord laat.
3. Door een fout in het computerprogramma dat de scores op de eindexamens verwerkt, kregen alle leerlingen 10 punten meer dan ze eigenlijk verdienden.
4. Sandra heeft tijdens de colleges één van de uitspraken van de docent fout opgeschreven. Daardoor beantwoordt zij een tentamenvraag die te maken heeft met deze informatie fout.
5. Een observator die de interventies van therapeuten-in-opleiding beoordeelt, geeft hogere scores aan vrouwelijke therapeuten dan aan mannelijke, ook al gedragen ze zich precies hetzelfde.

*Opgave 7*
Hein behaalde een IQ van 135 op een papier-en-pen-intelligentietest. Korte tijd later werd hij getest met een individuele intelligentietest, en hij behaalde een IQ van 110. Het verschil tussen 110 en 135 is te groot om verklaard te kunnen worden door meetfouten. Heins docent begrijpt niet hoe dit mogelijk is, en vraagt zich af welk van de twee testscores een betere meting is van Heins ware-score. Hoe kun je dit uitleggen, vertrekkende van de Klassieke Testtheorie.

*Opgave 8*
Welke van de volgende beweringen zijn waar in de Klassieke Testtheorie? Indien een bewering onwaar is, leg uit waarom dit zo is.

1. De correlatie tussen $X$ en $E$ is nul.
2. De correlatie tussen $T$ en $E$ is nul.
3. De correlatie tussen $T$ en $X$ is nul.
4. De verwachte waarde van $E$ is nul.
5. De variantie van $E$ is nul.
6. De varianties van $X$ en $T$ zijn gelijk.

*Opgave 9*
Met welke van de onderstaande uitspraken ben je het eens en met welke ben je het oneens? Indien je het oneens bent met een bewering, leg dan uit waarom.

1. Als twee tests een gelijke, hoge betrouwbaarheid hebben, zijn ze ook in redelijke mate parallel.
2. Als twee tests parallel zijn, hebben ze ook een redelijk hoge betrouwbaarheid.

---

### 2.2.5 Testverlenging en Betrouwbaarheid

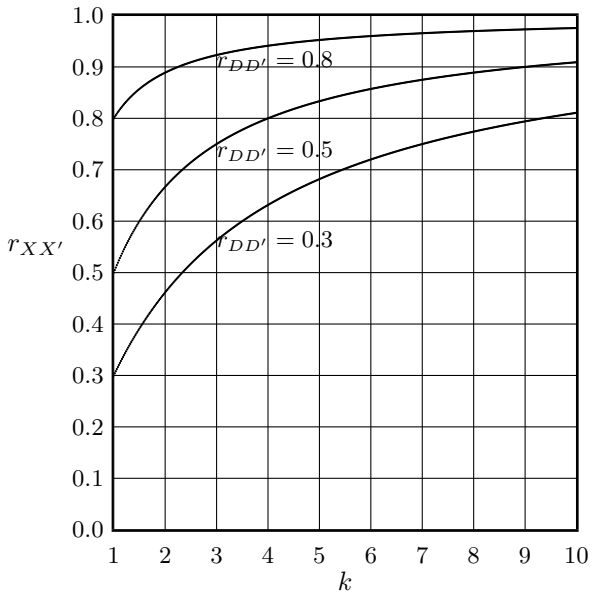
Stel dat men over twee parallelle tests beschikt, elk met een betrouwbaarheid van 0.70 (parallelle tests hebben altijd dezelfde betrouwbaarheid, want hun $\text{VAR}(T)$ en $\text{VAR}(X)$ zijn gelijk). Nu worden die twee tests samengevoegd; van twee tests wordt één test gemaakt. De score op de samengestelde test is gelijk aan de som van de scores op de twee deeltests. De betrouwbaarheid van deze score op de samengestelde test zal groter zijn dan de betrouwbaarheid van de scores op de deeltests. Het principe dat hierachter zit is hetzelfde dat maakt dat het gemiddelde van een steekproef des te korter bij de verwachte waarde (i.e., het gemiddelde in de populatie) ligt naarmate het op meer observaties gebaseerd is.

Er bestaat een formule waarmee men kan uitrekenen wat de betrouwbaarheid is van een test die bestaat uit twee parallelle deeltests, gegeven dat men weet wat de betrouwbaarheid is van deze deeltests. Deze formule is zelfs algemener, in de zin dat ze gebruikt kan worden voor het bepalen van de betrouwbaarheid van test die bestaat uit $k$ parallelle deeltests, waarbij $k$ alle mogelijk waarden kan aannemen (e.g., 2, 3, 5). Deze formule heet de *Spearman-Brown* (S-B) *formule* :

$$\boxed{r_{XX'} = \frac{k r_{DD'}}{1 + (k-1) r_{DD'}} \quad .}$$

In deze formule staat $r_{XX'}$ voor de betrouwbaarheid van de uit $k$ parallelle deeltests samengestelde test $X$, en $r_{DD'}$ voor de betrouwbaarheid van elk van de parallelle deeltests. De afleiding van de S-B formule wordt gegeven op pagina 124, maar alleen voor het speciale geval van twee parallelle deeltests ($k = 2$). De afleiding voor het algemene geval (een willekeurige $k$) verloopt analoog, maar maakt gebruik van een formule die in dit boek niet behandeld wordt.

Aan de vorm van de S-B formule is te zien dat de betrouwbaarheid toeneemt naarmate de test langer wordt ($k$ groter wordt). Hoe langer de test, hoe dichter de betrouwbaarheid bij 1 komt (perfecte betrouwbaarheid). In Figuur 2.6 wordt de betrouwbaarheid uitgezet tegen de verlengingsfactor $k$. Dit gebeurt voor drie verschillende waarden van $r_{DD'}$, de betrouwbaarheid van de deeltest. De curves worden vlakker naarmate $k$ groter wordt. Dit betekent dat het rendement van een testverlenging met een vast aantal items afneemt met de testlengte. Bijvoorbeeld, stel dat een test bestaat uit 10 items, en dat zijn betrouwbaarheid gelijk is aan 0.5. Als we deze test verlengen met 10 items, dan wordt zijn betrouwbaarheid 0.67 (dit is uit te rekenen met de S-B formule met $k = 2$ en $r_{DD'} = 0.5$). Door 10 items toe te voegen hebben we dus een rendement gehaald van 0.17 (=0.67-0.5). Verlengen we deze test nogmaals met 10 items, dan wordt zijn betrouwbaarheid 0.75 (uit te rekenen met de S-B formule met $k = 3$ en $r_{DD'} = 0.5$). Door deze 10 items toe te voegen hebben we een rendement gehaald van 0.08 (=0.75-0.67). Dit is minder dan de helft van het rendement dat we gehaald hebben met de eerste 10 items.




Figuur 2.6: Betrouwbaarheid van een samengestelde test als functie van de verlengingsfactor $k$ voor drie verschillende deeltestbetrouwbaarheden ($r_{DD'}$ =0.3, 0.5 en 0.8).

Het is dus mogelijk om een test te beschouwen als opgebouwd uit een aantal parallelle

deeltests. Voor een gegeven test, is er een maximum aantal parallelle deeltests. Dit komt omdat het item de kleinst mogelijke deeltest is. Dus, voor een test met $I$ items is de maximale waarde van $k$ gelijk aan $I$. Indien alle items parallel zijn (i.e., dezelfde ware itemscores, dezelfde itemmeetfoutvariantie, en ongecorreleerde itemmeetfouten), dan kan men de testbetrouwbaarheid berekenen met de S-B formule met $k$ gelijk aan $I$ en $r_{DD'}$ gelijk aan de itembetrouwbaarheid.

De verlengingsfactor $k$ in de S-B formule is niet noodzakelijk een geheel getal. De S-B formule geldt bijvoorbeeld ook als je voor $k$ de waarde 1.5 invult. Dan geeft de S-B formule de betrouwbaarheid van een test die verlengd wordt met de helft van zijn huidige aantal items. Nog belangrijker is dat $k$ ook kleiner dan 1 mag zijn. In dit geval is $k$ een inkortingsfactor. Zo kunnen we de betrouwbaarheid van een item bepalen. Bijvoorbeeld, stel dat een test uit 10 items bestaat, en zijn betrouwbaarheid is 0.7. Met $k = 0.1$ en $r_{DD'} = 0.7$, levert de S-B formule ons een itembetrouwbaarheid van 0.189. (Merk op dat, in dit geval, $r_{DD'}$ de *test*betrouwbaarheid is en $r_{XX'}$ de *item*betrouwbaarheid.) Dit gebruik van de S-B formule is alleen toegestaan als de items parallel zijn (i.e., met allemaal dezelfde ware-itemscore en dezelfde meetfoutvariantie). Deze voorwaarde van parallelle items is analoog aan het gebruik van de S-B formule voor test*verlenging* (met $k > 1$), hetgeen alleen toegestaan is als de toegevoegde deeltests parallel zijn.

### 2.2.6 Praktische Methoden voor het Bepalen van de Betrouwbaarheid Gebaseerd op Parallelliteit

We stellen in totaal vier methoden voor waarmee we de betrouwbaarheid van een test kunnen bepalen. De eerste drie worden in deze paragraaf behandeld, en zijn alledrie gebaseerd op het gegeven dat de betrouwbaarheid van een test gelijk is aan de correlatie tussen twee parallelle tests. De vierde methode (Cronbachs alpha) wordt in de volgende paragraaf behandeld en is niet gebaseerd op parallelliteit.

#### De Alternatieve-Vormmethode

Bij deze methode maakt men gebruik van twee equivalente vormen van de test. Die worden hier *alternatieve vormen* genoemd om consistent te blijven met de Engelse benaming *alternate form method*. Men berekent dan de correlatie tussen deze twee alternatieve vormen. Deze correlatie ís de betrouwbaarheid als de alternatieve vormen parallel zijn.

#### Hertesting

Bij deze methode neemt men dezelfde test twee keer af, en men berekent de correlatie tussen de twee sets van scores. Deze correlatie is de betrouwbaarheid als men de twee testafnamen kan beschouwen als twee parallelle tests. Dit impliceert onder andere dat de ware-scores constant moeten blijven tussen de twee afnamen. Als het gemeten kenmerk niet stabiel is in de tijd, dan mag deze correlatie niet beschouwd worden als de betrouwbaarheid. Een voorbeelden van een eerder instabiel kenmerk zijn emoties (angst, neerslachtigheid, ...). Wil men ervoor zorgen dat de ware-scores op deze tests zoveel mogelijk constant blijven, dan moet de tijd tussen de twee afnamen kort zijn. Dit

maakt het dan weer weinig aannemelijk dat er geen geheugeneffecten zijn. De hertestingsmethode is beter geschikt voor tests die stabiele kenmerken meten, zoals intelligentie en allerlei vaardigheden. Men vindt hier meestal correlaties tussen 0.70 en 0.95. Toch blijkt ook hier dat het tijdsinterval tussen de afnamen van invloed is: men vindt kleinere correlaties naarmate het interval groter is. Dit geeft aan dat ook deze kenmerken niet perfect stabiel zijn.

### Halvering

Deze methode is gebaseerd op de idee dat een test beschouwd kan worden als de samenvoeging van twee parallelle deeltests; de score op de volledige test is gelijk aan de som van twee deeltestscores. De methode wordt in vier stappen uitgevoerd:

1. Deel de test in twee helften. Eén manier om dit te doen is door de oneven items in één testhelft samen te nemen, en de even items in de andere.

2. Bereken van iedere persoon zijn scores op de twee testhelften.

3. Bereken de correlatie $r_{DD'}$ tussen hun scores op de twee testhelften. Vanwege de veronderstelling van parallelliteit, is deze $r_{DD'}$ de betrouwbaarheid van de halve test.

4. Corrigeer $r_{DD'}$ met behulp van de S-B formule met $k = 2$ tot de betrouwbaarheid $r_{XX'}$ van de gehele test:

$$\boxed{r_{XX'} = \frac{2r_{DD'}}{1 + r_{DD'}} \quad .}$$

In Tabel 2.3 wordt een voorbeeld gegeven van hoe men deze methode toepast op een test bestaande uit zes items.

Het opsplitsen van een test in twee helften kan op een groot aantal manieren gebeuren. Het is niet gegarandeerd dat elke opsplitsing in dezelfde $r_{DD'}$ resulteert (en dus ook dezelfde $r_{XX'}$). Gelijke $r_{DD'}$'s bij alle mogelijke opsplitsingen zijn alleen maar gegarandeerd als alle items parallel zijn. Niet-parallelle items hoeven echter geen probleem te zijn, want het enige wat de halveringsmethode vereist is dat de twee deeltests parallel zijn. Door op een aantal punten te letten kan men proberen de parallelliteit van de twee deeltests niet te schenden:

1. Parallelle deeltests hebben hetzelfde gemiddelde[4]. Aan deze voorwaarde is met een grote kans voldaan als we de items eerst ordenen naar gemiddelde itemscore, en dan het eerste item in de eerste deeltest opnemen, het tweede en het derde in de tweede deeltest, het vierde en het vijfde weer in de eerste, enzovoort.

[4]In de praktijk zijn deze gemiddelden nooit precies gelijk. Dit komt omdat, in de praktijk, deze gemiddelden altijd berekend worden op basis van een *steekproef* van personen uit één of andere populatie. Het is natuurlijk mogelijk om statistisch te toetsen of het verschil tussen de steekproefgemiddelden significant is (dit wil zeggen, groot genoeg om te besluiten dat er ook in de populatie een verschil is).

| Persoon | Item 1 | 2 | 3 | 4 | 5 | 6 | Subtest A (Oneven) | Subtest B (Even) | Totaal |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 |
| 3 | 1 | 0 | 1 | 1 | 1 | 0 | 3 | 1 | 4 |
| 4 | 1 | 1 | 1 | 1 | 1 | 1 | 3 | 3 | 6 |
| 5 | 1 | 1 | 1 | 1 | 1 | 1 | 3 | 3 | 6 |
| 6 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 |
| 7 | 0 | 0 | 1 | 1 | 1 | 0 | 2 | 1 | 3 |
| 8 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 |
| 9 | 1 | 0 | 1 | 1 | 1 | 0 | 3 | 1 | 4 |
| 10 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 3 | 3 |
| | Gemiddelde | | | | | | 1.6 | 1.3 | 2.9 |
| | Standaarddeviatie | | | | | | 1.28 | 1.19 | 2.02 |
| | $r_{AB} = 0.34$ | | | | | | | | |
| | $r_{XX'} = \frac{2(0.34)}{1+0.34} = 0.51$ | | | | | | | | |

Tabel 2.3: De berekening van de betrouwbaarheid van een 6-item test met behulp van de halveringsmethode.

2. Als de deeltests inhoudelijk van elkaar verschillen, dan zijn de ware-scores vrijwel zeker niet gelijk aan elkaar. Men kan dit proberen te voorkomen door de items op te delen in paren met dezelfde inhoud en dan elk item van een paar in één van de twee deeltests te stoppen.

Het resultaat van de procedure die hier beschreven werd zal in wat volgt de *halveringsbetrouwbaarheid* genoemd worden, ook indien de twee deeltests niet parallel zijn. Indien de twee deeltests wel parallel zijn, dan is de halveringsbetrouwbaarheid gelijk aan de werkelijke betrouwbaarheid.

### 2.2.7 Cronbachs Alpha

Cronbachs alpha is de meest gebruikte methode voor het bepalen van de betrouwbaarheid. Het is een formule die ons echter niet de betrouwbaarheid zelf geeft, maar een ondergrens van de betrouwbaarheid. Dit wil zeggen dat de betrouwbaarheid altijd groter of gelijk is aan Cronbachs alpha (verder, gewoon "alpha"). Dit is niet altijd een probleem. Bijvoorbeeld, als alpha gelijk is aan 0.95, dan weten we dat de betrouwbaarheid tussen 0.95 en 1.00 ligt, en dat is een vrij klein gebied. In de praktijk is alpha echter zelden 0.95 of hoger. Een alpha van 0.60 betekent dat de betrouwbaarheid tussen 0.60 en 1.00 ligt, en dat is wél een groot gebied.

#### De Formule

Alpha kan gebruikt worden indien de testscore bepaald wordt als de som van de itemscores. Concreet, elk item wordt afzonderlijk gescoord en de testscore is gelijk aan de

som van de afzonderlijke itemscores. Alpha wordt als volgt berekend:

$$\boxed{\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\mathrm{VAR}(X)}\right) \quad .}$$

In deze formule staat $k$ voor het aantal items in de test, en $\sigma_i^2$ (spreek uit: "sigma kwadraat ie") voor de variantie van de itemscores op het i-de item. De kern van de formule is de verhouding van de som van de itemvarianties, $\sum_{i=1}^{k}\sigma_i^2$, tot de variantie van de testscore, $\mathrm{VAR}(X)$; als deze verhouding klein is, dan is alpha groot. De term $k/(k-1)$ is een correctiefactor, en is van weinig belang voor een goed begrip van de formule.

Voor een goed begrip van alpha, is het nuttig om $\mathrm{VAR}(X)$ te schrijven in termen van de itemvarianties en -covarianties. We doen dit eerst voor een test met slechts twee items. De scores op het eerste item duiden we aan met $J_1$ en de scores op het tweede item met $J_2$. De testscore $X$ is gelijk aan de som $J_1 + J_2$. We maken nu gebruik van het feit dat de variantie van een som van twee variabelen gelijk is aan de som van de varianties van die variabelen plus twee keer de covariantie van die variabelen:

$$\mathrm{VAR}(X) = \mathrm{VAR}(J_1) + \mathrm{VAR}(J_2) + 2\mathrm{COV}(J_1, J_2) \quad .$$

(Deze formule wordt afgeleid op pagina 123.) Het blijkt handig te zijn als we de varianties $\mathrm{VAR}(J_1)$ en $\mathrm{VAR}(J_2)$ korter aanduiden met $\sigma_1^2$ en $\sigma_2^2$, en de covariantie $\mathrm{COV}(J_1, J_2)$ met $\sigma_{12}$. Dus,

$$\mathrm{VAR}(X) = \sigma_1^2 + \sigma_2^2 + 2\sigma_{12} \quad .$$

Omdat $\sigma_{12}$ gelijk is aan $\sigma_{21}$ (de covariantie is symmetrisch), is $2\sigma_{12}$ gelijk aan $\sigma_{12} + \sigma_{21}$. Dus,

$$\mathrm{VAR}(X) = \sigma_1^2 + \sigma_2^2 + \sigma_{12} + \sigma_{21} \quad .$$

Het rechterlid van deze formule wordt in Tabel 2.4 weergegeven als een twee-bij-twee tabel. De som van de itemvarianties en -covarianties in deze tabel is gelijk aan $\mathrm{VAR}(X)$.

Deze manier om $\mathrm{VAR}(X)$ uit te drukken als een som van de itemvarianties en -covarianties is niet alleen mogelijk met slechts twee items, maar ook met drie of willekeurig veel. Deze laatste twee situaties ($k = 3$ en $k$ is willekeurig) worden voorgesteld in Tabel 2.4 (b) en (c).

Omdat $\mathrm{VAR}(X)$ gelijk is aan de som van de itemvarianties en -covarianties in Tabel 2.4 (c), kan men zien dat de breuk in de formule van alpha van de vorm $\frac{a}{a+b}$ is, waarbij $a$ de som van de itemvarianties is, en $b$ de som van de itemcovarianties. De term $a$ is de som van de vetgedrukte symbolen op de diagonaal van Tabel 2.4 (c) en de term $b$ is de som van gewoon gedrukte symbolen boven en onder de diagonaal van Tabel 2.4 (c). We kunnen alpha dus als volgt schrijven:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{a}{a+b}\right) \quad .$$

Als $a$ klein is in vergelijking met $b$, dan is alpha groot. Dus, als $\mathrm{VAR}(X)$ vooral toe te schrijven is aan de itemcóvarianties, dan is alpha groot. In wat volgt, noemen we de bovenstaande formule de vereenvoudigde formule van alpha.

(a)

| Item | 1 | 2 |
|---|---|---|
| 1 | $\boldsymbol{\sigma_1^2}$ | $\sigma_{12}$ |
| 2 | $\sigma_{21}$ | $\boldsymbol{\sigma_2^2}$ |

(b)

| Item | 1 | 2 | 3 |
|---|---|---|---|
| 1 | $\boldsymbol{\sigma_1^2}$ | $\sigma_{12}$ | $\sigma_{13}$ |
| 2 | $\sigma_{21}$ | $\boldsymbol{\sigma_2^2}$ | $\sigma_{23}$ |
| 3 | $\sigma_{31}$ | $\sigma_{32}$ | $\boldsymbol{\sigma_3^2}$ |

(c)

| Item | 1 | 2 | $\cdots$ | $k$ |
|---|---|---|---|---|
| 1 | $\boldsymbol{\sigma_1^2}$ | $\sigma_{12}$ | $\cdots$ | $\sigma_{1k}$ |
| 2 | $\sigma_{21}$ | $\boldsymbol{\sigma_2^2}$ | $\cdots$ | $\sigma_{2k}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| k | $\sigma_{k1}$ | $\sigma_{k2}$ | $\cdots$ | $\boldsymbol{\sigma_k^2}$ |

Tabel 2.4: Varianties en covarianties van de itemscores weergegeven in een tabel, apart voor twee (a), drie (b), en $k$ (c) items.

We krijgen nog een beter zicht op alpha als we $\sigma_{ij}$ $(i, j = 1, \ldots, k)$ schrijven als functie van $r_{ij}$, de correlatie tussen item $i$ en item $j$:

$$\sigma_{ij} = r_{ij}\sigma_i\sigma_j \quad .$$

Hieruit volgt dat, voor vaste waarden van de standaarddeviaties $\sigma_i$ en $\sigma_j$, alpha groter wordt naarmate de inter-item correlaties groter worden.

### Wanneer is Alpha een Goede Benadering van de Betrouwbaarheid?

Alpha is altijd kleiner dan of gelijk aan de betrouwbaarheid. Om bruikbaar te zijn, mag alpha niet véél kleiner zijn dan de betrouwbaarheid. Om te bepalen wanneer alpha bruikbaar is, is de volgende regel nuttig: alpha is precies gelijk aan de betrouwbaarheid als alle items onderling parallel zijn. Ter herinnering, twee items zijn parallel indien (1) hun ware itemscores gelijk zijn, (2) ze dezelfde itemmeetfoutvariantie hebben, en (3) hun itemmeetfouten ongecorreleerd zijn. Om hier een beter zicht op te krijgen, bespreken we nu een aantal concrete situaties die van elkaar verschillen wat betreft de bruikbaarheid van alpha. Deze situaties zijn de volgende:

1. De items zijn perfect ónbetrouwbaar; de itemscores bestaan alleen uit meetfout.
2. De items zijn perfect betrouwbaar; de itemscores bevatten geen meetfout. Hier maken we een onderscheid tussen parallelle en niet-parallelle items.
3. De itembetrouwbaarheid ligt tussen de 0 en de 1. Ook hier maken we een onderscheid tussen parallelle en niet-parallelle items.

Bij de bespreking van deze drie situaties gaan we er vanuit dat de meetfouten ongecorreleerd zijn.

Ten eerste, stel dat **alle items perfect onbetrouwbaar** zijn. In dit geval, zijn alle itemcovarianties $\sigma_{ij}$ gelijk aan 0. Immers, de scores op perfect onbetrouwbare items bestaan alleen uit meetfout, en meetfout correleert met niets ($r_{ij} = 0$, hetgeen impliceert dat $\sigma_{ij} = 0$). Dus, de term $b$ in de vereenvoudigde formule van alpha is gelijk aan 0, en hieruit volgt dat ook alpha gelijk is aan 0. In de vorm van een afleiding:

$$\begin{aligned}\alpha =& \frac{k}{k-1}\left(1 - \frac{a}{a+b}\right)\\ =& \frac{k}{k-1}\left(1 - \frac{a}{a}\right)\\ =& \frac{k}{k-1}(1-1)\\ =& 0 \quad .\end{aligned}$$

De werkelijke betrouwbaarheid is ook gelijk aan 0. Immers, als alle itemscores alleen uit meetfout bestaan, bestaat ook de testscore $X$ alleen uit meetfout. Dus, indien de items perfect onbetrouwbaar zijn, dan is alpha precies gelijk aan de betrouwbaarheid van de test.

Ten tweede, stel dat **alle items perfect betrouwbaar** zijn. In dit geval, is ook de testscore $X$ perfect betrouwbaar. Immers, als de itemscores geen meetfout bevatten, bevat ook de testscore $X$ geen meetfout. We maken nu een onderscheid tussen parallelle en niet-parallelle items. Indien deze perfect betrouwbare items parallel zijn, dan correleren ze perfect met elkaar ($r_{ij} = 1$). Immers, (1) omdat ze perfect betrouwbaar zijn, bestaan de geobserveerde itemscores alleen uit ware-score, en (2) omdat ze parallel zijn, zijn hun ware-scores gelijk. Indien ze parallel zijn, dan hebben ze ook dezelfde variantie ($\sigma_i^2 = \sigma_j^2$), die we verder met het symbool $\sigma^2$ zullen aanduiden. Hieruit volgt dat alle covarianties $\sigma_{ij}$ gelijk zijn aan deze gemeenschappelijke variantie $\sigma^2$. Immers, $\sigma_{ij}$ is gelijk aan $r_{ij}\sigma_i\sigma_j$ en dit is gelijk aan $\sigma^2$ (omdat $r_{ij} = 1$ en $\sigma_i = \sigma_j = \sigma$). Hieruit volgt dat de term $a$ in de vereenvoudigde formule van alpha gelijk is aan $k\sigma^2$ (want er zijn $k$ items) en dat de term $b$ gelijk is aan $(k^2 - k)\sigma^2$ (want er zijn $k^2 - k$ covarianties). Alpha kan dus als volgt geschreven worden:

$$\begin{aligned}\alpha &= \frac{k}{k-1}\left(1 - \frac{k\sigma^2}{k\sigma^2 + (k^2-k)\sigma^2}\right)\\ &= \frac{k}{k-1}\left(1 - \frac{k\sigma^2}{k^2\sigma^2}\right)\\ &= \frac{k}{k-1}\left(1 - \frac{1}{k}\right)\\ &= \frac{k}{k-1}\left(\frac{k-1}{k}\right)\\ &= 1 \quad .\end{aligned}$$

Dus, indien de items perfect betrouwbaar en parallel zijn, dan is alpha precies gelijk aan de betrouwbaarheid van de test (namelijk, 1).

Indien deze perfect betrouwbare items níet parallel zijn, dan is alpha níet gelijk aan de betrouwbaarheid. Een voorbeeld van een test met perfect betrouwbare niet-parellelle items is een test met de volgende drie items: (1) het lichaamsgewicht, (2) het maandinkomen en (3) het aantal kinderen. Het is duidelijk dat deze drie items perfect betrouwbaar zijn; als men ze meerdere keren zou afnemen (met een digitale weegschaal voor het eerste item) dan zou men steeds dezelfde antwoorden krijgen. Hieruit volgt dat ook de testscore (de som van de drie itemscores) perfect betrouwbaar is. In het meest extreme geval, zijn de correlaties tussen alle paren van items gelijk aan 0, en voor het voorbeeld dat hierboven gegeven werd lijkt dit ook zeer goed mogelijk. Indien alle correlaties gelijk zijn aan 0 dan zijn ook alle covarianties gelijk aan 0 en daarom ook de term $b$ in de vereenvoudigde formule van alpha. Zoals reeds eerder werd aangetoond, is alpha in dit geval gelijk aan 0. Een alpha gelijk aan 0 is wel erg ver onder de betrouwbaarheid van 1! Deze sterke onderschatting van de betrouwbaarheid doet zich voor in alle gevallen waarin de items onderling laag correleren *doordat ze iets anders meten*.

Ten derde, stel dat **de betrouwbaarheid van de items tussen de 0 en de 1 ligt**. Als de items parallel zijn, dan is de correlatie tussen alle paren van items gelijk, alsook de variantie van elk van de items. Dit volgt uit de redenering in paragraaf 2.2.4 maar nu met "parallelle items" in plaats van "parallelle tests". Merk op dat de correlatie tussen twee parallelle items gelijk is aan de itembetrouwbaarheid. De gemeenschappelijke correlatie tussen een paar van parallelle items wordt aangeduid met $\rho$ (spreek uit: "rho") en de gemeenschappelijke variantie van de items met $\sigma^2$. De covarianties $\sigma_{ij}$ zijn gelijk aan $\rho\sigma^2$. Immers, $\sigma_{ij}$ is gelijk aan $r_{ij}\sigma_i\sigma_j$ en dit is gelijk aan $\rho\sigma^2$ (omdat $r_{ij} = \rho$ en $\sigma_i = \sigma_j = \sigma$). Uit het voorgaande volgt dat de term $a$ in de vereenvoudigde formule van alpha gelijk is aan $k\sigma^2$ en dat de term $b$ gelijk is aan $(k^2 - k)\rho\sigma^2$. Alpha kan dus als volgt geschreven worden:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{k\sigma^2}{k\sigma^2 + (k^2-k)\rho\sigma^2}\right) \quad .$$

Het rechterlid van deze gelijkheid kan uitgewerkt worden (zie pagina 125), met als resultaat de volgende formule:

$$\alpha = \frac{k\rho}{1 + (k-1)\rho} \quad .$$

Dit is de S-B formule voor de situatie met $k$ gelijk aan het aantal items en $\rho$ gelijk aan de itembetrouwbaarheid. Omdat de S-B formule ons de betrouwbaarheid van de samengestelde test geeft, hebben we aangetoond dat, indien de items parallel zijn, alpha precies gelijk is aan de betrouwbaarheid van de test.

We bekijken nu wat er gebeurt als de items níet parallel zijn. We beperken ons tot de situatie waarin deze niet-parallelliteit toe te schrijven is aan het feit dat de ware-itemscores niet gelijk zijn voor alle paren van items. In vergelijking met de situatie waarin deze ware-itemscores wél gelijk zijn, zullen de inter-item-correlaties $r_{ij}$ lager zijn dan de itembetrouwbaarheden. Omdat alpha kleiner is naarmate de inter-item-correlaties kleiner zijn, heeft dit tot gevolg dat alpha ónder de testbetrouwbaarheid komt te liggen. Deze onderschatting van de testbetrouwbaarheid zal des te groter zijn naarmate de items onderling lager correleren *doordat ze iets anders meten* (i.e., andere ware-itemscores hebben). Merk op dat het onderling laag correleren van de items op zích

geen probleem is voor de kwaliteit van alpha als benadering van de betrouwbaarheid; het is alleen een probleem als die lage correlaties het gevolg zijn van het feit dat de items iets anders meten. Indien de lage correlaties het gevolg zijn van een lage itembetrouwbaarheid, dan heeft dit geen gevolg voor de kwaliteit van alpha als benadering van de betrouwbaarheid.

Bij het gebruik van alpha in de praktijk stuit men op het probleem dat men nooit weet in welke mate de items hetzelfde meten (in hoeverre ze parallel zijn). Daarom weet men ook nooit hoe ver alpha onder de betrouwbaarheid ligt. Het enige dat men kan doen, is kijken naar de inhoud van de items en zich op basis daarvan een indruk vormen van de mate waarin de items hetzelfde meten. Als men de items te heterogeen vindt, dan verwerpt men alpha als bruikbare benadering van de betrouwbaarheid. Dit is geen volledig objectieve werkwijze, maar het is het beste wat we hebben.

De onwetendheid over hoe ver alpha onder de betrouwbaarheid ligt heeft ervoor gezorgd dat men alpha vaak opvat als maat voor een speciaal soort betrouwbaarheid: de *interne consistentie betrouwbaarheid*. Deze naam verwijst naar het feit dat alpha stijgt naarmate de items onderling hoger correleren. Dit blijkt het duidelijkst uit een vergelijking van de twee meest extreme situaties: indien alle inter-item-correlaties gelijk zijn aan 0 dan is ook alpha gelijk aan 0, en indien alle inter-item-correlaties gelijk zijn aan 1 dan is ook alpha gelijk aan 1. Door alpha op te vatten als maat voor een speciaal soort betrouwbaarheid is het oorspronkelijke probleem echter niet dichter bij een oplossing gekomen, want we weten nog steeds niet hoe ver alpha onder de gewone betrouwbaarheid ligt. Bovendien levert de naam "interne consistentie betrouwbaarheid" niets nieuws op: we weten niet wat deze speciale soort betrouwbaarheid precies is, en de uitspraak dat alpha gelijk is aan de interne consistentie betrouwbaarheid is dus niet te controleren.

---

*Opgave 10*

Met welke van de onderstaande stellingen ben je het eens en met welke ben je het oneens? Indien je het oneens bent met een stelling, leg dan uit waarom.

1. De betrouwbaarheid van een test met parallelle items kan uitgedrukt worden als een functie van het aantal items en de itembetrouwbaarheid.
2. Een test die uit meer items bestaat dan een andere test, is altijd betrouwbaarder. (Bij de beantwoording van deze vraag, ga ervan uit dat de items binnen een test paralllel zijn.)
3. Twee tests, A en B, bestaan elk uit parallelle items (parallel bínnen elk van de twee tests). De items van test A zijn betrouwbaarder dan die van test B. Hieruit volgt dat test A betrouwbaarder is dan test B.
4. Een test met parallelle items is altijd betrouwbaarder dan een test met eenzelfde aantal niet-parallelle items.

*Opgave 11*

Van een test voor ruimtelijk geheugen bestaan twee varianten, A en B. Variant A bestaat uit 25 items en variant B uit 50 items. Beide tests bestaan uit onderling parallelle items. De itembetrouwbaarheid is bij beide tests gelijk.

1. Welke test kun je het beste gebruiken als je iemands ruimtelijk geheugen wil testen? Motiveer je antwoord.
2. Beide testvarianten worden verlengd met 25 parallelle items. Bij welke testvariant zal de betrouwbaarheid het meest veranderen? Beargumenteer.

*Opgave 12*
Met welke van de onderstaande stellingen ben je het eens en met welke ben je het oneens? Indien je het oneens bent met een stelling, leg dan uit waarom.

1. Twee tests waarvan de correlatie tussen de testhelften (bínnen elk van de twee tests) gelijk is, hebben dezelfde halveringsbetrouwbaarheid.
2. Als de halveringsbetrouwbaarheid laag is dan wil dat niet noodzakelijk zeggen dat de testscores ook veel meetfouten bevatten.
3. Bij een gelijke correlatie tussen twee parallelle testhelften, zal de betrouwbaarheid van de test des te groter zijn naarmate deze testhelften uit minder items bestaan.

*Opgave 13*
Er werden twee nieuwe tests ontwikkeld, één voor ruimtelijk voorstellingsvermogen en één voor numerieke intelligentie. De test voor ruimtelijk voorstellingsvermogen bestaat uit 50 items en de test voor numerieke intelligentie uit 100. Beide tests bestaan uit onderling parallelle items. Na afname van beide tests bij een groep studenten, werden voor beide tests de somscores op de even en de oneven genummerde items berekend. Deze scores werden per test met elkaar gecorreleerd. Het bleek dat deze correlatie voor beide tests gelijk was aan 0.75.

1. Welke van beide tests is het meest betrouwbaar, de test voor ruimtelijk voorstellingsvermogen (bestaande uit 50 items) of de test voor numerieke intelligentie (bestaande uit 100 items)? Beargumenteer je antwoord.
2. Aan beide tests worden 25 parallelle items toegevoegd. Van welke test zal de betrouwbaarheid het meest veranderen? Beargumenteer je antwoord.

*Opgave 14*
Er werden twee nieuwe tests ontwikkeld, A en B. Test A bestaat uit 60 onderling parallelle items en test B uit 30 onderling parallelle items. Voor beide tests werd de betrouwbaarheid bepaald met de halveringsmethode (met toepassing van de S-B formule). Beide tests bleken een betrouwbaarheid te hebben van 0.80.

1. Voor welk van de tests was de correlatie tussen de somscores op de test*helften* het grootst? Motiveer je antwoord.
2. Beide tests worden verminderd met 10 items. Van welke test verandert de betrouwbaarheid het meest? Motiveer je antwoord.

*Opgave 15*
Er zijn twee nieuwe tests ontwikkeld: een test voor sociale vaardigheden (SVT) bestaande uit 50 onderling parallelle items en een test voor numerieke intelligentie (NIT) bestaande uit 100 onderling parallelle items. Deze tests werden afgenomen bij een groep personen. Eerst werd alpha berekend op de data van alle 150 items samen. Dit leverde een alpha van 0.60 op.

1. Wat zegt dit over de betrouwbaarheid van de samengevoegde test van 150 items? Leg uit.

Vervolgens werden voor de twee tests afzonderlijk somscores op de even en de oneven items berekend. Per test werd de correlatie tussen deze twee somscores berekend. Voor de SVT kwam de correlatie tussen de somscores op de 25 even en de 25 oneven items uit op 0.75. De correlatie tussen de somscores op de 50 even en de 50 oneven items van de NIT bleek eveneens 0.75 te bedragen.

2. Welke test heeft de grootste halveringsbetrouwbaarheid, de SVT of de NIT, of hebben ze een gelijke betrouwbaarheid, of valt hierover niets te zeggen? Leg uit.

3. De SVT werd uitgebreid met 25 parallelle SVT-items en de NIT met 25 parallelle NIT-items. Van welke test zal de betrouwbaarheid het meest veranderen, of zal de verandering gelijk zijn, of valt hierover niets te zeggen? Leg uit.

*Opgave 16*
Is het mogelijk dat een bepaalde test zowel een hoge hertestingsbetrouwbaarheid (bijvoorbeeld, 0.95) heeft, en een lage alpha (bijvoorbeeld, 0.40)? Zo ja, wat betekent dit dan voor datgene wat de test meet?

*Opgave 17*
Op basis van de uitkomsten van een vooronderzoek, besluit een onderzoeker dat het verstandiger is zijn vragenlijst op te splitsen in twee delen. Hij vraagt zich af wat er zal gebeuren met de alpha's van de beide delen, als ze vergeleken worden met de alpha bepaald op de volledige test.

1. Beargumenteer waarom de alpha's van de delen lager zouden kunnen zijn dan die van het geheel.
2. Beargumenteer waarom de alpha's van de delen hoger zouden kunnen zijn dan die van het geheel.

*Opgave 18*
Een onderzoeker beschikt over twee tests, A en B. Test A heeft een alpha van 0.80 en test B een van 0.85. Nu voegt hij de items van beide tests samen en stelt tot zijn verbazing vast dat de alpha van deze samengevoegde test slechts gelijk is aan 0.50. Hoe kun je dit verklaren?

*Opgave 19*
Een studente werkt aan haar afstudeerscriptie. Zij krijgt van haar begeleider een vragenlijst en een databestand met de antwoorden van 200 proefpersonen op de 40 items van deze vragenlijst. Deze vragenlijst wordt geacht de persoonlijkheidstrek doorzettingsvermogen te meten. De alpha van deze vragenlijst is 0.71. De studente is slechts geïnteresseerd in de 18 items die betrekking hebben op doorzettingsvermogen in de omgang met anderen (iemand ergens toe overhalen, proberen je zin te krijgen, ...). Deze 18 items hebben een alpha van 0.80. De begeleider is zeer verbaasd en beweert dat dit niet kan volgens de formule van Spearman-Brown. De begeleider beweert dat een deeltest altijd een lágere betrouwbaarheid heeft dan de volledige test. Ben jij het eens met de begeleider? Motiveer je antwoord.

*Opgave 20*
Een onderzoeker beschikt over een test met een betrouwbaarheid van 0.70.

1. Deze onderzoeker beweert dat, indien hij alpha zou berekenen, deze alpha groter zal zijn dan 0.70. Ben je het hiermee eens? Leg uit.
2. Stel dat je over de bijkomende informatie beschikt dat alle items van deze test parallel zijn. Zou je nu de vraag in de vorige deelopgave anders beantwoorden of maakt het niet uit. Leg uit.

---

### 2.2.8 Itemanalyse

Het komt regelmatig voor dat men bij de testconstructie vertrekt van een bepaalde set items, maar er daarvan een aantal verwijdert op basis van een itemanalyse. De idee bij itemanalyse is dat men van een set items alleen die items behoudt die voldoen aan een bepaald criterium. Een criterium dat vaak gebruikt wordt, is dat alle items een positieve

bijdrage moeten leveren aan de betrouwbaarheid. Dit betekent dat voor alle items in de definitieve test moet gelden dat als men dit item zou weglaten, de betrouwbaarheid van de test zou dalen. Stel dat we de betrouwbaarheid bepalen met alpha. Dan kunnen we met de volgende procedure een set items selecteren die aan dit criterium voldoen:

1. Bereken alpha voor de volledige itemset. Het aantal items in deze set duiden we aan met $k$.

2. Bereken $k$ alpha's, één voor elk van de itemsets die men krijgt door één van de $k$ items uit de volledige set weg te laten. SPSS noemt zo een alpha een *alpha-if-item-deleted*.

3. Vergelijk elke alpha-if-item-deleted met de alpha voor de volledige itemset. Als de alpha-if-item-deleted groter is dan de alpha voor de volledige itemset, dan verwijdert men dit item. Door het item weg te laten stijgt immers de betrouwbaarheid.

4. Op de overblijvende items berekent men opnieuw alpha. Het aantal items in deze set duiden we aan met $k$. We gaan nu opnieuw naar stap 2, en herhalen de procedure totdat er geen enkel item meer verwijderd kan worden.

We behandelen deze procedure omdat ze zo vaak gebruikt wordt. Ze is echter niet vrij van kritiek. De kritiek komt er in essentie op neer dat de motivering van de procedure inconsistent is. We zien deze inconsistentie als we redeneren vanuit twee uitgangspunten: (1) vanuit alpha als benadering van de betrouwbaarheid en (2) vanuit de mogelijkheid dat sommige items niet bijdragen aan de betrouwbaarheid. De eerste redenering is als volgt:

1. Alpha is bruikbaar als benadering van de betrouwbaarheid als de items allemaal ongeveer hetzelfde meten of, anders uitgedrukt, als zij vrijwel parallel zijn.

2. Als de items vrijwel parallel zijn, dan heeft het geen zin om items te verwijderen. Immers, weglating van parallelle items verlaagt de betrouwbaarheid (zie de S-B formule).

En de tweede redenering is als volgt:

1. Als sommige items niet bijdragen aan de betrouwbaarheid, dan is

   het zinvol om die items te verwijderen.

2. Als sommige items niet bijdragen aan de betrouwbaarheid, dan zijn ze ook niet parallel met (het merendeel van) de andere items. In dit geval is alpha niet bruikbaar als benadering van de betrouwbaarheid.

In het algemeen, zorgt de bovenstaande procedure ervoor dat de overblijvende items onderling gemiddeld hoger correleren dan de oorspronkelijke items. Dit komt omdat alpha groter wordt naarmate de inter-item-correlaties groter worden. Items die laag of negatief correleren met de andere items zorgen ervoor dat de som van de itemcovarianties (term $b$ in de vereenvoudigde formule van alpha) slechts weinig afneemt of zelfs toeneemt door het weglaten van zo'n item. De som van de itemvarianties (term $a$ in

de vereenvoudigde formule van alpha) neemt áltijd af als een item wordt weggelaten (onafhankelijk van de inter-item-correlaties). Hieruit volgt dat bij een item dat laag of negatief correleert met de andere items, de breuk $\frac{a}{a+b}$ kleiner wordt als het item weggelaten wordt uit de test. Stel bijvoorbeeld, dat de $a$ en de $b$ van de volledige itemset gelijk zijn aan, respectievelijk, 10 en 20, en dat de $a$ en de $b$ van de met één item gereduceerde itemset gelijk zijn aan, respectievelijk, 8 en 20. Dan verandert $\frac{a}{a+b}$ van 0.33 ($= \frac{10}{10+20}$) in 0.28 ($= \frac{8}{8+20}$) door weglating van dit item. Stel verder dat de volledige itemset uit 10 items bestaat. Dan verandert alpha van 0.74 ($= \frac{10}{9}[1-0.33]$) in 0.80 ($= \frac{9}{8}[1-0.28]$). Dit item zal dus weggelaten worden in de bovenstaande procedure, en daardoor stijgt de gemiddelde inter-item-correlatie.

Om een indicatie te krijgen van de mate waarin elk van de items correleert met de andere items, kunnen we de *item-rest-correlatie* berekenen. De item-rest-correlatie is de correlatie tussen het item en de somscore op de andere $k-1$ items ("de rest"). Het is vrijwel altijd zo dat een grote alpha-if-item-deleted samengaat met een kleine item-rest-correlatie. In SPSS wordt de item-rest-correlatie de *corrected item-total correlation* genoemd.

Men mag van zo'n itemanalyse geen wonderen verwachten. Door slechte items te verwijderen, zal de betrouwbaarheid wel stijgen, maar als de oorspronkelijke itemset zeer heterogeen is (lage inter-item-correlaties) en de itemscores bevatten veel meetfouten, dan zal een itemanalyse daar toch geen betrouwbare test van kunnen maken. Het enige wat de itemanalyse kan, is het verwijderen van een paar rotte appels (items) uit een kist met ook veel goede appels; een kist met alleen maar rotte appels kun je niet beter maken door er een aantal te verwijderen!

Deze vorm van itemanalyse is niet beperkt tot het selecteren van items die de betrouwbaarheid maximaliseren. Men kan ook items selecteren zodat ze een andere variabele (bijvoorbeeld, studiesucces) zo goed mogelijk voorspellen. In plaats van de alpha-if-item-deleted berekent men dan de *correlation-if-item-deleted*, de correlatie tussen de somscore op alle items *min één* en die andere variabele.

### 2.2.9 Het Restriction-of-Range Fenomeen

De betrouwbaarheid is niet alleen een kenmerk van de test maar ook van de populatie waarin die test wordt afgenomen. Meer in het bijzonder, de betrouwbaarheid hangt af van de spreiding van de ware-scores in de populatie. In het algemeen geldt dat in een homogene populatie (kleine ware-score-variantie) een test een *lagere* betrouwbaarheid zal hebben dan in een heterogene populatie (grote ware-score-variantie).

Dit fenomeen kan verduidelijkt worden aan de hand van een spreidingsdiagram van de scores op twee parallelle tests. In Figuur 2.7 wordt zo een spreidingsdiagram getoond. Met de dubbele pijlen worden varianties aangeduid. Pijl $a$ staat voor de meetfoutvariantie $\mathrm{VAR}(E)$: des te langer die pijl, des te groter $\mathrm{VAR}(E)$[5]. Pijl $c$ staat voor de testscorevariantie $\mathrm{VAR}(X)$ in de volledige groep. De betrouwbaarheid van de test in

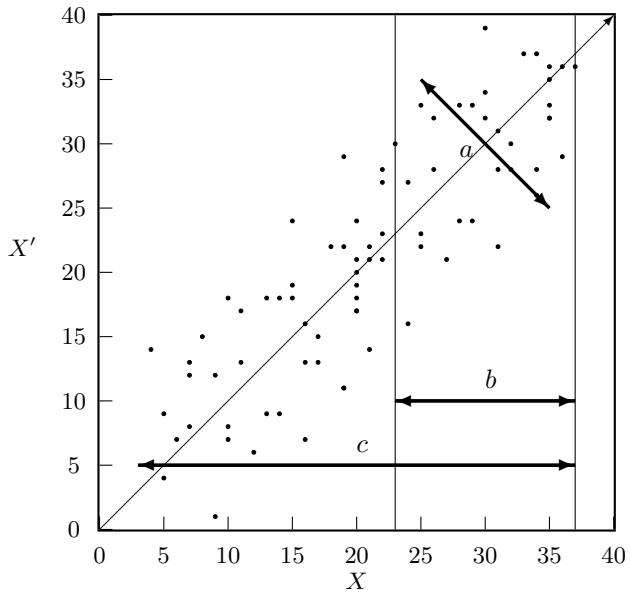
[5]Merk op dat pijl $a$ schuin staat. Dit komt omdat er zowel in $X$ (horizontale as) als in $X'$ (verticale as) meetfouten zitten. Dit is anders dan in de enkelvoudige regressieanalyse, want daar staat de onafhankelijke variabele op de horizontale en de afhankelijke variabele op de verticale as, en kijkt men naar de *verticale* spreiding van de punten rond de regressielijn. In de regressieanalyse kijkt men naar de verticale spreiding omdat alleen de afhankelijke variabele met fouten behept is (i.c., voorspellingsfouten).

deze groep wordt bepaald door de verhouding van VAR($E$) (de lengte van pijl $a$) en VAR($X$) (de lengte van pijl $c$). De manier waarop dit gebeurt wordt uitgedrukt door de volgende formule, die is afgeleid uit de definitie van betrouwbaarheid:

$$\text{Betrouwbaarheid} = 1 - \frac{\text{VAR}(E)}{\text{VAR}(X)} \quad .$$

Dus, de betrouwbaarheid stijgt naarmate pijl $c$ (VAR($X$)) langer is in vergelijking met pijl $a$ (VAR($E$)).

Bekijken we nu, in plaats van de volledige groep, de groep die een testscore[6] heeft tussen 23 en 37. De testscorevariantie in deze groep wordt aangeduid met pijl $b$. De meetfoutvariantie in deze geselecteerde (homogene) groep is echter hetzelfde als de meetfoutvariantie in de volledige (heterogene) groep (de spreiding van de punten rond de diagonaal is hetzelfde voor alle waarden van $X$). Pijl $a$ is veel korter in vergelijking met pijl $c$ dan in vergelijking met pijl $b$. Bijgevolg is de betrouwbaarheid kleiner in de geselecteerde groep dan in de volledige groep.




Figuur 2.7: Illustratie van het restriction-of-range fenomeen: spreidingsdiagram van de scores op twee parallelle tests met aanduiding van de meetfoutvariantie (pijl a) en de testscorevariantie in een heterogene (pijl c) en een homogene (pijl b) populatie.

Een tweede manier om het restriction-of-range fenomeen uit te leggen is aan de hand

[6]Selectie op basis van testscores is niet de enige manier om personen te selecteren. Men kan bijvoorbeeld ook selecteren op basis van opleiding. Als men alle hoogopgeleiden samen neemt, dan kan dat resulteren in een vergelijkbare selectie als in Figuur 2.7.

van de volgende formule, die is afgeleid uit de definitie van betrouwbaarheid:

$$\text{Betrouwbaarheid} = \frac{\text{VAR}(T)}{\text{VAR}(T) + \text{VAR}(E)} \quad .$$

Restriction-of-range betekent dat de ware score variantie $\text{VAR}(T)$ in de homogene groep kleiner is dan de ware score variantie in de heterogene groep. Dit zorgt ervoor dat de betrouwbaarheid in de homogene groep kleiner is dan de betrouwbaarheid in de heterogene groep. Dit volgt uit de bovenstaande formule in combinatie met het gegeven dat de meetfoutvariantie $\text{VAR}(E)$ gelijk is voor de homogene en de heterogene groep.

We kunnen dit illusteren met een voorbeeld. Een vragenlijst voor milieubewust gedrag heeft een ware score variantie van 70 in een representatieve steekproef uit de Nederlandse populatie. De variantie van de geobserveerde testscores in deze steekproef was gelijk aan 100. De meetfoutvariantie is dus gelijk aan 100-70=30. En de betrouwbaarheid in deze steekproef is gelijk aan $70/100 = 0.70$. We nemen deze vragenlijst nu ook af bij een representatieve steekproef studenten van de Faculteit Sociale Wetenschappen. In deze steekproef blijkt de ware score variantie slechts gelijk te zijn aan 10. Omdat de meetfoutvariantie in beide steekproeven gelijk is aan 30, is de variantie van de geobserveerde testscores gelijk aan $10 + 30 = 40$. De betrouwbaarheid in deze meer homogene steekproef is dan gelijk aan $10/40 = 0.25$. Dit is een groot verschil met de betrouwbaarheid in de meer heterogene steekproef.

Het restriction-of-range fenomeen kan ook met een tweede formule beschreven worden. Stel dat we één test in twee populaties afgenomen hebben, een heterogene en een homogene. De homogene populatie kan een onderdeel zijn van de heterogene populatie, maar dat hoeft niet. De testscores in de heterogene populatie duiden we aan met $X$, en de testscores in de homogene populatie met $Y$. Men kan het volgende aantonen (voor een afleiding, zie pagina 126):

$$1 - r_{YY'} = \frac{\text{VAR}(X)}{\text{VAR}(Y)}(1 - r_{XX'}) \quad .$$

Deze formule wordt het beste gelezen in termen van de ónbetrouwbaarheden van $X$ en $Y$, namelijk $1 - r_{XX'}$ en $1 - r_{YY'}$. Uit deze formule kunnen we dan aflezen dat de ónbetrouwbaarheid van $Y$ stijgt met dalende $\text{VAR}(Y)$. Dus, de betrouwbaarheid van $Y$ dáálT met dalende $\text{VAR}(Y)$. Anders uitgedrukt, hoe homogener de populatie, hoe lager de betrouwbaarheid.

We kunnen de bovenstaande formule illustreren met het voorbeeld van de vragenlijst voor milieubewust gedrag. Deze vragenlijst heeft een betrouwbaarheid van 0.70 in een representatieve steekproef uit de Nederlandse populatie. De variantie van de testscores in deze steekproef is gelijk aan 100. In een representatieve steekproef studenten van de Faculteit Sociale Wetenschappen is de variantie van de testscores slechts gelijk aan 40. Wat is dan de betrouwbaarheid van deze vragenlijst in deze steekproef? Het invullen van de bovenstaande getallen in de bovenstaande formule levert het volgende voor de ónbetrouwbaarheid (i.e., $1 - r_{YY'}$): $100/40(1 - 0.70) = 0.75$. Dus, de betrouwbaarheid in deze steekproef is slechts gelijk aan $1 - 0.75 = 0.25$.

### 2.2.10 De Standaardfout van Meting

In de praktijk is men soms geïnteresseerd in de ware-score van een persoon. De vraag is wat men kan zeggen over die niet-geobserveerde ware-score $T_p$ op basis van de geobserveerde testscore $X_p$. Volgens de Klassieke Testtheorie schommelt $X_p$ rond $T_p$. De variantie van deze schommelingen hebben we aangeduid met $\text{VAR}(E_p)$. We gaan er nu vanuit dat elke persoon dezelfde $\text{VAR}(E_p)$ heeft. Omdat $\text{VAR}(E_p)$ dan gelijk is aan $\text{VAR}(E)$ (het gemiddelde van alle $\text{VAR}(E_p)$'s in de populatie), kunnen we van de betrouwbaarheid gebruik maken om $\text{VAR}(E_p)$ te bepalen. Men kan namelijk het volgende aantonen (voor een afleiding, zie pagina 126):

$$\boxed{\text{SD}(E) = \text{SD}(X)\sqrt{1 - r_{XX'}} \quad .}$$

Wat hier staat is intuïtief aannemelijk: als de betrouwbaarheid toeneemt ($r_{XX'}$ komt korter bij 1) dan wordt $\text{SD}(E)$ kleiner. De standaarddeviatie van de meetfouten ($\text{SD}(E)$) wordt de *standaardfout van meting* genoemd.

De standaardfout van meting zegt ons hoe sterk $X_p$ schommelt rond $T_p$. Het is ook mogelijk om "de andere kant op te denken". De standaardfout van meting zegt ons dan iets over de aannemelijkheid dat $T_p$ in de buurt van $X_p$ ligt. Deze idee brengt ons bij het betrouwbaarheidsinterval. Een betrouwbaarheidsinterval is een interval rond $X_p$ dat met een bepaalde kans de ware-score $T_p$ bevat. Bijvoorbeeld:

1. Het interval $[X_p - 1\text{SD}(E), X_p + 1\text{SD}(E)]$ is een 68%- betrouwbaarheidsinterval omdat het in 68 procent van de gevallen de ware score $T_p$ bevat.

2. Het interval $[X_p - 1.96\text{SD}(E), X_p + 1.96\text{SD}(E)]$ is een 95%- betrouwbaarheidsinterval omdat het in 95 procent van de gevallen de ware-score $T_p$ bevat.

Hierbij zijn we er stilzwijgend vanuit gegaan dat $X_p$ normaal verdeeld is rond $T_p$. Als deze veronderstelling niet klopt, dan zijn de bovenstaande percentages (68 en 95) niet correct. In veel situaties is de normale verdeling echter een plausibele veronderstelling.

### 2.2.11 Correctie voor Attenuatie

In de praktijk is men vaak geïnteresseerd in de correlatie tussen twee tests. Als die correlatie laag is, dan kan dat aan twee factoren liggen:

1. De betrouwbaarheid van de tests. Als één van de twee tests onbetrouwbaar is of ze zijn beiden onbetrouwbaar, dan kunnen ze nooit hoog correleren. Onbetrouwbare testscores bestaan immers voor een groot gedeelte uit meetfouten, en meetfouten correleren met niets.

2. De ware-scores van de twee tests hebben een lage correlatie.

Wat men verder doet met een lage correlatie hangt af van de vraag waar men in geïnteresseerd is:

1. Als men de scores op één van beide tests wil gebruiken om de scores op de andere test te voorspellen, dan is de kous af: met een lage correlatie kan men niet goed voorspellen.

2. Als men vooral theoretisch geïnteresseerd is, dan is met die lage correlatie nog niet alles gezegd. Stel dat de onderzoeker op grond van zijn theorie verwacht dat de twee tests hoog moeten correleren. Nu, deze theorie gaat zeker niet over meetfouten. Dus, de verwachting van de onderzoeker is dat de ware-scores van twee tests hoog moeten correleren. Een lage geobserveerde correlatie is daarom niet noodzakelijk strijdig met zijn verwachting. Die kan namelijk ook het gevolg zijn van de lage betrouwbaarheid van de tests.

Om de theoretisch geïnteresseerde onderzoeker van dienst te zijn, werd daarom een formule afgeleid waarmee hij de correlatie tussen de ware-scores kan berekenen op basis van (1) de geobserveerde correlatie tussen de testscores, en (2) de betrouwbaarheden van de twee tests. De scores op de twee tests duiden we aan met $X$ en $Y$. De ware-scores en de meetfouten van $X$ duiden we aan met $T_X$ en $E_X$, en de ware-scores en de meetfouten van $Y$ met $T_Y$ en $E_Y$. Ons vertrekpunt is de basisformule van de Klassieke Testtheorie, toegepast op $X$ en $Y$:

$$\begin{aligned} X &= T_X + E_X \\ Y &= T_Y + E_Y \quad . \end{aligned}$$

En we maken de bijkomende veronderstelling dat de meetfouten $E_X$ en $E_Y$ ongecorreleerd zijn. Men kan nu aantonen dat de correlatie tussen de ware-scores $T_X$ en $T_Y$ als volgt berekend kan worden (voor een afleiding, zie pagina 127):

$$\boxed{r_{T_X T_Y} = \frac{r_{XY}}{\sqrt{r_{XX'}}\sqrt{r_{YY'}}} \quad .}$$

Dit is de formule voor correctie voor attenuatie. “Attenuatie” komt uit het Latijn, en betekent “afzwakking”. De formule corrigeert voor de afzwakking van de correlatie die het gevolg is van de meetfouten in $X$ en $Y$. De correctie bestaat erin dat de geobserveerde correlatie ($r_{XY}$) gedeeld wordt door het product van de wortels uit de betrouwbaarheden ($r_{XX'}$ en $r_{YY'}$). Deze betrouwbaarheden zijn altijd kleiner of gelijk aan 1, en hetzelfde geldt voor het product van hun vierkantswortels. Delen door een getal kleiner dan 1 komt neer op vermenigvuldigen met een getal groter dan 1. Daarom zal $r_{T_X T_Y}$ in het linkerlid van de attenuatieformule altijd extremer zijn dan $r_{XY}$.

Een veelgemaakte fout is dat men denkt dat men met de attenuatieformule de ware-scores kan berekenen. Dit is niet zo. Met de attenuatieformule kan men alleen de correlatie tussen de ware-scores berekenen.

Om gebruik te kunnen maken van de attenuatieformule moet men beschikken over de betrouwbaarheden $r_{XX'}$ en $r_{YY'}$. Voor het bepalen van deze betrouwbaarheden kan men gebruik maken van één van de procedures die in het voorgaande besproken werden (alternatieve-vormmethode, hertesting, halvering of alpha). Men kan echter geen harde garantie geven dat men met deze procedures ook echt de betrouwbaarheid bepaalt. Zoals we reeds aangegeven hebben, hangt dit af van het feit of bepaalde testvormen, testherhalingen, deeltests of items parallel zijn, en dit is niet met zekerheid vast te stellen. Men moet vooral voorzichtig zijn met het gebruik van alpha in combinatie met de attenuatieformule. Alpha is immers een óndergrens van de betrouwbaarheid, en zorgt daarom voor een óvercorrectie.

De attenuatieformule is nuttig als men geïnteresseerd is in de correlatie tussen de niet-observeerbare ware-scores $T_X$ en $T_Y$. Als men geïnteresseerd is in hoe de geobserveerde correlatie $r_{XY}$ totstandkomt, kan men de formule ook andersom gebruiken. Men drukt $r_{XY}$ dan uit als functie van $r_{T_X T_Y}$ en de betrouwbaarheden van de twee tests. Door de attenuatieformule te herschrijven krijgt men het volgende:

$$r_{XY} = r_{T_X T_Y} \sqrt{r_{XX'}} \sqrt{r_{YY'}} \quad .$$

Uit deze formule kunnen we het volgende afleiden:

1. De correlatie $r_{XY}$ is kleiner naarmate de twee tests minder betrouwbaar zijn.

2. De correlatie $r_{XY}$ kan nooit hoger worden dan $\sqrt{r_{XX'}}\sqrt{r_{YY'}}$, en hij is precies gelijk aan $\sqrt{r_{XX'}}\sqrt{r_{YY'}}$ als $T_X$ en $T_Y$ perfect met elkaar correleren, want dan is $r_{T_X T_Y}$ gelijk aan 1.

3. De correlatie van $X$ met een andere variabele $Y$ kan nooit hoger zijn dan $\sqrt{r_{XX'}}$, en hij is gelijk aan $\sqrt{r_{XX'}}$ als (1) $Y$ perfect betrouwbaar is ($Y = T_Y$) en (2) $T_X$ en $T_Y$ perfect met elkaar correleren. Dus, de wortel uit de betrouwbaarheid is de bovengrens van de correlatie. Wat omstandiger geformuleerd, de correlatie van een test met een andere variabele wordt begrensd door de betrouwbaarheid van de test.

   De maximale correlatie van $X$ met $Y$ wordt slechts in één geval bereikt: wanneer $Y$ perfect correleert met $T_X$. Dus, $X$ kan nooit hoger correleren met een andere variabele dan met haar eigen ware-score.

### 2.2.12 Berekenen van de Betrouwbaarheid met de SPSS-procedure Reliability

Nu volgt een korte beschrijving van de SPSS-procedure Reliability (Analysis), die gebruikt wordt om de betrouwbaarheid van een test te berekenen. Deze beschrijving leest men best vóór een computer waarop SPSS draait.

Men vindt Reliability onder *Scale* binnen het menu *Analyze*. In het dialoogvenster van Reliability kies je de items van de test door ze naar het vakje onder *Items* over te brengen. In dit dialoogvenster moet je een keuze maken uit vijf zogenaamde *models*: alpha, split-half, Guttman, parallel en strict parallel. De naam "model" is ongelukkig gekozen, want het gaat hier over vijf verschillende methoden om de betrouwbaarheid te berekenen. Hiervan werden in dit hoofdstuk alleen de eerste twee (alpha en halvering) besproken.

Met de keuze "Model: Alpha" krijgt men in de uitvoer eenvoudig Cronbachs alpha. Met de keuze "Model: Split-half" krijgt men, behalve de halveringsbetrouwbaarheid, ook informatie die in dit hoofdstuk niet besproken werd. Ten eerste, in de uitvoer wordt zowel de *equal-length Spearman-Brown* als de *unequal-length Spearman-Brown* gegeven. De *equal*-length Spearman-Brown is de halveringsbetrouwbaarheid zoals beschreven in dit hoofdstuk ("Spearman-Brown" verwijst naar de toepassing van de Spearman-Brown formule met $k = 2$). De *unequal*-length Spearman-Brown wordt berekend met een

formule die corrigeert voor het feit dat de twee testhelften mogelijk uit een verschillend aantal items bestaan, hetgeen kan zorgen voor een schending van de parallelliteit tussen de twee testhelften. Indien de twee testhelften uit een gelijk aantal items bestaan, zijn de equal-length en de unequal-length Spearman-Brown altijd gelijk. Ten tweede, met de keuze "Model: Split-half" krijgt men in de uitvoer ook altijd de zogenaamde *Guttman split-half* betrouwbaarheid. Dit is een andere manier om de betrouwbaarheid te berekenen op basis van een opsplitsing in twee testhelften. En ten derde, men krijgt in de uitvoer ook de Cronbachs alpha voor elk van de testhelften afzonderlijk. Het is niet duidelijk wat men met deze informatie kan doen.

Bij het berekenen van de halveringsbetrouwbaarheid met Reliability, kan men zelf bepalen hoe de items over de twee testhelften verdeeld worden, namelijk door de items in een bepaalde volgorde naar het vakje onder *Items* over te brengen. De procedure Reliability stopt immers de eerste helft van de items (item 1 t/m item 5 als er in totaal tien items zijn) in de eerste testhelft en de tweede helft van de items (item 6 t/m item 10 als er in totaal tien items zijn) in de tweede testhelft. Door de volgorde van de items te veranderen, zorgt men dus voor een andere verdeling van de items over de twee testhelften.

In het dialoogvenster van Reliability staat ook een knop *Statistics* waarmee men een aantal statistieken kan vragen die verband houden met de betrouwbaarheid. Men moet hier denken aan itemgemiddelden en -varianties, inter-item-correlaties en -covarianties, het schaalgemiddelde en de schaalvariantie na weglating van één item (optie *Scale if item deleted*), en een aantal statistieken uit de variantieanalyse.

---

*Opgave 21*
Twee onderzoekers deden elk een onderzoek naar de betrouwbaarheid van een test voor introversie-extraversie. De eerste onderzoeker nam de test af bij een groep personen waarvan men kan verwachten dat ze allemaal vrij extravert zijn (bijvoorbeeld, studenten aan een toneelacademie). De andere onderzoeker nam de test af bij een groep personen waarvan men kan verwachten dat zij allemaal vrij introvert zijn (bijvoorbeeld, de regelmatige bezoekers van een bibliotheek). Stel nu dat men de betrouwbaarheid berekent op beide groepen samen. Zal deze dan groter, kleiner of dezelfde zijn dan de betrouwbaarheden op basis van de afzonderlijke groepen? Motiveer je antwoord.

*Opgave 22*
Karel behaalde een score van 115 op een IQ-test. De betrouwbaarheid van deze test is 0.91, en de testscorevariantie van de groep waarop deze betrouwbaarheid bepaald werd, was 225. Bepaal een 95%-betrouwbaarheidsinterval voor Karels ware-score.

*Opgave 23*
Twee tests hebben dezelfde ware-scores. De ene test heeft een hoge betrouwbaarheid en de andere test heeft een lage. Van welke test zullen de scores het hoogst correleren met andere testscores? Beargumenteer je antwoord.

*Opgave 24*
Met welke van de onderstaande uitspraken ben je het eens, en met welke ben je het oneens? Leg uit waarom wel of waarom niet.

1. Als een test hoog correleert met een andere test, dan moet hij ook een hoge betrouwbaarheid hebben.

2. Een test die niet correleert met een bepaalde andere test moet ook een lage betrouwbaarheid hebben.

*Opgave 25*
Twee onderzoekers ontwikkelden ieder een vragenlijst om stressbestendigheid te meten. De vragenlijst van onderzoeker A bestaat uit 20 vragen en bleek bij een groep kaderleden van financiële instellingen een betrouwbaarheid van 0.70 te hebben. De vragenlijst van onderzoeker B bestaat uit 16 vragen en bleek bij een representatieve steekproef uit de Nederlandse populatie eveneens een betrouwbaarheid van 0.70 te hebben.

1. Hoe kun je onderzoeken of beide tests wel hetzelfde construct meten? Motiveer je antwoord.
2. Als je één van de twee tests wil gebruiken om de stressbestendigheid van een groep jonge werknemers te meten, welke test zou je dan kiezen? Motiveer je antwoord.

*Opgave 26*
Er werden twee nieuwe tests ontwikkeld, een intelligentietest (A) en een attitudevragenlijst (B). De intelligentietest had een betrouwbaarheid van 0.80 en de attitudevragenlijst een betrouwbaarheid van 0.60. Beide tests hadden een correlatie van 0.67 met het studiesucces. Met welk kenmerk vertoont het studiesucces de grootste samenhang, met intelligentie zoals gemeten door test A, met attitude zoals gemeten door test B, of is er geen verschil? Leg uit.

*Opgave 27*
De scores op een nieuw ontwikkelde intelligentietest correleerden even hoog met een bestaande numerieke intelligentietest als met een bestaande verbale intelligentietest. Voor beide tests was de correlatie duidelijk positief. Voor de numerieke test werd een betrouwbaarheid van 0.90 gerapporteerd, terwijl de betrouwbaarheid van de verbale test op 0.60 uitkwam.

1. Met welk kenmerk vertoont de in de nieuwe test gemeten vorm van intelligentie de grootste samenhang, met numerieke of met verbale intelligentie, of valt dit niet te zeggen? Leg uit.
2. Hoe hoog kan de correlatie tussen de bestaande numerieke en verbale intelligentietests maximaal zijn? Wat zouden we moeten concluderen als dit maximum gehaald werd? Leg uit.

*Opgave 28*
Een onderzoeker berekende de correlatie tussen de ware-scores op twee tests met behulp van de attenuatieformule. Hij bepaalde de betrouwbaarheden van de tests met alpha. De onderzoeker was verbaasd dat de correlatie groter was dan 1. De steekproef was zo groot dat het niet aan toevallige steekproeffluctuaties kon liggen. Hoe kun je dit vreemde resultaat verklaren?

*Opgave 29* COMPUTEROPDRACHT
Een klinisch psycholoog is geïnteresseerd in het fenomeen sociale angst; angst die zich voordoet binnen sociale situaties. Volgens sommige psychologen speelt sociale angst een rol bij subassertief gedrag. Voor zijn onderzoek maakte de psycholoog gebruik van een zelfbeoordelingsvragenlijst om de mate van sociale angst te bepalen. De vragenlijst bestaat uit 10 uitspraken (items). De proefpersoon moet bij elke uitspraak aangeven of hij/zij deze wel (score=1) of niet (score =0) op zichzelf van toepassing acht.

Het betreft de volgende uitspraken:

1. Het gebeurt vaak dat ik de persoon met wie ik spreek niet recht in de ogen durf te kijken.
2. Als ik in een groep ben, dan verkies ik het meestal om me op de achtergrond te houden.
3. Als ik in een vreemde stad ben, en de weg niet ken, durf ik niet iemand aan te spreken om me de weg te laten wijzen.

4. Ik ben vaak zo bevreesd om besmet te raken met bepaalde ziektes dat ik het contact met anderen liever mijd.
5. Als ik in een schoenen- of kledingzaak een aantal schoenen of kledingsstukken gepast heb, en er zijn er geen die mij bevallen, dan gebeurt het toch vaak dat ik ze koop, alleen maar omdat ik me verplicht voel ten opzichte van de verkoopster.
6. Ik vind het bijzonder vervelend om voor een groep te spreken.
7. Als een kennis me om een gunst vraagt, dan vind ik het erg moeilijk om te weigeren.
8. Ik vind anderen vaak zo saai dat ik maar liever alleen blijf.
9. Als iemand me onterecht ergens van beschuldigt, dan gebeurt het toch vaak dat ik maar zwijg.
10. Als iemand onvriendelijk is tegen mij, dan kruip ik meestal in mijn schulp.

De onderzoeker wilde weten of de zelfgerapporteerde graad van sociale angst in verband staat met het voorkomen van spanningshoofdpijn (een bepaald soort hoofdpijn, onder andere gekenmerkt door het gevoel van een knellende band rond het hoofd). Daartoe vroeg hij aan de proefpersonen om in een periode van twee maanden te tellen op hoeveel dagen zij last hadden van spanningshoofdpijn.

Aan dit gefingeerde onderzoek namen 300 proefpersonen deel. Alle gegevens zijn opgeslagen in het SPSS-bestand "soangst.sav". De antwoorden op de uitspraken staan onder de variabelenamen ITEM1 tot en met ITEM10 en het aantal dagen spanningshoofdpijn onder variabelennaam HOOFDPIJN.

1. Bereken alpha met de SPSS-procedure Reliability. Selecteer ook de optie *Scale if item deleted* en vraag om de inter-item-correlaties.
2. Hoe zie je in de uitvoer dat twee items niet goed passen in de schaal? Probeer een inhoudelijke verklaring te geven voor het feit dat juist deze twee items niet goed passen in de schaal.
3. Doe dezelfde analyse als bij de vorige opdracht, maar nu op de acht items die overblijven na verwijdering van de twee die niet goed passen. Wat concludeer je ten aanzien van de betrouwbaarheid van de gereduceerde schaal?
4. Bereken de halveringsbetrouwbaarheid van de volledige schaal (dus met alle tien items). Zet de optie *Scale if item deleted* uit en vraag niet meer om de inter-item-correlaties.
5. Doe dezelfde analyse nu nog een keer, alleen zorg er nu voor dat de schaal op een andere manier in twee groepen verdeeld wordt. Vergelijk de uitvoer van deze analyse met de uitvoer van de vorige analyse. Hoe kun je je resultaten verklaren? Is het überhaupt wel te verantwoorden dat je de betrouwbaarheid van deze schaal berekent met de halveringsmethode?
6. Bereken de correlatie tussen de score op de gereduceerde schaal en het aantal dagen spanningshoofdpijn. Hiertoe moet dus een nieuwe variabele gecreëerd worden die gelijk is aan de som van de acht items van de gereduceerde schaal. Gebruik hiervoor de procedure *Compute*.
7. Verwijder alle delen van de uitvoer die je niet wil bewaren. Druk de rest af.

## 2.3 Validiteit

We onderscheiden drie vormen van validiteit: inhouds-, construct- en criteriumvaliditeit. Deze drie vormen van validiteit zijn elk verbonden met een bepaalde opvatting van het gemeten psychologisch kenmerk:

1. Bij inhoudsvaliditeit gaat men ervan uit dat er een definitie is van het kenmerk.
2. Bij constructvaliditeit gaat men ervan uit dat het kenmerk een plaats heeft binnen een psychologische theorie. Op basis van die theorie verwacht men bepaalde verbanden met andere tests.
3. Bij criteriumvaliditeit is men alleen geïnteresseerd in de nuttigheidswaarde van het gemeten kenmerk. Men vraagt zich af of men er een andere variabele mee kan voorspellen.

Criteriumvaliditeit is van belang in situaties waarin men beslissingen moet nemen, bijvoorbeeld over het wel of niet toelaten van iemand tot een bepaalde opleiding, of het selecteren van iemand voor een bepaalde functie. Op het einde van deze paragraaf staan we hier even bij stil.

### 2.3.1 Inhoudsvaliditeit

Inhoudsvaliditeit is een begrip uit de onderwijskunde. Het begrip heeft betrekking op kennis- en vaardigheidstoetsen, zoals bijvoorbeeld de Eindtoets Basisonderwijs (de CITO-toets) en het centrale eindexamen. Voor dit soort toetsen bestaan er vaak duidelijke specificaties over het soort items waaruit de toets moet bestaan. Het is duidelijk dat de inhoud van een toets of een examen overeen moet stemmen met datgene wat in het onderwijs behandeld werd.

De inhoud van het onderwijs wordt vaak vastgelegd in de vorm van zogenaamde *eindtermen*. Eindtermen zijn omschrijvingen van wat een leerling allemaal moet kennen en/of kunnen na afloop van het onderwijs. Bijvoorbeeld, het vak Wiskunde B zou onder andere de volgende eindtermen kunnen bevatten:

1. Het kunnen bepalen van de afgeleiden van de trigoniometrische functies sinus, cosinus en tangens.
2. Het kunnen vinden van een onbepaalde integraal met behulp van de techniek partieel integreren.
3. Het kunnen berekenen van de oppervlakte onder een parabole en een hyperbole curve.
4. Het kunnen bepalen van de inverse van een $(2 \times 2)$-matrix.

Een toets of examen is inhoudelijk valide als de items een goede afspiegeling zijn van de eindtermen. Het is meestal zó dat niet alle eindtermen even belangrijk zijn. In dat geval spreken we van een inhoudelijk valide toets als de items een goede afspiegeling zijn van de eindtermen én hun relatieve belangrijkheid. Rekening houden met deze relatieve

belangrijkheid vraagt extra aandacht als, omwille van tijdsbeperkingen, niet voor elke eindterm een item geformuleerd kan worden. Dit is vrijwel steeds het geval. In zo'n situatie betekent inhoudsvaliditeit dat de toets relatief veel items bevat over onderwerpen die volgens de eindtermen belangrijk zijn, en weinig items over onderwerpen die minder belangrijk zijn. De inhoudsvaliditeit van een toets wordt dus níet uitgedrukt in een getal. Het is een oordeel over de mate van overeenstemming tussen de inhoud van de toets en de lijst van eindtermen.

In veel situaties beschikken wij niet over een lijst van eindtermen. Het bepalen van de inhoudsvaliditeit kan dan gebeuren op basis van (1) de inhoud van het gebruikte handboek of (2) het oordeel van experts. In het eerste geval kijkt men of de inhoud van de toets overeenstemt met de inhoud van het handboek en het relatieve belang dat aan de verschillende onderdelen gegeven wordt (bijvoorbeeld, zoals uitgedrukt in het aantal pagina's per onderwerp). In het tweede geval legt men de toets voor aan experts op het betreffende inhoudelijk gebied.

Bij het formuleren van items op basis van eindtermen of de inhoud van het handboek, is het vaak handig om een onderscheid te maken tussen onderwerpen en vaardigheden. De onderwerpen hebben betrekking op de inhoud van het geleerde en de vaardigheden hebben betrekking op wat men met het geleerde kan doen. Relevante vaardigheden zijn bijvoorbeeld reproduceren van informatie, uitrekenen van formules, en het toepassen van principes in nieuwe situaties. We kunnen dit illustreren aan de hand van het tentamen voor het vak Methoden en Technieken. Twee onderwerpen die in dit vak aan de orde komen zijn testverlenging met parallelle deeltests en Cronbachs alpha. En twee vaardigheden die relevant zijn voor dit vak zijn uitrekenen van formules en toepassen van principes. Het formuleren van items begint nu met het kruisen van de onderwerpen met de vaardigheden. Met twee onderwerpen en twee vaardigheden krijgt men dan een $(2 \times 2)$-tabel, zoals getoond wordt in Tabel 2.5. Zo'n tabel wordt een *toetsmatrijs* genoemd. In de cellen van de toetsmatrijs schrijft men de items die betrekking hebben op het onderwerp in de rij en die de vaardigheid in de kolom toetsen.

Het is duidelijk dat voor een echte toets de toetsmatrijs altijd uit meer dan twee rijen (onderwerpen) bestaat en dat het aantal kolommen (vaardigheden) ook groter kan zijn dan twee. Het is niet nodig dat alle cellen gevuld zijn, dat er in elke rij evenveel items staan, noch dat er in elke kolom evenveel items staan. Door de rijen en de kolommen onderling te laten verschillen wat betreft het aantal items geeft men aan welke onderwerpen en vaardigheden het belangrijkst zijn voor het vak.

### 2.3.2 Constructvaliditeit

Constructvaliditeit betreft de vraag of de test het psychologisch kenmerk (construct) meet dat de test geacht wordt te meten. De meest eenvoudige manier om dit te achterhalen is door te onderzoeken wat de samenhang is met andere tests die hetzelfde of een nauw verwant kenmerk meten. Bijvoorbeeld, de constructvaliditeit van een nieuwe intelligentietest kan onderzocht worden door de correlaties te berekenen met een aantal bestaande intelligentietests. Het is een teken van constructvaliditeit als deze correlaties hoog zijn. Deze vorm van constructvaliditeit (hoge correlaties met tests die hetzelfde meten) wordt *convergente validiteit* genoemd.

Hoge correlaties met andere tests die hetzelfde kenmerk meten zijn alleen dán een

| | Formules Uitrekenen | Principes Toepassen |
|---|---|---|
| Testverlenging met parallelle deeltests | Reken de Spearman-Brown formule uit voor testverdubbeling ($k = 2$) en een correlatie tussen de testhelften van 0.70. | Twee tests, één met 20 en één met 30 items, hebben dezelfde betrouwbaarheid. Welk test heeft de meest betrouwbare items? Waarom? |
| Cronbachs alpha | Reken Cronbachs alpha uit voor een test met 5 items, waarvan de item-varianties allemaal gelijk zijn aan 2, en de testscore-variantie gelijk is aan 40. | Een test wordt verdubbeld qua lengte. Dit gebeurt ofwel met items die niet correleren met de oorspronkelij -ke items, ofwel met items die sterk correleren met de oorspronkelijke items. Heeft dit gevolgen voor hoe goed Cronbachs alpha de betrouwbaarheid van de test benadert? Leg uit. |

Tabel 2.5: Toetsmatrijs voor twee onderwerpen en twee vaardigheden voor het vak Methoden en Technieken.

indicatie van constructvaliditeit als die andere tests ook werkelijk dit kenmerk meten. Met andere woorden, hoge correlaties met andere tests zijn alleen dán een indicatie van constructvaliditeit als men er zeker van is dat die andere tests constructvalide zijn. Als de constructvaliditeit van die andere tests zelf ook gebaseerd is op hoge correlaties met wéér andere tests, dan schuilt hier het gevaar van het doorschuiven van correlaties. Hiermee bedoelen we dat de constructvaliditeit van een bepaalde test uiteindelijk afhangt van de correlatie met een test uit een ver verleden waarvan men niet meer weet waarom zij meet wat zij geacht wordt te meten (constructvalide is).

De tegenhanger van de convergente validiteit is de *divergente validiteit*. Divergente validiteit betekent dat een test niet of laag correleert met andere tests die een ongerelateerd, irrelevant construct meten. Bijvoorbeeld, het is een teken van divergente validiteit dat een vragenlijst voor neuroticisme niet correleert met een intelligentietest. Dus, divergente validiteit heeft te maken met nulcorrelaties, en niet met negatieve correlaties, zoals wel eens gedacht wordt.

Het onderzoek naar convergente en divergente validiteit blijft niet beperkt tot correlaties tussen tests onderling. Men kan ook de correlaties onderzoeken tussen een bepaalde test- en een niet-test-variabele (i.e., een variabele die geen testscore is) zoals schooluitslagen of beoordelingen door anderen. Voor een intelligentietest, bijvoorbeeld, is het een indicatie van convergente validiteit als hij hoog correleert met schooluitslagen. We verwachten immers dat intelligente leerlingen over het algemeen hogere schooluitslagen behalen dan minder intelligente leerlingen. Deze correlatie met schooluitslagen mag echter niet zo hoog zijn als de correlatie met andere intelligentietests, want dit

zou erop wijzen dat de test geen intelligentie meet maar het presteren op school. Een ander voorbeeld is een vragenlijst voor extraversie. Van deze vragenlijst verwachten wij dat hij hoog zal correleren met een beoordeling door vrienden van de mate waarin iemand spraakzaam en levendig is. Een hoge correlatie met deze beoordelingen is dus een indicatie van convergente validiteit.

Bij convergente en divergente validiteit bekijkt men afzonderlijke correlaties. Correleert de test hoog met een bepaalde andere variabele, of correleert hij laag? Er zijn twee methoden voor het onderzoek naar constructvaliditeit waarbij men niet zozeer naar afzonderlijke correlaties kijkt, maar naar het *patroon* van correlaties tussen een groot aantal variabelen: (1) factoranalyse en (2) de *multitrait-multimethod* (MTMM) procedure.

Alhoewel de factoranalyse pas in hoofdstuk 5 behandeld wordt, geven we hier toch kort aan hoe deze techniek gebruikt wordt binnen het onderzoek naar de constructvaliditeit. Kort geformuleerd, het gebruik van de factoranalyse binnen het onderzoek naar de constructvaliditeit komt erop neer dat men

1. de correlaties berekent tussen een groot aantal variabelen, waaronder de test(s) waarvan men de constructvaliditeit onderzoekt, en

2. voorspellingen doet over de factoren waaruit de verschillende testscores zijn opgebouwd, en

3. kijkt of deze voorspellingen ook uitkomen (na een eventuele rotatie).

Zo kan men bijvoorbeeld de voorspelling maken dat de scores op een test voor verbale intelligentie opgebouwd zijn uit de factoren algemene intelligentie (de g-factor) en woordenschatkennis, en níet uit de factoren ruimtelijk voorstellingsvermogen en kortetermijngeheugen.

De MTMM procedure vertrekt van een aantal verschillende meetmethoden (vandaar "multimethod") om elk van een aantal verschillende kenmerken (vandaar "multitrait") te meten. Deze kenmerken zijn bijvoorbeeld de volgende: vriendelijkheid ($A$), gewetensvolheid ($B$) en neuroticisme ($C$). En de methoden waarmee we deze kenmerken meten zijn bijvoorbeeld: een vragenlijst ($a$), een zelfbeschrijving in termen van tien vrij te kiezen adjectieven ($b$), en een beoordeling van gedrag in een aantal gestandaardiseerde situaties ($c$). Voor elke persoon krijgen we dus negen scores: drie vriendelijkheidscores, één voor elk van de drie methoden en aangeduid met $A_a$, $A_b$ en $A_c$; drie gewetensvolheidscores, aangeduid met $B_a$, $B_b$ en $B_c$; en drie neuroticismescores, aangeduid met $C_a$, $C_b$ en $C_c$. Men berekent dan de correlaties tussen deze negen variabelen en schrijft ze in een matrix. In Tabel 2.6 wordt zo'n matrix getoond.

Op de diagonaal staan tussen haakjes de betrouwbaarheden. Deze betrouwbaarheden worden de *monotrait-monomethod* correlaties genoemd, omdat ze beschouwd kunnen worden als de correlaties tussen twee parallelle tests. De correlaties tussen tests van hetzelfde kenmerk maar op basis van verschillende methoden worden vet gedrukt. Deze correlaties worden *monotrait-heteromethod* correlaties genoemd. De correlaties tussen tests van verschillende kenmerken maar op basis van dezelfde methode worden schuin gedrukt. Deze correlaties worden *heterotrait-monomethod* correlaties genoemd. En de correlaties tussen tests van verschillende kenmerken op basis van verschillende methoden

worden in gewoon lettertype gedrukt. Deze correlaties worden *heterotrait-heteromethod* correlaties genoemd. De correlaties boven de diagonaal zijn weggelaten omdat ze hetzelfde zijn als de correlaties onder de diagonaal.

| | Kenmerk | Methode $a$ | | | Methode $b$ | | | Methode $c$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $A_a$ | $B_a$ | $C_a$ | $A_b$ | $B_b$ | $C_b$ | $A_c$ | $B_c$ | $C_c$ |
| Methode $a$ | $A_a$ | (.89) | | | | | | | | |
| | $B_a$ | *.51* | (.89) | | | | | | | |
| | $C_a$ | *.38* | *.37* | (.76) | | | | | | |
| Methode $b$ | $A_b$ | **.57** | .22 | .09 | (.93) | | | | | |
| | $B_b$ | .22 | **.57** | .10 | *.68* | (.94) | | | | |
| | $C_b$ | .11 | .11 | **.46** | *.59* | *.58* | (.84) | | | |
| Methode $c$ | $A_c$ | **.56** | .22 | .11 | **.67** | .42 | .33 | (.94) | | |
| | $B_c$ | .23 | **.58** | .12 | .43 | **.66** | .34 | *.67* | (.92) | |
| | $C_c$ | .11 | .11 | **.45** | .34 | .32 | **.58** | *.58* | *.60* | (.85) |

Tabel 2.6: Een hypothetische Multitrait-Multimethod-matrix.

Binnen de MTMM procedure hanteert men twee criteria voor constructvaliditeit, een streng en een zwak criterium. Het strenge criterium komt erop neer dat de inhoudelijke samenhang groter moet zijn dan de samenhang door een gemeenschappelijke meetmethode. Concreet betekent dit dat de monotrait-heteromethod correlaties (vet gedrukt) groter moeten zijn dan de heterotrait-monomethod correlaties (schuin gedrukt). Voor de correlaties in Tabel 2.6 is dat niet altijd het geval: (1) de monotrait-heteromethod correlaties van kenmerk $C$ zijn vaak lager dan de heterotrait-monomethod correlaties, en (2) de heterotrait-monomethod correlaties van methode $b$ en $c$ zijn ongeveer even groot als de monotrait-heteromethod correlaties.

Het zwakke criterium voor constructvaliditeit komt erop neer dat tests die hetzelfde kenmerk meten met verschillende methoden onderling hoger moeten correleren dan tests die verschillende kenmerken meten met verschillende methoden. Met andere woorden, de grootte van de correlaties moet afhangen van het feit of hetzelfde dan wel een verschillend kenmerk gemeten wordt. Dit betekent dat de monotrait-heteromethod correlaties (vet gedrukt) groter moeten zijn dan de heterotrait-heteromethod correlaties (normaal gedrukt). Voor de correlaties in Tabel 2.6 blijkt dit altijd het geval te zijn.

Correlaties tussen variabelen zijn niet de enige basis voor het onderzoek naar constructvaliditeit. Twee andere bronnen van informatie zijn (1) veranderingen in de testscores met de leeftijd (in de literatuur, *leeftijdsdifferentiatie* genoemd), en (2) het effect van interventies op de testscores. Van een aantal kenmerken, zoals intelligentie en woordenschatkennis, verwacht men dat zij toenemen met de leeftijd. Het is dan een teken van constructvaliditeit als ook de testscores toenemen met de leeftijd. Het is duidelijk dat dit een noodzakelijke maar geen voldoende voorwaarde is voor constructvaliditeit, want er zijn heel veel kenmerken die toenemen met de leeftijd (zoals lengte en gewicht, binnen een bepaald leeftijdsbereik).

Van sommige interventies verwacht men een effect op een bepaald kenmerk. Zo verwacht men van een rationeel-emotieve therapie dat de cliënt minder depressief wordt, van een sociale vaardigheidstraining dat de cliënt minder sociaal angstig wordt, en van

extra taalonderwijs dat de leesvaardigheid van de leerling toeneemt. Als een test een meting is van, respectievelijk, depressiviteit, sociale angst of leesvaardigheid, dan verwacht men dezelfde effecten op het niveau van de testscores. Het vaststellen van zo'n effect is dan evidentie voor de constructvaliditeit. Het nadeel van deze manier om de constructvaliditeit te onderzoeken, is dat het áfwezig zijn van zo'n effect niet eenduidig geïnterpreteerd kan worden. Het kan zijn dat de test het bedoelde kenmerk helemaal niet meet (de test is niet constructvalide), maar het kan ook zijn dat de interventie geen effect heeft.

### 2.3.3 Criteriumvaliditeit

Criteriumvaliditeit betreft de mate waarin de test een bepaalde andere variabele, het criterium, kan voorspellen. Om één of andere reden zijn de criteriumscores zelf niet beschikbaar, en daarom is het interessant om ze te voorspellen op basis van de testscores. Hierbij moeten we "voorspellen" ruim opvatten, in de zin dat de criteriumscores niet noodzakelijk in de toekomst, maar ook in het heden kunnen liggen. Wanneer de criteriumscores in de toekomst liggen, spreken we van *predictieve validiteit* (in het Engels, *predictive validity*), en wanneer ze in het heden liggen, spreken we van *gelijktijdige validiteit* (in het Engels, *concurrent validity*).

#### Predictieve Validiteit

Een voorbeeld van een situatie waarin predictieve validiteit van belang is, is de situatie waarin men wil voorspellen hoe goed iemand zal presteren in een bepaalde functie. Het bestaan van tests[7] op basis waarvan men toekomstig functioneren kan voorspellen, vormen het bestaansrecht van de vele centra voor personeelsselectie, zowel binnen als buiten de bedrijven. Voor een aantal functies bestaat er een duidelijke en objectieve maat voor hoe goed iemand presteert. Dit geldt bijvoorbeeld voor verkopers en vertegenwoordigers, waarvan de prestaties uitgedrukt kunnen worden in termen van verkoopcijfers. Voor andere functies is dat moeilijker, en maakt men meestal gebruik van beoordelingen door anderen. Vaak is het de directe chef die een beoordeling geeft van iemands functioneren. Voor leidinggevende functies zijn ook beoordelingen door ondergeschikten een belangrijke indicator van hoe goed iemand functioneert.

Een andere situatie waarin predictieve validiteit van belang is, is de situatie waarin men wil voorspellen hoe goed iemand zal presteren als hij een bepaalde opleiding volgt. Voor de hand liggende criteria zijn hier het gemiddelde tentamencijfer, het al dan niet succesvol beëindigen van de opleiding, en het aantal jaar dat men over de opleiding doet. Predictieve validiteit is vooral van belang als men een toelatingsexamen moet afleggen voor men aan de opleiding mag beginnen. Een toelatingsexamen met een hoge predictieve validiteit is kostenbesparend, want dan zal het merendeel van degenen die tot de opleiding toegelaten worden deze opleiding ook succesvol afronden.

[7] Tests in de ruime zin; dus ook de beoordeling op basis van het curriculum vitae, de beoordeling na een interview en ratings op basis van het gedrag in gestandaardiseerde situaties, zoals verzameld in zogenaamde "assessment centers".

### Gelijktijdige Validiteit

Gelijktijdige validiteit is van belang als de scores op de criteriumvariabele zeer moeilijk te verkrijgen zijn. Dit was aan de orde bij de ontwikkeling van de MMPI (Minnesota Multiphasic Personality Inventory). Dat was in een periode waarin het aantal psychiatrische ziekenhuisopnamen sterk steeg. De psychiatrische diagnose was zeer tijdrovend, en daarom zocht men naar een test waarmee men ongeveer dezelfde diagnose zou stellen. In de terminologie van deze paragraaf, men zocht naar een test met een hoge gelijktijdige validiteit voor de criteriumvariabele PSYCHIATRISCHE DIAGNOSE.

Een ander voorbeeld van een situatie waarin de criteriumvariabele zeer moeilijk te verkrijgen is, is bij een praktijktentamen voor autotechnicus. Als men *op de werkplek* wil nagaan of de leerling de belangrijkste defecten aan een auto correct kan diagnosticeren en herstellen, door hem effectief aan de auto's te laten werken, dan duurt zo'n tentamen zeker enkele dagen. Hier kan een schriftelijke toets van kennis van autotechniek voor een flinke tijdbesparing zorgen, tenminste als die test de "werkplek"-score goed kan voorspellen.

Zoals we dadelijk zullen zien, vereist het bepalen van de *predictieve* validiteit dat men (1) de testscores verzamelt, (2) wacht tot de criteriumscores beschikbaar zijn, en (3) het verband tussen test- en criteriumscores onderzoekt. Soms wil men een idee krijgen van de predictieve validiteit zonder te moeten wachten op de criteriumscores. Hier kan de gelijktijdige validiteit fungeren als vervanger voor de predictieve validiteit. Stel dat men kandidaten wil selecteren voor een opleiding tot medisch specialist. De test die men hiervoor wil gebruiken, kan men afnemen bij een groep artsen die hun opleiding reeds voltooid hebben of bij een groep die in de eindfase van hun opleiding zitten. Van deze groep kan men ongeveer tegelijkertijd de criteriumscores verzamelen (bijvoorbeeld, beoordelingen door het afdelingshoofd en de verpleging). Als de artsen de test op (ongeveer) dezelfde manier beantwoorden als zij gedaan zóuden hebben voor aanvang van hun opleiding, dan is de gelijktijdige validiteit die men op basis van deze gegevens kan bepalen vergelijkbaar met de predictieve validiteit. Dit geldt zeker niet voor alle tests. Bijvoorbeeld, als de "test" een beoordeling is na afloop van een interview, dan zullen de antwoorden van de arts beïnvloed zijn door zijn ervaringen in het ziekenhuis, hetgeen op zijn beurt zijn beoordeling zal beïnvloeden.

### De Validiteitscoëfficiënt

De basis voor de voorspelling is een steekproef waarin per persoon de scores op de test en het criterium zijn geobserveerd. In de context van personeelsselectie betekent dit bijvoorbeeld dat men zowel dient te beschikken over de scores op de situatieproeven van een assessment center, als over de scores op de criteriumvariabele. Op basis van deze gegevens onderzoekt men de samenhang tussen de test en het criterium en kan men een uitspraak doen over de kans dat iemand een bepaalde criteriumscore behaalt gegeven dat hij een bepaalde testscore heeft. Dergelijke uitspraken kan men niet alleen doen voor de personen op basis waarvan men deze samenhang onderzocht heeft, maar voor élke persoon waarvan men de testscore kent. Dit stelt ons in staat om voorspellingen te maken van de criteriumscores.

De accuraatheid van de voorspellingen hangt af van de sterkte van de samenhang

tussen test en criterium. Des te sterker de samenhang, des te accurater men kan voorspellen. Meestal wordt deze samenhang uitgedrukt in een correlatiecoëfficiënt tussen testscores $X$ en criteriumscores $Y$, aangeduid met $r_{XY}$. Deze correlatiecoëfficiënt wordt de *validiteitscoëfficiënt* genoemd.

In sommige gevallen is het criterium een categorische variabele, zoals bijvoorbeeld psychiatrische diagnose. In dit geval kan men de samenhang niet uitdrukken in een correlatiecoëfficiënt. Hier maakt men meestal gebruik van de *discriminantanalyse* om de aard en de sterkte van de samenhang te beschrijven.

### Correctie voor Attenuatie

In sommige gevallen is de validiteitscoëfficiënt laag, maar vindt men toch niet dat dit te wijten is aan de kwaliteit van de test. Dit is het geval als het criterium een lage betrouwbaarheid heeft. Een weinig betrouwbaar criterium bestaat vooral uit meetfouten, en het is onduidelijk met welke andere variabele deze meetfouten voorspeld kunnen worden. (Meetfouten zijn het gevolg van niet-gecontroleerde exogene factoren en endogene factoren die wisselen over de testafnamen.) We kunnen dit ook uitdrukken in een formule. We maken hierbij gebruik van de andersom geschreven attenuatieformule op pagina 100, die uitgaat van ongecorreleerde meetfouten in testscores $X$ en criteriumscores $Y$. Uit deze formule volgt dat

$$r_{XY} \leq \sqrt{r_{XX'}}\sqrt{r_{YY'}} \quad .$$

Dus, $r_{XY}$ is begrensd door het product van de wortels uit de betrouwbaarheden van $X$ en $Y$. Dit impliceert dat $r_{XY}$ nooit groter is dan de wortel uit $r_{YY'}$.

Om een idee te krijgen van wat de test waard zou zijn als het criterium een perfecte betrouwbaarheid zou hebben, kan men uitrekenen wat de correlatie is tussen de test en dat deel van het criterium dat geen meetfout bevat (i.e., $T_Y$, de ware score van $Y$). Deze correlatie kan men berekenen met de attenuatieformule. Omdat de ware score van test $X$ ($T_X$) verdwijnt uit de attenuatieformule, verdwijnt ook zijn betrouwbaarheid ($r_{XX'}$). We krijgen dan de volgende variant op de attenuatieformule:

$$r_{XT_Y} = \frac{r_{XY}}{\sqrt{r_{YY'}}} \quad .$$

Als voorbeeld van een situatie waarin deze formule gebruikt kan worden, stel dat je een voorspelling wil maken van iemands tevredenheid met zijn woning, drie jaar nadat hij in deze woning zijn intrek heeft genomen. Stel dat de vragenlijst waarmee deze tevredenheid gemeten wordt (i.e., de variabele $Y$) een betrouwbaarheid heeft van 0.64. En stel ook dat vragenlijst op basis waarvan de voorspelling gemaakt wordt (i.e., de variabele $X$) een correlatie heeft van 0.76 met de eerste vragenlijst (i.e., $r_{XY} = 0.76$). De correlatie tussen $X$ en $T_Y$, de ware-score-component van $Y$, is dan de volgende:

$$\begin{aligned} r_{XT_Y} =& \frac{0.76}{\sqrt{0.64}} \\ =& 0.95 \quad . \end{aligned}$$

Dit is bijna gelijk aan de perfecte correlatie van 1. Dus, de vragenlijst $X$ is een bijna perfecte voorspeller van de ware criteriumscore $T_Y$. Maar de vragenlijst $X$ is duidelijk

geen perfecte voorspeller van de geobserveerde criteriumscore $Y$ (want $r_{XY} = 0.76$); deze geobserveerde criteriumscore bevat immers ook een behoorlijk gedeelte meetfout, en die is met geen enkele andere variabele te voorspellen.

De formule waarin $r_{XY}$ begrensd wordt door het product van de wortels uit de betrouwbaarheden van $X$ en $Y$ kan men ook gebruiken om te bepalen wat de maximale validiteitscoëfficiënt is voor een bepaalde test. We krijgen deze maximale $r_{XY}$ als het criterium perfect betrouwbaar is ($r_{YY'} = 1$). In dit geval is $r_{XY}$ begrensd door de wortel uit $r_{XX'}$, de betrouwbaarheid van de test. Dit impliceert dat een onbetrouwbare test nooit een hoge criteriumvaliditeit kan hebben. Bijvoorbeeld, een test met een betrouwbaarheid van 0.64 kan ten hoogste een criteriumvaliditeit van 0.80 ($= \sqrt{0.64}$) hebben.

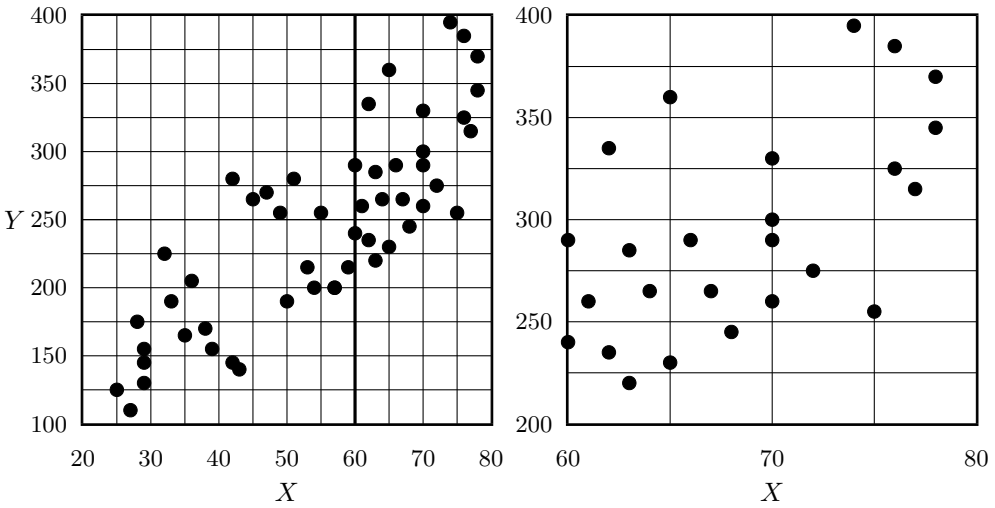
### Restriction-of-Range

Predictieve validiteit is meestal van belang in situaties waar men aan selectie doet. Het bepalen van de validiteitscoëfficiënt gebeurt echter op een steekproef waarin zowel de test- als de criteriumscores verzameld worden, maar waar geen selectie uitgevoerd wordt. Dit is nodig omdat de validiteitscoëfficiënt alleen correct berekend kan worden als men voor iedereen in de steekproef de criteriumscores kan bepalen (eindexamencijfers, verkoopcijfers, ...) en dit laatste is meestal alleen mogelijk als er geen selectie is. Dit plaatst ons in de vreemde situatie dat we, om de validiteitscoëfficiënt te kunnen berekenen, testscores moeten verzamelen die we daarna niet mogen gebruiken om kandidaten te selecteren.

Het naast zich neerleggen van de testscores (i.e., ze niet gebruiken voor selectie) is in sommige praktijksituaties niet vol te houden. Men berekent dan een validiteitscoëfficiënt op de test- en criteriumscores van alléén de personen die geselecteerd werden. Deze validiteitscoëfficiënt zal altijd lager zijn dan de validiteitscoëfficiënt berekend op de volledige groep. Het restriction-of-range fenomeen is hiervoor verantwoordelijk. In de paragraaf over betrouwbaarheid spraken we reeds over een restriction-of-range fenomeen en daar betrof het een verlaging van de betrouwbaarheid als functie van de homogeneïteit van de populatie. In deze paragraaf bespreken we een tweede restriction-of-range fenomeen, en hier heeft het betrekking op een verlaging van de criteriumvaliditeit als functie van de sterkte van de selectie.

We kunnen het restriction-of-range fenomeen beschrijven aan de hand van het spreidingsdiagram in Figuur 2.8. Op de horizontale as staat de testscore, en op de verticale de criteriumscore. In het linker gedeelte van de figuur staat de spreidingsdiagram die geobserveerd zóu zijn indien er niet geselecteerd werd op testscore. Er is een duidelijk positief verband tussen test en criterium. De kritische testscore bij de selectie was 60. Indien men hoger scoort dan 60, dan wordt men geselecteerd, en anders niet. In het rechter gedeelte van de figuur staat de spreidingsdiagram die men observeert na selectie. De punten in dit diagram zijn precies dezelfde als de punten rechts van de vette verticale lijn in het linker diagram. In deze geselecteerde gegevens is er geen verband meer tussen test en criterium.

Een onderzoeker die de validiteitscoëfficiënt $r_{XY}$ wil bepalen voor de groep waarop geen selectie is toegepast en alleen maar beschikt over de validiteitscoëfficiënt voor de geselecteerde groep, is niet noodzakelijk verloren. Er bestaat namelijk een formule waarmee




Figuur 2.8: Een illustratie van het restriction-of-range fenomeen: links het verband tussen $X$ en $Y$ zonder selectie en rechts het verband na selectie.

je de validiteitscoëfficiënt voor de groep waarop geen selectie is toegepast kunt berekenen op basis van (1) de validiteitscoëfficiënt voor de geselecteerde groep, en (2) de varianties van de testscores in de twee groepen. Het zou ons te ver voeren om deze formule hier uit de doeken te doen. Men vindt deze formule, en de veronderstellingen waarop ze gebaseerd is, in meer gespecialiseerde boeken zoals dat van Anastasi (1982).

### 2.3.4 Criteriumvaliditeit en Beslissingen

We vertrekken van de situatie waarin de criteriumvariabele in twee stukken verdeeld wordt: één stuk boven een zogenaamde *slaagnorm*, en één stuk eronder. Alle personen die zich boven de slaagnorm bevinden zijn geslaagd, en degenen die zich onder deze norm bevinden zijn niet geslaagd. Deze situatie is typisch voor onderwijssystemen waarin slagen of niet-slagen bepaald wordt door één enkele score. We hebben een vergelijkbare situatie in het Nederlandse universitair systeem als we "slagen" opvatten als het voldoen aan de prestatiebeursnorm na één jaar studie. Het criterium is dan het aantal studiepunten dat men gehaald heeft en de slaagnorm is de prestatiebeursnorm.

Het doel van een selectieprocedure is het percentage slagers te laten toenemen. Dit is mogelijk als we over een selectievariabele beschikken waarmee we het criterium kunnen voorspellen. Deze selectievariabele is meestal een testscore, en dit is ook de situatie die we hier beschouwen. Net als de criteriumvariabele, delen we ook de testscore in twee stukken: een stuk boven en stuk onder een zogenaamde *selectienorm*. Alle personen die zich boven de selectienorm bevinden worden toegelaten, en degenen die zich onder dit punt bevinden worden afgewezen. Als de test positief correleert met het criterium, dan zal het percentage geslaagden in de groep die toegelaten wordt, groter zijn dan in de groep waarop geen selectie is toegepast.

We kunnen dit illustreren aan de hand van het spreidingsdiagram dat getoond wordt

in Figuur 2.9. Het diagram is opgedeeld in vier velden:

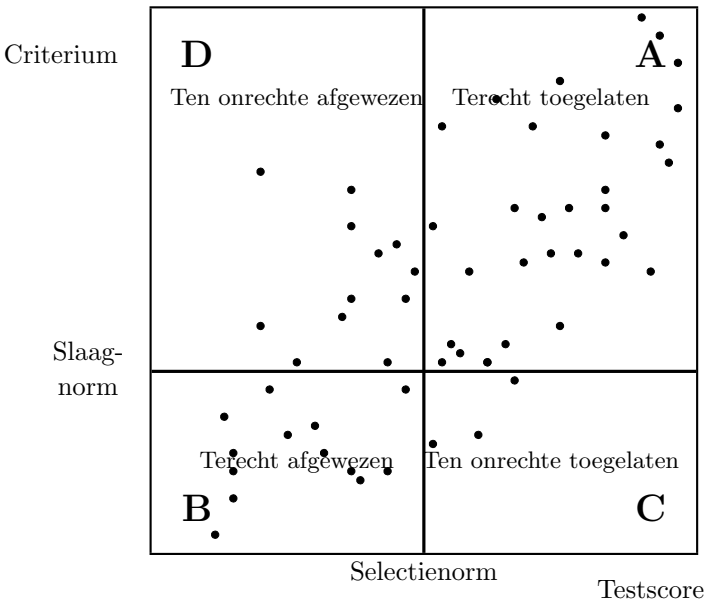
A: de personen die terecht toegelaten werden, want ze scoren boven de slaagnorm.

B: de personen die terecht afgewezen werden, want ze scoren onder de slaagnorm.

C: de personen die onterecht toegelaten werden (ook *valse positieven* genoemd).

D: de personen die onterecht afgewezen werden (ook *valse negatieven* genoemd).

De winst die geboekt wordt door te selecteren op basis van de testscore bestaat erin dat het percentage geslaagden in de groep die toegelaten wordt ($\frac{A}{A+C}$) groter is dan in de volledige groep ($\frac{A+D}{A+B+C+D}$).




Figuur 2.9: Illustratie van de toename in het percentage geslaagden na selectie.

We introduceren nu een aantal termen die in de literatuur veelvuldig gebruikt worden:

1. *Base-rate.* Dit is het percentage geslaagden zonder selectie.

$$\boxed{\text{Base-rate} = \frac{A+D}{A+B+C+D} \quad .}$$

De base-rate wordt bepaald door twee factoren: (1) de hoogte van de slaagnorm en (2) de scores op de criteriumvariabele. De base-rate stijgt (i.e., een groter percentage geslaagden zonder selectie) naarmate (1) de slaagnorm daalt en (2) de scores op de criteriumvariabele stijgen.

2. *Selectieratio.* Dit is het percentage toegelaten personen.

$$\boxed{\text{Selectieratio} = \frac{A+C}{A+B+C+D} \quad .}$$

Analoog aan de base-rate, wordt ook de selectieratio bepaald door twee factoren: (1) de hoogte van de selectienorm en (2) de scores op de test. De selectieratio stijgt (i.e., een groter percentage geselecteerden) naarmate (1) de selectienorm daalt en (2) de scores op de test stijgen.

3. *Succesratio.* Dit is het percentage geslaagden in de groep van toegelaten personen.

$$\boxed{\text{Succesratio} = \frac{A}{A+C} \quad .}$$

De succesratio wordt bepaald door drie factoren: (1) de base-rate, (2) de selectieratio, en (3) de correlatie tussen test en criterium. Naarmate de correlatie stijgt, wordt de puntenwolk smaller en vallen er minder personen in veld C.

4. *Selectiewinst* (in het Engels, *incremental validity*). Dit is de *toename* van het percentage geslaagden door gebruik te maken van de test. Als we selecteren op basis van de test, dan is het percentage geslaagden gelijk aan de succesratio. Als we niet selecteren, dan is het percentage geslaagden gelijk aan de base-rate.

$$\boxed{\begin{aligned} \text{Selectiewinst} &= \text{Succesratio} - \text{Base-rate} \\ &= \frac{A}{A+C} - \frac{A+D}{A+B+C+D} \quad . \end{aligned}}$$

Wanneer we in de praktijk moeten beslissen of het zinvol is om een selectieprocedure op te zetten, dan is de selectiewinst nuttige informatie. De selectiewinst zegt ons immers hoe groot de winst is (in termen van het percentage geslaagden) die we maken door te selecteren op basis van de testscores. Daarom is het belangrijk om na te gaan van welke factoren de selectiewinst afhankelijk is. Deze factoren zijn (1) de base-rate, (2) de selectieratio, en (3) de correlatie tussen test en criterium, de validiteitscoëfficiënt.

Taylor en Russell (1939) waren één van de eersten die op een systematische manier onderzochten hóe deze factoren de selectiewinst bepalen. Zij maakten een aantal tabellen waarin je de succesratio kan aflezen. Er is één Taylor-Russell-tabel per base-rate (één voor 0.05, één voor 0.10, één voor 0.20, ...). In Tabel 2.7 staat een deel van de Taylor-Russell tabel voor een base-rate gelijk aan 0.50. In deze tabel wordt de succesratio gegeven als functie van (1) de selectieratio en (2) de validiteitscoëfficiënt. Je vindt de selectiewinst door van de succesratio's 0.50 af te trekken.

Omdat men de base-rate nooit kent, zijn de Taylor-Russell-tabellen niet rechtstreeks bruikbaar voor de selectiepraktijk. Ze zijn echter buitengewoon nuttig om een inzicht te krijgen in de manier waarop de base-rate, de selectieratio, en de validiteitscoëfficiënt de succesratio bepalen. Men krijgt dit inzicht door te zoeken naar patronen in deze tabellen. Er zitten twee opvallende patronen in de succesratio's in Tabel 2.7. Ten eerste,

| | | Selectieratio | | | | | |
|---|---|---|---|---|---|---|---|
| Base-rate | $r_{XY}$ | 0.05 | 0.20 | 0.40 | 0.60 | 0.80 | 0.95 |
| 50 procent | 0.00 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 |
| | 0.10 | 0.58 | 0.56 | 0.54 | 0.53 | 0.51 | 0.50 |
| | 0.30 | 0.74 | 0.67 | 0.62 | 0.58 | 0.54 | 0.51 |
| | 0.50 | 0.88 | 0.78 | 0.70 | 0.63 | 0.57 | 0.52 |
| | 0.70 | 0.98 | 0.90 | 0.80 | 0.70 | 0.60 | 0.53 |
| | 0.90 | 1.00 | 0.99 | 0.92 | 0.78 | 0.62 | 0.53 |

Tabel 2.7: Deel van de Taylor-Russell-tabel voor een base-rate gelijk aan 0.50.

de succesratio neemt toe met de validiteitscoëfficiënt $r_{XY}$. Dit mag geen verbazing wekken, want naarmate $r_{XY}$ groter wordt, komen er meer personen in veld A, en minder in veld C. Ten tweede, de succesratio neemt toe naarmate de selectieratio kleiner wordt. Met andere woorden, naarmate men strenger selecteert (minder personen toelaat) stijgt het percentage geslaagden in de geselecteerde groep. Dit resultaat heeft een groot praktisch belang. Het betekent dat men met een laag correlerende test toch een behoorlijke winst kan boeken, als men maar streng genoeg selecteert. Bijvoorbeeld, met een validiteitscoëfficiënt gelijk aan 0.30 en een selectieratio van 0.05, is de selectiewinst toch 0.24 (=0.74-0.50). Dit is goed nieuws voor selectiepsychologen in tijden waarin er een overschot is op de arbeidsmarkt. Voor elke openstaande functie is er dan een groot aantal kandidaten, en kan men de selectienorm zeer hoog leggen en toch nog iemand vinden die eraan voldoet.

We hebben de invloed onderzocht van de validiteitscoëfficiënt en de selectieratio op de selectiewinst. Nu onderzoeken we de invloed van de base-rate. Daartoe vergelijken we de Taylor-Russell tabel voor een base-rate van 0.50 (zie Tabel 2.7) met de tabellen voor een base-rate van 0.10 en 0.90 (zie Tabel 2.8). Om de selectiewinst te bepalen moet men 0.10 aftrekken van de succesratio's in het bovenste deel van Tabel 2.8, en 0.90 van de succesratio's in het onderste deel. We zien dan een opvallend patroon in de selectiewinsten: bij een vaste validiteitscoëfficiënt en een vaste selectieratio, is de selectiewinst bijna altijd het grootst bij een base-rate van 0.50, en neemt af naarmate de base-rate korter bij 0.00 of 1.00 ligt. Met andere woorden, de meeste winst is te boeken als 50 procent van de volledige groep boven de slaagnorm scoort. Bijvoorbeeld, bij een validiteitscoëfficiënt van 0.70 en een selectieratio van 0.40, is de selectiewinst 0.30 bij een base-rate van 0.50, 0.12 bij een base-rate van 0.10, en 0.09 bij een base-rate van 0.90. In zeldzame gevallen is het niet zo dat de meeste winst te boeken is bij een base-rate van 0.50. Dit is bijvoorbeeld het geval bij een validiteitscoëfficiënt van 0.90 en een selectieratio van 0.05. Hier is de meeste winst te boeken bij een base-rate van 0.10.

De lage selectiewinst bij extreme base-rates is vaak een belemmering in toegepast klinisch onderzoek. De base-rate van belangrijke psychologische kenmerken is vaak erg laag. Dit geldt bijvoorbeeld voor psychopathologische stoornissen zoals schizofrenie en depressie, en afwijkend gedrag zoals alcoholverslaving en kindermishandeling. Er wordt vaak gepleit voor een vroegtijdige screening op deze stoornissen. Dit wordt ondersteund met het feit dat therapie een grotere kans op succes heeft naarmate de stoornis

| | | Selectieratio | | | | | |
|---|---|---|---|---|---|---|---|
| Base-rate | $r_{XY}$ | 0.05 | 0.20 | 0.40 | 0.60 | 0.80 | 0.95 |
| 10 procent | 0.00 | 0.10 | 0.10 | 0.10 | 0.10 | 0.10 | 0.10 |
| | 0.10 | 0.14 | 0.13 | 0.12 | 0.11 | 0.11 | 0.10 |
| | 0.30 | 0.25 | 0.19 | 0.15 | 0.13 | 0.12 | 0.10 |
| | 0.50 | 0.39 | 0.26 | 0.19 | 0.15 | 0.12 | 0.11 |
| | 0.70 | 0.58 | 0.35 | 0.22 | 0.16 | 0.12 | 0.11 |
| | 0.90 | 0.86 | 0.46 | 0.25 | 0.17 | 0.12 | 0.11 |
| 90 procent | 0.00 | 0.90 | 0.90 | 0.90 | 0.90 | 0.90 | 0.90 |
| | 0.10 | 0.93 | 0.92 | 0.92 | 0.91 | 0.91 | 0.90 |
| | 0.30 | 0.98 | 0.96 | 0.95 | 0.94 | 0.92 | 0.91 |
| | 0.50 | 1.00 | 0.99 | 0.97 | 0.96 | 0.94 | 0.92 |
| | 0.70 | 1.00 | 1.00 | 0.99 | 0.98 | 0.96 | 0.93 |
| | 0.90 | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 | 0.94 |

Tabel 2.8: Delen van de Taylor-Russell-tabellen voor base-rates gelijk aan 0.10 en 0.90.

minder lang bestaat. Tenzij men over zeer valide tests beschikt (een hoge $r_{XY}$), zal de selectiewinst van zo'n vroegtijdige screening laag zijn, en de reden hiervoor is de lage base-rate.

In het voorgaande hebben we het gehad over de selectiewinst. Die heeft echter alleen te maken met het percentage geslaagden in de toegelaten groep. Als dit percentage maar hoog genoeg was (i.c., veel hoger dan de base-rate), dan vonden we de selectieprocedure verantwoord. We hielden echter geen rekening met het percentage geslaagden in de afgewezen groep (veld D). Met recht kunnen we deze groep zien als de gedupeerden van de selectieprocedure. Sommige instanties zijn in een positie om zich weinig aan te trekken van deze onterecht afgewezen kandidaten. Dit geldt bijvoorbeeld bij personeelsselectie in bedrijven. Hier is het eerste doel het maximaliseren van de bedrijfswinst en hiervoor is alleen de selectiewinst van belang. Andere instanties kunnen de onterecht afgewezen kandidaten niet zomaar naast zich neerleggen. Dit geldt bijvoorbeeld voor de overheid (i.c., het Ministerie van Onderwijs) als die overweegt om een toelatingsexamen voor de universiteit in te voeren. De kans bestaat dat de samenleving het onaanvaardbaar vindt dat kandidaten voor de universiteit onterecht afgewezen worden. Omdat onze tests nooit perfect zullen correleren met de uitslagen aan de universiteit, zal er nooit een toelatingsexamen komen zolang de samenleving vindt dat één enkele onterecht afgewezen student al onaanvaardbaar is.

---

*Opgave 30*

Een arbeidspsychologe bij een grote bank kreeg van haar directie de opdracht een selectieprocedure te ontwikkelen voor pas afgestudeerde informatici. Op advies van een collega las zij de handleiding bij de zogenaamde *Toets voor Informaticakennis in het Bankwezen.* Zij las hierin dat deze toets een goede inhoudsvaliditeit bezit voor het vak Financiële Informatica dat in de meeste informatica-opleidingen onderwezen wordt. Verder werden er geen validiteitsgegevens verstrekt.

1. Wat kan men bedoelen met "inhoudsvaliditeit voor het vak Financiële Informatica", en

op welke wijze kan men dit onderzoeken?

2. Denk je dat de psychologe tevreden is met de validiteitsinformatie in de handleiding? Motiveer je antwoord.

*Opgave 31*
Moet de docent van het vak Methoden en Technieken zich zorgen maken als

1. de tentamencijfers voor dit vak niet correleren met een intelligentietest?
2. hij op het tentamen geen vragen gesteld heeft over betrouwbaarheid?
3. de tentamencijfers niet correleren met de hoeveelheid hulp die de studenten nodig hebben bij het analyseren van hun scriptiegegevens (voor de studenten die het tentamen meerdere keren gedaan hebben, neem je voor het tentamencijfer het laatstbehaalde cijfer).

*Opgave 32*
Beschrijf vier manieren waarop we de constructvaliditeit kunnen onderzoeken van een vragenlijst waarmee men de houding van autochtonen tegenover buitenlanders meet.

*Opgave 33*
Een selectiepsycholoog beweert dat hij voor zijn werk geen constructvalide tests nodig heeft. Voor hem is alleen ... van belang. Wat moet er op de plaats van de puntjes staan, en hoe kun je deze opvatting van de selectiepsycholoog verduidelijken?

*Opgave 34*
Een A&O-psycholoog onderzocht het verband tussen intelligentietestscores en een maat voor jobsucces (i.c., het gemiddeld aantal promoties per vijf jaar) bij een groep topmanagers. Hij vond een correlatie van 0.08. Hieruit besloot hij dat intelligentie (zoals gemeten door de intelligentietest) vrijwel niet gerelateerd is aan jobsucces (zoals hier vastgesteld). Is deze conclusie terecht? Motiveer je antwoord.

*Opgave 35*
Er werden twee nieuwe tests ontwikkeld, een intelligentietest (A) en een attitudevragenlijst (B). Beide tests werden afgenomen bij een groep eindexamenleerlingen. De intelligentietest had een betrouwbaarheid van 0.80 en de attitudevragenlijst een betrouwbaarheid van 0.60. Vier jaar later bleek dat de twee tests in gelijke, hoge, mate de studieresultaten van de geteste (toenmalige) scholieren voorspelden: beide tests hadden een correlatie van 0.52 met de studieresultaten.

1. Welke test heeft een grotere criteriumvaliditeit, de intelligentietest, de attitudevragenlijst, of is deze bij beide gelijk? Motiveer je antwoord.

Een andere onderzoeker wilde de bovenstaande resultaten repliceren. Hij nam test A en B af bij eindexamenleerlingen van een gymnasium. Vier jaar later bekeek hij de studieresultaten van dezelfde mensen. Hij stelde de studieresultaten op dezelfde manier vast als in het eerdere onderzoek. De correlatie met de studieresultaten was 0.27 voor de intelligentietest en 0.35 voor de attitudevragenlijst.

2. Hoe kun je deze lage correlaties verklaren?
3. Hoe kun je verklaren dat de intelligentietest een lagere correlatie heeft met de studieresultaten dan de attitudevragenlijst?

*Opgave 36*
Een onderzoeker heeft twee vragenlijsten ontwikkeld: één vragenlijst over gezonde eetgewoonten en één over lichamelijke conditie. Beide vragenlijsten werden afgenomen bij een groep jonge hartpatiënten (tussen de 35 en de 45 jaar). Binnen deze groep heeft de eetgewoontenvragenlijst een betrouwbaarheid van 0.60 en de conditievragenlijst een betrouwbaarheid van 0.85. Deze

vragenlijsten werden gebruikt als voorspeller van de mate van hartslagaderverkalking op 60-jarige leeftijd. (Van de groep jonge hartpatiënten was er tussentijds niemand overleden.) Beide vragenlijsten bleken een correlatie van 0.35 te hebben met de mate van hartslagaderverkalking.

1. Welke test heeft de grootste criteriumvaliditeit, de eetgewoontenvragenlijst, de conditievragenlijst, of hebben ze alletwee dezelfde criteriumvaliditeit? Leg uit.
2. Met welk kenmerk vertoont de mate van hartslagaderverkalking de grootste samenhang, met gezonde eetgewoonten, zoals gemeten door de eetgewoontenvragenlijst, of met lichamelijke conditie, zoals gemeten door de conditievragenlijst? Leg uit.
3. De Gemeentelijke Gezondheidsdienst van een grote stad maakt gebruik van dezelfde vragenlijsten om een voorspelling te maken van hartslagaderverkalking binnen een representatieve steekproef uit de populatie van 35-tot-45-jarigen. Verwacht je dat de correlatie van de conditievragenlijst met de mate van hartslagaderverkalking groter, kleiner, of gelijk aan 0.35 (de correlatie gevonden door de eerste onderzoeker) zal zijn? Leg uit.

*Opgave 37*
Een privéschool voor diamantbewerking hanteert een toelatingsexamen voor de eerstejaars leerlingen. Na elk jaar wordt de correlatie berekend tussen dit toelatingsexamen en het gemiddelde cijfer op de vakken van het eerste jaar. Men slaagt voor het eerste jaar als het gemiddelde cijfer hoger is dan 5.5. De correlatie tussen toelatingsexamen en het gemiddelde cijfer bleek slechts 0.18 te zijn.

1. Betekent dit dat het toelatingsexamen niet erg geschikt is als selectievariabele (i.e., de variabele waarop men selecteert)? Motiveer je antwoord.
2. Wat wordt in deze context bedoeld met de term “base-rate”?

*Opgave 38*
Een medewerker van het Ministerie van Onderwijs is zeer enthousiast wanneer hij ontdekt dat er een test bestaat die, indien gebruikt als toelatingsexamen voor de universiteit, zorgt voor een succesratio van 100 procent. De minister deelt zijn enthousiasme echter niet. Teken een figuur met de velden “Terecht toegelaten”, “Ten onrechte toegelaten”, “Ten onrechte afgewezen” en “Terecht afgewezen” en gebruik deze figuur om (1) te laten zien wat de medewerker bedoelt met “een succesratio van 100 procent” en (2) waarom de minister toch niet enthousiast is over dit gegeven.

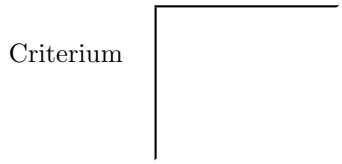
*Opgave 39*
Een instelling voor de behandeling van jonge criminelen stelde vast dat, na ontslag, heel wat jongeren toch weer opgepakt werden voor één of ander misdrijf. De directie besloot om in de toekomst slechts deze jongeren uit de instelling te ontslaan waarvan men kon verwachten dat zij geen of weinig misdrijven zouden plegen, en zeker geen ernstige misdrijven. Hiervoor was het nodig om een voorspelling te maken van het aantal en de ernst van de misdrijven die een jongere pleegt na ontslag uit de instelling. Om de mogelijkheid tot voorspelling te onderzoeken startte de directie een onderzoek waarin zij een groot aantal vragenlijsten afnam van elke jongere op het moment dat hij of zij de instelling verliet. Tien jaar na ontslag werd voor elke jongere van destijds het strafblad opgezocht en geteld hoe vaak hij of zij in contact geweest was met justitie en hoe ernstig de misdrijven waren. Tot verrassing van de directie bleek dat geen enkel van de vragenlijsten significant correleerde met het aantal of de ernst van de misdrijven. Alle vragenlijsten waren erg betrouwbaar, evenals de inhoud van het strafblad. Een medewerker van het Ministerie van Justitie leest over deze resultaten en beweert dat het nu aangetoond is dat criminaliteit niet te voorspellen is op basis van de vragenlijsten die in dit onderzoek gebruikt werden. Ben je het eens met de medewerker van het ministerie? Motiveer je antwoord.

*Opgave 40*
Waar of onwaar, en indien onwaar, leg uit: een positieve selectiewinst is mogelijk als de proportie terecht toegelaten kandidaten (ten opzichte van het totaal aantal toegelaten kandidaten) gelijk is aan de base-rate.

*Opgave 41*
Twee selectiepsychologen hebben een gesprek over hun werk. Zij zijn verbaasd over het feit dat, alhoewel zij beiden een test gebruiken met een criteriumvaliditeit van 0.70, één van hen een succesratio van 0.80 bereikt en de ander slechts een succesratio van 0.60. Bovendien selecteren ze elk even streng; ze hanteren elk een selectieratio van 0.40. Hoe kun je dit verklaren en mag je hieruit besluiten dat de ene selectiepsycholoog in vergelijking met de andere meer baat h bij de informatie in de testscores? Motiveer je antwoord.

*Opgave 42*
In veel handboeken wordt het spreidingsdiagram in Figuur 2.9 voorgesteld zond met een ellips, zoals in Figuur 2.10. Deze ellips omcirkelt het gebied waarb gedeelte van de puntjes ligt (bijvoorbeeld, 95 procent). De lijnen door slaagnorm delen deze ellips in vier stukken. Deze stukken worden in de a, b, c en d.

1. Wat gebeurt er met deze figuur (i.c., de ellips, de lijnen als de base-rate toeneemt?
2. Wat gebeurt er met deze figuur als de selectier
3. Wat gebeurt er met deze figuur als de crite




Figuur 2.10: Puntenwolk voorgesteld door een ellips, met aanduiding van de selectienorm en de slaagnorm. Ten behoeve van opgave 42.

*Opgave 43*

Stel je een selectieprocedure voor waar 100 personen aan deelnemen. In veld A en D samen zitten 60 personen, en in veld B en D samen zitten er ook 60. Stel dat selecteren in deze situatie helemaal geen zin heeft. Wat zijn dan de aantallen personen in A, B, C en D?

*Opgave 44*
Maak een figuur van een selectieprocedure met een negatieve selectiewinst. Duid de vier velden aan met A, B, C en D. Laat zien hoe je de selectiewinst bepaalt. Hoe kun je van deze testscores gebruik maken om toch een positieve selectiewinst te krijgen?

*Opgave 45*
In deze opgave gaan we ervan uit dat de testscore en het criterium dichotome variabelen zijn; zij kunnen slechts twee waarden aannemen. Dit verschilt van de situatie die in het stuk *Criteriumvaliditeit en Beslissingen* beschreven werd. De gelijkenis met de situatie met een continue testscore en een continu criterium is echter voldoende groot om de opgave te kunnen maken.

In een bepaalde gemeente in Nederland is er de afgelopen jaren een sterke toename van het aantal zedendelicten. Omdat de schuldigen maar niet gepakt kunnen worden, overweegt de recherche om alle volwassen mannen uit die gemeente te onderwerpen aan een test met een leugendetector. Een leugendetector is een apparaat dat allerlei reacties van je lichaam registreert (hartslag, ademhalingsritme, zweetproductie door de huid, ...) terwijl je ondervraagd wordt. Uit onderzoek blijkt dat de leugendetector vrij goed in staat is om leugenaars te detecteren. Concreet, de leugendetector ontmaskert 95 procent van alle leugenaars.

In de volgende drie vragen wordt naar je mening gevraagd. Er is geen objectief goed of fout antwoord. Waar het om gaat, is te ontdekken hoe je mening (eventueel) verandert afhankelijk van hoe je naar de situatie kijkt. Ga ervan uit dat de leugendetector de laatste hoop is van de recherche; zij hebben geen andere methode achter de hand.

1. Op basis van de informatie die je nu hebt (95 procent van alle leugenaars wordt ontmaskerd), zou jij de leugendetector gebruiken?
2. Stel dat de leugendetector ook 1 procent van degenen die de waarheid spreken ten onrechte aanwijst als leugenaar. (In werkelijkheid ligt dat percentage veel hoger, maar we houden ons hier bij 1 procent omdat dit ons voorbeeld overtuigender maakt.) Op basis van de informatie die je nu hebt, zou jij de leugendetector gebruiken?

| | Veroordeeld | Vrijgesproken | |
|---|---|---|---|
| Schuldig | | | 100 |
| Onschuldig | | | 9900 |
| | | | 10000 |

Tabel 2.9: Tabel met de aantallen terecht en ten onrechte schuldig bevonden mannen. Ten behoeve van opgave 45.

3. Stel dat er 10000 volwassen mannen in de gemeente wonen. Deze worden allemaal aan een test met de leugendetector onderworpen. Stel ook dat 100 mannen zich de afgelopen tien jaar schuldig gemaakt hebben aan een ernstig zedendelict. Alle 10000 mannen ontkennen dat ze de afgelopen tien jaar een zedendelict gepleegd hebben. Bepaal het aantal mannen dat ten onrechte schuldig bevonden wordt. Is dit aantal groter of kleiner dan het aantal mannen dat terecht schuldig bevonden wordt? Op basis van de informatie die je nu

hebt, zou jij de leugendetector gebruiken? (Voor het bepalen van het aantal mannen dat terecht en ten onrechte schuldig bevonden wordt, kun je gebruik maken van Tabel 2.9.)

---

## 2.4 Appendix bij Hoofdstuk 2

### Covariantie en Correlatie

In de afleidingen van een aantal formules in dit hoofdstuk wordt gebruik gemaakt van het begrip "covariantie" en van de relatie tussen de covariantie en de correlatie. Daarom gaan we er hier wat dieper op in.

De covariantie tussen twee variabelen is een maat voor hun samenhang. Ook de correlatie is een maat voor de samenhang tussen twee variabelen. Het verschil tussen beide is dat de correlatie een gestandaardiseerde maat is voor de samenhang, en de covariantie niet; de correlatie neemt alleen waarden aan tussen -1 (perfect negatieve samenhang) en +1 (perfect positieve samenhang).

We geven nu de formule voor de covariantie tussen $Y$ en $Z$:

$$\mathrm{COV}(Y,Z) = \frac{1}{n}\sum_{p=1}^{n}(Y_p - \overline{Y})(Z_p - \overline{Z}) \quad .$$

In deze formule gebruiken we $p$ als index voor de observaties en staat $n$ voor het aantal $(Y_p, Z_p)$-paren. Aan deze formule kan men zien dat $\mathrm{COV}(Y,Z)$ een maat is voor de samenhang tussen $Y$ en $Z$:

1. Als grote $Y_p$'s (i.c., groter dan $\overline{Y}$) vooral samengaan met grote $Z_p$'s (i.c., groter dan $\overline{Z}$), en kleine $Y_p$'s met kleine $Z_p$'s, dan zijn de termen in het rechterlid vooral positief (want positief $\times$ positief en negatief $\times$ negatief zijn beide positief) en is $\mathrm{COV}(Y,Z)$ ook positief.

2. Als grote $Y_p$'s (i.c., groter dan $\overline{Y}$) vooral samengaan met kleine $Z_p$'s (i.c., kleiner dan $\overline{Z}$), en kleine $Y_p$'s met grote $Z_p$'s, dan zijn de termen in het rechterlid vooral negatief (want positief $\times$ negatief en negatief $\times$ positief zijn beide negatief) en is $\mathrm{COV}(Y,Z)$ ook negatief.

De correlatie tussen $Y$ en $Z$ ($r_{YZ}$) is een gestandaardiseerde versie van $\mathrm{COV}(Y,Z)$. We standaardiseren $\mathrm{COV}(Y,Z)$ door hem te delen door het product van de standaarddeviaties van $Y$ en $Z$ ($\mathrm{SD}(Y)$ en $\mathrm{SD}(Z)$):

$$r_{YZ} = \frac{\mathrm{COV}(Y,Z)}{\mathrm{SD}(Y)\mathrm{SD}(Z)} \quad .$$

Hieruit volgt dat $\mathrm{COV}(Y,Z)$ als volgt geschreven kan worden:

$$\mathrm{COV}(Y,Z) = r_{YZ}\mathrm{SD}(Y)\mathrm{SD}(Z) \quad .$$

Het effect van de standaardisering is dat $r_{YZ}$ alleen waarden kan aannemen tussen $-1$ en $+1$. Daar staat tegenover dat het bereik van $\mathrm{COV}(X,Y)$ afhangt van $\mathrm{SD}(Y)$ en $\mathrm{SD}(Z)$.

## De Variantie van de Som van Twee Variabelen

We zijn geïnteresseerd in de variantie van een variabele $X$ (niet noodzakelijk een testscore) die geschreven kan worden als de som van twee variabelen $Y$ en $Z$. Dus,

$$X = Y + Z \quad . \tag{2.1}$$

We vertrekken van de formule voor $\text{VAR}(X)$:

$$\text{VAR}(X) = \frac{1}{n}\sum_{p=1}^{n}[X_p - \overline{X}]^2 \quad .$$

In deze formule staat $n$ voor het aantal observaties en $\overline{X}$ voor het gemiddelde van $X$. Uit het substitueren van (2.1) volgt dan:

$$\text{VAR}(X) = \frac{1}{n}\sum_{p=1}^{n}[(Y_p + Z_p) - (\overline{Y} + \overline{Z})]^2 \quad .$$

Het rechterlid van deze formule kunnen we als volgt uitwerken:

$$\begin{aligned}
\text{VAR}(X) &= \frac{1}{n}\sum_{p=1}^{n}[(Y_p - \overline{Y}) + (Z_p - \overline{Z})]^2 \\
&= \frac{1}{n}\sum_{p=1}^{n}[(Y_p - \overline{Y})^2 + (Z_p - \overline{Z})^2 + 2(Y_p - \overline{Y})(Z_p - \overline{Z})] \\
&= \frac{1}{n}\sum_{p=1}^{n}(Y_p - \overline{Y})^2 + \frac{1}{n}\sum_{p=1}^{n}(Z_p - \overline{Z})^2 + 2\frac{1}{n}\sum_{p=1}^{n}(Y_p - \overline{Y})(Z_p - \overline{Z}) \quad .
\end{aligned}$$

Rechts van het laatste gelijkheidsteken herkennen we achtereenvolgens $\text{VAR}(Y)$, $\text{VAR}(Z)$ en $2\text{COV}(Y, Z)$. Dus, $\text{VAR}(X)$ kan als volgt geschreven worden:

$$\text{VAR}(X) = \text{VAR}(Y) + \text{VAR}(Z) + 2\text{COV}(Y, Z) \quad . \tag{2.2}$$

In woorden, staat er in (2.2) dat de variantie van een som van twee termen (hier, $Y$ en $Z$) gelijk is aan de som van de varianties van de termen plus twee keer de covariantie van die termen.

## Afleiding van $\textbf{VAR}(X) = \textbf{VAR}(T) + \textbf{VAR}(E)$

Omdat $X$ gelijk is aan de som van $T$ en $E$, kan men gebruik maken van (2.2) om $\text{VAR}(X)$ als volgt te herschrijven:

$$\text{VAR}(X) = \text{VAR}(T) + \text{VAR}(E) + 2\text{COV}(T, E) \quad .$$

De covariantie in het rechterlid van deze formule is gelijk aan nul. Dit komt omdat (1) $r_{TE}$ gelijk is aan nul, en (2) $\text{COV}(T, E)$ gelijk is aan $r_{TE}\text{SD}(T)\text{SD}(E)$. Ons resultaat is nu het volgende:

$$\text{VAR}(X) = \text{VAR}(T) + \text{VAR}(E) \quad .$$

## Afleiding van de Formule van de Correlatie tussen Twee Parallelle Tests

De correlatie $r_{XX'}$ is als volgt gedefinieerd:

$$r_{XX'} = \frac{\mathrm{COV}(X, X')}{\mathrm{SD}(X)\mathrm{SD}(X')} \quad .$$

Omdat $\mathrm{VAR}(X)$ gelijk is aan $\mathrm{VAR}(X')$, is ook $\mathrm{SD}(X)$ gelijk aan $\mathrm{SD}(X')$. De noemer van het rechterlid is daarom gelijk aan $\mathrm{VAR}(X)$. De teller kunnen we als volgt uitwerken:

$$\begin{aligned}
\mathrm{COV}(X, X') &= \frac{1}{n}\sum_{p=1}^{n}[X_p - \overline{X}][X'_p - \overline{X'}] \\
&= \frac{1}{n}\sum_{p=1}^{n}[(T_p - \overline{T}) + (E_p - \overline{E})][(T'_p - \overline{T'}) + (E'_p - \overline{E'})] \\
&= \frac{1}{n}\sum_{p=1}^{n}(T_p - \overline{T})(T'_p - \overline{T'}) + \frac{1}{n}\sum_{p=1}^{n}(T_p - \overline{T})(E'_p - \overline{E'}) + \\
&\frac{1}{n}\sum_{p=1}^{n}(E_p - \overline{E})(T'_p - \overline{T'}) + \frac{1}{n}\sum_{p=1}^{n}(E_p - \overline{E})(E'_p - \overline{E'}) \quad .
\end{aligned}$$

Omdat de tests parallel zijn, kunnen $T'_p$ en $\overline{T'}$ vervangen worden door $T_p$ en $\overline{T}$. Dan kan $\mathrm{COV}(X, X')$ als volgt verder uitgewerkt worden:

$$\mathrm{COV}(X, X') = \mathrm{VAR}(T) + \mathrm{COV}(T, E') + \mathrm{COV}(E, T) + \mathrm{COV}(E, E') \quad .$$

Omdat hun bijbehorende correlaties gelijk zijn aan 0 ($r_{TE'}$, $r_{ET}$ en $r_{EE'}$), zijn ook de covarianties in het rechterlid allemaal gelijk aan 0. Dus, $\mathrm{COV}(X, X')$ is gelijk aan $\mathrm{VAR}(T)$, zodat $r_{XX'}$ als volgt geschreven kan worden:

$$r_{XX'} = \frac{\mathrm{VAR}(T)}{\mathrm{VAR}(X)} \quad .$$

## Afleiding van de Spearman-Brown-Formule voor $k = 2$

De scores op de twee parallelle deeltests duiden we aan met $D$ en $D'$. De ware-scores op deeltests duiden we aan met $T_D$, dit om het onderscheid te maken met de ware-score op de samengestelde test, die we aanduiden met $T_X$. De meetfouten op de parallelle deeltests duiden we aan met $E_D$ en $E_{D'}$, en de meetfout op de samengestelde test met $E_X$. We passen de basisformule van de Klassieke Testtheorie nu toe op de twee parallelle deeltests:

$$D = T_D + E_D \tag{2.3}$$

$$D' = T_D + E_{D'} \quad . \tag{2.4}$$

Merk op dat de twee ware-scores gelijk zijn (twee keer $T_D$). Dit is zo omdat de deeltests parallel zijn. Omdat $X$ gelijk is aan $(D + D')$, volgt uit (2.3)-(2.4):

$$X = 2T_D + (E_D + E_{D'}) \quad .$$

Passen we de basisformule van de Klassieke Testtheorie nu ook toe op $X$, dan zien we dat $T_X$ gelijk is aan $2T_D$ en $E_X$ aan $(E_D + E_{D'})$. Hieruit volgt dat we $\text{VAR}(T_X)$ als volgt kunnen schrijven:

$$\begin{aligned} \text{VAR}(T_X) &= \text{VAR}(2T_D) \\ &= 4\text{VAR}(T_D) \quad . \end{aligned}$$

Voor het bepalen van $\text{VAR}(E_X)$ vertrekken we van het feit dat $E_X$ gelijk is aan de som van $E_D$ en $E_{D'}$. We maken dan gebruik van formule (2.2) die uitdrukt dat de variantie van een som van twee termen gelijk is aan de som van de varianties van de termen plus twee keer de covariantie van die termen:

$$\text{VAR}(E_X) = \text{VAR}(E_D) + \text{VAR}(E_{D'}) + 2\text{COV}(E_D, E_{D'}) \quad .$$

Omdat $E_D$ en $E_{D'}$ ongecorreleerd zijn en dezelfde variantie hebben, kunnen we $\text{VAR}(E_X)$ als volgt schrijven:

$$\text{VAR}(E_X) = 2\text{VAR}(E_D) \quad .$$

Het belangrijke punt hier, is dat bij verdubbeling van de testlengte de ware-score-variantie $\text{VAR}(T_X)$ vier keer zo groot wordt en de meetfoutvariantie $\text{VAR}(E_X)$ maar twee keer. Hierdoor stijgt de betrouwbaarheid bij testverlenging. Toepassing van de formule voor de correlatie tussen twee parallelle tests levert ons:

$$\begin{aligned} r_{XX'} &= \frac{\text{VAR}(T_X)}{\text{VAR}(X)} \\ &= \frac{\text{VAR}(T_X)}{\text{VAR}(T_X) + \text{VAR}(E_X)} \\ &= \frac{4\text{VAR}(T_D)}{4\text{VAR}(T_D) + 2\text{VAR}(E_D)} \quad . \end{aligned}$$

Nu delen we teller en noemer door $\text{VAR}(D)$:

$$\begin{aligned} r_{XX'} &= \frac{4r_{DD'}}{4r_{DD'} + 2(1 - r_{DD'})} \\ &= \frac{2r_{DD'}}{1 + r_{DD'}} \quad . \end{aligned}$$

Dit is de S-B formule voor $k = 2$.

## De Formule van Alpha als de Items Parallel Zijn.

Uit de tekst volgt dat alpha als volgt geschreven kan worden:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{k\sigma^2}{k\sigma^2 + (k^2 - k)\rho\sigma^2}\right) \quad .$$

Het rechterlid van deze gelijkheid kan als volgt uitgewerkt worden:

$$\begin{aligned}\alpha &= \frac{k}{k-1}\left(1-\frac{1}{1+(k-1)\rho}\right)\\ &= \frac{k}{k-1}\left(\frac{1+(k-1)\rho-1}{1+(k-1)\rho}\right)\\ &= \frac{k}{k-1}\left(\frac{(k-1)\rho}{1+(k-1)\rho}\right)\\ &= \frac{k\rho}{1+(k-1)\rho} \quad ,\end{aligned}$$

de S-B formule voor de situatie met $k$ gelijk aan het aantal items en $\rho$ gelijk aan de itembetrouwbaarheid.

## Afleiding van de Formule van Restriction-of-Range

De ware-scores en de meetfouten van $X$ duiden we aan met $T_X$ en $E_X$, en de ware-scores en de meetfouten van $Y$ met $T_Y$ en $E_Y$. Ons vertrekpunt is de variantie-ontbinding op pagina 79. Het rechterlid en het linkerlid van deze formule delen we door $\text{VAR}(X)$. Het resultaat kunnen we als volgt uitwerken:

$$\begin{aligned}1 &= \frac{\text{VAR}(T_X)}{\text{VAR}(X)} + \frac{\text{VAR}(E_X)}{\text{VAR}(X)}\\ 1 &= r_{XX'} + \frac{\text{VAR}(E_X)}{\text{VAR}(X)}\\ 1 - r_{XX'} &= \frac{\text{VAR}(E_X)}{\text{VAR}(X)}\\ \text{VAR}(X)(1-r_{XX'}) &= \text{VAR}(E_X) \quad . \qquad (2.5)\end{aligned}$$

Voor $Y$ kunnen we eenzelfde formule afleiden:

$$\text{VAR}(Y)(1-r_{YY'}) = \text{VAR}(E_Y) \quad . \qquad (2.6)$$

We gaan er nu vanuit dat $\text{VAR}(E_X)$ gelijk is aan $\text{VAR}(E_Y)$. (Deze veronderstelling is analoog aan de veronderstelling dat de spreiding van de punten rond de diagonaal in Figuur 2.7 hetzelfde is voor alle waarden van $X$.) We kunnen dan het linkerlid van (2.5) en (2.6) aan elkaar gelijk stellen. Hieruit volgt dan:

$$1 - r_{YY'} = \frac{\text{VAR}(X)}{\text{VAR}(Y)}(1-r_{XX'}) \quad .$$

## Afleiding van de Formule van de Standaardfout van Meting

Ons vertrekpunt is de variantie-ontbinding op pagina 79. Het rechterlid en het linkerlid van de variantie-ontbinding delen we door $\text{VAR}(X)$. Het resultaat kunnen we als volgt

uitwerken:

$$\begin{aligned}
1 &= \frac{\mathrm{VAR}(T)}{\mathrm{VAR}(X)} + \frac{\mathrm{VAR}(E)}{\mathrm{VAR}(X)} \\
1 &= r_{XX'} + \frac{\mathrm{VAR}(E)}{\mathrm{VAR}(X)} \\
1 - r_{XX'} &= \frac{\mathrm{VAR}(E)}{\mathrm{VAR}(X)} \\
\mathrm{VAR}(E) &= \mathrm{VAR}(X)(1 - r_{XX'}) \\
\mathrm{SD}(E) &= \mathrm{SD}(X)\sqrt{1 - r_{XX'}} \quad ,
\end{aligned}$$

waarbij de laatste stap volgt uit het nemen van de vierkantswortel aan de linker- en de rechterkant van het gelijkheidsteken.

## Afleiding van de Formule voor Correctie voor Attenuatie

De correlatie $r_{T_X T_Y}$ is als volgt gedefinieerd:

$$r_{T_X T_Y} = \frac{\mathrm{COV}(T_X, T_Y)}{\mathrm{SD}(T_X)\mathrm{SD}(T_Y)} \quad . \tag{2.7}$$

En de geobserveerde correlatie $r_{XY}$ is als volgt gedefinieerd:

$$r_{XY} = \frac{\mathrm{COV}(X, Y)}{\mathrm{SD}(X)\mathrm{SD}(Y)} \quad . \tag{2.8}$$

We gaan nu de formule voor $r_{XY}$ omschrijven naar de formule voor $r_{T_X T_Y}$. Eerst laten we zien dat $\mathrm{COV}(X, Y)$ gelijk is aan $\mathrm{COV}(T_X, T_Y)$. Dit gebeurt op dezelfde manier als in de afleiding van de formule voor de correlatie tussen twee parallelle tests, waar we lieten zien dat $\mathrm{COV}(X, X')$ gelijk is aan $\mathrm{COV}(T_X, T_X) = \mathrm{VAR}(T_X)$.

$$\begin{aligned}
\mathrm{COV}(X, Y) &= \mathrm{COV}[(T_X + E_X), (T_Y + E_Y)] \\
&= \mathrm{COV}(T_X, T_Y) + \mathrm{COV}(T_X, E_Y) + \mathrm{COV}(E_X, T_Y) + \mathrm{COV}(E_X, E_Y) \qquad (2.9) \\
&= \mathrm{COV}(T_X, T_Y) \quad .
\end{aligned}$$

De laatste stap volgt uit het feit dat de meetfouten $E_X$ en $E_Y$ niet met andere variabelen correleren. Nu schrijven we de noemer van (2.8), $\mathrm{SD}(X)\mathrm{SD}(Y)$, als functie van de noemer van (2.7), $\mathrm{SD}(T_X)\mathrm{SD}(T_Y)$. Dit kan omdat $\mathrm{SD}(X)$ en $\mathrm{SD}(T_X)$ op de volgende manier aan elkaar gerelateerd zijn:

$$\begin{aligned}
\mathrm{SD}(X)\sqrt{r_{XX'}} &= \mathrm{SD}(X)\sqrt{\frac{\mathrm{VAR}(T_X)}{\mathrm{VAR}(X)}} \\
&= \mathrm{SD}(X)\frac{\mathrm{SD}(T_X)}{\mathrm{SD}(X)} \\
&= \mathrm{SD}(T_X) \quad . \qquad (2.10)
\end{aligned}$$

Op dezelfde manier kan men laten zien dat voor $\mathrm{SD}(Y)$ en $\mathrm{SD}(T_Y)$ het volgende geldt:

$$\mathrm{SD}(Y)\sqrt{r_{YY'}} = \mathrm{SD}(T_Y) \quad . \tag{2.11}$$

Delen we nu linker- en rechterlid van (2.8) door $\sqrt{r_{XX'}}\sqrt{r_{YY'}}$, dan krijgen we het volgende:

$$\begin{aligned}
\frac{r_{XY}}{\sqrt{r_{XX'}}\sqrt{r_{YY'}}} &= \frac{\mathrm{COV}(X,Y)}{\mathrm{SD}(X)\sqrt{r_{XX'}}\mathrm{SD}(Y)\sqrt{r_{YY'}}} \\
&= \frac{\mathrm{COV}(T_X,T_Y)}{\mathrm{SD}(T_X)\mathrm{SD}(T_Y)} \\
&= r_{T_X T_Y} \quad ,
\end{aligned}$$

waarbij we in de tweede stap gebruikmaken van (2.9), (2.10) en (2.11).

# Hoofdstuk 3

# Variantie- en Covariantieanalyse

## Inhoudsopgave


**3.1 Inleiding . . . . . 130**

3.1.1 Voorbeeld: Mobiliteit en Sociale Activiteit . . . . . 130

3.1.2 Experimenteel en Niet-Experimenteel Onderzoek . . . . . 131

3.1.3 De Toewijzing aan de Condities . . . . . 132

3.1.4 Experimentele Controle van de Condities . . . . . 135

**3.2 Variantieanalyse met Eén Factor . . . . . 137**

3.2.1 Voorbeeld: Mobiliteit en Sociale Activiteit . . . . . 137

3.2.2 De Notatie voor de Gegevens . . . . . 139

3.2.3 Het Rekenwerk . . . . . 140

3.2.4 De F-Toets . . . . . 143

3.2.5 Het Variantieanalysemodel . . . . . 145

3.2.6 De Effectkwadratensom . . . . . 148

3.2.7 De Errorkwadratensom, Totale Kwadratensom, en de Kwadratensomontbinding . . . . . 152

3.2.8 Berekening en Interpretatie van de F-Statistiek . . . . . 154

3.2.9 De Proportie Verklaarde Variantie . . . . . 156

**3.3 Variantieanalyse met Twee Factoren . . . . . 160**

3.3.1 Voorbeeld 1: Mobiliteit en Sociale Activiteit . . . . . 160

3.3.2 Wat zijn Ongebalanceerde Designs en Wanneer Krijgt men Ermee te Maken? . . . . . 161

3.3.3 Voorbeeld 2: Pesten in de Klas . . . . . 164

3.3.4 De Notatie voor de Gegevens . . . . . 169

3.3.5 De Statistische Toetsing . . . . . 174

3.3.6 Het Rekenwerk in een Gebalanceerd Design . . . . . 178

3.3.7 Het Rekenwerk in een Ongebalanceerd Design . . . . . 184

3.3.8 F-statistieken en Kwadratensommen . . . . . 190
3.3.9 Interactie tussen Factoren . . . . . 208
3.3.10 Het Rekenwerk in een Gebalanceerd Design (Met Interactie) . 211
3.3.11 Het Rekenwerk in een Ongebalanceerd Design (Met Interactie) 217
3.3.12 F-statistieken en Kwadratensommen (Met Interactie) . . . . 221
3.3.13 Het Nut van het Toevoegen van Factoren . . . . . 237
3.3.14 Een Relativering van de Kracht van Statistische Uitzuivering 238
3.3.15 Designs met Lege Cellen . . . . . 239
3.3.16 De Variantieanalyse met Meer dan Twee Factoren . . . . . 240
3.3.17 Uitvoeren van een Variantieanalyse met de SPSS-procedure GLM-Univariate . . . . . 240
**3.4 Uitzuivering voor een Kwantitatieve Variabele met Covariantieanalyse . . . . . 252**
3.4.1 Voorbeeld: Pesten in de Klas . . . . . 252
3.4.2 Uitgezuiverde Verschillen en Gemiddelden . . . . . 258
3.4.3 De Nulhypothese voor het Uitgezuiverde Populatieverschil . . 260
3.4.4 De F-statistieken . . . . . 262
3.4.5 De Effectkwadratensommen . . . . . 262
3.4.6 De Errorkwadratensom en de Kwadratensom-ontbinding . . . 267
3.4.7 Nogmaals De F-Statistieken . . . . . 267
3.4.8 Overzichtstabellen . . . . . 269
3.4.9 Uitvoeren van een Covariantieanalyse met de SPSS-procedure GLM-Univariate . . . . . 271
3.4.10 Twee Redenen om Uit te Zuiveren voor een Covariaat . . . . 272
3.4.11 Wanneer Stel je de Covariaatscores Vast? . . . . . 274
3.4.12 Een Relativering van de Kracht van Statistische Uitzuivering 275
**3.5 Designs met een Voor- en een Nameting . . . . . 285**
3.5.1 Het Design met Eén Groep . . . . . 285
3.5.2 Het Design met Twee Groepen . . . . . 287
**3.6 Appendix bij Hoofdstuk 3 . . . . . 290**

## 3.1 Inleiding

### 3.1.1 Voorbeeld: Mobiliteit en Sociale Activiteit

Om de theorie in dit hoofdstuk te illustreren gebruiken we als voorbeeld een fictief onderzoek naar het effect van mobiliteit (het zich kunnen verplaatsen) op de sociale activiteit van bejaarden. De onderzoekster was geïnteresseerd in de contacten van bejaarden met personen buiten familie en gezin. Bijvoorbeeld, deelname aan verenigingsleven, op visite gaan bij kennissen, of samen met vrienden een tentoonstelling bezoeken. Dit is de afhankelijke variabele in dit onderzoek, en hij wordt verder aangeduid als SOCIALE ACTIVITEIT. De mate van sociale activiteit werd vastgesteld als het gemiddeld aantal uren

sociale contacten per week. Dit werd bepaald op basis van een soort dagboek waarin de bejaarde noteerde hoeveel tijd hij op een dag gespendeerd had aan sociale activiteiten.

De onafhankelijke variabele in dit onderzoek is de mobiliteit van de bejaarden, verder aangeduid als MOBILITEIT. De bejaarden in dit onderzoek beschikten niet over eigen vervoer. Hierdoor kon de onderzoekster zelf de bejaarden indelen in een hoog en een laag mobiele groep. Zij deed dit door de groep bejaarden in de ene groep gebruik te laten maken van een speciale taxiservice die hen gratis vervoerde van en naar vrienden, bijeenkomsten, voorstellingen, en dergelijke. De bejaarden in de andere groep konden hier geen gebruik van maken.

Op het einde van het onderzoek leverden de bejaarden hun tijdsbestedingsdagboeken in bij de onderzoekster. Voor elke bejaarde werd dan het gemiddeld aantal uren sociale contacten per week bepaald. Deze gemiddelden per bejaarde werden dan gemiddeld over de bejaarden in elk van de twee groepen. De onderzoekster stelde vast dat de bejaarden in de groep met gratis taxi gemiddeld 9.9 uur per week besteedden aan sociale activiteiten en de bejaarden in de andere groep slechts 6.4 uur.

### 3.1.2 Experimenteel en Niet-Experimenteel Onderzoek

Het onderzoeken van het effect van een onafhankelijke variabele (in ons voorbeeld, MOBILITEIT, het al dan niet gebruik kunnen maken van een gratis taxi) gebeurt meestal met behulp van een experiment. In paragraaf 1.2, *Effectonderzoek*, werd gesteld dat manipulatie het essentiële kenmerk is van een experiment. Er werden toen drie vormen van manipulatie besproken: (1) manipulatie door elke proefpersoon aan een van de condities toe te wijzen, (2) manipulatie van de condities zelf en (3) manipulatie door elke proefpersoon aan alle condities toe te wijzen. De functie van manipulatie is het controleren van storende variabelen: storende persoonsvariabelen door middel van zuiver toevallige toewijzing aan de condities en toewijzing van elke proefpersoon aan álle condities, en storende situatievariabelen door middel van manipulatie van de condities zelf (experimentele controle van de condities).

In dit hoofdstuk, richten we ons op het niet-experimenteel effectonderzoek. In niet-experimenteel effectonderzoek is manipulatie onmogelijk en kan daarom ook niet gebruikt worden om storende variabelen te controleren. De term “experimenteel” omvat twee aspecten: (1) toewijzing aan condities is mogelijk en kan dus gebruikt worden om storende persoonsvariabelen te controleren (door middel van zuiver toevallige toewijzing aan de condities), en (2) experimentele controle van de condities is mogelijk en kan dus gebruikt worden om storende situatievariabelen te controleren (door middel van experimentele controle van de condities). Het is mogelijk dat een effectonderzoek experimenteel is voor wat betreft slechts een van deze twee aspecten en niet voor wat betreft het andere. Bijvoorbeeld, een effectonderzoek kan experimenteel zijn voor wat betreft de controle van storende persoonsvariabelen (i.c., als de proefpersonen op zuiver toevallige wijze aan de condities worden toegewezen) maar niet voor wat betreft de controle van storende situatievariabelen (i.c., als het niet mogelijk is om de condities zelf te manipuleren). Dit is meestal het geval in een veldexperiment. Ook het omgekeerde is mogelijk (controle van storende situatievariabelen maar niet van storende persoonsvariabelen). Dit gebeurt bijvoorbeeld in een laboratoriumexperiment waarin de proefleider de proefpersonen niet op zuiver toevallige wijze toewijst aan de condities. In dit hoofdstuk

stellen we enkele methoden voor die kunnen fungeren als alternatief voor manipulatie. Controle door middel van manipulatie is echter bijna altijd te verkiezen boven controle met behulp van deze alternatieve methoden.

In de rest van deze inleiding, bespreken we eerst het belang van de toewijzing aan condities en stellen kort de alternatieve methoden voor die men kan gebruiken als toewijzing aan condities onmogelijk is. Hierbij beperken we ons tot het tussen-subjectenexperiment. Dit betekent dat elke proefpersoon slechts in één conditie geobserveerd wordt. Daarna richten we ons op het belang van experimentele controle van de condities en stellen kort de alternatieve methoden voor die men kan gebruiken als men geen controle heeft op bepaalde storende situatievariabelen.

### 3.1.3 De Toewijzing aan de Condities

#### Het Belang van de Toewijzing aan de Condities

Toewijzing aan de condities is zo belangrijk omdat men meestal wil weten of er een causaal verband is tussen de afhankelijke en de onafhankelijke variabele. We spreken van een causaal verband als het in isolatie variëren van de onafhankelijke variabele (dit wil zeggen, zonder dat enige andere variable mee varieert) voldoende is voor het optreden van variatie in de afhankelijke variabele. Een andere manier om dit uit te drukken is de volgende: een verband is causaal als het verschil tussen de *groepen* toe te schrijven is aan het verschil tussen de *condities*.

Onder "condities" verstaan we de verschillende omstandigheden waaraan de groepen zijn blootgesteld. Bijvoorbeeld, in het onderzoek met de bejaarden zijn de twee condities "met gratis taxi" en "zonder gratis taxi". De twee groepen kunnen echter op meer punten verschillen dan alleen de condities waartoe ze behoren. Stel dat de bejaarden, voorafgaand aan het onderzoek, onderling inventariseerden hoe groot ieders voorkeur was voor de conditie met gratis taxi (bijvoorbeeld, aan te geven met een getal tussen 1 en 5). En zij kwamen overeen dat de 20 bejaarden met de sterkste voorkeur voor deze gratis taxiservice er ook effectief gebruik van mochten maken. De andere bejaarden kwamen terecht in de conditie zonder gratis taxi. Nu is het zeer aannemelijk dat de sociaal meest actieve bejaarden de grootste voorkeur hadden voor de conditie met gratis taxi, en er daardoor ook in terecht gekomen zijn. Het gevolg hiervan is dat deze groep bejaarden meer sociale activiteiten zal ondernemen dan de andere groep, ook al is er geen effect van de condities. We weten dus niet in welke mate een bepaald geobserveerd verschil in sociale activiteit toe te schrijven is aan de condities of aan verschillen tussen de groepen die bestaan los van de condities. De variabelen waarop de twee groepen van elkaar verschillen, worden storende persoonsvariabelen genoemd.

Het is belangrijk om te zien dat deze verschillen tussen de groepen (i.e., de verschillen die bestaan los van de condities) *systematisch* zijn. Met "systematisch" bedoelen we dat deze verschillen blijven bestaan als we het onderzoek zouden herhalen. Bijvoorbeeld, als we het onderzoek met de bejaarden zouden herhalen, met dezelfde manier van toewijzen aan condities, dan zal de groep in de conditie met gratis taxi opnieuw uit sociaal meer actieve bejaarden bestaan. Dit komt omdat hun behoefte aan sociale activiteiten hen in deze conditie gebracht heeft (door het uitspreken van hun voorkeur). Een statistische toetsing is dan ook geen oplossing voor het probleem van systematische verschillen tussen

de groepen. Bij een statistische toetsing wordt immers alleen rekening gehouden met een *toevallige* verschillende samenstelling van de groepen.

Als men wil dat het verschil tussen de gemiddelden in de twee groepen alleen toe te schrijven is aan de condities, dan moet men ervoor zorgen dat deze groepen niet systematisch van elkaar verschillen. Men kan dit op twee manieren doen: *matching* en *randomisering*. Bij matching gaat men ervan uit dat men alle relevante kenmerken van de proefpersonen kent, dit wil zeggen, alle kenmerken die samenhangen met de afhankelijke variabele. Men zorgt er dan voor dat de groepen gelijk samengesteld zijn wat betreft deze kenmerken. Matching op álle relevante kenmerken is vrijwel nooit te realiseren in de praktijk, omdat men er vrijwel nooit zeker van kan zijn dat men alle relevante kenmerken kent. Matching in zijn zuivere vorm (niet in combinatie met randomisering) kan niet garanderen dat er geen systematische verschillen bestaan tussen de groepen; men kan immers altijd relevante kenmerken over het hoofd gezien hebben.

Bij randomisering hoeft men helemaal niets te weten over de kenmerken die samenhangen met de afhankelijke variabele. Randomisering betekent dat men de proefpersonen op zuiver toevallige wijze aan de condities toewijst. Dit kan bijvoorbeeld door voor elke bejaarde een muntje op te werpen en op basis van de uitkomst (kruis of munt) de bejaarde aan de conditie met of de conditie zonder gratis taxi toe te wijzen.

De kracht van randomisering is erin gelegen dat de toewijzing aan condities met geen enkele andere variabele samenhangt. Iets zuiver toevalligs hangt namelijk met níets samen. Het toewijzen van de bejaarden op basis van hun voorkeur voor een van beide condities is duidelijk geen gerandomiseerde toewijzing. Deze toewijzing hangt namelijk samen met hun behoefte aan sociale activiteit (een storende persoonsvariabele): hoe groter deze behoefte, hoe sterker de voorkeur voor een gratis taxi. Indien de toewijzing gerandomiseerd was (bijvoorbeeld, op basis van een worp met een muntje), dan zouden de sociaal actieve en de sociaal minder actieve bejaarden niet systematisch ongelijk verdeeld zijn over de twee condities. Het kan natuurlijk wel zijn dat deze verdeling over de condities toevallig ongelijk is. Dit is echter niet zo erg want (1) de kans is klein dat een dergelijke ongelijke verdeling zich bij een herhaling van het onderzoek nog eens zal voordoen, en (2) bij de statistische toetsing wordt rekening gehouden met toevallige ongelijke verdelingen.

Randomisering is de enige methode die ons kan garanderen dat er geen systematische verschillen zijn tussen de twee groepen. Dit verklaart de grote populariteit van het gerandomiseerde experiment.

### Als Toewijzing aan de Condities niet Mogelijk is

Randomisering en matching kunnen alleen toegepast kan worden als de onderzoeker zelf de proefpersonen kan toewijzen aan de condities. Vaak is dit onmogelijk. Het kan bijvoorbeeld niet als we te maken hebben met onafhankelijke variabelen zoals geslacht, nationaliteit en sociale klasse. Een onderzoeker kan immers niet bepalen wat iemands geslacht is, wat zijn of haar nationaliteit is en tot welke sociale klasse hij of zij behoort. Dit zijn *natuurlijke* onafhankelijke variabelen, omdat zij niet gemanipuleerd kunnen worden door de onderzoeker.

Er zijn ook variabelen die in principe wel gemanipuleerd kunnen worden, maar waarbij dit zeer moeilijk is of ethisch onverantwoord. Dit is bijvoorbeeld het geval in een

onderzoek naar het effect van marihuana op psychoses bij schizofrenen, want men kan de schizofrenen niet zomaar indelen in een groep die marihuana moet gebruiken en een groep die dat niet mag. Hetzelfde geldt voor een onderzoek naar het effect van het wel of niet hebben van een partner (verder aangeduid als de variabele WEL/GEENPARTNER) op de sociale activiteit van bejaarden, want men kan de bejaarden niet zomaar indelen in een groep met en een groep zonder partner. Toewijzing aan condities is hier niet mogelijk, en zuiver toevallige toewijzing dus ook niet.

Men spreekt van een observationeel onderzoek als de proefpersonen niet door de onderzoeker aan de condities toegewezen worden, maar gewoon geobserveerd wordt in welke conditie zij zich bevinden. We hebben te maken met een observationeel onderzoek als men het effect wil onderzoeken van WEL/GEENPARTNER op SOCIALE ACTIVITEIT bij bejaarden. Hier observeert men de conditie waarin de bejaarde zich bevindt (met/zonder partner).

Met een observationeel onderzoek kan men geen harde uitspraken doen over een causaal verband tussen de afhankelijke en de onafhankelijke variabele. Het is immers niet uit te sluiten dat er storende persoonsvariabelen zijn waarop de groepen van elkaar verschillen. Bijvoorbeeld, bejaarden zonder partner kunnen gemiddeld minder sociaal actief zijn dan bejaarden met partner, zonder dat er een causaal verband is tussen WEL/GEENPARTNER en SOCIALE ACTIVITEIT. Bejaarden zonder partner zijn misschien gemiddeld ouder dan bejaarden met partner en hebben daarom (door een slechtere gezondheid, e.d.) ook minder sociale contacten. Dus, WEL/GEENPARTNER en SOCIALE ACTIVITEIT kunnen beide een verband hebben met de storende persoonsvariabele LEEFTIJD.

Vaak wil men op basis van een observationeel onderzoek toch een causale uitspraak doen ("A heeft een effect op Y."). Alhoewel men een dergelijke causale uitspraak nooit helemaal kan rechtvaardigen, kan men toch gebruik maken van een aantal methoden waarmee men zo een uitspraak meer aannemelijk kan maken. Deze methoden zijn de volgende:

1. *Statistische controle* van de invloed van storende persoonsvariabelen. Dit noemt men ook vaak statistische *uitzuivering*, of uitzuivering zondermeer. Bijvoorbeeld, stel dat leeftijd, naast het al dan niet hebben van een partner, samenhangt met sociale activiteit. Dan is het mogelijk om het verschil in sociale activiteit tussen de groep met en de groep zonder partner uit te zuiveren voor de bijdrage van de variabele leeftijd aan dit verschil. Er zijn twee technieken waarmee men kan toetsen of het uitgezuiverde verschil statistisch significant is: (1) de *variantieanalyse*[1] *voor ongebalanceerde designs*, die men kan gebruiken als de storende variabele categorisch[2] is (zoals geslacht, persoonlijkheidstype, burgerlijke stand, ...), en (2) de

[1] Deze techniek wordt vaak aangeduid met zijn Engelse afkorting *ANOVA* (van *analysis of variance*), ook in Nederlandse teksten.

[2] Een categorische variabele geeft aan tot welke categorie een persoon behoort. Een categorische variabele wordt soms ook een *kwalitatieve* variabele genoemd. De term "kwalitatieve variabele" heeft het nadeel dat hij de indruk wekt dat hiermee iets gezegd wordt over de kwaliteit van een persoon. Dit is niet zo, want met "kwalitatief" bedoelen we hier alleen dat deze variabele aangeeft tot welke categorie een persoon behoort en het enige wat hierbij telt is dat elke persoon tot één en slechts één categorie behoort. Het doet er dus niet toe of deze categorieën betrekking hebben op een bepaalde kwaliteit van de persoon of niet.

*covariantieanalyse*[3], die men kan gebruiken als de storende variabele kwantitatief[4] is (zoals leeftijd, een testscore, ...).

2. Gebruikmaken van een onderzoeksdesign met *voor- en nameting*. Dit onderzoeksdesign houdt dit in dat men de afhankelijke variabele één keer registreert vóór de behandeling (interventie, training, ...) en één keer erna. Men berekent dan de verschilscores (voor min na) voor elke proefpersoon en kijkt of de groepen gemiddeld van elkaar verschillen op deze verschilscores. De idee achter deze werkwijze is dat men, door het berekenen van verschilscores, corrigeert voor eventuele verschillen tussen de groepen die reeds bestaan vóór de behandeling.

   Dit onderzoeksdesign lijkt op een binnen-subjectenexperiment maar is het niet. Het is geen experiment omdat er geen sprake is van manipulatie; de proefpersonen worden immers niet toegewezen aan de condities. Daarom is contrabalancering ook onmogelijk; het is inherent aan een voor- en een nameting dat zij alleen in deze volgorde geobserveerd kunnen worden. Het onderzoeksdesign met een voor- en een nameting wordt een *quasi-experiment* genoemd.

### 3.1.4 Experimentele Controle van de Condities

#### Het Belang van Experimentele Controle van de Condities

Men is vaak geïnteresseerd in één welbepaald aspect van de condities die men met elkaar vergelijkt. In paragraaf 1.2 gaven we in dit verband het voorbeeld van een onderzoek naar het effect van EEN- VERSUS TWEEVERDIENER op TEVREDENHEID MET DE RELATIE bij een groep werkenden. In dit onderzoek was men geïnteresseerd in het effect van EEN- VERSUS TWEEVERDIENER op zich en niet van een andere variabele die ermee samenhangt. Het is echter zeer goed mogelijk dat de condities (in dit voorbeeld, een- en tweeverdiener) op meerdere aspecten van elkaar verschillen, zoals bijvoorbeeld de variabele OPLEIDINGSNIVEAU PARTNER. Deze andere aspecten worden storende situatievariabelen genoemd. Het is dan mogelijk dat het geobserveerde verband tussen de onafhankelijke en de afhankelijke variabele (blijkend uit een verschil tussen de condities) niet toe te schrijven is aan het aspect waar men in geïnteresseerd is, maar wel aan die storende situatievariabelen.

De oplossing voor het probleem van storende situatievariabelen is experimentele controle van de condities. Experimentele controle van de condities houdt in dat men de verschillende condities vergelijkbaar maakt wat betreft de mogelijke storende situatievariabelen. Voor het onderzoek naar TEVREDENHEID MET DE RELATIE zou dit bijvoorbeeld inhouden dat men ervoor zorgt dat de partners in de eenverdienersconditie gemiddelde hetzelfde opleidingsniveau hebben als de partners in de tweeverdienersconditie. Hoe men dit concreet moet doen, is niet onmiddellijk duidelijk, en het toont aan

[3] Deze techniek wordt vaak aangeduid met zijn Engelse afkorting *ANCOVA* (van *analysis of covariance*), ook in Nederlandse teksten.

[4] We noemen een variabele "kwantitatief" als hij een bepaalde graad of hoeveelheid van iets uitdrukt. De waarden van een kwantitatieve variabele zijn altijd geordend (i.e., van twee personen met een verschillende score op deze variabele ligt er één boven de ander). De waarden van een kwantitatieve variabele zijn altijd getallen. De waarden van een categorische variabele, daarentegen, kunnen ook met letters aangeduid worden.

waar het probleem ligt bij experimentele controle: experimentele controle van de condities is slechts mogelijk als de onderzoeker zelf de aard van de condities kan bepalen. In veel onderzoek buiten het laboratorium is dit niet mogelijk; daar moet men de condities nemen zoals ze zijn.

### Als Experimentele Controle van de Condities niet Mogelijk is

Als experimentele controle van de condities niet mogelijk is, dan kan men gebruik maken van statistische controle van (uitzuivering voor) storende situatievariabelen. Voor het voorbeeld over het effect van EEN- VERSUS TWEEVERDIENER op TEVREDENHEID MET DE RELATIE bij een groep werkenden betekent dit dat men uitzuivert voor de bijdrage van de storende situatievariabele OPLEIDINGSNIVEAU PARTNER aan het verschil in tevredenheid tussen de twee groepen werkenden (een- en tweeverdieners). Dit impliceert dat men voor elke proefpersoon registreert wat het opleidingsniveau is van zijn of haar partner. Omdat het hier gaat om een kwántitatieve storende variabele, gebeurt de toetsing van het uitgezuiverde verschil met behulp van de covariantieanalyse. Als de storende variabele categorisch zou zijn (bijvoorbeeld, type afdeling), dan zou men hiervoor gebruik maken van de variantieanalyse voor ongebalanceerde designs.

In de volgende paragrafen behandelen we de bovengenoemde technieken, variantieanalyse voor ongebalanceerde designs, covariantieanalyse, en het design met een voor- en een nameting. De variantieanalyse voor ongebalanceerde designs is ingewikkelder dan de variantieanalyse die steevast als eerste behandeld wordt in het statistiek- en methodenonderwijs, de variantieanalyse voor gebalanceerde designs. In dit boek, worden gebalanceerde en ongebalanceerde designs samen behandeld. Een lezer met een basiskennis van variantieanalyse, zal bepaalde begrippen en formules herkennen, maar alles wat specifiek is voor ongebalanceerde designs zal wellicht nieuw zijn. Na de variantieanalyse, wordt de covariantieanalyse behandeld en daarna de analyse van data uit een design met een voor- en een nameting.

Voorafgaand aan dit alles, behandelen we de variantieanalyse met één factor. Dit is de meeste eenvoudige vorm van variantieanalyse. Uitzuivering voor storende variabelen is hier niet aan de orde. De functie van de paragraaf over variantieanalyse met één factor is voorbereiding op de latere paragrafen over variantieanalyse met twee factoren en covariantieanalyse.

---

*Opgave 1*
Een docent wil weten of een bijspijkercursus wiskunde een effect heeft op de tentamencijfers voor het vak Statistiek. Hij selecteert daartoe een groep van 20 personen die at random gekozen wordt uit de bijspijkercursus en andere groep van 20 personen die at random gekozen wordt uit de groep die wel het statistiektentamen gaan doen, maar niet de bijspijkercursus volgen.

1. Is dit een experiment of een observationeel onderzoek? Leg uit.
2. Wat is het voordeel van een gerandomiseerd experiment boven een observationeel onderzoek?

*Opgave 2*
Een onderzoeker is geïnteresseerd in de effectiviteit van twee nieuwe therapieën ter bestrijding van vereenzaming bij bejaarden. Hij selecteert daartoe eerst een groep mensen die hoog scoren

op een schaal voor vereenzaming. Vervolgens verdeelt hij deze mensen at random over de twee therapievormen.

1. Is hier sprake van experimentele controle? Zo ja, waarvoor wordt hier gecontroleerd?
2. Naar welke populatie kun je generaliseren op basis van dit onderzoek?

Is dit een experiment of een observationeel onderzoek? Leg uit.

*Opgave 3*
Hieronder vind je een lijst van onderzoeksvragen. Voor sommige van deze vragen kan men zowel een experimenteel als een observationeel onderzoek opzetten. Voor de anderen kan men alleen een observationeel onderzoek opzetten. Geef voor elke onderzoeksvraag aan hoe een experimenteel en een observationeel onderzoek eruit zou kunnen zien. Voor sommige onderzoeksvragen moet je een experimenteel én een observationeel onderzoek beschrijven, en voor de anderen moet je alleen een observationeel onderzoek beschrijven.

1. Medicijn X zorgt ervoor dat men 's nachts langer slaapt.
2. Mensen die elke dag een half uur mediteren irriteren zich minder aan de mensen in hun omgeving.
3. Allochtonen vertonen meer crimineel gedrag dan autochtonen.
4. Mensen met een partner hebben minder last van hart- en vaatziekten dan mensen zonder partner.
5. Werknemers met een managementfunctie hebben een grotere kans om overspannen te worden dan werknemers die in het productieproces staan.

---

## 3.2 Variantieanalyse met Eén Factor

### 3.2.1 Voorbeeld: Mobiliteit en Sociale Activiteit

We nemen hier hetzelfde voorbeeld als in de vorige paragraaf. Er is slechts één verschil: er waren dríe mobiliteitscondities in plaats van twee:

1. De gratis-taxiconditie. De bejaarden in deze conditie konden gratis gebruik maken van een speciale taxi-service die hen vervoerde van en naar vrienden, bijeenkomsten, voorstellingen, ....
2. De vervoer-op-bestellingconditie. De bejaarden in deze conditie konden geen gebruik maken van een gratis taxi-service die onmiddellijk voor hen klaar stond, maar moesten in het begin van elke week doorgeven op welke dagen en uren zij vervoer nodig hadden. Dit vervoer was wel gratis.
3. De conditie zonder gratis vervoer. De bejaarden in deze conditie werd geen enkele vervoersfaciliteit aangeboden.

Op het einde van het onderzoek werd voor elke bejaarde het gemiddeld aantal uren sociale contacten per week berekend. Deze gegevens staan in Tabel 3.1. Merk op dat het aantal bejaarden in de drie groepen niet gelijk is.

| Gratis taxi | | Vervoer op bestelling | | Zonder gratis vervoer | |
|---|---|---|---|---|---|
| Proef-persoon | Aantal uren sociale contacten | Proef-persoon | Aantal uren sociale contacten | Proef-persoon | Aantal uren sociale contacten |
| 1 | 11 | 1 | 6.2 | 1 | 8.4 |
| 2 | 10 | 2 | 9 | 2 | 6 |
| 3 | 7.5 | 3 | 7.3 | 3 | 3 |
| 4 | 11.6 | 4 | 6 | 4 | 9 |
| 5 | 9 | 5 | 7 | 5 | 7.6 |
| 6 | 11.4 | 6 | 5.8 | 6 | 5 |
| 7 | 7.4 | 7 | 6 | 7 | 5.3 |
| 8 | 11.3 | 8 | 8.7 | 8 | 4 |
| | | 9 | 6.1 | 9 | 8.7 |
| | | | | 10 | 7 |
| Gem.: | 9.9 | Gem.: | 6.9 | Gem.: | 6.4 |

Tabel 3.1: Gegevens uit een hypothetisch onderzoek naar het effect van MOBILITEIT op SOCIALE ACTIVITEIT bij bejaarden.

De gemiddelden in de drie groepen verschillen van elkaar (9.9, 6.9 en 6.4). Op het eerst zicht, zou men hieruit besluiten dat MOBILITEIT wel degelijk een effect heeft op SOCIALE ACTIVITEIT. Echter, om dit te kunnnen besluiten moet men uitsluiten dat de verschillen tussen de groepsgemiddelden ook toevallig kunnen zijn. Met "toevallig" bedoelen we dat deze verschillen best kunnen verdwijnen, of zelfs omkeren, als we het onderzoek zouden herhalen met een nieuwe groep bejaarden. De vraag of de gevonden verschillen toevallig kunnen zijn kan men beantwoorden met een statistische toets, de zogenaamde F-toets, die een onderdeel is van de variantieanalyse.

In de variantieanalyse wordt een onafhankelijke variabele meestal een *factor* genoemd en hier doen wij dat ook. Deze keuze is niet helemaal gelukkig, want het creëert wel eens verwarring met de term factor zoals die gebruikt wordt in de factoranalyse. Daar heeft hij echter een andere betekenis.

De variantieanalyse is een techniek voor een *kwantitatieve* afhankelijke variabele en een *categorische* onafhankelijke variabele. Kwantitatieve variabelen drukken een bepaalde graad of hoeveelheid uit. In het voorbeeldonderzoek is de kwantitatieve afhankelijke variabele het aantal uren sociale contacten per week. Andere voorbeelden van kwantitatieve variabelen zijn reactietijden, testscores, en beoordelingen op een 10-puntenschaal. Categorische predictoren, daarentegen, drukken alleen uit dát er een verschil is, maar de getallen waarmee dit gebeurt drukken geen hoeveelheid uit. In het voorbeeldonderzoek is de categorische onafhankelijke variabele MOBILITEIT, een variabele met drie niveaus. Men kan deze drie niveaus aanduiden met de getallen 1, 2, en 3. Dit betekent echter niet dat de conditie met vervoer op bestelling precies in het midden tussen de twee andere condities ligt. Andere voorbeelden van categorische variabelen zijn geslacht, nationaliteit, politieke voorkeur, en studierichting.

### 3.2.2 De Notatie voor de Gegevens

De notatie voor de gegevens wordt getoond in Tabel 3.2. Het aantal groepen wordt aangeduid met $K$. In het voorbeeldonderzoek is $K$ gelijk aan 3. Als index voor de groepen gebruiken we $k$ ($k = 1, \ldots, K$) en als index voor de personen binnen een groep gebruiken we $i$. De observatie voor de i-de persoon in de k-de groep (in het voorbeeldonderzoek, het gemiddeld aantal uren sociale contacten per week) duiden we aan met $y_{ki}$ en het gemiddelde in de k-de groep met $\overline{y}_k$. Het aantal observaties in de k-de groep duiden we aan met $n_k$. In het voorbeeldonderzoek is $n_1$ gelijk aan 8, $n_2$ gelijk aan 9 en $n_3$ gelijk aan 10. Het aantal observaties in alle $K$ groepen samen duiden we aan met $N$ ($N = n_1 + \ldots + n_K$).

| Persoon | Groep 1 | Groep 2 | $\cdots$ | Groep $K$ |
|---|---|---|---|---|
| 1 | $y_{11}$ | $y_{21}$ | $\cdots$ | $y_{K1}$ |
| 2 | $y_{12}$ | $y_{22}$ | $\cdots$ | $y_{K2}$ |
| 3 | $y_{13}$ | $y_{23}$ | $\cdots$ | $y_{K3}$ |
| 4 | $y_{14}$ | $y_{24}$ | $\cdots$ | $y_{K4}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $n_1$ | $y_{1n_1}$ | | | |
| $\vdots$ | | $\vdots$ | $\vdots$ | $\vdots$ |
| $n_K$ | | | | $y_{Kn_K}$ |
| $\vdots$ | | $\vdots$ | $\vdots$ | |
| $n_2$ | | $y_{2n_2}$ | | |
| Gem.: | $\overline{y}_1$ | $\overline{y}_2$ | $\cdots$ | $\overline{y}_K$ |

Tabel 3.2: Datamatrix voor een onderzoek met $K$ groepen (met een verschillend aantal observaties per groep).

Behalve zoals in Tabel 3.2, bestaat er nog een tweede manier om de gegevens te ordenen. Deze wordt getoond in Tabel 3.3. In de eerste kolom staat de afhankelijke variabele. Hier staan de $K$ kolommen met $y_{ki}$'s uit Tabel 3.2 allemaal onder elkaar. In de tweede kolom staat de factor. Elk van de niveaus van deze factor wordt aangeduid met een getal tussen 1 en $K$. In de meeste computerprogramma's worden de gegevens op deze manier geordend.

De kern van een variantieanalyse is de berekening van een F-statistiek. In de volgende paragraaf, *Het Rekenwerk*, laten we zien hoe deze F-statistiek berekend wordt voor de gegevens van het voorbeeldonderzoek in Tabel 3.1. In de daarna volgende paragrafen geven we de formules die nodig zijn om deze F-statistiek ook voor andere gegevens uit te rekenen. Bij het bestuderen van de algemene formules is het goed om terug te bladeren naar de concrete berekeningen van het voorbeeldonderzoek. Dan zie je waar de algemene formules in een concreet onderzoek voor staan.

| Afhankelijke Variabele | Factor |
|---|---|
| $y_{11}$ | 1 |
| $y_{12}$ | 1 |
| $\vdots$ | $\vdots$ |
| $y_{1n_1}$ | 1 |
| $y_{21}$ | 2 |
| $y_{22}$ | 2 |
| $\vdots$ | $\vdots$ |
| $y_{2n_2}$ | 2 |
| $\vdots$ | $\vdots$ |
| $y_{K1}$ | $K$ |
| $y_{K2}$ | $K$ |
| $\vdots$ | $\vdots$ |
| $y_{Kn_K}$ | $K$ |

Tabel 3.3: Datamatrix voor een onderzoek met $K$ groepen met ordening van de gegevens in twee kolommen.

### 3.2.3 Het Rekenwerk

De berekening van de F-statistiek is gebaseerd op voorspelling van observaties. Dit voorspellen moet men niet te letterlijk nemen, want de observaties zijn er reeds en hoeven dus niet meer voorspeld te worden. Men kan "voorspelling" hier het beste opvatten als de waarde die men aan een observatie zou geven indien men niet meer over de individuele observaties zou beschikken maar wel nog over de groepsgemiddelden.

Voor de berekening van de F-statistiek hebben we voor elke observatie twee voorspellingen nodig: (1) de voorspelling onder de veronderstelling dat er geen verschil is tussen de groepen, en (2) de voorspelling onder de veronderstelling dat er wel een verschil is tussen de groepen. Onder de veronderstelling dat er geen verschil is tussen de groepen, ligt het voor de hand om voor elke observatie het gemiddelde van alle observaties te voorspellen. Voor de gegevens in Tabel 3.1 is dit gemiddelde gelijk aan 7.6 (na afronding). Dus, voor elke observatie maken we de voorspelling 7.6. En onder de veronderstelling dat er wel een verschil is tussen de groepen, ligt het voor de hand om voor elke observatie het groepsgemiddelde te voorspellen. Voor de gegevens in Tabel 3.1 zijn de groepsgemiddelden gelijk aan 9.9, 6.9 en 6.4 (voor, respectievelijk, gratis taxi, vervoer op bestelling, en zonder gratis vervoer). Dus, voor de observaties in de gratis-taxigroep maken we de voorspelling 9.9, voor de observaties in de vervoer-op-bestellinggroep maken we de voorspelling 6.9, en voor de observaties in de groep zonder gratis vervoer maken we de voorspelling 6.4.

In de volgende stap van de berekening wordt het verschil tussen deze twee voorspellin-

gen genomen en dit verschil wordt gekwadrateerd. Deze gekwadrateerde verschillen zijn des te groter naarmate de groepsgemiddelden meer van elkaar verschillen. De berekeningen voor het voorbeeldonderzoek worden gegeven in Tabel 3.4. De observaties in de eerste groep worden aangeduid met $y_{1i}$, en $i$ loopt van 1 tot $n_1 = 8$. De observaties in de tweede en de derde groep worden aangeduid met $y_{2i}$ en $y_{3i}$ en $i$ loopt nu van 1 tot $n_2 = 9$ (voor de tweede groep) en van 1 tot $n_3 = 10$ (voor de derde groep). De gemiddelden in de drie groepen worden aangeduid met, respectievelijk, $\overline{y}_1$, $\overline{y}_2$, en $\overline{y}_3$. En het gemiddelde van alle 27 $(= 8+9+10)$ observaties samen wordt aangeduid met $\overline{y}_{\sqcup}$ (spreek uit: "$y$ gemiddeld bakje[5]"). De gekwadrateerde verschillen tussen de twee voorspellingen worden aangeduid met $(\overline{y}_1 - \overline{y}_{\sqcup})^2$, $(\overline{y}_2 - \overline{y}_{\sqcup})^2$, en $(\overline{y}_3 - \overline{y}_{\sqcup})^2$ voor, respectievelijk, de eerste, de tweede, en de derde groep.

| | Gratis taxi | | Vervoer op bestelling | | Zonder gratis vervoer | | |
|---|---|---|---|---|---|---|---|
| | $y_{1i}$ | $(\overline{y}_1 - \overline{y}_{\sqcup})^2$ | $y_{2i}$ | $(\overline{y}_2 - \overline{y}_{\sqcup})^2$ | $y_{3i}$ | $(\overline{y}_3 - \overline{y}_{\sqcup})^2$ | |
| | 11 | $(9.9 - 7.6)^2$ | 6.2 | $(6.9 - 7.6)^2$ | 8.4 | $(6.4 - 7.6)^2$ | |
| | 10 | $(9.9 - 7.6)^2$ | 9 | $(6.9 - 7.6)^2$ | 6 | $(6.4 - 7.6)^2$ | |
| | 7.5 | $(9.9 - 7.6)^2$ | 7.3 | $(6.9 - 7.6)^2$ | 3 | $(6.4 - 7.6)^2$ | |
| | 11.6 | $(9.9 - 7.6)^2$ | 6 | $(6.9 - 7.6)^2$ | 9 | $(6.4 - 7.6)^2$ | |
| | 9 | $(9.9 - 7.6)^2$ | 7 | $(6.9 - 7.6)^2$ | 7.6 | $(6.4 - 7.6)^2$ | |
| | 11.4 | $(9.9 - 7.6)^2$ | 5.8 | $(6.9 - 7.6)^2$ | 5 | $(6.4 - 7.6)^2$ | |
| | 7.4 | $(9.9 - 7.6)^2$ | 6 | $(6.9 - 7.6)^2$ | 5.3 | $(6.4 - 7.6)^2$ | |
| | 11.3 | $(9.9 - 7.6)^2$ | 8.7 | $(6.9 - 7.6)^2$ | 4 | $(6.4 - 7.6)^2$ | |
| | | | 6.1 | $(6.9 - 7.6)^2$ | 8.7 | $(6.4 - 7.6)^2$ | |
| | | | | | 7 | $(6.4 - 7.6)^2$ | |
| Som | | **42.32** | | **4.41** | | **14.4** | **<u>61.13</u>** |

Tabel 3.4: Berekening van de effectkwadratensom voor de data uit het voorbeeldonderzoek. De som van de gekwadrateerde verschillen per groep werd vet gedrukt en de som hiervan over de groepen werd vet gedrukt en onderlijnd.

Voor de berekening van de F-statistiek moeten we de som nemen van alle gekwadrateerde verschillen in Tabel 3.4. Deze som wordt de *effectkwadratensom* genoemd. We nemen eerst de som voor elk van de groepen afzonderlijk en dan nemen we de som over de groepen. Voor de eerste groep, hebben we 8 keer $(9.9 - 7.6)^2$, voor de tweede groep, 9 keer $(6.9 - 7.6)^2$, en voor de derde groep, 10 keer $(6.4 - 7.6)^2$. Dus,

$$\begin{aligned}\text{Effectkwadratensom} &= 8 \times (9.9 - 7.6)^2 + 9 \times (6.9 - 7.6)^2 + 10 \times (6.4 - 7.6)^2 \\ &= 42.32 + 4.41 + 14.4 \\ &= 61.13 \quad .\end{aligned}$$

Merk op dat deze effectkwadratensom groter is naarmate de groepsgemiddelden meer van elkaar verschillen.

[5]Het gemiddelde van alle observaties op één hoop gegooid, in één *bakje*.

Voor de berekening van de F-statistiek hebben we ook de zogenaamde *errorkwadratensom* nodig. Deze errorkwadratensom is gebaseerd op de verschillen tussen de observaties ($y_{1i}$, $y_{2i}$, en $y_{3i}$) en de voorspellingen onder de veronderstelling dat er een verschil is tussen de groepen ($\overline{y}_1$, $\overline{y}_2$, en $\overline{y}_3$). Deze verschillen worden gekwadrateerd en opgeteld, net zoals bij de effectkwadratensom. Deze berekeningen worden weergegeven in Tabel 3.5. In deze tabel worden de gekwadrateerde verschillen aangeduid met $(y_{1i} - \overline{y}_1)^2$, $(y_{2i} - \overline{y}_2)^2$, en $(y_{3i} - \overline{y}_3)^2$. We nemen eerst de som voor elk van de groepen afzonderlijk en dan nemen we de som over de groepen:

$$\begin{aligned}
\text{Errorkwadratensom} &= (11-9.9)^2 + (10-9.9)^2 + \ldots + (11.3-9.9)^2 + \qquad \text{(groep 1)} \\
&\quad (6.2-6.9)^2 + (9-6.9)^2 + \ldots + (6.1-6.9)^2 + \qquad \text{(groep 2)} \\
&\quad (8.4-6.4)^2 + (6-6.4)^2 + \ldots + (7-6.4)^2 \qquad \text{(groep 3)} \\
&= 21.14 + 11.78 + 38.5 \\
&= 71.42 \quad .
\end{aligned}$$

| | Gratis taxi | | Vervoer op bestelling | | Zonder gratis vervoer | | |
|---|---|---|---|---|---|---|---|
| | $y_{1i}$ | $(y_{1i} - \overline{y}_1)^2$ | $y_{2i}$ | $(y_{2i} - \overline{y}_2)^2$ | $y_{3i}$ | $(y_{3i} - \overline{y}_3)^2$ | |
| | 11 | $(11-9.9)^2$ | 6.2 | $(6.2-6.9)^2$ | 8.4 | $(8.4-6.4)^2$ | |
| | 10 | $(10-9.9)^2$ | 9 | $(9-6.9)^2$ | 6 | $(6-6.4)^2$ | |
| | 7.5 | $(7.5-9.9)^2$ | 7.3 | $(7.3-6.9)^2$ | 3 | $(3-6.4)^2$ | |
| | 11.6 | $(11.6-9.9)^2$ | 6 | $(6-6.9)^2$ | 9 | $(9-6.4)^2$ | |
| | 9 | $(9-9.9)^2$ | 7 | $(7-6.9)^2$ | 7.6 | $(7.6-6.4)^2$ | |
| | 11.4 | $(11.4-9.9)^2$ | 5.8 | $(5.8-6.9)^2$ | 5 | $(5-6.4)^2$ | |
| | 7.4 | $(7.4-9.9)^2$ | 6 | $(6-6.9)^2$ | 5.3 | $(5.3-6.4)^2$ | |
| | 11.3 | $(11.3-9.9)^2$ | 8.7 | $(8.7-6.9)^2$ | 4 | $(4-6.4)^2$ | |
| | | | 6.1 | $(6.1-6.9)^2$ | 8.7 | $(8.7-6.4)^2$ | |
| | | | | | 7 | $(7-6.4)^2$ | |
| Som | | **21.14** | | **11.78** | | **38.5** | **71.42** |

Tabel 3.5: Berekening van de errorkwadratensom voor de data uit het voorbeeldonderzoek.

De volgende stap is de berekening van de *gemiddelde* kwadratensommen. Anders dan je misschien zou denken, worden deze gemiddelde kwadratensommen niet berekend door de gewone kwadratensommen (de effect- en de errorkwadratensom) te delen door het aantal observaties. De gemiddelde effectkwadratensom wordt berekend door de effectkwadratensom te delen door het aantal groepen min 1. In het voorbeeldonderzoek zijn er drie groepen, en dus moeten we delen door 2:

$$\begin{aligned}
\text{Gemiddelde effectkwadratensom} &= \frac{61.13}{2} \\
&= 30.56 \quad .
\end{aligned}$$

En de gemiddelde errorkwadratensom wordt berekend door de errorkwadratensom te delen door het aantal observaties min het aantal groepen. In het voorbeeldonderzoek

zijn er 27 observaties en drie groepen, en dus moeten we delen door 24:

$$\begin{aligned}\text{Gemiddelde errorkwadratensom} &= \frac{71.42}{24} \\ &= 2.98 \quad .\end{aligned}$$

De laatste stap in de berekening is het delen van de gemiddelde effectkwadratensom door de gemiddelde errorkwadratensom. Dit levert ons de F-statistiek:

$$\begin{aligned}F &= \frac{\text{Gemiddelde effectkwadratensom}}{\text{Gemiddelde errorkwadratensom}} \\ &= \frac{30.56}{2.98} \\ &= 10.26 \quad .\end{aligned}$$

Deze F-statistiek geeft een antwoord op de vraag of de verschillen tussen de groepsgemiddelden toevallig zijn: hoe groter de F-statistiek, hoe minder aannemelijk het is dat de verschillen tussen de groepsgemiddelden toevallig zijn. De onvermijdelijke vraag is dan hoe groot de F-statistiek precies moet zijn om te kunnen besluiten dat de verschillen tussen de gemiddelden niet toevallig zijn. Deze vraag kunnen we beantwoorden door gebruik te maken van het feit dat de F-statistiek een bekende verdeling heeft indien de verschillen tussen de groepsgemiddelden daadwerkelijk toevallig zijn. Dit is de zogenaamde $\mathcal{F}$-verdeling (de *ronde* F-verdeling). Men kan uitrekenen wat de kans is onder deze $\mathcal{F}$-verdeling om een F-statistiek te observeren die nog extremer is dan de geobserveerde. Voor het voorbeeldonderzoek is de kans om een F-statistiek te observeren die groter is dan 10.26, gesteld dat de verschillen tussen de groepsgemiddelden toevallig zijn, gelijk aan 0.0006. Deze kans is kleiner dan 0.05 (de kritische kans die in de meeste onderzoeken gehanteerd wordt) en daarom besluiten we dat de verschillen tussen de gemiddelden niet toevallig zijn.

### 3.2.4 De F-Toets

We gebruiken de F-toets om te bepalen of de gemiddelden in de groepen (de $\overline{y}_k$'s) significant van elkaar verschillen. Het significant zijn van deze verschillen kan op de volgende equivalente manieren opgevat worden:

1. De verschillen tussen de gemiddelden kunnen niet door toeval verklaard worden.
2. De verschillen tussen de gemiddelden zijn groot genoeg om te concluderen dat ook de corresponderende verwachte waarden van elkaar verschillen.

Een verwachte waarde is een gemiddelde van een oneindig aantal observaties. In de variantieanalyse corresponderen de observaties meestal met personen. De verwachte waarde in de k-de groep wordt dan gevonden door een oneindig aantal personen te observeren in deze groep en het gemiddelde te nemen van hun scores. Deze verwachte waarde wordt aangeduid met $\mu_k$.

De verwachte waarde is iets theoretisch omdat er voor zijn berekening een oneindig aantal observaties nodig zijn, en dit komt in de praktijk niet voor. In de praktijk hebben

we alleen te maken met steekproeven uit populaties. Het oneindig aantal personen in de k-de groep, gebruikt om de verwachte waarde $\mu_k$ te definiëren, wordt ook de k-de populatie genoemd. Het eindige aantal personen in de k-de groep dat feitelijk geobserveerd werd, en waarop het gemiddelde berekend wordt, is een steekproef uit deze populatie.

Het doel van de F-toets is om op basis van de steekproefgemiddelden $\overline{y}_k$ een uitspraak te doen over de verschillen tussen de verwachte waarden $\mu_k$. We zeggen ook dat we de F-toets gebruiken om de volgende *nulhypothese* te toetsen:

$$\mu_1 = \mu_2 = \cdots = \mu_K \quad .$$

In woorden, staat er in deze nulhypothese dat er geen verschillen zijn tussen de verwachte waarden in de $K$ populaties.

De F-toets maakt gebruikt van de F-statistiek, die als volgt gedefinieerd is:

$$\boxed{F = \frac{MS_A}{MS_{Error|A}} \quad .}$$

De teller en de noemer zijn gemiddelde kwadratensommen (in het Engels, *mean squares*): in de teller staat de gemiddelde *effect*kwadratensom (in het Engels, *effect mean squares*) en in de noemer staat de gemiddelde *error*kwadratensom (in het Engels, *error mean squares*). De letter A duidt de factor aan en het subscript $Error|A$ staat voor "fouten (error) na voorspelling op basis van factor A". Deze F-statistiek is een maat voor de *on*aannemelijkheid van de nulhypothese: hoe groter de F-statistiek, hoe minder aannemelijk de nulhypothese.

De gemiddelde kwadratensommen worden als volgt berekend:

$$\boxed{\begin{aligned} MS_A &= \frac{SS_A}{K-1} \\ MS_{Error|A} &= \frac{SS_{Error|A}}{N-K} \quad . \end{aligned}}$$

De tellers van deze breuken, $SS_A$ en $SS_{Error|A}$ (de $SS$ komt van het Engelse *sum of squares*), zijn (gewone) kwadratensommen en hun noemers zijn de daarmee corresponderende *vrijheidsgraden*. Dus, het aantal vrijheidsgraden van $SS_A$ is gelijk aan $K-1$, het aantal groepen min 1, en het aantal vrijheidsgraden van $SS_{Error|A}$ is gelijk aan $N-K$, het totaal aantal observaties min het aantal groepen. We komen later nog terug op de berekening van de vrijheidsgraden.

De grootte van de effectkwadratensom $SS_A$ hangt af van de verschillen tussen de groepsgemiddelden $\overline{y}_k$: hoe meer de groepsgemiddelden van elkaar verschillen, hoe groter de effectkwadratensom. Dit zorgt ervoor dat de F-toets de volgende plausibele eigenschap heeft: hoe meer de groepsgemiddelden van elkaar verschillen, hoe groter de F-statistiek, en hoe minder aannemelijk de nulhypothese dat alle verwachte waarden $\mu_k$ gelijk zijn.

De F-statistiek hangt ook af van zijn noemer, de gemiddelde errorkwadratensom $MS_{Error|A}$, en deze hangt op zijn beurt af van $SS_{Error|A}$, de gewone errorkwadratensom.

De grootte van $SS_{Error|A}$ hangt af van de spreiding van de observaties binnen elk van de groepen: hoe dichter de observaties bínnen elk van de groepen bij elkaar liggen, hoe kleiner de errorkwadratensom. Dit zorgt ervoor dat de F-toets de volgende eigenschap heeft: hoe dichter de observaties binnen elk van de groepen bij elkaar liggen, hoe groter de F-statistiek, en hoe minder aannemelijk de nulhypothese dat alle verwachte waarden $\mu_k$ gelijk zijn. Dit is een plausibele eigenschap, want de spreiding van de observaties binnen de groepen bepaalt mede de verschillen tussen de groepsgemiddelden: als de nulhypothese waar is (de verwachte waarden $\mu_k$ verschillen niet van elkaar) en de spreiding binnen de groepen is klein, dan verwachten we ook kleine verschillen tussen de groepsgemiddelden $\overline{y}_k$. Hieruit volgt dat, bij gegeven verschillen tussen de groepsgemiddelden, de nulhypothese minder aannemelijk is naarmate de spreiding binnen de groepen kleiner is (en dus ook de errorkwadratensom).

In de volgende paragrafen bespreken we de berekening van de effect- en de errorkwadratensommen. Bij deze berekening wordt gebruik gemaakt van voorspellingen op basis van het zogenaamde *variantieanalysemodel*. We bespreken eerst dit model, dan de effectkwadratensom, en daarna de errorkwadratensom.

### 3.2.5 Het Variantieanalysemodel

Men kan de variantieanalyse op meerdere manieren uitleggen. In de vorige paragraaf, *Het Rekenwerk*, werd de variantieanalyse uitgelegd vertrekkende van de berekeningen. We gaan nu een stap verder en leggen uit waarom de variantieanalyse geschikt is voor zijn doel, namelijk het toetsen van de nulhypothese dat de verwachte waarden in de $K$ groepen (de $\mu_k$'s) gelijk zijn. De belangrijkste ingrediënt van deze uitleg is het *variantieanalysemodel*. Strikt genomen, is een model een gedachtenspinsel; het drukt onze verwachting uit over hoe de observaties, de $y_{ki}$'s, afhangen van de verwachte waarden, de $\mu_k$'s:

$$\boxed{y_{ki} = \mu_k + e_{ki} \quad .}$$

De term $e_{ki}$ is de *error* (ruis, foutenscore) van de i-de persoon in de k-de groep. De $e_{ki}$'s zijn verantwoordelijk voor het feit dat de scores bínnen een groep van elkaar verschillen. De grootte van de $e_{ki}$'s wordt onder andere bepaald door interindividuele verschillen en meetfouten. Het klinkt misschien vreemd dat interindividuele verschillen beschouwd worden als ruis, maar binnen de variantieanalyse is dat zo. Hier is men geïnteresseerd in de verschillen tussen de groepen en alle andere verschillen beschouwt men als ruis.

Het variantieanalysemodel is nuttig omdat we het kunnen gebruiken om precies díe informatie uit de observaties te halen die relevant is voor de nulhypothese. Zoals we in de volgende paragraaf zullen zien, wordt deze informatie samengevat in de effectkwadratensom. Bij de berekening van de effectkwadratensom worden de voorspellingen op basis van het variantieanalysemodel vergeleken met de voorspellingen op basis van een ander model dat veronderstelt dat de nulhypothese waar is (i.e., geen verschil tussen de $K$ verwachte waarden $\mu_k$). Als de $K$ $\mu_k$'s gelijk zijn, dan kunnen we ze vervangen door $\mu$, en dan krijgen we het volgende model:

$$\boxed{y_{ki} = \mu + e_{ki} \quad .}$$

**Parameters**
De term "parameter" wordt, afhankelijk van de context, in meerdere betekenissen gebruikt. In dit boek wordt het in de volgende betekenis gebruikt: een parameter is een onbekend getal dat op een systematische wijze de data beïnvloedt. Bijvoorbeeld, de verwachte waarde in de k-de groep, aangeduid met $\mu_k$, is een parameter. Deze verwachte waarde is onbekend en hij bepaalt of men gemiddeld genomen grote of kleine getallen zal observeren in de k-de groep; naarmate $\mu_k$ groter is, zal men gemiddeld genomen grotere getallen observeren. De effectparameter $\alpha_k$ bepaalt, net zoals $\mu_k$, ook of men gemiddeld genomen grote of kleine getallen zal observeren. Anders dan $\mu_k$, geeft $\alpha_k$ aan of de getallen in de k-de groep gemiddeld genomen boven of onder het gemiddelde van alle groepen samen ligt; als $\alpha_k$ positief is, dan liggen de getallen in de k-de groep gemiddeld genomen boven het gemiddelde van alle groepen samen en als $\alpha_k$ negatief is, dan geldt het omgekeerde.
Het is meestal mogelijk om, op basis van de data, een schatting te maken van de onbekende parameters. In wat volgt, zullen we laten zien dat de verwachte waarde $\mu_k$ geschat kan worden met het groepsgemiddelde $\overline{y}_k$. De effectparameter $\alpha_k$ kan geschat worden met het verschil $\overline{y}_k - \overline{y}_{\sqcup}$, het groepsgemiddelde min het gemiddelde van alle observaties samen.

Dit model noemen we ook het *referentiemodel* omdat het de situatie voorstelt waarin de nulhypothese waar is. Het variantieanalysemodel ($y_{ki} = \mu_k + e_{ki}$) stelt de situatie voor waarin de nulhypothese zowel waar als onwaar kan zijn, want de verwachte waarden $\mu_k$ kunnen in principe alle mogelijke waarden hebben.

In de volgende paragraaf, *De Effectkwadratensom*, wordt uitgelegd hoe we het variantieanalyse- en het referentiemodel kunnen gebruiken om precies díe informatie uit de observaties te halen die relevant is voor de nulhypothese. Voor we hiermee beginnen, laten we eerst nog een tweede formulering van het variantieanalysemodel zien. In deze tweede formulering maken we gebruik van zogenaamde *effectparameters*. Het nut van de effectparameterformulering van het variantieanalysemodel is dat het ons in staat stelt om, in een variantieanalyse met twee of meer factoren, het interactie-effect statistisch te toetsen. Voor de variantieanalyse met slechts één factor hebben we de effectparameterformulering niet nodig. Hij wordt hier alleen voorgesteld omdat hij in de volgende paragraaf (*Variantieanalyse met Twee Factoren*) nuttig blijkt te zijn.

Het uitgangspunt van de effectparameterformulering is dat we $\mu_k$ schrijven als de som van een zogenaamde *globale* verwachte waarde $\mu$ en een effectparameter $\alpha_k$. Deze globale verwachte waarde $\mu$ is het gemiddelde van de $K$ groepsspecifieke verwachte waarden $\mu_k$:

$$\mu = \frac{\sum_{k=1}^{K} \mu_k}{K} \quad .$$

En de effectparameter $\alpha_k$ wordt als volgt bepaald:

$$\boxed{\alpha_k = \mu_k - \mu \quad .}$$

De effectparameter is positief als de verwachte waarde in de k-de groep groter is dan het gemiddelde van de verwachte waarden van de $K$ groepen samen. Ervan uitgaande dat de verschillen tussen de gemiddelden in Tabel 3.1 niet toevallig zijn, zal de effectparameter $\alpha_k$ positief zijn voor de groep met gratis taxi en negatief voor de groep zonder gratis vervoer.

**Somtekens**
In de formule van de globale verwachte waarde $\mu$ staat de volgende uitdrukking:

$$\sum_{k=1}^{K} \mu_k \quad .$$

Het symbool $\sum$ wordt een *somteken* genoemd, en het wordt gebruikt om een som met veel termen op een korte manier op te schrijven. Concreet,

$$\sum_{k=1}^{K} \mu_k = \mu_1 + \mu_2 + \ldots + \mu_K \quad ,$$

een som van $K$ termen. De uitdrukking $\sum_{k=1}^{K} \mu_k$ wordt gelezen als "de som van de $\mu_k$'s van $k = 1$ tot $K$".

Uit de definitie van de effectparameter $\alpha_k$ volgt dat $\mu_k$ gelijk is aan $\mu+\alpha_k$. Door $\mu_k$ in het variantieanalysemodel te vervangen door $\mu+\alpha_k$ krijgen we het variantieanalysemodel in termen van effectparameters:

$$\boxed{y_{ki} = \mu + \alpha_k + e_{ki} \quad .}$$

Het feit dat $\mu$ gelijk is aan het gemiddelde van de $\mu_k$'s zorgt ervoor dat er een restrictie ligt op de effectparameters $\alpha_k$, namelijk dat de som van deze effectparameters gelijk moet zijn aan 0. Dit kan men eenvoudig aantonen door de som te nemen over $k$ van het rechterlid van de definitie van de effectparameter $\alpha_k$ (zie pagina 146):

$$\begin{aligned} \sum_{k=1}^{K} \alpha_k &= \sum_{k=1}^{K} \mu_k - \sum_{k=1}^{K} \mu \\ &= K\mu - K\mu \\ &= 0 \quad . \end{aligned}$$

(In de stap van de eerste naar de tweede regel maken we gebruik van het feit dat de globale verwachte waarde $\mu$ gelijk is aan het gemiddelde van de $K$ groepsspecifieke verwachte waarden $\mu_k$.) We kunnen dit illustreren met een getallenvoorbeeld. Stel dat de $\mu_k$'s in een onderzoek met vier groepen de volgende zijn: 12, 6, 14 en 8. Dan is de globale verwachte waarde $\mu$ gelijk aan $(12+6+14+8)/4 = 10$. De effectparameters $\alpha_k$ zijn dan de volgende:

$$\begin{aligned} \alpha_1 &= \mu_1 - \mu = 12 - 10 = 2 \qquad & \alpha_2 &= \mu_2 - \mu = 6 - 10 = -4 \\ \alpha_3 &= \mu_3 - \mu = 14 - 10 = 4 \qquad & \alpha_4 &= \mu_4 - \mu = 8 - 10 = -2 \quad . \end{aligned}$$

Merk op dat de som van de vier effectparameters gelijk is aan 0. Het feit dat de som van de $K$ effectparameters $\alpha_k$ gelijk is aan 0 betekent dat er slechts $K-1$ *vrije* $\alpha_k$'s zijn: als $K-1$ $\alpha_k$'s bekend zijn, dan is ook de K-de $\alpha_k$ bekend, want deze is gelijk aan *min* de som van de eerste $K-1$ $\alpha_k$'s.

De term "effect" in "effectparameter" verwijst niet naar het *causale* effect van de factor. In de terminologie van de variantieanalyse verstaat men onder "effect" een verschil tussen verwachte waarden, bijvoorbeeld, tussen $\mu_1$ en $\mu_2$ of tussen $\mu_1$ en de globale verwachte waarde $\mu$. Het is goed mogelijk dat er een verschil is tussen de verwachte waarden in de verschillende niveaus van een factor maar dat dit verschil niet het gevolg is van een causaal effect van deze factor. Bijvoorbeeld, indien de bejaarden uit het voorbeeldonderzoek zelf hun conditie mochten kiezen, dan is het aannemelijk dat vooral de bejaarden met een grote behoefte aan sociaal contact in de gratis-taxi- en de vervoer-op-bestellingconditie terecht zullen komen. Een verschil tussen de verwachte waarden in deze twee condities en de verwachte waarde in de conditie zonder gratis vervoer kan dan het resultaat zijn van de invloed van de storende variabele BEHOEFTE AAN SOCIAAL CONTACT en niet van een causaal effect van de factor MOBILITEIT.

### 3.2.6 De Effectkwadratensom

We leggen nu uit hoe we het variantieanalysemodel kunnen gebruiken om precies díe informatie uit de observaties te halen die relevant is voor de nulhypothese. Deze informatie wordt samengevat in de effectkwadratensom. De basisgedachte bij de berekening van de effectkwadratensom is dat we de voorspellingen op basis van het variantieanalysemodel vergelijken met de voorspellingen op basis van het referentiemodel dat veronderstelt dat de nulhypothese waar is. Voor het gemak, schrijven we hier beide modellen onder elkaar:

$$\boxed{\begin{array}{rll} \text{Variantieanalysemodel:} & y_{ki} & = \mu_k + e_{ki} \\ \text{Referentiemodel:} & y_{ki} & = \mu + e_{ki} \quad . \end{array}}$$

Om het effect van de factor A te evalueren maken we twee voorspellingen van de observaties, één op basis van het variantieanalysemodel en één op basis van het referentiemodel, en vergelijken die met elkaar. Een voorspelling van de observatie $y_{ki}$ duiden we aan met $\widehat{y}_{ki}$ ("$y_{ki}$ *dakje*"). Als deze voorspelling gemaakt wordt op basis van het variantieanalysemodel dan wordt zij aangeduid met $\widehat{y}^A_{ki}$, waarin de letter $A$ aangeeft dat het variantieanalysemodel het effect van de factor A specificeert (d.m.v. de effectparameters $\alpha_k$). En als deze voorspelling gemaakt wordt op basis van het referentiemodel, dan wordt hij aangeduid met $\widehat{y}^{niks}_{ki}$, waarin *niks* staat voor "geen effect". De effectkwadratensom wordt nu als volgt berekend:

$$\boxed{SS_A = \sum_{Obs} (\widehat{y}^A_{ki} - \widehat{y}^{niks}_{ki})^2 \quad ,}$$

de som van de gekwadrateerde verschillen tussen de voorspellingen op basis van het model met factor A en het model zonder factor A, en dit over alle observaties. Deze som over alle observaties kunnen we wat preciezer uitdrukken door expliciet te verwijzen naar de rijen (met index $i$) en de kolommen (met index $k$) van Tabel 3.2:

$$SS_A = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (\widehat{y}^A_{ki} - \widehat{y}^{niks}_{ki})^2 \quad .$$

**Dubbele Somtekens**
In de formule van de effectkwadratensom $SS_A$ gebruiken we het dubbele somteken

$$\sum_{k=1}^{K}\sum_{i=1}^{n_k} \quad .$$

Voor een goed begrip van dit dubbele somteken kijkt men het beste naar Tabel 3.6. Het eerste somteken (met index $k = 1, \ldots, K$) correspondeert met de kolommen van deze tabel en het tweede somteken (met index $i = 1, \ldots, n_k$) met de rijen. Omdat het aantal observaties per groep verschillend is, staat er een verschillend aantal getallen in elk van de kolommen van Tabel 3.2. En daarom loopt de index van het tweede somteken van 1 tot $n_k$, een getal dat afhangt van k, de kolomindex. (In Tabel 3.6 bevat groep 2 de meeste proefpersonen, maar dit is willekeurig; in een ander onderzoek kan dat best een andere groep zijn.)
Het dubbele somteken is niets anders dan een precieze manier om uit te drukken dat we de som nemen over alle observaties. Het dubbele somteken kan daarom als volgt verkort worden:

$$\sum_{Obs} \quad ,$$

waarbij *Obs* aangeeft dat we de som nemen over alle observaties. Soms is deze verkorte notatie handiger dan het dubbele somteken, maar soms ook niet.
In paragraaf 3.3 hebben we een drievoudig somteken nodig om over alle observaties te sommeren. Dit drievoudige somteken zullen we op dezelfde manier afkorten.

Om een concrete voorstelling te krijgen van de som in deze formule, is het goed om de termen $(\widehat{y}_{ki}^{A} - \widehat{y}_{ki}^{niks})^2$ in een tabel te zetten. Dit gebeurt in Tabel 3.6, die er precies uitziet als Tabel 3.2 maar met andere getallen in de cellen. Men krijgt de som in de bovenstaande formule door alle getallen in de cellen van deze tabel bij elkaar op te tellen.

Om de effectkwadratensom $SS_A$ te berekenen moeten we de voorspellingen $\widehat{y}_{ki}^{A}$ en $\widehat{y}_{ki}^{niks}$ kennen. Deze voorspellingen worden bepaald op basis van het *kleinste-kwadratencriterium*. Het kleinste-kwadratencriterium houdt in dat de voorspellingen zo bepaald worden dat de gekwadrateerde afwijkingen tussen de observaties $y_{ki}$ en hun voorspellingen ($\widehat{y}_{ki}^{A}$ resp. $\widehat{y}_{ki}^{niks}$) zo klein mogelijk zijn. Hoe dat precies in zijn werk gaat wordt uitgelegd in de Appendix op pagina 290 (voor $\widehat{y}_{ki}^{A}$; de uitleg voor voor $\widehat{y}_{ki}^{niks}$ is analoog). Op basis van het kleinste-kwadratencriterium is de voorspelling $\widehat{y}_{ki}^{A}$ gelijk aan $\overline{y}_k$, het gemiddelde in de k-de groep, en is de voorspelling $\widehat{y}_{ki}^{niks}$ gelijk aan $\overline{y}_{\sqcup}$, het gemiddelde van alle $N$ observaties:

$$\overline{y}_{\sqcup} = \frac{\sum_{k=1}^{K}\sum_{i=1}^{n_k} y_{ki}}{N} \quad .$$

Hier en in de rest van dit hoofdstuk, wordt $\overline{y}_{\sqcup}$ ("$y$ gemiddeld bakje") een *gewogen gemiddelde* genoemd.

We kunnen nu een nieuwe formule geven van $SS_A$ door $\widehat{y}_{ki}^{A}$ en $\widehat{y}_{ki}^{niks}$ in de eerder gegeven formule te vervangen door $\overline{y}_k$ en $\overline{y}_{\sqcup}$. Hieruit volgt:

$$SS_A = \sum_{k=1}^{K}\sum_{i=1}^{n_k}(\overline{y}_k - \overline{y}_{\sqcup})^2 \quad .$$

| Persoon | Groep 1 | Groep 2 | $\cdots$ | Groep $K$ |
|---|---|---|---|---|
| 1 | $(\widehat{y}^A_{11} - \widehat{y}^{niks}_{11})^2$ | $(\widehat{y}^A_{21} - \widehat{y}^{niks}_{21})^2$ | $\cdots$ | $(\widehat{y}^A_{K1} - \widehat{y}^{niks}_{K1})^2$ |
| 2 | $(\widehat{y}^A_{12} - \widehat{y}^{niks}_{12})^2$ | $(\widehat{y}^A_{22} - \widehat{y}^{niks}_{22})^2$ | $\cdots$ | $(\widehat{y}^A_{K2} - \widehat{y}^{niks}_{K2})^2$ |
| 3 | $(\widehat{y}^A_{13} - \widehat{y}^{niks}_{13})^2$ | $(\widehat{y}^A_{23} - \widehat{y}^{niks}_{23})^2$ | $\cdots$ | $(\widehat{y}^A_{K3} - \widehat{y}^{niks}_{K3})^2$ |
| 4 | $(\widehat{y}^A_{14} - \widehat{y}^{niks}_{14})^2$ | $(\widehat{y}^A_{24} - \widehat{y}^{niks}_{24})^2$ | $\cdots$ | $(\widehat{y}^A_{K4} - \widehat{y}^{niks}_{K4})^2$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $n_1$ | $(\widehat{y}^A_{1n_1} - \widehat{y}^{niks}_{1n_1})^2$ | | | |
| $\vdots$ | | $\vdots$ | $\vdots$ | $\vdots$ |
| $n_K$ | | | | $(\widehat{y}^A_{Kn_K} - \widehat{y}^{niks}_{Kn_K})^2$ |
| $\vdots$ | | $\vdots$ | $\vdots$ | |
| $n_2$ | | $(\widehat{y}^A_{2n_2} - \widehat{y}^{niks}_{2n_2})^2$ | | |

Tabel 3.6: Tabel met de gekwadrateerde verschillen tussen de voorspellingen op basis van het model met en het model zonder het effect van factor A.

**Het Gewogen Gemiddelde**
Het gemiddelde in de formule van $\overline{y}_{\sqcup}$ is een *gewogen* gemiddelde omdat het rechterlid van deze formule als volgt herschreven kan worden:

$$\begin{aligned}\overline{y}_{\sqcup} &= \frac{\sum_{k=1}^{K}\left[\sum_{i=1}^{n_k} y_{ki}\right]}{N} \\ &= \frac{\sum_{k=1}^{K} n_k \overline{y}_k}{N} \quad ,\end{aligned}$$

hetgeen volgt uit de definitie van $\overline{y}_k$:

$$\overline{y}_k = \frac{\sum_{i=1}^{n_k} y_{ki}}{n_k} \quad ,$$

door links en rechts van het gelijkheidsteken te vermenigvuldigen met $n_k$. Het rechterlid van de bovenstaande formule voor $\overline{y}_{\sqcup}$ is een gewogen gemiddelde omdat de $\overline{y}_k$'s gewogen worden met de aantallen observaties in de groepen (de $n_k$'s).

Deze som kan vereenvoudigd worden. Dat ziet men onmiddellijk als de termen $(\widehat{y}_{ki}^{A} - \widehat{y}_{ki}^{niks})^2$ in Tabel 3.6 vervangen worden door $(\overline{y}_k - \overline{y}_{\sqcup})^2$. Dit gebeurt in Tabel 3.7. Men ziet onmiddellijk dat elke kolom hetzelfde getal bevat ($(\overline{y}_1 - \overline{y}_{\sqcup})^2$ in de eerste kolom, $(\overline{y}_2 - \overline{y}_{\sqcup})^2$ in de tweede, ...). Met andere woorden, $(\overline{y}_k - \overline{y}_{\sqcup})^2$ hangt niet af van $i$. Daarom kan $SS_A$ als volgt geschreven worden:

$$\boxed{SS_A = \sum_{k=1}^{K} n_k(\overline{y}_k - \overline{y}_{\sqcup})^2 \quad .}$$

Uit deze formule blijkt dat $SS_A$ groter wordt naarmate de groepsgemiddelden $\overline{y}_k$ meer van elkaar verschillen. De kleinst mogelijke effectkwadratensom is 0 en hij wordt verkregen als alle groepsgemiddelden $\overline{y}_k$ gelijk zijn aan elkaar. In dat geval is het gewogen gemiddelde $\overline{y}_{\sqcup}$ immers ook gelijk aan elk van deze groepsgemiddelden, en zijn de verschillen $(\overline{y}_k - \overline{y}_{\sqcup})$ gelijk aan 0. Naarmate de groepsgemiddelden meer van elkaar verschillen, verschillen zij ook meer van het globale gemiddelde en daardoor wijken de verschillen $(\overline{y}_k - \overline{y}_{\sqcup})$ meer af van 0.

| Persoon | Groep 1 | Groep 2 | $\cdots$ | Groep $K$ |
|---|---|---|---|---|
| 1 | $(\overline{y}_1 - \overline{y}_{\sqcup})^2$ | $(\overline{y}_2 - \overline{y}_{\sqcup})^2$ | $\cdots$ | $(\overline{y}_K - \overline{y}_{\sqcup})^2$ |
| 2 | $(\overline{y}_1 - \overline{y}_{\sqcup})^2$ | $(\overline{y}_2 - \overline{y}_{\sqcup})^2$ | $\cdots$ | $(\overline{y}_K - \overline{y}_{\sqcup})^2$ |
| 3 | $(\overline{y}_1 - \overline{y}_{\sqcup})^2$ | $(\overline{y}_2 - \overline{y}_{\sqcup})^2$ | $\cdots$ | $(\overline{y}_K - \overline{y}_{\sqcup})^2$ |
| 4 | $(\overline{y}_1 - \overline{y}_{\sqcup})^2$ | $(\overline{y}_2 - \overline{y}_{\sqcup})^2$ | $\cdots$ | $(\overline{y}_K - \overline{y}_{\sqcup})^2$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $n_1$ | $(\overline{y}_1 - \overline{y}_{\sqcup})^2$ | | | |
| $\vdots$ | | $\vdots$ | $\vdots$ | $\vdots$ |
| $n_K$ | | | | $(\overline{y}_K - \overline{y}_{\sqcup})^2$ |
| $\vdots$ | | $\vdots$ | $\vdots$ | |
| $n_2$ | | $(\overline{y}_2 - \overline{y}_{\sqcup})^2$ | | |
| Som: | $n_1(\overline{y}_1 - \overline{y}_{\sqcup})^2$ | $n_2(\overline{y}_2 - \overline{y}_{\sqcup})^2$ | $\cdots$ | $n_K(\overline{y}_K - \overline{y}_{\sqcup})^2$ |

Tabel 3.7: Tabel met de gekwadrateerde verschillen tussen de voorspellingen op basis van het model met het effect van factor A ($\overline{y}_k$) en de voorspellingen op basis van het model zonder effect ($\overline{y}_{\sqcup}$).

In het voorbeeldonderzoek zijn de steekproefgemiddelden $\overline{y}_1$, $\overline{y}_2$, en $\overline{y}_3$ gelijk aan, respectievelijk, 9.9, 6.9, en 6.4. En het gewogen gemiddelde $\overline{y}_{\sqcup}$ is gelijk aan 7.6. De effectkwadratensom $SS_A$ kan dan als volgt berekend worden:

$$\begin{aligned} SS_A &= 8 \times (9.9 - 7.6)^2 + 9 \times (6.9 - 7.6)^2 + 10 \times (6.4 - 7.6)^2 \\ &= 61.13 \quad . \end{aligned}$$

### 3.2.7 De Errorkwadratensom, Totale Kwadratensom, en de Kwadratensomontbinding

#### De Errorkwadratensom

De errorkwadratensom $SS_{Error|A}$ is een maat voor de fouten die we maken als we de observaties voorspellen op basis van het variantieanalysemodel. Bij de berekening van deze errorkwadratensom vergelijken we de observaties $y_{ki}$ met de voorspellingen $\widehat{y}_{ki}^A$:

$$\boxed{SS_{Error|A} = \sum_{Obs} (y_{ki} - \widehat{y}_{ki}^A)^2 \quad ,}$$

de som over alle observaties van de gekwadrateerde verschillen tussen de observaties en de voorspellingen op basis van het variantieanalysemodel met factor A. Deze formule kan wat preciezer geschreven worden door de som te nemen over de indices $k$ en $i$:

$$SS_{Error|A} = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (y_{ki} - \widehat{y}_{ki}^A)^2 \quad .$$

De voorspellingen $\widehat{y}_{ki}^A$ zijn gelijk aan de groepsgemiddelden $\overline{y}_k$. We krijgen dan de volgende formule voor $SS_{Error|A}$:

$$\boxed{SS_{Error|A} = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (y_{ki} - \overline{y}_k)^2 \quad .}$$

Om een concrete voorstelling te krijgen van de som in de bovenstaande formule, is het goed om de termen $(y_{ki} - \overline{y}_k)^2$ in een tabel te zetten. Dit gebeurt in Tabel 3.8. In tegenstelling tot Tabel 3.7, zijn alle getallen in Tabel 3.8 in principe verschillend van elkaar. Voor de gegevens uit het voorbeeldonderzoek is $SS_{Error|A}$ gelijk aan 71.42.

#### De Totale Kwadratensom en de Kwadratensomontbinding

De totale kwadratensom, aangeduid met $SS_{Totaal}$, is een speciale errorkwadratensom: het is de errorkwadratensom voor het model zonder het effect van A (i.e., het referentiemodel). Voor elk van de observaties $y_{ki}$ is de voorspelling het globale gemiddelde $\overline{y}_{\sqcup}$.

| Persoon | Groep 1 | Groep 2 | $\cdots$ | Groep $K$ |
|---|---|---|---|---|
| 1 | $(y_{11}-\overline{y}_1)^2$ | $(y_{21}-\overline{y}_2)^2$ | $\cdots$ | $(y_{K1}-\overline{y}_K)^2$ |
| 2 | $(y_{12}-\overline{y}_1)^2$ | $(y_{22}-\overline{y}_2)^2$ | $\cdots$ | $(y_{K2}-\overline{y}_K)^2$ |
| 3 | $(y_{13}-\overline{y}_1)^2$ | $(y_{23}-\overline{y}_2)^2$ | $\cdots$ | $(y_{K3}-\overline{y}_K)^2$ |
| 4 | $(y_{14}-\overline{y}_1)^2$ | $(y_{24}-\overline{y}_2)^2$ | $\cdots$ | $(y_{K4}-\overline{y}_K)^2$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $n_1$ | $(y_{1n_1}-\overline{y}_1)^2$ | | | |
| $\vdots$ | | $\vdots$ | $\vdots$ | $\vdots$ |
| $n_K$ | | | | $(y_{Kn_K}-\overline{y}_K)^2$ |
| $\vdots$ | | $\vdots$ | $\vdots$ | |
| $n_2$ | | $(y_{2n_2}-\overline{y}_2)^2$ | | |
| Som: | $\sum_{i=1}^{n_1}(y_{1i}-\overline{y}_1)^2$ | $\sum_{i=1}^{n_2}(y_{2i}-\overline{y}_2)^2$ | $\cdots$ | $\sum_{i=1}^{n_K}(y_{Ki}-\overline{y}_K)^2$ |

Tabel 3.8: Tabel met de gekwadrateerde verschillen tussen de observaties ($y_{ki}$) en de voorspellingen op basis van het model met het effect van factor A ($\overline{y}_k$).

Daarom wordt $SS_{Totaal}$ als volgt berekend:

$$\boxed{SS_{Totaal} = \sum_{k=1}^{K}\sum_{i=1}^{n_k}(y_{ki}-\overline{y}_{\sqcup})^2 \quad .}$$

De totale kwadratensom is een maat voor de spreiding in de observaties: als de observaties $y_{ki}$ onderling veel van elkaar afwijken dan wijken zij ook veel af van hun gemiddelde $\overline{y}_{\sqcup}$. De totale kwadratensom is de teller van de formule voor de variantie: de variantie is gelijk aan $SS_{Totaal}$ gedeeld door $N$, het aantal observaties. We komen hier nog op terug.

Voor de gegevens uit het voorbeeldonderzoek wordt $SS_{Totaal}$ als volgt berekend:

$$\begin{aligned} SS_{Totaal} &= (11-7.6)^2 + (10-7.6)^2 + \ldots + (11.3-7.6)^2 + \qquad \text{(groep 1)} \\ &\quad (6.2-7.6)^2 + (9-7.6)^2 + \ldots + (6.1-7.6)^2 + \qquad \text{(groep 2)} \\ &\quad (8.4-7.6)^2 + (6-7.6)^2 + \ldots + (7-7.6)^2 \qquad \text{(groep 3)} \\ &= 132.55 \quad . \end{aligned}$$

De totale kwadratensom wordt hier besproken omdat hij gerelateerd is aan de effectkwadratensom $SS_A$ en de errorkwadratensom $SS_{Error|A}$. De totale kwadratensom is

gelijk aan de som van de effect- en de errorkwadratensom:

$$SS_{Totaal} = SS_A + SS_{Error|A} \quad .$$

De opsplitsing van $SS_{Totaal}$ in twee andere kwadratensommen wordt een *kwadratensom-ontbinding* genoemd. We kunnen deze kwadratensom-ontbinding controleren met de gegevens uit het voorbeeldonderzoek: $132.55 = 61.13 + 71.42$. Deze kwadratensom-ontbinding drukt uit dat de spreiding in de observaties, zoals gemeten door $SS_{Totaal}$, uit twee delen bestaat: (1) de spreiding die toe te schrijven is aan factor A, zoals gemeten door de effectkwadratensom $SS_A$, en (2) de spreiding die niet toe te schrijven is aan factor A, zoals gemeten door de errorkwadratensom $SS_{Error|A}$.

De kwadratensommen $SS_{Error|A}$ en $SS_A$ worden in andere boeken vaak met andere symbolen aangeduid. Zo lees je in plaats van $SS_{Error|A}$ ook vaak $SS_{Within}$. Met het subscript *Within* wordt aangeduid dat deze kwadratensom een functie is van de variantie bínnen (in het Engels, *within*) de groepen. Dit zien we in de formule van de errorkwadratensom $SS_{Error|A}$: deze kwadratensom wordt bepaald door de afwijkingen tussen de observaties (de $y_{ki}$'s) en hun bijbehorende groepsgemiddelden (de $\overline{y}_k$'s). En in plaats van $SS_A$ lees je ook vaak $SS_{Between}$ of $SS_{Groepen}$. Met de subscripten *Between* en *Groepen* wordt aangeduid dat deze kwadratensom een functie is van de variantie tússen (in het Engels, *between*) de groepen. Dit zien we in de formule van de effectkwadratensom $SS_A$: deze kwadratensom wordt bepaald door de verschillen tussen de groepsgemiddelden (de $\overline{y}_k$'s) en het gemiddelde van alle observaties samen ($\overline{y}_{\sqcup}$).

De totale kwadratensom $SS_{Totaal}$ wordt in het Engels soms *total sum of squares* genoemd, maar soms ook *corrected total sum of squares*. Het adjectief "corrected" dient om aan te geven dat van de observaties $y_{ki}$ eerst het gemiddelde $\overline{y}_{\sqcup}$ wordt afgetrokken (in het Engels, *corrected for the mean*), en daarna pas het kwadraat wordt genomen. Iemand die de term "corrected total sum of squares" gebruikt, verstaat onder de gewone (uncorrected) total sum of squares de som van de gekwadrateerde observaties $y_{ki}$. In dit boek verstaan we onder "totale kwadratensom" altijd de corrected variant.

### 3.2.8 Berekening en Interpretatie van de F-Statistiek

De F-statistiek wordt als volgt berekend:

$$\boxed{F = \frac{MS_A}{MS_{Error|A}} \quad .}$$

De gemiddelde kwadratensommen in teller en noemer van deze formule worden als volgt berekend:

$$\boxed{\begin{aligned} MS_A &= \frac{SS_A}{K-1} \\ MS_{Error|A} &= \frac{SS_{Error|A}}{N-K} \quad . \end{aligned}}$$

In de noemers van deze gemiddelde kwadratensommen staan de corresponderende vrijheidsgraden. Deze vrijheidsgraden hangen af van (1) het aantal vrije parameters in het

model met en het model zonder effect van A en (2) het aantal observaties. Concreet, het aantal vrijheidsgraden van $SS_A$ wordt bepaald op basis van de volgende regel:

$$\boxed{\begin{array}{c}\text{aantal vrije parameters in}\\ \text{het model met A}\end{array}} \quad - \quad \boxed{\begin{array}{c}\text{aantal vrije parameters in}\\ \text{het model zonder A}\end{array}} \quad .$$

Het aantal vrije[6] parameters in het model met A, is $K$, want er zijn $K$ vrije $\mu_k$'s. Het model zonder A, heeft slechts één parameter: $\mu$. Het aantal vrijheidsgraden van $SS_A$ is dus $K-1$, het aantal groepen min 1. Het aantal vrijheidsgraden van $SS_{Error|A}$ wordt bepaald op basis van de volgende regel:

$$\boxed{\text{aantal observaties}} \quad - \quad \boxed{\begin{array}{c}\text{aantal vrije parameters in}\\ \text{het model met A}\end{array}} \quad .$$

Het aantal observaties is $N$ en het aantal vrije parameters in het model met A is $K$. Het aantal vrijheidsgraden van $SS_{Error|A}$ is dus $N-K$.

Onder de nulhypothese dat alle $\mu_k$'s gelijk zijn ($\mu_1 = \mu_2 = \cdots = \mu_K$) heeft de F-statistiek een $\mathcal{F}$-verdeling met $K-1$ tellervrijheidsgraden en $N-K$ noemervrijheidsgraden. Strikt genomen geldt dit alleen als ook voldaan is aan de meer technische veronderstellingen dat de $e_{ki}$'s onafhankelijk normaal verdeeld zijn met dezelfde variantie in elk van de groepen. Maar ook als er niet voldaan is aan deze meer technische veronderstellingen, dan nog is de $\mathcal{F}$-verdeling in de meeste praktijksituaties een zeer goede benadering.

De *F-toets* houdt in dat de kans berekend wordt onder deze $\mathcal{F}$-verdeling om een F-statistiek te observeren die groter of gelijk is aan de geobserveerde F-statistiek[7]. Deze kans wordt de *p-waarde* genoemd. Als deze kans kleiner is dan 0.05 of 0.01 dan vinden we deze gebeurtenis zo zeldzaam dat het niet aannemelijk is dat de nulhypothese waar is. De nulhypothese ($\mu_1 = \mu_2 = \cdots = \mu_K$) wordt dan verworpen.

Hoe groter de F-statistiek, hoe kleiner de p-waarde. De F-statistiek, op haar beurt, is groter naarmate $SS_A$ groter is. En $SS_A$, tenslotte, is groter naarmate de groepsgemiddelden (de $\overline{y}_k$'s) meer van elkaar verschillen. Dus, de nulhypothese zal eerder verworpen worden naarmate de groepsgemiddelden meer van elkaar verschillen.

Een handige interpretatie van de F-statistiek is de volgende: de teller van de F-statistiek geeft aan wat men toetst, en de noemer geeft aan hoe gevoelig men toetst. De teller van de F-statistiek ($MS_A$) is een functie van de verschillen tussen de $\overline{y}_k$'s. Dus, met de F-statistiek toetst men of deze gemiddelden significant van elkaar verschillen. De noemer van de F-statistiek ($MS_{Error|A}$) hangt af van de afwijkingen tussen de individuele

[6] We spreken hier over *vrije* parameters en niet over parameters zondermeer, omdat de effectparameterformulering van dit model $K+1$ parameters bevat ($K$ effectparameters $\alpha_k$ en één globale verwachte waarde $\mu$), waarvan er slechts $K$ vrij zijn. Op de effectparameters $\alpha_k$ ligt immers de restrictie dat hun som gelijk is aan 0. Dit impliceert dat slechts $K-1$ effectparameters vrij zijn: als $K-1$ effectparameters gekend zijn, dan is ook de resterende $K$-de effectparameter gekend, want de $K$ effectparameters tellen op tot 0.

[7] Een alternatieve procedure is dat men de geobserveerde F-statistiek vergelijkt met een kritieke waarde onder deze $\mathcal{F}$-verdeling (meestal, het 95-ste quantiel). Deze alternatieve procedure leidt tot dezelfde beslissing (i.c., het verwerpen of aanhouden van de nulhypothese) als de eerste. In dit boek concentreren wij ons op de eerste procedure.

observaties en de gemiddelden. Naarmate deze afwijkingen kleiner zijn, is ook $MS_{Error|A}$ kleiner, en is een kleinere $MS_A$ voldoende om een significante F-statistiek te krijgen. Met andere woorden, de toets wordt gevoeliger.

De gegevens die nodig zijn om een F-toets uit te voeren worden meestal gegeven in een *ANOVA-tabel*. In Tabel 3.9 wordt de ANOVA-tabel gegeven voor het voorbeeldonderzoek. De layout van deze tabel is dezelfde als die van de ANOVA-tabellen in de uitvoer van het computerprogramma SPSS. Het aantal vrijheidsgraden wordt in deze tabel aangeduid met *df* (in het Engels, *degrees of freedom*). Men vindt de p-waarde in de kolom *Sig.* (van het Engelse *significance*). Omdat de p-waarde kleiner is dan 0.05 verwerpen we de nulhypothese en besluiten dat er een significant verschil is tussen de groepen.

Dependent Variable: SOCIALE ACTIVITEIT

| Source | Sum of Squares | df | Mean Squares | F | Sig. |
|---|---|---|---|---|---|
| MOBILITEIT | 61.13 | 2 | 30.56 | 10.26 | 0.0006 |
| Error | 71.42 | 24 | 2.98 | | |
| Corrected Total | 132.55 | 26 | | | |

Tabel 3.9: ANOVA-tabel voor de data uit het onderzoek naar het effect van MOBILITEIT op SOCIALE ACTIVITEIT bij bejaarden.

Men mag het belang van de p-waarde van een F-toets niet overschatten. Soms is deze p-waarde zeer klein ("zeer significant", wordt wel eens gezegd) ondanks het feit dat de verschillen tussen de $\overline{y}_k$'s niet bijzonder groot zijn. Deze situatie kan zich voordoen als het aantal observaties ($N$) erg groot is. De verklaring voor dit fenomeen is dat de teller van de F-statistiek ($MS_A$) meestal sterk toeneemt met het aantal observaties en de noemer ($MS_{Error|A}$) helemaal niet. Dit kunnen we illustreren met een getallenvoorbeeld. Stel dat we een onderzoek deden met 3 groepen ($K = 3$) en in totaal 20 proefpersonen ($N = 20$). De effectkwadratensom $SS_A$ is 30 en de errorkwadratensom $SS_{Error|A}$ is 85. De gemiddelde kwadratensommen, $MS_A$ en $MS_{Error|A}$, zijn dan gelijk aan, respectievelijk, 15 en 5, en de F-statistiek is gelijk aan 3. De bijbehorende p-waarde is 0.077. Stel nu dat we het aantal proefpersonen vervijfvoudigen. $SS_A$ en $MS_A$ zullen dan ongeveer vijf keer zo groot worden. Dit komt omdat $\overline{y}_{\sqcup}$ en de $\overline{y}_k$'s in de formule van de effectkwadratensom vrij weinig zullen veranderen en alleen de aantallen observaties in de groepen ($n_k$) vijf keer zo groot worden. Ook $SS_{Error|A}$ zal ongeveer vijf keer zo groot worden, maar $MS_{Error|A}$ zal vrij weinig veranderen. Dit laatste komt omdat het aantal vrijheidsgraden van $MS_{Error|A}$ ($N - K$) ook bijna vijf keer zo groot wordt. We krijgen dan: $SS_A = 150$, $K = 3$, $MS_A = 75$, $SS_{Error|A} = 425$, $N = 100$, $N - K = 97$, en $MS_{Error|A} = 4.38$. De F-statistiek is dan gelijk aan 17.12 en de bijbehorende p-waarde is 0.00000043, hetgeen veel kleiner ("significanter") is dan bij 20 proefpersonen.

### 3.2.9 De Proportie Verklaarde Variantie

Het is een goede strategie om bij de data-analyse niet alleen te kijken naar de p-waarde van de F-toets. Behalve in de vraag of de gemiddelden significant van elkaar verschillen,

is men meestal ook geïnteresseerd in de vraag hoe belangrijk de factor is voor de verklaring van de variantie in de afhankelijke variabele. De *proportie verklaarde variantie* (PVV) is een maat waarmee we deze belangrijkheid kunnen uitdrukken.

De *te verklaren variantie* is de variantie van de afhankelijke variabele over alle groepen. Een variantie wordt berekend door het verschil te nemen tussen de observaties (hier, de $y_{ki}$'s) en hun gemiddelde (hier, $\overline{y}_{\sqcup}$), deze verschillen te kwadrateren, op te tellen, en ze te delen door het aantal observaties (hier, $N$):

$$\begin{aligned}\text{Te Verklaren Variantie} &= \frac{1}{N}\sum_{Obs}(y_{ki} - \overline{y}_{\sqcup})^2 \\ &= \frac{1}{N} SS_{Totaal} \quad .\end{aligned}$$

Omdat $SS_{Totaal}$ gelijk is aan $(SS_A + SS_{Error|A})$, kan de te verklaren variantie als volgt geschreven worden:

$$\begin{aligned}\text{Te Verklaren Variantie} &= \frac{1}{N}(SS_A + SS_{Error|A}) \\ &= \frac{1}{N} SS_A + \frac{1}{N} SS_{Error|A} \quad .\end{aligned}$$

In de deze formule wordt de te verklaren variantie opgesplitst in de *verklaarde variantie* ($\frac{1}{N}SS_A$) en de *onverklaarde variantie* ($\frac{1}{N}SS_{Error|A}$). De verklaarde variantie geeft aan welk deel van de te verklaren variantie toe te schrijven is aan de verschillen tussen de groepsgemiddelden. En de onverklaarde variantie geeft aan welk deel van de te verklaren variantie nog overblijft als men de observaties voorspelt met de groepsgemiddelden.

De PVV is de verhouding van de verklaarde variantie ten opzichte van de te verklaren variantie:

$$\text{PVV} = \frac{\frac{1}{N}SS_A}{\frac{1}{N}SS_{Totaal}} \quad .$$

Hieruit volgt:

$$\boxed{\text{PVV} = \frac{SS_A}{SS_{Totaal}} \quad .}$$

Voor het voorbeeldonderzoek is de PVV gelijk aan 61.13/132.55 = 0.46. Dus, 46 procent van de variantie in SOCIALE ACTIVITEIT wordt verklaard door MOBILITEIT.

De PVV is een uitstekende maat voor de belangrijkheid van een factor als voorspeller/verklaarder van de variantie in de afhankelijke variabele. Toch moeten we ook het belang van de PVV relativeren. Een nadeel van de PVV is dat hij afhangt van de keuze van de niveaus van de factor die men in het onderzoek opneemt. Vaak is deze keuze arbitrair. Bijvoorbeeld, in het onderzoek naar de sociale activiteit van bejaarden had men de groep met gratis taxi kunnen laten wegvallen. Dan zou de PVV gelijk zijn aan 0.02 in plaats van 0.46. Dit komt omdat het gemiddelde van de groep met gratis taxi (9.9) meer verschilt van de gemiddelden van de twee andere groepen (6.9 en 6.4)

dan het verschil tussen deze laatste twee onderling. Toch spreekt men in beide gevallen van dé PVV door de variabele MOBILITEIT. In het algemeen, zal de PVV groter zijn naarmate de niveaus van de factor sterker van elkaar verschillen.

Soms is het ook zo dat men alleen geïnteresseerd is in het resultaat van de F-toets en niet in de grootte van de PVV. Dit is het geval als de onderzoeker alleen geïnteresseerd is in het feit óf een bepaald effect zich voordoet en niet in de grootte ervan. Bijvoorbeeld, stel dat een geheugenpsycholoog wil weten of het gelijktijdig visueel en auditief aanbieden van een zin ervoor zorgt dat deze zin beter onthouden wordt dan wanneer die zin alleen visueel aangeboden wordt. Als er een statistisch significant verschil bestaat tussen de gecombineerd visueel-auditieve conditie en de zuiver visuele, hoe klein ook, dan zegt dit iets over de cognitieve verwerking van zinnen. Als het doel van het onderzoek er alleen in bestaat om een hypothese te toetsen over het hoe van deze cognitieve verwerking, dan doet de grootte van het effect er verder niet meer toe.

Meestal is de beste strategie bij de data-analyse om zowel rekening te houden met de PVV als met de p-waarde van de F-toets. De PVV geeft aan hoe groot de bijdrage is van de factor aan de verklaring van de variantie, en de p-waarde zegt ons of deze bijdrage statistisch significant is.

---

*Opgave 4*
In Tabel 3.10 staan de gegevens van een onderzoek met drie groepen.

1. Welke voorspelling zou je maken voor de score van een persoon in Groep 1, ervan uitgaande dat de groepen van elkaar verschillen?
2. Welke voorspelling zou je maken voor de score van een persoon in Groep 1, ervan uitgaande dat de groepen niet van elkaar verschillen?
3. Bereken de F-statistiek waarmee je kunt toetsen of de groepen significant van elkaar verschillen.

| Groep 1 | | Groep 2 | | Groep 3 | |
|---|---|---|---|---|---|
| Proef-persoon | Score | Proef-persoon | Score | Proef-persoon | Score |
| 1 | 15 | 1 | 6 | 1 | 11 |
| 2 | 17 | 2 | 5 | 2 | 10 |
| 3 | 22 | 3 | 9 | 3 | 13 |
| 4 | 18 | 4 | 12 | 4 | 9 |
| 5 | 19 | 5 | 4 | 5 | 6 |
| 6 | 11 | 6 | 9 | 6 | 15 |
| 7 | 18 | 7 | 8 | 7 | 11 |
| 8 | 20 | 8 | 10 | 8 | 9 |
| | | 9 | 6 | 9 | 5 |
| | | 10 | 9 | | |
| Gem.: | 17.5 | Gem.: | 7.8 | Gem.: | 9.88 |

Tabel 3.10: Tabel bij opgave 4.

*Opgave 5*
In een onderzoek met vijf groepen zijn de $\mu_k$'s de volgende: 22, 15, 28, 12 en 23. Wat is de globale verwachte waarde $\mu$ en wat zijn de effectparameters $\alpha_k$?

*Opgave 6*
Een klinisch psycholoog heeft de effectiviteit onderzocht van vier therapieën bij de behandeling van depressie. Twaalf proefpersonen werden at random toegewezen aan de vier therapievormen zodanig dat er drie proefpersonen zaten in elk van de vier condities: psycho-analytische (PA), rationeel-emotieve (RE), ervaringsgerichte (E), en gedragstherapie (G). Na de beëindiging van de therapie werd een depressievragenlijst afgenomen. De scores op deze test staan in Tabel 3.11. Hogere scores op deze test duiden op een ernstiger graad van depressie.

| PA | RE | E | G |
|---|---|---|---|
| 21 | 13 | 15 | 16 |
| 25 | 17 | 19 | 20 |
| 26 | 18 | 20 | 24 |

Tabel 3.11: Tabel bij opgave 6.

1. Is er een significant verschil tussen de gemiddelden van de vier condities? Bereken de kwadratensommen, de F-statistiek en de p-waarde. Deze p-waarde kun je aflezen uit een tabel van de $\mathcal{F}$-verdeling.
2. Wat is de PVV voor deze gegevens?
3. Stel dat de gegevens verdubbeld worden. Dit wil zeggen, in conditie PA komen de observaties 21, 25 en 26 erbij, in conditie RE de observaties 13, 17 en 18, in conditie E de observaties 15, 19 en 20, en in conditie G de observaties 16, 20 en 24. Wat zijn nu de kwadratensommen, de F-statistiek, de p-waarde en de PVV? Wat kun je hieruit besluiten over de invloed van de steekproefomvang op de p-waarde en de PVV?

*Opgave 7*
In Tabel 3.12 staan gegevens uit acht onderzoeken, elk met drie groepen. Het aantal proefpersonen per groep is gelijk. Je kunt dit aantal berekenen door $N$ te delen door 3. In deze tabel staan alle benodigde ingrediënten voor de berekening van de F-statistiek. Bereken deze F-statistieken en geef aan op welke manier zij afhangen van (1) de $\overline{y}_k$'s, (2) $MS_{Error|A}$ en (3) $N$.

| | $\overline{y}_1$ | $\overline{y}_2$ | $\overline{y}_3$ | $MS_{Error\|A}$ | $N$ |
|---|---|---|---|---|---|
| 1 | 8 | 10 | 12 | 10 | 15 |
| 2 | 4 | 10 | 16 | 10 | 15 |
| 3 | 8 | 10 | 12 | 40 | 15 |
| 4 | 4 | 10 | 16 | 40 | 15 |
| 5 | 8 | 10 | 12 | 10 | 60 |
| 6 | 4 | 10 | 16 | 10 | 60 |
| 7 | 8 | 10 | 12 | 40 | 60 |
| 8 | 4 | 10 | 16 | 40 | 60 |

Tabel 3.12: Tabel bij opgave 7.

*Opgave 8*
In Tabel 3.13 staan gegevens uit vier onderzoeken, elk met vijf groepen. De variantie van de afhankelijke variabele ($y$) was gelijk aan 100 in elk van deze vier onderzoeken. Bereken, voor elk van deze onderzoeken, de PVV, de F-statistiek, en de bijbehorende p-waarde. De F-statistiek kan berekend worden als je weet dat (1) de variantie van $y$ gelijk is aan $\frac{1}{N}SS_{Totaal}$, (2) dat de verklaarde variantie gelijk is $\frac{1}{N}SS_A$ en (3) dat $SS_{Error|A}$ gelijk is aan $SS_{Totaal} - SS_A$. Van

welke factoren (de verklaarde variantie en/of het aantal observaties $N$) hangen de PVV en de F-statistiek af?

| | Verklaarde Variantie | $N$ |
|---|---|---|
| 1 | 20 | 10 |
| 2 | 20 | 50 |
| 3 | 80 | 10 |
| 4 | 80 | 50 |

Tabel 3.13: Tabel bij opgave 8.

---

## 3.3 Variantieanalyse met Twee Factoren

### 3.3.1 Voorbeeld 1: Mobiliteit en Sociale Activiteit

We beginnen met een getallenvoorbeeld dat verder in dit hoofdstuk gebruikt zal worden om allerlei berekeningen te illustreren. Voor de inhoudelijke achtergrond van dit getallenvoorbeeld bouwen we verder op het voorbeeld uit de vorige paragraaf. Stel dat de onderzoekster niet alleen geïnteresseerd is in de rol van MOBILITEIT (een situatievariabele) maar ook in die van de BURGERLIJKE STAND van de bejaarden (een persoonsvariabele). Wat betreft BURGERLIJKE STAND, beperkt de onderzoekster zich tot de volgende drie types bejaarden: (1) bejaarden met een partner, (2) bejaarden waarvan de partner onlangs is overleden en (3) bejaarden die hun hele leven vrijgezel zijn geweest. (Gescheiden bejaarden en bejaarden waarvan de partner reeds lange tijd overleden is namen niet deel aan dit onderzoek.) De onderzoekster zocht contact met dertig bejaarden, tien van elk type. Binnen elk type (met partner, partner overleden en vrijgezel) werden op zuiver toevallige wijze vijf bejaarden toegewezen aan de conditie met gratis taxi en vijf aan de conditie zonder gratis taxi. De conditie met vervoer op bestelling, die weinig verschilde van de conditie zonder gratis vervoer, wordt hier dus buiten beschouwing gelaten.

Op basis van de factoren MOBILITEIT en BURGERLIJKE STAND kan men de bejaarden in twee keer drie groepen indelen. De gegevens van dit onderzoek staan in Tabel 3.14. We duiden SOCIALE ACTIVITEIT aan met SA en de proefpersonen binnen elke groep met de letter $i$. In de tabel staan de *celgemiddelden*, de *marginale gemiddelden*, en het *globale gemiddelde*. De celgemiddelden zijn de gemiddelden binnen elk van de zes groepen, de marginale gemiddelden zijn de gemiddelden binnen elk van de niveaus van de twee onafhankelijke variabelen, en het globale gemiddelde is het gemiddelde van álle observaties. In wat volgt, spreken we over cellen in plaats van over groepen. Hiermee maken we duidelijk dat het over een speciaal soort groepen gaat: groepen die gevormd worden door twee factoren met elkaar te kruisen.

In elke cel van Tabel 3.14 bevindt zich hetzelfde aantal proefpersonen (namelijk, 5). Dit is een belangrijk gegeven, want óf het aantal proefpersonen per cel gelijk is of niet

heeft implicaties voor wat men toetst met de variantieanalyse. Een onderzoeksdesign met een gelijk aantal observaties per cel wordt een *gebalanceerd design* genoemd en een onderzoeksdesign met een ongelijk aantal observaties per cel een *ongebalanceerd design*. Het onderscheid tussen een gebalanceerd en een ongebalanceerd design en de implicaties ervan voor de resultaten van de variantieanalyse is een van de belangrijkste onderwerpen van dit hoofdstuk.

| | Met partner | | Partner verloren | | Vrijgezel | | Marg. Gem. |
|---|---|---|---|---|---|---|---|
| | $i$ | SA | $i$ | SA | $i$ | SA | |
| Met gratis taxi | 1 | 10 | 1 | 10 | 1 | 12 | |
| | 2 | 7 | 2 | 13 | 2 | 10 | |
| | 3 | 9 | 3 | 8 | 3 | 12 | |
| | 4 | 8 | 4 | 15 | 4 | 13 | |
| | 5 | 9 | 5 | 13 | 5 | 10 | |
| | Celgem. | 8.6 | Celgem. | 11.8 | Celgem. | 11.4 | 10.6 |
| | $i$ | SA | $i$ | SA | $i$ | SA | |
| Zonder gratis taxi | 1 | 7 | 1 | 4 | 1 | 11 | |
| | 2 | 6 | 2 | 8 | 2 | 12 | |
| | 3 | 7 | 3 | 5 | 3 | 13 | |
| | 4 | 8 | 4 | 7 | 4 | 11 | |
| | 5 | 6 | 5 | 5 | 5 | 10 | |
| | Celgem. | 6.8 | Celgem. | 5.8 | Celgem. | 11.4 | 8 |
| Marg. Gem. | | 7.7 | | 8.8 | | 11.4 | 9.3 |

Tabel 3.14: Gegevens uit een hypothetisch onderzoek naar het effect van MOBILITEIT en BURGERLIJKE STAND op SOCIALE ACTIVITEIT bij bejaarden.

Voor we verder ingaan op het belang van ongebalanceerde designs, staan we even stil bij het design van het voorbeeldonderzoek. In dit design worden twéé factoren gevarieerd (MOBILITEIT én BURGERLIJKE STAND). Hierdoor is het mogelijk om te onderzoeken of het effect van de ene factor (bijvoorbeeld, MOBILITEIT) verschilt naargelang het niveau van de andere factor (in ons voorbeeldonderzoek, BURGERLIJKE STAND). Dit fenomeen wordt *interactie* genoemd. Het fenomeen interactie is ook een van de belangrijkste onderwerpen van dit hoofdstuk. De behandeling van het fenomeen interactie wordt uitgesteld tot na de behandeling van de ongebalanceerde designs.

### 3.3.2 Wat zijn Ongebalanceerde Designs en Wanneer Krijgt men Ermee te Maken?

#### Wat zijn Ongebalanceerde Designs?

Een design met twee of meer factoren kan gebalanceerd of ongebalanceerd zijn. Of we te maken hebben met een gebalanceerd of een ongebalanceerd design, hangt af van het aantal observaties per cel. Een design met een gelijk aantal observaties per cel wordt in dit boek gebalanceerd genoemd en een design met een ongelijk aantal observaties

per cel wordt ongebalanceerd genoemd. Er bestaan ook designs met een ongelijk aantal observaties per cel die door andere auteurs gebalanceerd genoemd worden. Het gaat hier dan wel over designs met een bepaald patroon in de aantallen observaties per cel. Rekening houden met dit patroon zou de uiteenzetting aanzienlijk complexer maken. En daar staat tegenover dat de implicaties voor de praktijk (van het wel of geen rekening houden met het genoemde patroon in de aantallen observaties) beperkt zijn. Daarom beschouwen we in dit boek "gebalanceerd" als synoniem voor "met een gelijk aantal observaties per cel".

Ongebalanceerde designs komen vaak voor en de reden is dat we in de praktijk vaak geconfronteerd worden met storende variabelen. Om te corrigeren voor de invloed van een storende variabele, wordt deze vaak als extra factor opgenomen in de variantieanalyse, en dit resulteert dan in een ongebalanceerd design.

### Experimenteel en Observationeel Onderzoek

In een ideale wereld zouden we, indien gewenst, elke onderzoeksvraag beantwoorden met behulp van een experiment. Met een experiment is het immers mogelijk om storende variabelen te controleren. Bijvoorbeeld, een onderzoeker wilde weten of de deelname aan een afslankingsprogramma zorgt voor blijvend gewichtsverlies. Om deze vraag te beantwoorden, werd een groep zwaarlijvigen die deelnam aan dit afslankingsprogramma (de behandelgroep) vergeleken met een groep zwaarlijvigen die geen inspanning deed om af te slanken (de controlegroep). Om het effect van het afslankingsprogramma te evalueren vergeleek de onderzoeker het gemiddelde lichaamsgewicht van de twee groepen, drie jaar na afloop van het afslankingsprogramma.

Voor de interpretatie van de resultaten van dit onderzoek is het van essentieel belang om te weten op welke manier de proefpersonen terecht zijn gekomen in één van de twee condities (afslankingsprogramma of controleconditie). Indien de proefpersonen op zuiver toevallige wijze werden toegewezen aan de condities, dan garandeert dit dat de twee groepen niet systematisch van elkaar verschillen op één of andere persoonsvariabele. Met andere woorden, er is geen storende persoonsvariabele. Dit is de sterkte van het experiment.

Stel nu dat de onderzoeker de proefpersonen niet op zuiver toevallige wijze heeft toegewezen aan de condities; hij observeerde gewoon wie er koos voor het behandelprogramma. In dit geval hebben we te maken met een observationeel onderzoek. Omdat de proefpersonen zelf kozen voor deelname of geen deelname, is het zeer goed mogelijk dat de personen die deelnamen aan het afslankingsprogramma op een systematische manier verschillen van de personen die er niet aan deelnamen. Bijvoorbeeld, stel dat men de proefpersonen in de volgende categorieën kan indelen: (1) gemotiveerd om af te slanken en met realistische verwachtingen, (2) gemotiveerd om af te slanken en met ónrealistische verwachtingen, en (3) ongemotiveerd. Het is zeer goed mogelijk dat de controlegroep in vergelijking met de behandelgroep veel meer ongemotiveerden bevat. En indien nu juist de ongemotiveerden het slecht doen wat betreft gewichtsverlies (vaak geen gewichtsverlies en soms zelfs gewichtstoename), dan is het niet mogelijk om een uitspraak te doen over het effect van het afslankingsprogramma op basis van een vergelijking van de behandel- en de controlegroep. Kort samengevat, de aard van de motivatie (gemotiveerd-realistisch, gemotiveerd-onrealistisch of ongemotiveerd), verder aangeduid

als MOTIVATIETYPE, is een storende persoonsvariabele.

Het is mogelijk om uit te zuiveren voor de storende invloed van MOTIVATIETYPE door deze variabele als extra factor op te nemen in de variantieanalyse. Dit resulteert in een ongebalanceerd design omdat de verdeling van de proefpersonen over de drie niveaus van MOTIVATIETYPE anders is in de behandel- dan in de controlegroep. Concreet, in de controlegroep zullen meer ongemotiveerde proefpersonen zitten dan in de behandelgroep.

Storende persoonsvariabelen zijn niet het enige probleem van observationeel onderzoek. Een ander probleem zijn storende *situatie*variabelen. We krijgen hiermee te maken als de experimentele controle gebrekkig is. Bijvoorbeeld, in een onderzoek naar het effect van NACHTARBEID op ARBEIDSTEVREDENHEID bij een groep verplegers kan het zo zijn dat de verplegers die veel nachtarbeid verrichten vooral op bepaalde afdelingen werken (bijvoorbeeld, intensive care, neonatologie) en minder op andere (bijvoorbeeld, operatiekamers, inwendige ziekten). Met andere woorden, er is een verband tussen de variabelen AFDELINGSTYPE en NACHTARBEID. De variabele AFDELINGSTYPE is een storende variabele indien er, behalve het genoemde verband met NACHTARBEID, ook een verband is met de variabele ARBEIDSTEVREDENHEID. In dit geval, kan het zijn dat een eventueel verband tussen NACHTARBEID en ARBEIDSTEVREDENHEID toe te schrijven is aan de storende situatievariabele AFDELINGSTYPE.

Corrigeren voor de invloed van een storende situatievariabele gebeurt op dezelfde manier als voor een storende persoonsvariabele: deze storende variabele wordt als extra factor opgenomen in de variantieanalyse en dit resulteert dan in een ongebalanceerd design. Voor ons voorbeeld betekent dit dat de verdeling van de proefpersonen (i.c., de verplegers) over de niveaus van AFDELINGSTYPE anders is in de groep die veel nachtarbeid verricht dan in de groep die weinig nachtarbeid verricht; de verplegers in de conditie met veel nachtarbeid werken vooral op andere afdelingen dan de verplegers met weinig nachtarbeid. Bijvoorbeeld, het aantal verplegers in de conditie [veel nachtarbeid, intensive care] zal wellicht verschillen van het aantal verplegers in de conditie [weinig nachtarbeid, intensive care].

### Als Proefpersonen niet Komen Opdagen

Verreweg de meeste ongebalanceerde designs die men in de praktijk tegenkomt zijn die waarin één of meer factoren storende variabelen zijn waarvoor men wil uitzuiveren. Er is echter nog een tweede manier waarop een ongebalanceerd design kan ontstaan, namelijk wanneer een aantal proefpersonen niet komt opdagen en het onmogelijk is om nog extra proefpersonen te ronselen. Dan kan het best zo zijn dat niet alle condities evenveel proefpersonen bevatten. Deze situatie (een ongebalanceerd design doordat een aantal proefpersonen niet komt opdagen) is voor de praktijk echter veel minder belangrijk dan de eerste situatie (een ongebalanceerd design met storende variabelen als factoren).

We moeten hier een onderscheid maken tussen proefpersonen die niet komen opdagen en proefpersonen die uitvallen. Proefpersonen die uitvallen hebben wel degelijk aan het onderzoek deelgenomen, maar om één of andere reden werd er voor hen geen score geobserveerd. Het kan bijvoorbeeld zijn dat zij afhaakten omdat zij niet langer wílden deelnemen aan het onderzoek. Als het uitvalpercentage verschillend is voor de verschillende condities, dan is het risico groot dat de resultaten van de variantieanalyse verstoord zijn. Onder “verstoord” verstaan we hier dat deze resultaten anders zijn dan

de resultaten die we geobserveerd zouden hebben indien er geen uitval geweest zou zijn (bijvoorbeeld, als de proefpersonen toch doorgezet zouden hebben). We krijgen zo'n verstoring als de proefpersonen die uitvallen juist zeer hoge of zeer lage scores op de afhankelijke variabele gehad zouden hebben *indien ze niet uitgevallen zouden zijn*. Het belangrijke punt hier, is dat uitzuivering voor de andere factoren ons hier niet helpt. De verstoring in de resultaten van de variantieanalyse blijft hier bestaan, met of zonder uitzuivering.

### 3.3.3 Voorbeeld 2: Pesten in de Klas

#### Inleiding

We geven nu een tweede getallenvoorbeeld dat verder in de hoofdstuk gebruikt zal worden om allerlei berekeningen te illustreren. Anders dan het eerste getallenvoorbeeld dat op pagina 160 besproken werd, geven we nu een voorbeeld van een óngebalanceerd design. De inhoudelijke achtergrond van dit voorbeeld is het groeiende probleem van leerlingen die gepest worden door hun klasgenoten. De directie van een basisschool heeft, in samenwerking met de plaatselijke RIAGG, een training in sociale weerbaarheid opgezet. Het doel van deze training is leerlingen bij te brengen hoe zij door een andere manier van reageren pesterijen kunnen vermijden. De mogelijkheid om deel te nemen aan deze training werd aangeboden aan leerlingen die zich tot hun leraar richtten, of door hun leraar aangesproken werden omdat zij gepest werden. De leerlingen konden vrij kiezen of zij aan de training deelnamen of niet.

Op het einde van het schooljaar wilde de directie onderzoeken of de training een effect heeft gehad. Daartoe vergeleek zij de groep die had deelgenomen aan de training met de groep die dat niet had gedaan, maar aan wie de training wel aangeboden was. De gegevens voor dit onderzoek werden verzameld met behulp van een vragenlijst. In deze vragenlijst moest een leerling onder andere aangeven of en hoe vaak hij of zij de afgelopen week gepest was door elk van de klasgenoten. Voor elke leerling werd een *pestscore* berekend door deze aantallen op te tellen over de verschillende klasgenoten.

De gemiddelde pestscore in de groep die aan de training heeft deelgenomen was 5.28. En de gemiddelde pestscore in de groep die niet aan de training heeft deelgenomen was 7.53. Dit verschil van 2.25 ($= 7.53 - 5.28$) punten was statistisch significant. Dit werd vastgesteld met de F-toets van de variantieanalyse met één factor, beschreven op pagina 154. (Men krijgt exact dezelfde p-waarde met de T-toets voor het verschil tussen twee groepen.) Dus, er lijkt een positief effect te zijn van de sociale weerbaarheidstraining.

#### De Rol van de Storende Variabele TYPE LERAAR

We leggen nu uit hoe we aan de gemiddelden voor de groep met en zonder training (5.28 en 7.53) zijn gekomen. We doen dit op een, op het eerste gezicht, omslachtige manier: we introduceren een storende situatievariabele, TYPE LERAAR, en laten zien dat er, bij de berekening van deze gemiddelden, geen rekening gehouden werd met deze variabele. Concreet, er waren drie types leraren: actieve, reactieve en passieve, en zij verschilden van elkaar wat betreft de manier waarop zij omgingen met pesten:

1. Actieve leraren bespraken zonder rechtstreekse aanleiding het probleem met hun leerlingen.

2. Reactieve leraren stelden het probleem alleen aan de orde als duidelijk werd dat een van hun leerlingen gepest werd.

3. Passieve leraren namen geen enkel initiatief, en beperkten zich tot het doorverwijzen van leerlingen naar de sociale weerbaarheidstraining.

Er was een verband tussen de variabele TYPE LERAAR (die aangeeft bij welk type leraar een leerling in de klas zit) en de variabele TRAINING (die aangeeft of een leerling wel of niet deelneemt aan de training): leerlingen van actieve leraren zijn het meest geneigd om aan de training deel te nemen, leerlingen van reactieve leraren een stuk minder en leerlingen van passieve leraren het minst. Dit kan men zien in Tabel 3.15 waar het aantal deelnemers aan de training uitgesplitst wordt naar het type leraar. Van de leerlingen met een actieve leraar neemt 80 procent ($= 28/35 \times 100$) deel aan de training, van de leerlingen met een reactieve leraar 43 procent ($= 14/30 \times 100$) en van de leerlingen met een passieve leraar slechts 32 procent ($= 8/25 \times 100$).

| | Actieve Leraar | Reactieve Leraar | Passieve Leraar | Som |
|---|---|---|---|---|
| Met Training | 28 | 14 | 8 | 50 |
| Zonder Training | 7 | 16 | 17 | 40 |
| Som | 35 | 30 | 25 | 90 |

Tabel 3.15: Aantal leerlingen in de zes cellen van het voorbeeldonderzoek.

Er was ook een verband tussen TYPE LERAAR en de afhankelijke variabele PESTSCORE: leerlingen van actieve leraren werden minder gepest dan leerlingen van reactieve leraren en die op hun beurt weer minder dan leerlingen van passieve leraren. Dit kan men zien in Tabel 3.16 waarin de gemiddelde pestscores gegeven worden voor elk van de zes cellen. Deze gemiddelden worden *celgemiddelden* genoemd. In de randen van de tabel worden ook de *marginale gemiddelden* gegeven. Deze marginale gemiddelden zijn gemiddelden van celgemiddelden (bijvoorbeeld, $(3 + 6 + 12)/3 = 7$ en $(6 + 6)/2 = 6$). Deze marginale gemiddelden zijn ongewogen gemiddelden want er wordt geen rekening gehouden met het aantal leerlingen in elk van de cellen. Uit de verschillen tussen de marginale kolomgemiddelden blijkt dat er een verband is tussen TYPE LERAAR en PESTSCORE.

Merk op dat er geen verband is tussen de variabele TRAINING en de variabele PESTSCORE: gemiddeld over de drie types leraren is er geen verschil tussen de leerlingen die wel en niet deelgenomen hebben aan de training, want beide groepen hebben een ongewogen gemiddelde van 7. De vraag is nu hoe we aan de eerste gemiddelden (5.28 en 7.53) gekomen zijn. We krijgen deze gemiddelden als we geen rekening houden met de variabele TYPE LERAAR. Het gemiddelde 5.28 is het gemiddelde van de 50 pestscores

| | Actieve Leraar | Reactieve Leraar | Passieve Leraar | Gemiddelde |
|---|---|---|---|---|
| Met Training | 3 | 6 | 12 | 7 |
| Zonder Training | 5 | 6 | 10 | 7 |
| Gemiddelde | 4 | 6 | 11 | 7 |

Tabel 3.16: Gemiddelde pestscores in de zes cellen van het voorbeeldonderzoek.

in de groep met training en het gemiddelde 7.53 is het gemiddelde van de 40 pestscores in de groep zonder training. We krijgen het gemiddelde 5.28 door de som te nemen van de 50 pestscores in de eerste groep en dan te delen door 50. Ook al zijn de individuele pestscores niet gegeven, we kunnen dit gemiddelde ook berekenen op basis van de gegevens in Tabel 3.15 en 3.16. Bij deze berekening maken we gebruik van het volgende:

1. De som van de pestscores in de groep met training is gelijk aan de som van drie andere sommen: de sommen van de individuele pestscores in de cellen met een actieve, een reactieve en een passieve leraar.

2. De som van de pestscores in een bepaalde cel is gelijk aan het celgemiddelde maal het aantal personen in die cel.

Het gemiddelde in de groep met training wordt dan als volgt berekend:

$$\frac{(28 \times 3) + (14 \times 6) + (8 \times 12)}{28 + 14 + 8} = \frac{264}{50} = 5.28 \quad .$$

Dit gemiddelde is een gewogen gemiddelde omdat de celgemiddelden gewogen worden met het aantal leerlingen in elk van die cellen. Het gewogen gemiddelde in de groep zonder training wordt op dezelfde manier berekend.

De gewogen gemiddelden worden getoond in Tabel 3.17. In de cellen van deze tabel staan de aantallen uit Tabel 3.15 samen met de celgemiddelden uit Tabel 3.16. Alle gemiddelden staan vet gedrukt en de aantallen normaal gedrukt. In de randen van deze tabel staan de gewogen marginale gemiddelden. Behalve de gewogen marginale rijgemiddelden, staan in deze tabel ook de gewogen marginale kolomgemiddelden. Bijvoorbeeld, het gewogen marginale gemiddelde van de derde kolom wordt als volgt berekend:

$$\frac{(8 \times 12) + (17 \times 10)}{8 + 17} = \frac{266}{25} = 10.64 \quad .$$

Uit de vergelijking van Tabel 3.16 en 3.17 blijkt dat er geen verschil is tussen de ongewogen marginale gemiddelden van de groep met en de groep zonder training, maar dat dit wel het geval is voor de gewogen marginale gemiddelden. Dit komt omdat in het gewogen marginale gemiddelde van de groep met training de kleine celgemiddelden (3 en 6) een groter gewicht krijgen dan het grote celgemiddelde (12). In de groep zonder training gebeurt het omgekeerde (het grote celgemiddelde krijgt een groter gewicht dan de

| | Actieve Leraar | Reactieve Leraar | Passieve Leraar | Gemiddelde |
|---|---|---|---|---|
| Met Training | **3** 28 | **6** 14 | **12** 8 | **5.28** |
| Zonder Training | **5** 7 | **6** 16 | **10** 17 | **7.53** |
| Gemiddelde | **3.4** | **6** | **10.64** | **6.28** |

Tabel 3.17: Aantal leerlingen en gemiddelde pestscores in de zes cellen van het voorbeeldonderzoek.

kleine celgemiddelden). Door de weging met het aantal observaties in de cellen manifesteert het effect van de tweede factor zich indirect in de gewogen marginale gemiddelden van de eerste factor.

**Waarom is het Verschil tussen de Ongewogen Marginale Gemiddelden "Beter"?**

We hebben gezien dat het effect van TYPE LERAAR zich indirect manifesteert in het verschil tussen de gewogen marginale gemiddelden voor *met* en *zonder* training. Als men het effect van TRAINING wil evalueren zonder dat het effect van TYPE LERAAR zich hierin manifesteert, dan moet men het verschil tussen de ongewogen marginale gemiddelden nemen. In deze zin is het verschil tussen de ongewogen marginale gemiddelden beter.

We laten nu zien waarom het effect van TYPE LERAAR zich niet manifesteert in het verschil tussen de ongewogen marginale gemiddelden voor met en zonder training. Het vertrekpunt is dat we het effect van TRAINING evalueren *binnen elk type leraar*. Binnen de groep van actieve leraren bepalen we het effect van TRAINING als het verschil tussen de celgemiddelden 5 and 3, binnen de groep van reactieve leraren als het verschil tussen 6 en 6, en binnen de groep van passieve leraren als het verschil tussen 10 en 12. Omdat we verschillen berekenen bínnen elk type leraar, kan het niet zo zijn dat het effect van TYPE LERAAR zich onrechtstreeks manifesteert in deze verschillen. We hebben dus drie verschillen tussen celgemiddelden, één voor elk van de drie types leraren. Om te komen tot één enkele aanduiding van het verschil tussen de groep met en de groep zonder training, neemt men het gemiddelde van deze drie verschillen:

$$\frac{(5-3)+(6-6)+(10-12)}{3} \quad .$$

Dit gemiddelde verschil kan men herschrijven als het verschil tussen de ongewogen gemiddelden:

$$\begin{aligned}\frac{(5-3)+(6-6)+(10-12)}{3} &= \frac{(5+6+10)}{3} - \frac{(3+6+12)}{3} \\ &= 7-7 \quad .\end{aligned}$$

Dus, het gemiddelde van de verschillen binnen de drie niveaus van TYPE LERAAR is gelijk aan het verschil tussen de ongewogen gemiddelden.

Het verschil tussen de ongewogen gemiddelden wordt ook het *uitgezuiverde* of *gecorrigeerde* verschil tussen de gemiddelden genoemd: het verschil tussen de groep met en de groep zonder training wordt uitgezuiverd of gecorrigeerd voor TYPE LERAAR. Uitzuiveren voor TYPE LERAAR houdt in dat men het effect van TRAINING evalueert bínnen elk type leraar. Verder zegt men ook dat bij de berekening van de ongewogen gemiddelden *rekening gehouden wordt* met de andere variabele (hier, TYPE LERAAR), terwijl dat niet gebeurt bij de berekening van de gewogen gemiddelden. Men moet hier voorzichtig zijn met de terminologie: rekening houden met een andere variabele gebeurt met een gemiddelde waarin géén rekening gehouden wordt met de aantallen observaties per cel (i.c., het ongewogen gemiddelde).

### Grafische Voorstelling

TYPE LERAAR is een storende variabele omdat deze variabele samenhangt met de afhankelijke variabele (PESTSCORE) en met de variabele waarvan men het effect wil onderzoeken (TRAINING). Dit wordt grafisch voorgesteld in Figuur 3.1. De samenhang met PESTSCORE blijkt uit Tabel 3.16: uit de ongewogen marginale gemiddelden voor de drie niveaus van TYPE LERAAR blijkt dat leerlingen van actieve leraren minder gepest werden dan leerlingen van reactieve leraren en die op hun beurt weer minder dan leerlingen van passieve leraren. De samenhang met TRAINING blijkt uit Tabel 3.15: de verhouding tussen de celfrequenties van de groep met en de groep zonder training verschilt naargelang het type leraar (i.c., 28/7, 14/16 en 8/17). Het gevolg hiervan is dat de groep met training vooral bestaat uit leerlingen met actieve leraren en de groep zonder training vooral uit leerlingen met passieve en reactieve leraren. De storende variabele TYPE LERAAR verklaart het verband tussen TRAINING en de PESTSCORE, een verband dat blijkt uit het verschil tussen de gewogen gemiddelden. Dit verschil kan niet toegeschreven worden aan een causaal verband tussen beide, want als we uitzuiveren voor TYPE LERAAR (door het berekenen van ongewogen gemiddelden) dan verdwijnt het verschil. Daarom loopt er geen pijl van TRAINING naar PESTSCORE.

### Verschillende Patronen van Verschillen tussen Gewogen en Ongewogen Gemiddelden

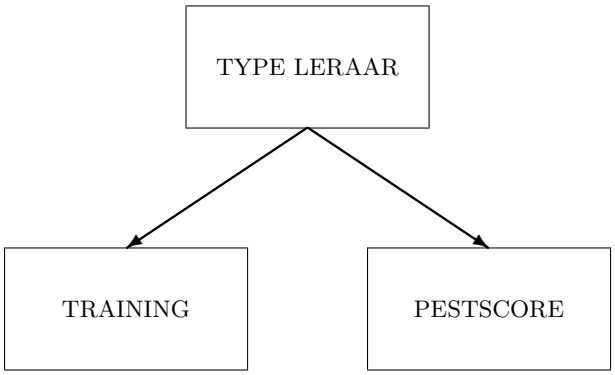
In het voorbeeldonderzoek is er een verschil tussen de gewogen gemiddelden en geen verschil tussen de ongewogen gemiddelden. Dit is niet altijd zo. De volgende patronen kunnen zich voordoen:

1. Het verschil tussen de ongewogen gemiddelden is kleiner dan het verschil tussen de gewogen gemiddelden. (In het meest extreme geval, zoals in het voorbeeldonderzoek, is er geen verschil tussen de ongewogen gemiddelden.)

2. Het verschil tussen de ongewogen gemiddelden is groter dan het verschil tussen de gewogen gemiddelden.

3. Het verschil tussen de ongewogen gemiddelden is in een andere richting dan het verschil tussen de gewogen gemiddelden.

4. De gewogen en de ongewogen gemiddelden zijn gelijk.




Figuur 3.1: Grafische voorstelling van de invloed van een storende variabele (TYPE LERAAR) op het verband tussen de afhankelijke variabele (PESTSCORE) en de variabele waarvan men het effect wil onderzoeken (TRAINING).

In de opgaven bij deze paragraaf worden voorbeelden gegeven van elk van deze vier patronen.

Wat ook het patroon is van de verschillen tussen ongewogen en gewogen gemiddelden, als het verschil tussen de ongewogen gemiddelden ongelijk is aan het verschil tussen de gewogen gemiddelden, dan is de variabele waarvoor we uitzuiveren een storende variabele (i.e., een variabele die zowel samenhangt met de afhankelijke variabele als de variabele waarvan men het effect wil onderzoeken). Dus, we spreken niet alleen van een storende variabele als er een verschil is tussen de gewogen gemiddelden en geen verschil tussen de ongewogen (i.e., na uitzuivering), maar in alle gevallen waarin de verschillen tussen de twee types gemiddelden ongelijk zijn.

### 3.3.4 De Notatie voor de Gegevens

#### De Factoren

De eerste factor duiden we aan met A. De niveaus van A duiden we aan met de index $k$ en het aantal niveaus met $K$. In het gebalanceerde voorbeeldonderzoek is factor A MOBILITEIT en in het ongebalanceerde is het TRAINING. In beide gevallen is $K$ gelijk aan 2. De tweede factor duiden we aan met B, de niveaus van B duiden we aan met de index $l$ en het aantal niveaus met $L$. In het gebalanceerde voorbeeldonderzoek is factor B BURGERLIJKE STAND en in het ongebalanceerde is het TYPE LERAAR. In beide gevallen is $L$ gelijk aan 3.

#### De Ruwe Gegevens

De ruwe gegevens kunnen geordend worden in een $K \times L$-matrix. Zo'n matrix wordt getoond in Tabel 3.18. Elke cel van deze matrix correspondeert met één groep, gevormd door een combinatie van niveaus van de twee factoren. In cel $(k, l)$ van deze matrix staan de observaties van het k-de niveau van A en het l-de niveau van B. We gebruiken

$i$ als index voor de observaties binnen een cel en de i-de observatie binnen cel $(k, l)$ duiden we aan met $y_{kli}$. De gegevens in Tabel 3.18 vormen een gebalanceerd design omdat het aantal observaties per cel gelijk is. Dit aantal duiden we aan met $n$. Voor het gebalanceerde voorbeeldonderzoek (zie Tabel 3.14) is $n = 5$. In een ongebalanceerd design is het aantal observaties per cel verschillend en dit aantal wordt dan aangeduid met $n_{kl}$, waarbij de index $kl$ verwijst naar de cel $(k, l)$. Voor het ongebalanceerde voorbeeldonderzoek (zie Tabel 3.17) is $n_{11} = 28$, $n_{12} = 14$, $n_{13} = 8$, $n_{21} = 7$, $n_{22} = 16$ en $n_{23} = 17$.

| | | B | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | $\cdots$ | L | M. G. |
| A | 1 | $y_{111}$<br>$y_{112}$<br>$\vdots$<br>$y_{11n}$ | $y_{121}$<br>$y_{122}$<br>$\vdots$<br>$y_{12n}$ | $\cdots$<br>$\cdots$<br>$\cdots$<br>$\cdots$ | $y_{1L1}$<br>$y_{1L2}$<br>$\vdots$<br>$y_{1Ln}$ | |
| | | $\overline{y}_{11}$ | $\overline{y}_{12}$ | $\cdots$ | $\overline{y}_{1L}$ | $\overline{y}_{1\bullet}$ |
| | 2 | $y_{211}$<br>$y_{212}$<br>$\vdots$<br>$y_{21n}$ | $y_{221}$<br>$y_{222}$<br>$\vdots$<br>$y_{22n}$ | $\cdots$<br>$\cdots$<br>$\cdots$<br>$\cdots$ | $y_{2L1}$<br>$y_{2L2}$<br>$\vdots$<br>$y_{2Ln}$ | |
| | | $\overline{y}_{21}$ | $\overline{y}_{22}$ | $\cdots$ | $\overline{y}_{2L}$ | $\overline{y}_{2\bullet}$ |
| | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| | K | $y_{K11}$<br>$y_{K12}$<br>$\vdots$<br>$y_{K1n}$ | $y_{K21}$<br>$y_{K22}$<br>$\vdots$<br>$y_{K2n}$ | $\cdots$<br>$\cdots$<br>$\cdots$<br>$\cdots$ | $y_{KL1}$<br>$y_{KL2}$<br>$\vdots$<br>$y_{KLn}$ | |
| | | $\overline{y}_{K1}$ | $\overline{y}_{K2}$ | $\cdots$ | $\overline{y}_{KL}$ | $\overline{y}_{K\bullet}$ |
| | M. G. | $\overline{y}_{\bullet 1}$ | $\overline{y}_{\bullet 2}$ | $\cdots$ | $\overline{y}_{\bullet L}$ | $\overline{y}_{\bullet\bullet}$ |

Tabel 3.18: Datamatrix voor een onderzoek met twee factoren (A en B) en een gelijk aantal observaties per cel. "M. G." staat voor "marginaal gemiddelde".

Het gemiddelde in de cel $(k, l)$ duiden we aan met $\overline{y}_{kl}$. Dit gemiddelde wordt als volgt berekend:

$$\overline{y}_{kl} = \frac{\sum_{i=1}^{n_{kl}} y_{kli}}{n_{kl}} \quad .$$

In een ongebalanceerd design verschilt het aantal observaties waarover het celgemiddelde $\overline{y}_{kl}$ berekend wordt naargelang de cel. In een gebalanceerd design is het aantal

observaties $n_{kl}$ gelijk voor alle cellen en kunnen we de notatie vereenvoudigen door $n_{kl}$ te vervangen door $n$, het gemeenschappelijke aantal observaties per cel.

We bekijken nu de gemiddelden voor één bepaald niveau van factor A of factor B. Deze gemiddelden worden marginale gemiddelden genoemd. Er zijn twee soorten marginale gemiddelden: gewogen en ongewogen marginale gemiddelden. We behandelen eerst de ongewogen marginale gemiddelden.

### De Ongewogen Marginale Gemiddelden

Het ongewogen marginale gemiddelde voor het k-de niveau van A duiden we aan met $\overline{y}_{k\bullet}$. Dit ongewogen marginale gemiddelde is het gewone gemiddelde van de $L$ celgemiddelden $\overline{y}_{kl}$ binnen het k-de niveau van A:

$$\overline{y}_{k\bullet} = \frac{\sum_{l=1}^{L} \overline{y}_{kl}}{L} \quad .$$

In het gebalanceerde voorbeeldonderzoek zijn $\overline{y}_{1\bullet}$ en $\overline{y}_{2\bullet}$ gelijk aan, respectievelijk, 10.6 en 8. En in het ongebalanceerde voorbeeldonderzoek zijn zij beide gelijk aan 7.

Het ongewogen marginale gemiddelde voor het l-de niveau van B duiden we aan met $\overline{y}_{\bullet l}$. (Ongewogen gemiddelden worden aangeduid met een zwart bolletje ($\bullet$) op de plaats van de index waarover het gemiddelde wordt genomen.) Dit ongewogen marginale gemiddelde is het gewone gemiddelde van de $K$ celgemiddelden $\overline{y}_{kl}$ binnen het l-de niveau van B:

$$\overline{y}_{\bullet l} = \frac{\sum_{k=1}^{K} \overline{y}_{kl}}{K} \quad .$$

In het gebalanceerde voorbeeldonderzoek zijn $\overline{y}_{\bullet 1}$, $\overline{y}_{\bullet 2}$ en $\overline{y}_{\bullet 3}$ gelijk aan, respectievelijk, 7.7, 8.8 en 11.4. En in het ongebalanceerde voorbeeldonderzoek zijn zij gelijk aan, respectievelijk, 4, 6 en 11. De marginale gemiddelden $\overline{y}_{k\bullet}$ en $\overline{y}_{\bullet l}$ zijn ongewogen omdat er geen rekening gehouden wordt met het aantal observaties in de cellen.

### De Gewogen Marginale Gemiddelden

We behandelen nu de gewogen marginale gemiddelden. Het is handig om hierbij te vertrekken van een tabel waarin men, anders dan Tabel 3.18, ook de gegevens van een ongebalanceerd design kan weergeven. Tabel 3.19 is zo een tabel. In de binnenste cellen van deze tabel (de cellen links van en boven de dubbele lijnen) staan nu niet de observaties $y_{kli}$ maar wel het celgemiddelde $\overline{y}_{kl}$ en het aantal observaties $n_{kl}$. De ongewogen marginale gemiddelden $\overline{y}_{k\bullet}$ en $\overline{y}_{\bullet l}$ staan in de cellen rechts van en onder de dubbele lijnen. Zoals hierboven beschreven, zijn de ongewogen marginale gemiddelden $\overline{y}_{k\bullet}$ en $\overline{y}_{\bullet l}$ gewone gemiddelden (over de kolommen, respectievelijk, de rijen) van de celgemiddelden $\overline{y}_{kl}$.

Het gewogen marginale gemiddelde voor het k-de niveau van A wordt aangeduid met $\overline{y}_{k\sqcup}$. Dit gewogen marginale gemiddelde wordt als volgt berekend:

$$\overline{y}_{k\sqcup} = \frac{\sum_{l=1}^{L} \sum_{i=1}^{n_{kl}} y_{kli}}{\sum_{l=1}^{L} n_{kl}} \quad .$$

| | | B | | | | | | Som | M. G. | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | | 2 | | $\cdots$ | L | | OG | G |
| A | 1 | $\overline{y}_{11}$ | $n_{11}$ | $\overline{y}_{12}$ | $n_{12}$ | $\cdots$ $\cdots$ | $\overline{y}_{1L}$ $n_{1L}$ | $n_{1+}$ | $\overline{y}_{1\bullet}$ | $\overline{y}_{1\sqcup}$ |
| | 2 | $\overline{y}_{21}$ | $n_{21}$ | $\overline{y}_{22}$ | $n_{22}$ | $\cdots$ $\cdots$ | $\overline{y}_{2L}$ $n_{2L}$ | $n_{2+}$ | $\overline{y}_{2\bullet}$ | $\overline{y}_{2\sqcup}$ |
| | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| | K | $\overline{y}_{K1}$ | $n_{K1}$ | $\overline{y}_{K2}$ | $n_{K2}$ | $\cdots$ $\cdots$ | $\overline{y}_{KL}$ $n_{KL}$ | $n_{K+}$ | $\overline{y}_{K\bullet}$ | $\overline{y}_{K\sqcup}$ |
| Som | | | $n_{+1}$ | | $n_{+2}$ | $\cdots$ | $n_{+L}$ | $N$ | | |
| M. G. | OG | $\overline{y}_{\bullet 1}$ | | $\overline{y}_{\bullet 2}$ | | $\cdots$ | $\overline{y}_{\bullet L}$ | | $\overline{y}_{\bullet\bullet}$ | |
| | G | $\overline{y}_{\sqcup 1}$ | | $\overline{y}_{\sqcup 2}$ | | $\cdots$ | $\overline{y}_{\sqcup L}$ | | | $\overline{y}_{\sqcup\sqcup}$ |

Tabel 3.19: Datamatrix voor een onderzoek met twee factoren (A en B) en een ongelijk aantal observaties per cel. Per cel wordt alleen het gemiddelde ($\overline{y}_{kl}$) en het aantal observaties ($n_{kl}$) gegeven. "M. G." staat voor "marginaal gemiddelde", "OG" voor "ongewogen" en "G" voor "gewogen".

In de teller van deze formule wordt de som genomen van alle observaties binnen het k-de niveau van A en in de noemer staat het totaal aantal observaties binnen dit niveau van A. Dit aantal obervaties duiden we aan met $n_{k+}$. We maken nu gebruik van het feit dat $\sum_{i=1}^{n_{kl}} y_{kli}$ gelijk is aan $n_{kl}\overline{y}_{kl}$. Hieruit volgt dat het gewogen marginale gemiddelde $\overline{y}_{k\sqcup}$ als volgt geschreven kan worden:

$$\overline{y}_{k\sqcup} = \frac{\sum_{l=1}^{L} n_{kl}\overline{y}_{kl}}{n_{k+}} \quad .$$

We zien nu waar de benaming *gewogen* marginale gemiddelde vandaan komt: de celgemiddelden $\overline{y}_{kl}$ worden gewogen met het aantal observaties in elk van de cellen ($n_{kl}$). In het ongebalanceerde voorbeeldonderzoek is $\overline{y}_{1\sqcup} = 5.28$ en $\overline{y}_{2\sqcup} = 7.53$.

Op een analoge manier berekent men de gewogen gemiddelden binnen de verschillende niveaus van B. Deze gewogen gemiddelden duiden we aan met $\overline{y}_{\sqcup l}$. (Gewogen gemiddelden worden aangeduid met een bakje ($\sqcup$) op de plaats van de index waarover het gemiddelde wordt genomen.) Men berekent het gewogen marginale gemiddelde $\overline{y}_{\sqcup l}$ als volgt:

$$\begin{aligned}\overline{y}_{\sqcup l} &= \frac{\sum_{k=1}^{K}\sum_{i=1}^{n_{kl}} y_{kli}}{\sum_{k=1}^{K} n_{kl}} \\ &= \frac{\sum_{k=1}^{K} n_{kl}\overline{y}_{kl}}{n_{+l}} \quad .\end{aligned}$$

In het ongebalanceerde voorbeeldonderzoek is $\overline{y}_{\sqcup 1} = 3.4$, $\overline{y}_{\sqcup 2} = 6$ en $\overline{y}_{\sqcup 3} = 10.64$.

In een gebalanceerd design zijn de gewogen marginale gemiddelden gelijk aan hun ongewogen tegenhangers. Bijvoorbeeld, in het gebalanceerde voorbeeldonderzoek zijn de gewogen marginale gemiddelden voor met en zonder gratis taxi (i.e., $\overline{y}_{1\sqcup}$ en $\overline{y}_{2\sqcup}$) gelijk aan, respectievelijk, 10.6 en 8, de corresponderende ongewogen marginale gemiddelden. Dit kunnen we ook laten zien vertrekkende van de formule voor het gewogen marginale gemiddelde voor het k-de niveau van factor A:

$$\begin{aligned}\overline{y}_{k\sqcup} &= \frac{\sum_{l=1}^{L}\sum_{i=1}^{n} y_{kli}}{nL} \\ &= \frac{\sum_{l=1}^{L} n\overline{y}_{kl}}{nL} \quad .\end{aligned}$$

Het product $n\overline{y}_{kl}$ in deze formule is gelijk aan de som van alle observaties in cel $(k, l)$ en het product $nL$ is gelijk aan het aantal observaties binnen het k-de niveau van A. Omdat $n$ niet afhangt van de index $l$ kan het vóór het somteken gebracht worden en valt het weg tegen de $n$ in de noemer:

$$\begin{aligned}\overline{y}_{k\sqcup} &= \frac{n\sum_{l=1}^{L}\overline{y}_{kl}}{nL} \\ &= \frac{\sum_{l=1}^{L}\overline{y}_{kl}}{L} \quad .\end{aligned}$$

Dit is de formule van $\overline{y}_{k\bullet}$, het corresponderende ongewogen marginale gemiddelde.

Voor de gewogen marginale gemiddelden voor het l-de niveau van B ($\overline{y}_{\sqcup l}$) kan men op dezelfde manier aantonen dat, in een gebalanceerd design, deze gemiddelden gelijk zijn aan de corresponderende ongewogen marginale gemiddelden $\overline{y}_{\bullet l}$. Dit kan men controleren aan de hand van de gegevens van het gebalanceerde voorbeeldonderzoek: de gewogen marginale gemiddelden voor de drie niveaus van BURGERLIJKE STAND (i.e., $\overline{y}_{\sqcup 1}$, $\overline{y}_{\sqcup 2}$ en $\overline{y}_{\sqcup 3}$) zijn gelijk aan 7.7, 8.8 en 11.4, de corresponderende ongewogen marginale gemiddelden.

### Het Globale Gemiddelde

We bekijken nu het gemiddelde van álle observaties (dus niet het gemiddelde binnen een bepaald niveau van factor A of B). Dit wordt verder het *globale gemiddelde* genoemd. Net zoals voor de marginale gemiddelden, is er een ongewogen en een gewogen globale gemiddelde. Het ongewogen globale gemiddelde duiden we aan met $\overline{y}_{\bullet\bullet}$. Dit ongewogen globale gemiddelde is het gewone gemiddelde van alle $K \times L$ celgemiddelden $\overline{y}_{kl}$:

$$\overline{y}_{\bullet\bullet} = \frac{\sum_{k=1}^{K}\sum_{l=1}^{L}\overline{y}_{kl}}{KL} \quad .$$

In het gebalanceerde voorbeeldonderzoek is $\overline{y}_{\bullet\bullet} = 9.3$ en in het ongebalanceerde voorbeeldonderzoek is $\overline{y}_{\bullet\bullet} = 7$.

Het gewogen globale gemiddelde duiden we aan met $\overline{y}_{\sqcup\sqcup}$. Dit gewogen globale gemiddelde is het gemiddelde van alle observaties:

$$\overline{y}_{\sqcup\sqcup} = \frac{\sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{i=1}^{n_{kl}} y_{kli}}{\sum_{k=1}^{K}\sum_{l=1}^{L} n_{kl}} \quad .$$

In de teller van deze formule wordt de som genomen van alle observaties en in de noemer staat het totaal aantal observaties. Het totaal aantal observaties wordt aangeduid met $N$. Door gebruik te maken van het feit dat $\sum_{i=1}^{n_{kl}} y_{kli}$ gelijk is aan $n_{kl}\overline{y}_{kl}$ kan $\overline{y}_{\sqcup\sqcup}$ als volgt geschreven worden:

$$\overline{y}_{\sqcup\sqcup} = \frac{\sum_{k=1}^{K}\sum_{l=1}^{L} n_{kl}\overline{y}_{kl}}{N} \quad .$$

In het ongebalanceerde voorbeeldonderzoek is $\overline{y}_{\sqcup\sqcup} = 6.28$.

Net zoals voor de gewogen marginale gemiddelden, geldt ook voor het gewogen *globale* gemiddelde dat, in een gebalanceerd design, dit gewogen globale gemiddelde gelijk is aan zijn ongewogen tegenhanger. Bijvoorbeeld, in het gebalanceerde voorbeeldonderzoek is het gewogen globale gemiddelde gelijk aan 9.3, het corresponderende ongewogen globale gemiddelde. Dit kunnen we ook laten zien vertrekkende van de formule voor het gewogen globale gemiddelde, en wel op dezelfde manier als die waarop we in het bovenstaande de gelijkheid hebben gedemonstreerd van $\overline{y}_{k\sqcup}$ en $\overline{y}_{k\bullet}$ in een gebalanceerd design.

**Een Alternatieve Manier om de Gegevens te Ordenen**

Behalve zoals in Tabel 3.18, bestaat er nog een tweede manier om de gegevens te ordenen. Deze wordt getoond in Tabel[8] 3.20. In de eerste kolom staat de afhankelijke variabele. De gegevens in de $K \times L$ cellen (met in elke cel telkens $n$ $y_{kli}$'s) staan hier allemaal onder elkaar (eerst cel $(1,1)$, dan cel $(1,2)$, ..., tot en met cel $(K,L)$). In de tweede kolom staat factor A. Elk van de niveaus van deze variabele wordt aangeduid met een getal tussen 1 en $K$. En in de derde kolom staat factor B. Elk van de niveaus van deze variabele wordt aangeduid met een getal tussen 1 en $L$. In de meeste computerprogramma's worden de gegevens op deze manier geordend.

### 3.3.5 De Statistische Toetsing

Men is bijna nooit alleen geïnteresseerd in de grootte van het verschil tussen de gemiddelden maar wil ook weten of dit verschil statistisch significant is. Met andere woorden, men wil weten of dit verschil door toeval te verklaren is. Dit is begrijpelijk, want in de praktijk verschillen de gemiddelden bijna altijd van elkaar, en is het de vraag of dit verschil door toeval te verklaren is.

Als men een verschil tussen gemiddelden statistisch significant noemt, dan bedoelt men dat dit verschil groot genoeg is om met een behoorlijke zekerheid te besluiten dat er ook een verschil is tussen de corresponderende verwachte waarden (i.e., een verschil tussen de gemiddelden in de populatie). Indien er een verschil is tussen deze verwachte

[8] In deze tabel staan de gegevens van een gebalanceerd design, maar hij is eenvoudig aan te passen voor een ongebalanceerd design: het aantal observaties per cel, $n$, moet vervangen worden door $n_{kl}$, het celspecifieke aantal observaties.

| Afhankelijke Variabele | $A$ | $B$ |
|---|---|---|
| $y_{111}$ | 1 | 1 |
| $y_{112}$ | 1 | 1 |
| $\vdots$ | $\vdots$ | $\vdots$ |
| $y_{11n}$ | 1 | 1 |
| $y_{121}$ | 1 | 2 |
| $y_{122}$ | 1 | 2 |
| $\vdots$ | $\vdots$ | $\vdots$ |
| $y_{12n}$ | 1 | 2 |
| $\vdots$ | $\vdots$ | $\vdots$ |
| $y_{1L1}$ | 1 | L |
| $y_{1L2}$ | 1 | L |
| $\vdots$ | $\vdots$ | $\vdots$ |
| $y_{1Ln}$ | 1 | L |
| $\vdots$ | $\vdots$ | $\vdots$ |
| $y_{K11}$ | K | 1 |
| $y_{K12}$ | K | 1 |
| $\vdots$ | $\vdots$ | $\vdots$ |
| $y_{K1n}$ | K | 1 |
| $y_{K21}$ | K | 2 |
| $y_{K22}$ | K | 2 |
| $\vdots$ | $\vdots$ | $\vdots$ |
| $y_{K2n}$ | K | 2 |
| $\vdots$ | $\vdots$ | $\vdots$ |
| $y_{KL1}$ | K | L |
| $y_{KL2}$ | K | L |
| $\vdots$ | $\vdots$ | $\vdots$ |
| $y_{KLn}$ | K | L |

Tabel 3.20: Datamatrix voor een onderzoek met twee factoren (A en B) met ordening van de gegevens in drie kolommen.

waarden dan is het verschil tussen de gemiddelden niet door het toeval te verklaren. Stel dat men geïnteresseerd is in het verschil tussen de gewogen gemiddelden $\overline{y}_{k\sqcup}$. De hiermee corresponderende verwachte waarden duiden we aan met $\mu_{k\sqcup}$, en ze zijn als volgt gedefinieerd:

$$\mu_{k\sqcup} = \frac{\sum_{l=1}^{L} n_{kl}\mu_{kl}}{n_{k+}} \quad .$$

In deze formule staat $\mu_{kl}$ voor de verwachte waarde in cel $(k, l)$ (het gemiddelde in de populatie die behoort tot het k-de niveau van A en het l-de niveau van B), $n_{kl}$ voor het aantal observaties in deze cel, en $n_{k+}$ voor het aantal observaties binnen het k-de niveau van factor A. Dus, de gewogen verwachte waarde $\mu_{k\sqcup}$ is de verwachte waarde in de populatie die (1) behoort tot het k-de niveau van A (bijvoorbeeld, met training) en (2) voor een proportie gelijk aan $n_{kl}/n_{k+}$ bestaat uit personen die behoren tot het l-de niveau van B (bijvoorbeeld, 56 procent met een actieve leraar, 28 procent met een reactieve en 16 procent met een passieve). Met behulp van een statistische toets willen we een antwoord geven op de vraag of deze $K$ gewogen verwachte waarden van elkaar verschillen. Met andere woorden, we willen de volgende nulhypothese toetsen:

$$\boxed{\mu_{1\sqcup} = \mu_{2\sqcup} = \cdots = \mu_{K\sqcup} \quad ,}$$

de gelijkheid van de gewogen verwachte waarden $\mu_{k\sqcup}$.

Omdat de gewogen verwachte waarden gewogen gemiddelden zijn van de $\mu_{kl}$'s kan het zijn dat een effect van B zich indirect manifesteert in deze gewogen verwachte waarden, namelijk als gevolg van een andere weging van de $\mu_{kl}$'s voor de verschillende niveaus van A (i.e., waarden van k). Zo zijn in ons voorbeeld de gewichten voor de groep met training 28/50, 14/50 en 8/50, en voor de groep zonder training 7/40, 16/40 en 17/40. Dus, in de groep met training zitten vooral leerlingen met een actieve leraar en in de groep zonder training zitten vooral leerlingen met een reactieve of passieve leraar.

Ook als men geïnteresseerd is in het verschil tussen de ongewogen gemiddelden $\overline{y}_{k\bullet}$, stelt zich de vraag of dit verschil statistisch significant is. De hiermee corresponderende verwachte waarden duiden we aan met $\mu_{k\bullet}$, en ze zijn als volgt gedefinieerd:

$$\mu_{k\bullet} = \frac{\sum_{l=1}^{L} \mu_{kl}}{L} \quad .$$

Met behulp van een statistische toets willen we een antwoord geven op de vraag of deze $K$ $\mu_{k\bullet}$'s van elkaar verschillen. Met andere woorden, we willen de volgende nulhypothese toetsen:

$$\boxed{\mu_{1\bullet} = \mu_{2\bullet} = \cdots = \mu_{K\bullet} \quad ,}$$

de gelijkheid van de ongewogen verwachte waarden $\mu_{k\bullet}$. De verschillen tussen de ongewogen verwachte waarden $\mu_{k\bullet}$ hangen uitsluitend af van verschillen tussen de niveaus van A bínnen elk van de niveaus van B. Dit kunnen we op dezelfde manier laten zien als voor de ongewogen gemiddelden (zie pagina 167). Het belang hiervan is dat een effect van de storende variabele B zich niet kan manifesteren in verschillen tussen de ongewogen verwachte waarden $\mu_{k\bullet}$.

Er zijn dus twee nulhypotheses die betrekking hebben op verschillen tussen de niveaus van factor A: de eerste heeft betrekking op gewogen verwachte waarden en de tweede

op ongewogen verwachte waarden. Parallel aan deze twee nulhypotheses over verschillen tussen de niveaus van factor A zijn er ook twee nulhypotheses over verschillen tussen de niveaus van factor B: één nulhypothese heeft betrekking op gewogen verwachte waarden (i.c., $\mu_{\sqcup 1} = \mu_{\sqcup 2} = \cdots = \mu_{\sqcup L}$) en één op ongewogen verwachte waarden (i.c., $\mu_{\bullet 1} = \mu_{\bullet 2} = \cdots = \mu_{\bullet L}$). Men wil niet altijd beide sets van nulhypotheses toetsen, want men is vaak slechts in één van de twee factoren geïnteresseerd. De andere factor is dan meestal een storende variabele waarvoor men wil uitzuiveren; in het ongebalanceerde voorbeeldonderzoek is dit factor B, TYPE LERAAR. Men is immers niet in een storende variabele op zich geïnteresseerd maar wel in een andere variabele (in het voorbeeldonderzoek, de variabele TRAINING) die men wil uitzuiveren voor de invloed van de storende variabele.

In een gebalanceerd design, zijn de gewogen verwachte waarden $\mu_{k\sqcup}$ (voor factor A) en $\mu_{\sqcup l}$ (voor factor B) gelijk aan de corresponderende ongewogen verwachte waarden $\mu_{k\bullet}$ en $\mu_{\bullet l}$. Dit kan men op dezelfde manier laten zien als de gelijkheid van de gewogen gemiddelden $\overline{y}_{k\sqcup}$ en $\overline{y}_{\sqcup l}$ (voor, respectievelijk, factor A en factor B) aan de corresponderende ongewogen gemiddelden $\overline{y}_{k\bullet}$ en $\overline{y}_{\bullet l}$ (zie pagina 173 voor de gelijkheid van $\overline{y}_{k\sqcup}$ en $\overline{y}_{k\bullet}$).

Alle nulhypotheses (over gewogen en ongewogen verwachte waarden binnen de niveaus van factor A en B) worden getoetst met een F-statistiek van de volgende algemene vorm:

$$\boxed{F = \frac{MS_{Effect|Ref}}{MS_{Error|Model}} \quad .}$$

In deze formule kunnen we voor *Effect* zowel het effect van A als het effect van B invullen. Dit effect moet men altijd nader specificeren door aan te geven wat het zogenaamde *referentiemodel* (aangeduid met *Ref*) is. De functie van dit referentiemodel, dat we later meer gedetailleerd zullen bespreken, is aan te geven met welke andere effecten rekening gehouden wordt. Als in dit referentiemodel geen effect gespecificeerd wordt, dan wordt ook met geen enkele andere factor rekening gehouden. In dit geval toetst men met de F-statistiek een nulhypothese die betrekking heeft op gewogen verwachte waarden (de gewogen verwachte waarden van de niveaus van A bij *Effect*=A en de gewogen verwachte waarden van de niveaus van B bij *Effect*=B). En als in dit referentiemodel het effect van de andere factor gespecificeerd wordt (factor B bij *Effect*=A en factor A bij *Effect*=B) dan wordt met deze factor rekening gehouden en toetst men met de F-statistiek een nulhypothese die betrekking heeft op ongewogen verwachte waarden. Dit is in een notendop een van de belangrijkste punten van dit hoofdstuk. We komen hier nog uitgebreid op terug.

De noemer van de formule van de F-statistiek, de gemiddelde errorkwadratensom $MS_{Error|Model}$, heeft dezelfde functie als in de variantieanalyse met één factor: hij bepaalt de gevoeligheid waarmee men toetst. Een verschil met de variantieanalyse met één factor is dat men nu een keuze heeft uit meerdere errorkwadratensommen, afhankelijk van het model dat men gebruikt om de observaties te voorspellen. Men herkent deze verschillende errorkwadratensommen aan wat er ingevuld wordt voor *Model* in het subscript van $MS_{Error|Model}$. Het spreekt voor zich dat men het model kiest met de kleinste $MS_{Error|Model}$, want dit resulteert in de meest gevoelige toets. Ook hier komen we nog uitgebreid op terug.

In de volgende paragrafen laten we zien hoe, voor de gegevens van de twee voorbeeldonderzoeken, de F-statistieken berekend worden. In de daarna volgende paragrafen geven we de formules die nodig zijn om de F-statistieken ook voor andere gegevens uit te rekenen. Voorlopig laten we de interactie tussen de factoren buiten beschouwing; vanaf pagina 211 bespreken we de aanpassingen die nodig zijn om met het interactie-effect rekening te houden.

### 3.3.6 Het Rekenwerk in een Gebalanceerd Design

**Het Effect van MOBILITEIT**

Net zoals bij de variantieanalyse met één factor, is de berekening van de F-statistieken gebaseerd op voorspelling van observaties. Voor de berekening van de F-statistiek voor het effect van MOBILITEIT hebben we voor elke observatie twee voorspellingen nodig: (1) de voorspelling onder de veronderstelling dat er geen verschil is tussen de twee niveaus van de factor MOBILITEIT (i.c., de groep met en de groep zonder gratis taxi) en (2) de voorspelling onder de veronderstelling dat er wel een verschil is tussen deze twee niveaus. We berekenen een effectkwadratensom door de verschillen tussen deze twee voorspellingen te nemen (observatie per observatie), ze te kwadrateren en op te tellen (over de observaties). En we berekenen een errorkwadratensom door de verschillen tussen de observaties en de voorspellingen onder punt (2) te nemen, ze te kwadrateren en op te tellen. Tot dusver, geen verschil met de variantieanalyse met één factor.

Echter, anders dan bij de variantieanalyse met één factor, beschikken we hier over een tweede factor, BURGERLIJKE STAND, waarvan we gebruik kunnen maken bij de voorspellingen die in de vorige alinea genoemd werden. Concreet, er zijn in totaal vier mogelijke veronderstellingen waaronder men voorspellingen kan maken:

1. Geen effect van MOBILITEIT, geen effect van BURGERLIJKE STAND.
2. Effect van MOBILITEIT, geen effect van BURGERLIJKE STAND.
3. Geen effect van MOBILITEIT, effect van BURGERLIJKE STAND.
4. Effect van MOBILITEIT, effect van BURGERLIJKE STAND.

Om het effect van MOBILITEIT statistisch te toetsen kunnen we (1) de voorspellingen onder 1. en 2. met elkaar vergelijken (geen effect van BURGERLIJKE STAND) of (2) de voorspellingen onder 3. en 4. (wel een effect van BURGERLIJKE STAND). We werken nu beide mogelijkheden verder uit en zullen zien dat wat men toetst met de F-statistiek hetzelfde is voor beide mogelijkheden, maar dat de gevoeligheid van de toets groter is als we rekening houden met het effect van BURGERLIJKE STAND.

**Zonder Rekening te Houden met het Effect van BURGERLIJKE STAND** We laten hier de factor BURGERLIJKE STAND gewoon buiten beschouwing en doen alsof we alleen met de factor MOBILITEIT te maken hebben. In dit geval, zijn de berekeningen hetzelfde als in de variantieanalyse met één factor. Voor deze berekeningen maken we gebruik van de gemiddelden in Tabel 3.21. In deze tabel staat gedeeltelijk dezelfde informatie als in Tabel 3.14. Nu, de resultaten van de berekeningen kunnen we als volgt samenvatten:

1. Onder de veronderstelling dat er geen verschil is tussen de twee niveaus van de factor MOBILITEIT (i.c., de groep met en de groep zonder gratis taxi), is de voorspelling gelijk aan 9.3, het gemiddelde van alle 30 observaties.

2. Onder de veronderstelling dat er wel een verschil is tussen de twee niveaus de factor MOBILITEIT, zijn de voorspellingen gelijk aan de marginale gemiddelden in de groep met en de groep zonder gratis taxi. Deze marginale gemiddelden zijn gelijk aan, respectievelijk, 10.6 en 8.

3. De effectkwadratensom wordt als volgt berekend:

$$\begin{aligned}\text{Effectkwadratensom} &= 15 \times (10.6 - 9.3)^2 + 15 \times (8 - 9.3)^2 \\ &= 25.35 + 25.35 \\ &= 50.7 \quad .\end{aligned}$$

   Deze effectkwadratensom heeft 1 (= 2 − 1) vrijheidsgraad en daarom is de gemiddelde effectkwadratensom ook gelijk aan 50.7.

4. De errorkwadratensom is 181.6. Deze errorkwadratensom heeft 28 (= 30 − 2) vrijheidsgraden en daarom is de gemiddelde errorkwadratensom 6.49 (= 181.6/28). Voor de berekening van de errorkwadratensom heeft men de getallen in Tabel 3.14 nodig. Concreet, voor elk van de getallen in deze tabel berekent men het verschil met het bijbehorende marginale gemiddelde (10.6 of 8), kwadrateert deze verschillen, en telt ze op. Dus, net zoals in de variantieanalyse met één factor.

5. De F-statistiek is 7.81 (= 50.7/6.49). Bij deze F-statistiek hoort een p-waarde van 0.0093. Daarom besluiten we dat de verschillen tussen de gemiddelden niet toevallig zijn.

| | Met partner | Partner verloren | Vrijgezel | Marg. Gem. |
|---|---|---|---|---|
| Met gratis taxi | 8.6 | 11.8 | 11.4 | 10.6 |
| Zonder gratis taxi | 6.8 | 5.8 | 11.4 | 8 |
| Marg. Gem. | 7.7 | 8.8 | 11.4 | 9.3 |

Tabel 3.21: Celgemiddelden en marginale gemiddelden voor de gegevens uit het onderzoek naar het effect van MOBILITEIT en BURGERLIJKE STAND op SOCIALE ACTIVITEIT bij bejaarden.

**Rekening Houdend met het Effect van BURGERLIJKE STAND** Rekening houden met het effect van BURGERLIJKE STAND houdt in dat we nu voorspellingen maken onder de veronderstelling dat er verschillen zijn tussen de drie niveaus van deze

factor. Deze veronderstelling combineren we met elk van twee andere veronderstellingen: (1) dat er geen verschil is tussen de twee niveaus van de factor MOBILITEIT (i.c., de groep met en de groep zonder gratis taxi), en (2) dat er wel een verschil is tussen deze twee niveaus. We maken dus voorspellingen onder de volgende veronderstellingen:

1. Er is alleen een verschil tussen de drie niveaus van BURGERLIJKE STAND (met partner, partner verloren en vrijgezel).
2. Er is een verschil tussen de drie niveaus van BURGERLIJKE STAND én een verschil tussen de twee niveaus van MOBILITEIT.

Voor de voorspelling onder de eerste veronderstelling (alleen een verschil tussen de drie niveaus van BURGERLIJKE STAND) maken we gebruik van de volgende regel:

$$\boxed{\text{voorspelling}} = \boxed{\begin{array}{c}\text{globale}\\ \text{gemiddelde}\end{array}} + \boxed{\begin{array}{c}\text{effect van}\\ \text{BURGERLIJKE}\\ \text{STAND}\end{array}} \quad .$$

Het globale gemiddelde is 9.3. En het effect van BURGERLIJKE STAND voor de groep met partner, de groep partner verloren en de groep vrijgezel is, respectievelijk, $7.7 - 9.3$, $8.8 - 9.3$ en $11.4 - 9.3$. Het effect voor een bepaalde groep is gelijk aan het verschil tussen het marginale gemiddelde van die groep en het globale gemiddelde. Nemen we nu de som van het globale gemiddelde en het effect voor een bepaalde groep (marginale gemiddelde min globale gemiddelde), dan krijgen we het marginale gemiddelde van die groep. Dus, de voorspellingen voor de drie groepen zijn gelijk aan, respectievelijk, 7.7, 8.8 en 11.4.

Voor de voorspelling onder de tweede veronderstelling (ook een verschil tussen de groep met en de groep zonder gratis taxi) maken we gebruik van de volgende regel:

$$\boxed{\text{voorspelling}} = \boxed{\begin{array}{c}\text{globale}\\ \text{gemiddelde}\end{array}} + \boxed{\begin{array}{c}\text{effect van}\\ \text{MOBILITEIT}\end{array}} + \boxed{\begin{array}{c}\text{effect van}\\ \text{BURGERLIJKE}\\ \text{STAND}\end{array}} \quad .$$

Het globale gemiddelde en het effect van BURGERLIJKE STAND worden op dezelfde manier berekend als bij de vorige regel (alleen een verschil tussen de niveaus van BURGERLIJKE STAND). Het effect van MOBILITEIT voor de groep met en de groep zonder gratis taxi is, respectievelijk, $10.6 - 9.3$, en $8 - 9.3$. Het effect voor een bepaalde groep is gelijk aan het verschil tussen het marginale gemiddelde van die groep en het globale gemiddelde.

De voorspellingen op basis van de bovenstaande regel kunnen vereenvoudigd worden. Bijvoorbeeld, de voorspelling voor de groep met gratis taxi (effect= $10.6 - 9.3$) en met partner (effect= $7.7 - 9.3$) kan als volgt geschreven worden:

$$\begin{aligned} 9.3 + (10.6 - 9.3) + (7.7 - 9.3) &= 10.6 + 7.7 + 9.3 - 9.3 - 9.3 \\ &= 10.6 + 7.7 - 9.3 \quad . \end{aligned}$$

We zien dat, na vereenvoudiging, de voorspelling gelijk is aan de volgende som: het marginale gemiddelde van van de groep met gratis taxi, plus het marginale gemiddelde

van de groep met partner, min het globale gemiddelde. Voor de andere cellen in het $2 \times 3$-design vinden we een gelijkaardige som. Deze sommen verschillen van elkaar wat betreft de marginale gemiddelden die in elke som gecombineerd worden. Bijvoorbeeld, voor de cel vrijgezellen zonder gratis taxi is de voorspelling gelijk aan de volgende som: het marginale gemiddelde van de groep zonder gratis taxi, plus het marginale gemiddelde van de groep vrijgezel, min het globale gemiddelde. In het algemeen, kan men deze som als volgt formuleren:

$$\boxed{\text{voorspelling}} = \boxed{\begin{array}{c}\text{marginale}\\ \text{gemiddelde}\\ \text{voor}\\ \text{MOBILITEIT}\end{array}} + \boxed{\begin{array}{c}\text{marginale}\\ \text{gemiddelde voor}\\ \text{BURGERLIJKE}\\ \text{STAND}\end{array}} - \boxed{\begin{array}{c}\text{globale}\\ \text{gemiddelde}\end{array}} .$$

De toepassing van deze som levert de getallen in de derde kolom van Tabel 3.22.

| | | Veronderstelling waaronder de voorspelling gemaakt wordt | | |
|---|---|---|---|---|
| | | Wel een verschil tussen met en zonder gratis taxi | Geen verschil tussen met en zonder gratis taxi | Gekwadrateerd verschil tussen de voorspellingen |
| Met gratis taxi | Met partner | $10.6 + 7.7 - 9.3$ | 7.7 | $(10.6 - 9.3)^2$ |
| | Partner overleden | $10.6 + 8.8 - 9.3$ | 8.8 | $(10.6 - 9.3)^2$ |
| | Vrijgezel | $10.6 + 11.4 - 9.3$ | 11.4 | $(10.6 - 9.3)^2$ |
| Zonder gratis taxi | Met partner | $8 + 7.7 - 9.3$ | 7.7 | $(8 - 9.3)^2$ |
| | Partner overleden | $8 + 8.8 - 9.3$ | 8.8 | $(8 - 9.3)^2$ |
| | Vrijgezel | $8 + 11.4 - 9.3$ | 11.4 | $(8 - 9.3)^2$ |

Tabel 3.22: Voorspellingen onder de veronderstelling van wel en geen effect van MOBILITEIT en hun gekwadrateerde verschil. Bij deze voorspellingen is rekening gehouden met het effect van de factor BURGERLIJKE STAND.

In de vijfde en laatste kolom van Tabel 3.22 staan de gekwadrateerde verschillen tussen de twee voorspellingen; de voorspelling onder de veronderstelling dat er een verschil is tussen de groep met en de groep zonder gratis taxi (in de derde kolom van Tabel 3.22) en de voorspelling onder de veronderstelling dat dit niet het geval is (in de vierde kolom van Tabel 3.22). Het belangrijke punt hier, is dat deze gekwadrateerde verschillen precies gelijk zijn aan de gekwadrateerde verschillen die we ook tegenkomen in de berekening van de effectkwadratensom op pagina 179: de effectkwadratensom van MOBILITEIT onder de veronderstelling dat er *geen* effect is van de factor BURGERLIJKE STAND. Dus, de effectkwadratensom van MOBILITEIT is hetzelfde als men wel of geen rekening houdt met het effect van BURGERLIJKE STAND. Verder zullen we zien dat dit alleen het geval is in een gebalanceerd design; in een ongebalanceerd design hangt de effectkwadratensom wél af van het feit of men wel of geen rekening houdt met het effect van de andere factor.

De berekening van de errorkwadratensom wordt weergegeven in Tabel 3.23. De gekwadrateerde verschillen tussen de observaties en de voorspellingen in de groep met gratis taxi en met partner worden aangeduid met $(y_{11i} - \widehat{y}_{11i}^{A,B})^2$. Hierin staat $\widehat{y}_{11i}^{A,B}$ voor de voorspelde observatie voor deze cel. Met de $A$ en de $B$ in het superscript van $\widehat{y}_{11i}^{A,B}$ wordt aangeduid dat we een voorspelling maken op basis van zowel factor A als factor B. Deze notatie wordt in een van de volgende paragrafen meer in detail uitgelegd. In het bovenstaande hebben we laten zien dat $\widehat{y}_{11i}^{A,B}$ gelijk is aan $(10.6 + 7.7 - 9.3) = 9$. De gekwadrateerde verschillen voor de andere cellen worden op een analoge manier aangeduid: $(y_{12i} - \widehat{y}_{12i}^{A,B})^2$, $(y_{13i} - \widehat{y}_{13i}^{A,B})^2$, ....

De errorkwadratensom is 109.4. Dit is 72.2 minder dan 181.6, de errorkwadratensom na voorspelling op basis van alleen het effect van MOBILITEIT (zie pagina 179). Dus, door bij de voorspelling ook rekening te houden met het effect van BURGERLIJKE STAND (en niet alleen met het effect van MOBILITEIT), wordt de errorkwadratensom kleiner. Verder zullen we zien dat dit altijd het geval is, zowel bij een gebalanceerd als bij een ongebalanceerd design.

| | Met partner | | Partner verloren | | Vrijgezel | | Som |
|---|---|---|---|---|---|---|---|
| | $y_{11i}$ | $(y_{11i} - \widehat{y}_{11i}^{A,B})^2$ | $y_{12i}$ | $(y_{12i} - \widehat{y}_{12i}^{A,B})^2$ | $y_{13i}$ | $(y_{13i} - \widehat{y}_{13i}^{A,B})^2$ | |
| Met gratis taxi | 10 | $(10-9)^2$ | 10 | $(10-10.1)^2$ | 12 | $(12-12.7)^2$ | |
| | 7 | $(7-9)^2$ | 13 | $(13-10.1)^2$ | 10 | $(10-12.7)^2$ | |
| | 9 | $(9-9)^2$ | 8 | $(8-10.1)^2$ | 12 | $(12-12.7)^2$ | |
| | 8 | $(8-9)^2$ | 15 | $(15-10.1)^2$ | 13 | $(13-12.7)^2$ | |
| | 9 | $(9-9)^2$ | 13 | $(13-10.1)^2$ | 10 | $(10-12.7)^2$ | |
| Som | | **6** | | **45.25** | | **15.65** | **66.9** |
| | $y_{21i}$ | $(y_{21i} - \widehat{y}_{21i}^{A,B})^2$ | $y_{22i}$ | $(y_{22i} - \widehat{y}_{22i}^{A,B})^2$ | $y_{23i}$ | $(y_{23i} - \widehat{y}_{23i}^{A,B})^2$ | |
| Zonder gratis taxi | 7 | $(7-6.4)^2$ | 4 | $(4-7.5)^2$ | 11 | $(11-10.1)^2$ | |
| | 6 | $(6-6.4)^2$ | 8 | $(8-7.5)^2$ | 12 | $(12-10.1)^2$ | |
| | 7 | $(7-6.4)^2$ | 5 | $(5-7.5)^2$ | 13 | $(13-10.1)^2$ | |
| | 8 | $(8-6.4)^2$ | 7 | $(7-7.5)^2$ | 11 | $(11-10.1)^2$ | |
| | 6 | $(6-6.4)^2$ | 5 | $(5-7.5)^2$ | 10 | $(10-10.1)^2$ | |
| Som | | **3.6** | | **25.25** | | **13.65** | **42.5** |
| | | | | | | | **109.4** |

Tabel 3.23: Berekening van de errorkwadratensom onder de veronderstelling dat er zowel een effect is van MOBILITEIT als van BURGERLIJKE STAND.

We berekenen nu de gemiddelde kwadratensommen. De gemiddelde effectkwadratensom wordt berekend door de effectkwadratensom te delen door zijn aantal vrijheidsgraden, het aantal niveaus van de factor MOBILITEIT min 1. Dus, de gemiddelde effectkwadratensom is gelijk aan $50.7/(2-1) = 50.7$. Analoog, wordt ook de gemiddelde errorkwadratensom berekend door de errorkwadratensom te delen door zijn aantal vrijheidsgraden. Dit aantal vrijheidsgraden wordt als volgt berekend:

$$\boxed{\text{aantal observaties}} - \left( \boxed{\text{aantal niveaus van MOBILITEIT}} + \boxed{\text{aantal niveaus van BURGERLIJKE STAND}} - 1 \right) .$$

Deze som is gelijk aan $30 - (2 + 3 - 1) = 26$. Dus,

$$\begin{aligned} \text{Gemiddelde errorkwadratensom} &= \frac{109.4}{26} \\ &= 4.21 \quad . \qquad \text{(na afronding)} \end{aligned}$$

In de laatste stap, wordt opnieuw de F-statistiek berekend door de gemiddelde effectkwadratensom te delen door de gemiddelde errorkwadratensom:

$$\begin{aligned} F &= \frac{50.7}{4.21} \\ &= 12.04 \quad . \qquad \text{(na afronding)} \end{aligned}$$

Deze F-statistiek geeft een antwoord op de vraag of het verschil tussen de marginale gemiddelden van de groepen met en zonder gratis taxi (respectievelijk, 10.6 en 8) toevallig is. Dit is precies de vraag die ook beantwoord wordt met de F-statistiek op pagina 179, die berekend werd zonder rekening te houden met het effect van BURGERLIJKE STAND. Dit is in overeenstemming met het feit dat de tellers van deze twee F-statistieken precies gelijk zijn. De teller van een F-statistiek geeft immers aan wat men ermee toetst.

Het verschil tussen de laatste F-statistiek en die op pagina 179 (respectievelijk, 12.04 en 7.81) komt omdat hun gemiddelde errorkwadratensommen verschillend zijn: de gemiddelde errorkwadratensom onder de veronderstelling dat er alleen een effect is van MOBILITEIT (i.e., 6.49) is groter dan de gemiddelde errorkwadratensom onder de veronderstelling dat er ook een effect is van BURGERLIJKE STAND (i.e., 4.21). Hoe kleiner de gemiddelde errorkwadratensom, hoe gevoeliger men toetst. Dus, er wordt gevoeliger getoetst als men ook rekening houdt met de andere factor, BURGERLIJKE STAND. Dit blijkt ook uit de p-waarde van de laatste F-statistiek. Deze is gelijk aan 0.0018, hetgeen kleiner is dan 0.0093, de p-waarde van de F-statistiek op pagina 179, waarin geen rekening gehouden wordt met het effect van BURGERLIJKE STAND.

### Het Effect van BURGERLIJKE STAND

De F-statistieken voor het effect van BURGERLIJKE STAND worden op een analoge manier berekend als die voor het effect van MOBILITEIT. De effectkwadratensom van BURGERLIJKE STAND is hetzelfde als men wel of geen rekening houdt met het effect van MOBILITEIT. In beide gevallen vinden we het volgende:

$$\begin{aligned} \text{Effectkwadratensom} &= 10 \times (7.7 - 9.3)^2 + 10 \times (8.8 - 9.3)^2 + 10 \times (11.4 - 9.3)^2 \\ &= 25.6 + 2.5 + 44.1 \\ &= 72.2 \quad . \end{aligned}$$

De gemiddelde effectkwadratensom is gelijk aan de bovenstaande effectkwadratensom gedeeld door zijn aantal vrijheidsgraden, het aantal niveaus van de factor BURGERLIJKE STAND min 1. De factor BURGERLIJKE STAND heeft drie niveaus en dus moeten we delen door 2. De gemiddelde effectkwadratensom is dus gelijk aan $72.2/2 = 36.1$.

De *error*kwadratensom verandert wél als men rekening houdt met het effect van MOBILITEIT (in vergelijking met de situatie waarin men dit niet doet): als men geen

rekening houdt met het effect van MOBILITEIT, is de errorkwadratensom 160.1, en als men er wel rekening mee houdt is de errorkwadratensom 109.4. Deze laatste errorkwadratensom hebben we reeds berekend, en wel in Tabel 3.23 op pagina 182, want daar werden de voorspellingen gemaakt onder de veronderstelling dat beíde factoren een effect hebben. Men toetst gevoeliger met de errorkwadratensom waarin rekening gehouden wordt met het effect van MOBILITEIT (i.e., 109.4) want deze is kleiner.

Merk op dat het verschil tussen de twee *error*kwadratensommen ($160.1 - 109.4 = 50.7$) precies gelijk is aan de *effect*kwadratensom van MOBILITEIT. Dus, door rekening te houden met het effect van MOBILITEIT verkleint de *error*kwadratensom met een hoeveelheid gelijk aan de *effect*kwadratensom van MOBILITEIT. Dit geldt ook voor de twee errorkwadratensommen die gebruikt worden voor de toetsing van het effect van MOBILITEIT: door rekening te houden met het effect van BURGERLIJKE STAND verkleint de errorkwadratensom van 181.6 naar 109.4 met een hoeveelheid gelijk aan de effectkwadratensom van BURGERLIJKE STAND ($181.6 - 109.4 = 72.2$). Deze gelijkheid geldt algemeen, en dus niet alleen voor dit voorbeeldonderzoek. We komen hier later nog op terug.

De F-statistieken voor BURGERLIJKE STAND zijn gelijk aan 6.09 en 8.57 voor, respectievelijk, de situatie waarin men geen rekening houdt met het effect van MOBILITEIT en de situatie waarin men dit wel doet. Deze F-statistieken corresponderen met de p-waarden 0.0066 en 0.0014. Dus, in beide gevallen besluiten we dat de verschillen tussen de gemiddelden van de drie niveaus van de factor BURGERLIJKE STAND niet toevallig zijn.

#### Overzichtstabel

In Tabel 3.24 geven we een overzicht van de kwadratensommen en de F-statistieken voor het toetsen van de effecten van MOBILITEIT en BURGERLIJKE STAND. We doen dit apart voor de situatie waarin er geen rekening wordt gehouden wordt met het effect van de andere factor en de situatie waarin hiermee wel rekening wordt gehouden.

### 3.3.7 Het Rekenwerk in een Ongebalanceerd Design

#### Het Effect van TRAINING

Net zoals in een gebalanceerd design, maken we ook hier een onderscheid tussen (1) voorspellingen waarin geen rekening wordt gehouden met het effect van de andere factor, TYPE LERAAR, en (2) voorspellingen waarin hier wel rekening mee wordt gehouden. We zullen laten zien dat, anders dan in een gebalanceerd design, zowel de error- als de effectkwadratensom veranderen als men rekening houdt met het effect van de andere factor (in vergelijking met de situatie waarin men er geen rekening mee houdt).

**Zonder Rekening te Houden met het Effect van TYPE LERAAR** We laten hier de factor TYPE LERAAR gewoon buiten beschouwing en doen alsof we alleen met de factor TRAINING te maken hebben. In dit geval, zijn de berekeningen hetzelfde als in de variantieanalyse met één factor. De resultaten van deze berekeningen kunnen we als volgt samenvatten:

| Effect van MOBILITEIT |
|---|
| 1. Zonder rekening te houden met BURGERLIJKE STAND |
| Vergelijking: MOBILITEIT vs. niets<br>Effectkwadratensom: 50.7<br>Errorkwadratensom: 181.6<br>F-statistiek: 7.81 |
| 2. Rekening houdend met BURGERLIJKE STAND |
| Vergelijking: (MOBILITEIT, BURGERLIJKE STAND) vs. BURGERLIJKE STAND<br>Effectkwadratensom: 50.7<br>Errorkwadratensom: 109.4<br>F-statistiek: 12.04 |
| Effect van BURGERLIJKE STAND |
| 1. Zonder rekening te houden met MOBILITEIT |
| Vergelijking: BURGERLIJKE STAND vs. niets<br>Effectkwadratensom: 72.2<br>Errorkwadratensom: 160.1<br>F-statistiek: 6.09 |
| 2. Rekening houdend met MOBILITEIT |
| Vergelijking: (BURGERLIJKE STAND, MOBILITEIT) vs. MOBILITEIT<br>Effectkwadratensom: 72.2<br>Errorkwadratensom: 109.4<br>F-statistiek: 8.57 |

Tabel 3.24: Overzichtstabel van de kwadratensommen en de F-statistieken, voor het toetsen van de effecten van MOBILITEIT en BURGERLIJKE STAND, en dit voor twee situaties: (1) de situatie waarin er geen rekening gehouden wordt met het effect van de andere factor, en (2) de situatie waarin hiermee wel rekening wordt gehouden.

1. Onder de veronderstelling dat er geen verschil is tussen de twee niveaus van de factor TRAINING (i.c., de groep met en de groep zonder training), is de voorspelling gelijk aan 6.28, het gewogen globale gemiddelde (zie Tabel 3.17).

2. Onder de veronderstelling dat er wel een verschil is tussen de twee niveaus de factor TRAINING, zijn de voorspellingen gelijk aan de marginale gemiddelden in de groep met en de groep zonder training. Deze marginale gemiddelden zijn gelijk aan, respectievelijk, 5.28 en 7.53 (zie Tabel 3.17).

3. De effectkwadratensom wordt als volgt berekend:

$$\begin{aligned}\text{Effectkwadratensom} &= 50 \times (5.28 - 6.28)^2 + 40 \times (7.53 - 6.28)^2 \\ &= 50 + 62.5 \\ &= 112.5 \quad .\end{aligned}$$

   Deze effectkwadratensom heeft 1 $(= 2 - 1)$ vrijheidsgraad en daarom is de gemiddelde effectkwadratensom ook gelijk aan 112.5.

4. De errorkwadratensom is 893.556. Deze errorkwadratensom heeft 88 $(= 90-2)$ vrijheidsgraden en daarom is de gemiddelde errorkwadratensom 10.15 $(= 893.556/88)$. We geven deze errorkwadratensom zonder motivering omdat de individuele observaties in dit onderzoek niet gegeven zijn. De berekening is echter hetzelfde als in de variantieanalyse met één factor.

5. De F-statistiek is 11.08 $(= 112.5/10.15)$. Bij deze F-statistiek hoort een p-waarde van 0.0013. Daarom besluiten we dat de verschillen tussen de gemiddelden niet toevallig zijn.

**Rekening Houdend met het Effect van TYPE LERAAR** We maken nu voorspellingen onder de veronderstelling dat er verschillen zijn tussen de drie niveaus van TYPE LERAAR. Deze veronderstelling combineren we met twee andere veronderstellingen: (1) dat er geen verschil is tussen de twee niveaus van de factor TRAINING (i.c., de groep met en de groep zonder training), en (2) dat er wel een verschil is tussen deze twee niveaus. We maken dus voorspellingen onder de volgende veronderstellingen:

1. Er is alleen een verschil tussen de drie niveaus van TYPE LERAAR (actief, reactief en passief).

2. Er is een verschil tussen de drie niveaus van TYPE LERAAR én een verschil tussen de twee niveaus van TRAINING.

Voor de voorspelling onder de eerste veronderstelling (alleen een verschil tussen de drie niveaus van TYPE LERAAR) maken we gebruik van de volgende regel:

$$\boxed{\text{voorspelling}} = \boxed{\begin{array}{c}\text{globale}\\ \text{gemiddelde}\end{array}} + \boxed{\begin{array}{c}\text{effect van}\\ \text{TYPE LERAAR}\end{array}} \quad .$$

De gemiddelden die we nodig hebben om deze voorspelling te berekenen vinden we in Tabel 3.17. Nu, het globale gemiddelde is 6.28. En het effect van TYPE LERAAR voor

de groep met een actieve, de groep met een reactieve, en de groep met een passieve leraar is, respectievelijk, $3.4-6.28$, $6-6.28$ en $10.64-6.28$. Het effect voor een bepaalde groep is gelijk aan het verschil tussen het marginale gemiddelde van die groep en het globale gemiddelde. Nemen we nu de som van het globale gemiddelde en het effect voor een bepaalde groep, dan krijgen we als voorspelling het marginale gemiddelde van die groep. Dus, de voorspellingen voor de drie groepen zijn gelijk aan, respectievelijk, 3.4, 6 en 10.64.

Voor de voorspelling onder de tweede veronderstelling (ook een verschil tussen de groep met en de groep zonder training) maken we gebruik van de volgende regel:

$$\boxed{\text{voorspelling}} = \boxed{\begin{array}{c}\text{globale}\\ \text{gemiddelde}\end{array}} + \boxed{\begin{array}{c}\text{effect van}\\ \text{TRAINING}\end{array}} + \boxed{\begin{array}{c}\text{effect van}\\ \text{TYPE LERAAR}\end{array}} .$$

Anders dan in een gebalanceerd design, beschikken we hier niet over eenvoudige formules voor het effect van TRAINING en het effect van TYPE LERAAR. Deze effecten kunnen we echter wel berekenen met de computer. Als we deze effecten dan invullen in de bovenstaande formule, dan krijgen we de volgende voorspellingen: (1) 3.396, 5.991 en 10.628 voor leerlingen die de training volgden en, respectievelijk, een actieve, een reactieve en een passieve leraar hadden, en (2) 3.4133, 6.0083 en 10.6453 voor leerlingen die geen training volgden en, respectievelijk, een actieve, een reactieve en een passieve leraar hadden.

De berekening van de effectkwadratensom van TRAINING wordt weergegeven in Tabel 3.25. De gekwadrateerde verschillen tussen de twee voorspellingen (met en zonder effect van TRAINING) worden aangeduid met $(\widehat{y}_{11}^{A,B} - \overline{y}_{\sqcup 1})^2$, $(\widehat{y}_{12}^{A,B} - \overline{y}_{\sqcup 2})^2$, .... De effectkwadratensom is gelijk aan 0.0055. Dit is 112.4945 minder dan 112.5, de effectkwadratensom van TRAINING zonder rekening te houden met het effect van TYPE LERAAR. Het grote verschil tussen deze twee effectkwadratensommen is toe te schrijven aan het feit dat de gekwadrateerde verschillen tussen de twee voorspellingen (met en zonder effect van TRAINING) veel kleiner zijn als men rekening houdt met het effect van TYPE LERAAR dan wanneer men dit niet doet. Als men bij de voorspelling rekening houdt met het effect van TYPE LERAAR, dan doet het er nog maar weinig toe of men ook nog rekening houdt met het effect van TRAINING.

| | Actieve Leraar | | Reactieve Leraar | | Passieve Leraar | | Som |
|---|---|---|---|---|---|---|---|
| Met Training | $n_{11}$ | $(\widehat{y}_{11}^{A,B} - \overline{y}_{\sqcup 1})^2$ | $n_{12}$ | $(\widehat{y}_{12}^{A,B} - \overline{y}_{\sqcup 2})^2$ | $n_{13}$ | $(\widehat{y}_{13}^{A,B} - \overline{y}_{\sqcup 3})^2$ | |
| | 28 | $(3.396-3.4)^2$ | 14 | $(5.991-6)^2$ | 8 | $(10.628-10.64)^2$ | |
| Som | | **0.0004** | | **0.0011** | | **0.0011** | **0.0026** |
| Zonder Training | $n_{21}$ | $(\widehat{y}_{21}^{A,B} - \overline{y}_{\sqcup 1})^2$ | $n_{22}$ | $(\widehat{y}_{22}^{A,B} - \overline{y}_{\sqcup 2})^2$ | $n_{23}$ | $(\widehat{y}_{23}^{A,B} - \overline{y}_{\sqcup 3})^2$ | |
| | 7 | $(3.4133-3.4)^2$ | 16 | $(6.0083-6)^2$ | 17 | $(10.6453-10.64)^2$ | |
| Som | | **0.0012** | | **0.0011** | | **0.0005** | **0.0028** |
| | | | | | | | **<u>0.0055</u>** |

Tabel 3.25: Berekening van de effectkwadratensom voor TRAINING onder de veronderstelling dat er een effect is van TYPE LERAAR.

De effectkwadratensom die we in Tabel 3.25 berekend hebben, wordt ook een *uitgezuiverde* kwadratensom genoemd: het is de effectkwadratensom van TRAINING uitgezuiverd voor TYPE LERAAR. Met deze effectkwadratensom toetst men de nulhypothese dat de ongewogen verwachte waarden (i.e., de $\mu_{k\bullet}$'s) gelijk zijn. Het valt op dat deze effectkwadratensom bijna gelijk is aan 0. Deze uitgezuiverde effectkwadratensom ligt zo kort bij 0 omdat de ongewogen gemiddelden in de steekproef (i.e., $\overline{y}_{1\bullet}$ en $\overline{y}_{2\bullet}$) precies gelijk zijn. Dat deze uitgezuiverde effectkwadratensom toch niet precies gelijk is aan 0 komt omdat er, in de steekproef, een interactie is tussen de twee factoren; indien deze interactie er niet zou zijn, dan zou deze effectkwadratensom precies gelijk zijn aan 0. We komen hier nog op terug.

Omdat we voor het ongebalanceerde voorbeeldonderzoek niet de individuele observaties gegeven hebben, geven we de errorkwadratensom zonder verdere motivering: hij is gelijk aan 238.1545. Dit is 655.4015 minder dan 893.556, de errorkwadratensom indien men geen rekening houdt met het effect van TYPE LERAAR. De gemiddelde errorkwadratensom is gelijk aan deze errorkwadratensom gedeeld door zijn aantal vrijheidsgraden, het aantal observaties min het aantal niveaus van factor TRAINING, min het aantal niveaus van factor TYPE LERAAR, plus 1:

$$90 - (2 + 3 - 1) = 90 - 2 - 3 + 1 = 86 \quad .$$

De gemiddelde errorkwadratensom wordt dus als volgt berekend:

$$\begin{aligned} \text{Gemiddelde errorkwadratensom} &= \frac{238.1545}{86} \\ &= 2.77 \quad . \qquad \text{(na afronding)} \end{aligned}$$

In de laatste stap, wordt opnieuw de F-statistiek berekend door de gemiddelde effectkwadratensom ($0.0055/1 = 0.0055$) te delen door de gemiddelde errorkwadratensom:

$$\begin{aligned} F &= \frac{0.0055}{2.77} \\ &= 0.002 \quad . \qquad \text{(na afronding)} \end{aligned}$$

Met deze F-statistiek kan men de nulhypothese toetsen dat er geen verschil is tussen de ongewogen marginale verwachte waarden $\mu_{k\bullet}$. De p-waarde van deze F-statistiek is gelijk aan 0.964, hetgeen groter is dan 0.05, en daarom behouden we deze nulhypothese.

Het verschil tussen de laatste F-statistiek en die op pagina 186 (respectievelijk, 0.002 en 11.08) is zowel toe te schrijven aan de gemiddelde effectkwadratensom in de teller als aan de gemiddelde errorkwadratensom in de noemer: beide gemiddelde kwadratensommen zijn kleiner indien men rekening houdt met het effect van TYPE LERAAR dan wanneer men dit niet doet. We weten dat de teller van een F-statistiek aangeeft wat men ermee toetst en de noemer hoe gevoelig men dit doet. Dus, met de laatste F-statistiek toetst men iets anders dan met de F-statistiek op pagina 186 (namelijk, het verschil tussen de ongewogen in plaats van het verschil tussen de gewogen marginale gemiddelden) en men toetst dit gevoeliger.

### Het Effect van TYPE LERAAR

Omdat we later de effectkwadratensommen van TYPE LERAAR nog nodig hebben, berekenen we nu de F-statistieken voor het effect van deze factor. We zijn echter niet geïnteresseerd in de F-statistiek waarmee men het verschil tussen de ongewogen gemiddelden van de verschillende niveaus van TYPE LERAAR toetst. De onderzoeksvraag die centraal staat in het voorbeeldonderzoek is "Heeft de training een effect?". Voor de beantwoording van deze onderzoeksvraag moet men de effectkwadratensom van de onafhankelijke variabele TRAINING uitzuiveren voor de storende variabele TYPE LERAAR. Met uitzuiveren in de omgekeerde volgorde (de effectkwadratensom van de storende variabele TYPE LERAAR uitzuiveren voor de onafhankelijke variabele TRAINING) wordt geen interessante onderzoeksvraag beantwoord.

De F-statistieken voor het effect van TYPE LERAAR worden op vrijwel dezelfde manier berekend als die voor het effect van TRAINING. De effectkwadratensom van TYPE LERAAR hangt af van het feit of men wel of geen rekening houdt met het effect van TRAINING. Als men geen rekening houdt met het effect van TRAINING, dan wordt de effectkwadratensom van TYPE LERAAR als volgt berekend:

$$\begin{aligned}\text{Effectkwadratensom} &= 35 \times (3.4 - 6.28)^2 + 30 \times (6 - 6.28)^2 + 25 \times (10.64 - 6.28)^2 \\ &= 290.304 + 2.352 + 475.24 \\ &= 767.896 \quad .\end{aligned}$$

Als men wel rekening houdt met het effect van TRAINING, dan is de effectkwadratensom van TYPE LERAAR gelijk aan 655.4015. Voor deze effectkwadratensom bestaat geen eenvoudige formule, maar met behulp van de computer kan men hem snel uitrekenen. We zien dat, door rekening te houden met het effect van TRAINING, de effectkwadratensom van TYPE LERAAR verkleint van 767.896 naar 655.4015.

Ook de errorkwadratensom is anders als men rekening houdt met het effect van TRAINING, zij het dat het verschil minimaal is: als men geen rekening houdt met het effect van TRAINING is de errorkwadratensom 238.16, en als men er wel rekening mee houdt is de errorkwadratensom 238.1545. Deze laatste errorkwadratensom werd ook gebruikt voor het toetsen van het verschil tussen de ongewogen marginale gemiddelden voor de twee niveaus van de factor TRAINING. Deze laatste errorkwadratensom is kleiner dan de eerste, zij het dat het verschil heel erg klein is (i.c., $238.16 - 238.1545 = 0.0055$). Dit verschil is precies gelijk aan de effectkwadratensom van TRAINING, en met name, de voor TYPE LERAAR uitgezuiverde effectkwadratensom. Dus, door rekening te houden met het effect van TRAINING verkleint de errorkwadratensom met een hoeveelheid gelijk aan de voor TYPE LERAAR uitgezuiverde effectkwadratensom van TRAINING. Dit geldt ook voor de twee errorkwadratensommen die gebruikt worden voor de toetsing van het effect van TRAINING: door rekening te houden met het effect van TYPE LERAAR verkleint de errorkwadratensom van 893.556 naar 238.1545 met een hoeveelheid gelijk aan de voor TRAINING uitgezuiverde effectkwadratensom van TYPE LERAAR ($893.556 - 238.1545 = 655.4015$).

De F-statistieken voor TYPE LERAAR zijn gelijk aan 140.256 en 118.304 voor, respectievelijk, de situatie waarin men niet uitzuivert voor TRAINING en de situatie waarin men dit wel doet. Deze F-statistieken corresponderen beide met een extreem lage p-waarde (kleiner dan 0.00000000001). Op basis van de p-waarde van de eerste F-statistiek (geen uitzuivering voor TRAINING) verwerpen we de nulhypothese dat er geen

verschillen zijn tussen de gewogen marginale verwachte waarden voor de drie niveaus van TYPE LERAAR (i.e., de $\mu_{\sqcup l}$'s). En op basis van de p-waarde van de tweede F-statistiek (wel uitzuivering voor TRAINING) verwerpen we de nulhypothese dat er geen verschillen zijn tussen de ongewogen marginale verwachte waarden voor de drie niveaus van TYPE LERAAR (i.e., de $\mu_{\bullet l}$'s).

**Overzichtstabel**

In Tabel 3.26 geven we een overzicht van de kwadratensommen en de F-statistieken voor het toetsen van de effecten van TRAINING en TYPE LERAAR. We doen dit apart voor de situatie waarin er geen rekening wordt gehouden wordt met het effect van de andere factor en de situatie waarin hiermee wel rekening wordt gehouden.

| Effect van TRAINING |
|---|
| 1. Zonder rekening te houden met TYPE LERAAR |
| Vergelijking: TRAINING vs. niets<br>Effectkwadratensom: 112.5<br>Errorkwadratensom: 893.556<br>F-statistiek: 11.08 |
| 2. Rekening houdend met TYPE LERAAR |
| Vergelijking: (TRAINING, TYPE LERAAR) vs. TYPE LERAAR<br>Effectkwadratensom: 0.0055<br>Errorkwadratensom: 238.1545<br>F-statistiek: 0.002 |
| Effect van TYPE LERAAR |
| 1. Zonder rekening te houden met TRAINING |
| Vergelijking: TYPE LERAAR vs. niets<br>Effectkwadratensom: 767.896<br>Errorkwadratensom: 238.16<br>F-statistiek: 140.256 |
| 2. Rekening houdend met TRAINING |
| Vergelijking: (TYPE LERAAR, TRAINING) vs. TRAINING<br>Effectkwadratensom: 655.4015<br>Errorkwadratensom: 238.1545<br>F-statistiek: 118.304 |

Tabel 3.26: Overzichtstabel van de kwadratensommen en de F-statistieken voor het toetsen van de effecten van TRAINING en TYPE LERAAR, en dit voor twee situaties: (1) de situatie waarin er geen rekening gehouden wordt met het effect van de andere factor, en (2) de situatie waarin hiermee wel rekening wordt gehouden.

### 3.3.8 F-statistieken en Kwadratensommen

We geven nu de formules die nodig zijn om de F-statistieken ook voor andere gegevens uit te rekenen. Bij het bestuderen van deze algemene formules is het goed om terug te bladeren naar de pagina's met de concrete berekeningen voor de voorbeeldonderzoeken.

### De F-statistieken

Alle nulhypotheses worden getoetst met de algemene F-statistiek:

$$\boxed{F = \frac{MS_{Effect|Ref}}{MS_{Error|Model}} \quad .}$$

In deze formule kunnen we voor *Effect* zowel het effect van A als het effect van B invullen. Dit effect moet men altijd nader specificeren door aan te geven wat het zogenaamde referentiemodel is (aangeduid met *Ref*). Indien dit referentiemodel het model zonder effecten is, dan toetst men met deze F-statistiek een nulhypothese die betrekking heeft op gewogen verwachte waarden (de gewogen verwachte waarden van de niveaus van A bij *Effect*=A en de gewogen verwachte waarden van de niveaus van B bij *Effect*=B). Dus, in dit geval wordt niet uitgezuiverd voor een andere factor. Indien het referentiemodel het model is met de andere factor (factor B bij *Effect*=A en factor A bij *Effect*=B) dan toetst men met deze F-statistiek een nulhypothese die betrekking heeft op ongewogen verwachte waarden. In dit geval, wordt dus wel uitgezuiverd voor een andere factor. We komen hier nog op terug.

De noemer van de F-statistiek, de gemiddelde errorkwadratensom $MS_{Error|Model}$, bepaalt de gevoeligheid waarmee men toetst. Afhankelijk van het model dat men gebruikt om de observaties te voorspellen, krijgt men een andere errorkwadratensom. Men herkent deze verschillende errorkwadratensommen aan wat er ingevuld wordt voor *Model* in het subscript van $MS_{Error|Model}$. Het spreekt voor zich dat men het model kiest met de kleinste $MS_{Error|Model}$, want dit resulteert in de meest gevoelige toets. Ook hier komen we nog op terug.

### De Effectkwadratensommen

De effectkwadratensom staat in de teller van de F-statistiek en bepaalt wat men toetst met deze F-statistiek. De effectkwadratensom wordt aangeduid met het symbool $SS_{Effect|Ref}$ en wordt als volgt berekend:

$$\boxed{SS_{Effect|Ref} = \sum_{Obs} (\widehat{y}_{kli}^{Ref+Effect} - \widehat{y}_{kli}^{Ref})^2 \quad ,}$$

waarin de som loopt over alle observaties. In het rechterlid van deze formule worden de voorspellingen op basis van twee verschillende modellen met elkaar vergeleken: (1) het referentiemodel samen met het effect ($\widehat{y}_{kli}^{Ref+Effect}$) en (2) het referentiemodel zonder het effect ($\widehat{y}_{kli}^{Ref}$). Met *Effect* geven we de factor waarvan we het effect willen evalueren, en met *Ref* geven we aan voor welke factor we uitzuiveren.

We beginnen met de effectkwadratensom van A, ook kortweg de kwadratensom van A genoemd. In de bovenstaande algemene formule van de effectkwadratensom vervangen we *Effect* door A en krijgen dan we de volgende formule van de kwadratensom van A:

$$SS_{A|Ref} = \sum_{Obs} (\widehat{y}_{kli}^{Ref+A} - \widehat{y}_{kli}^{Ref})^2 \quad .$$

Met het subscript $A|Ref$ wordt aangegeven dat de kwadratensom van A afhangt van het referentiemodel. Er zijn twee mogelijke referentiemodellen: (1) het model zonder effecten en (2) het één-factoriële model met B. Er zijn dan ook twee verschillende effectkwadratensommen: (1) de *type I* kwadratensom, die berekend wordt door de voorspellingen op basis van het model met alleen A te vergelijken met de voorspellingen op basis van het model zonder effecten (het referentiemodel), en (2) de *type II* kwadratensom, die berekend wordt door de voorspellingen op basis van het model met A én B te vergelijken met de voorspellingen op basis van het model met alleen B (het referentiemodel). Met de type I kwadratensom wordt er niet uitgezuiverd voor factor B, en met de type II kwadratensom gebeurt dit wel. De type I kwadratensom wordt gebruikt om het verschil tussen de gewogen marginale gemiddelden te toetsen, en de type II kwadratensom om het verschil tussen de ongewogen marginale gemiddelden te toetsen.

Het verhaal voor de effectkwadratensom van B is hetzelfde, maar met de rollen van A en B omgewisseld. In wat volgt concentreren we ons op de effectkwadratensom van A.

**De Type I Effectkwadratensom** De type I effectkwadratensom wordt berekend met het model zonder effecten als referentiemodel. Dit model zonder effecten is het volgende:

$$\boxed{y_{kli} = \mu + e_{kli} \quad .}$$

De ruisterm $e_{kli}$ heeft een verwachte waarde gelijk aan 0 in elke cel $(k, l)$. Dit model is precies hetzelfde als het referentiemodel voor de variantieanalyse met één factor. De verwachte waarde van de observaties $y_{kli}$ in cel $(k, l)$ (i.e., het gemiddelde in de populatie van alle personen die tot deze cel behoren) is $\mu$. Deze verwachte waarde is onafhankelijk van $k$ en van $l$. Met andere woorden, volgens dit model hangen de verwachte waarden in de cellen noch van factor A noch van factor B af.

Bij de berekening van de type I kwadratensom wordt dit model zonder effecten vergeleken met het één-factoriële model met A:

$$\boxed{y_{kli} = \mu + \alpha_k + e_{kli} \quad .}$$

Dit model is precies hetzelfde als het variantieanalysemodel met één factor. De verwachte waarde van de observaties $y_{kli}$ in cel $(k, l)$ is $\mu+\alpha_k$. Deze verwachte waarde is afhankelijk van $k$ en onafhankelijk van $l$. Dus, volgens dit model hangen de verwachte waarden in de cellen alleen af van factor A.

De type I kwadratensom van A wordt berekend door de voorspellingen op basis van het model met alleen A, aangeduid met $\widehat{y}^{A}_{kli}$, te vergelijken met de voorspellingen op basis van het model zonder effecten, aangeduid met $\widehat{y}^{niks}_{kli}$. In termen van de ingrediënten van de algemene formule van de effectkwadratensom:

$$\begin{aligned} \widehat{y}^{Ref}_{kli} &= \widehat{y}^{niks}_{kli} \qquad \text{en} \\ \widehat{y}^{Ref+A}_{kli} &= \widehat{y}^{A}_{kli} \quad . \end{aligned}$$

De voorspellingen $\widehat{y}_{kli}^{niks}$ en $\widehat{y}_{kli}^{A}$ worden gemaakt op basis van het kleinste-kwadratencriterium. Toepassing van dit criterium levert het volgende:

$$\widehat{y}_{kli}^{niks} = \overline{y}_{\sqcup\sqcup}, \quad \text{het gemiddelde van alle } N \text{ observaties}$$
$$\widehat{y}_{kli}^{A} = \overline{y}_{k\sqcup}, \quad \text{het gemiddelde van de observaties in het k-de niveau van A}$$

De formule van de type I kwadratensom van A is dan de volgende:

$$SS_A = \sum_{Obs} (\overline{y}_{k\sqcup} - \overline{y}_{\sqcup\sqcup})^2 \quad ,$$

waarin $SS_{A|niks}$ is vervangen door de verkorte notatie $SS_A$.

We kunnen $SS_A$ wat preciezer uitdrukken door de som over alle observaties te schrijven in termen van drie somtekens: één over de rijen van de datamatrix (met index $k$), één over de kolommen (met index $l$), en één over de observaties binnen de cellen (met index $i$). We krijgen dan de volgende formule van $SS_A$:

$$SS_A = \sum_{k=1}^{K} \sum_{l=1}^{L} \sum_{i=1}^{n_{kl}} (\overline{y}_{k\sqcup} - \overline{y}_{\sqcup\sqcup})^2 \quad .$$

Omdat $(\overline{y}_{k\sqcup} - \overline{y}_{\sqcup\sqcup})^2$ niet afhangt van $l$ en $i$, en er binnen het k-de niveau van factor A $n_{k+}$ observaties zijn, volgt hieruit:

$$\boxed{SS_A = \sum_{k=1}^{K} n_{k+} (\overline{y}_{k\sqcup} - \overline{y}_{\sqcup\sqcup})^2 \quad .}$$

Deze formule van de type I kwadratensom $SS_A$ is zowel geldig in een gebalanceerd als in een ongebalanceerd design. In een gebalanceerd design kan deze formule ook als volgt geschreven worden:

$$SS_A = \sum_{k=1}^{K} Ln (\overline{y}_{k\bullet} - \overline{y}_{\bullet\bullet})^2 \quad .$$

Dit volgt uit de vorige formule van $SS_A$ omdat, in een gebalanceerd design, (1) het aantal observaties in het k-de niveau van A ($n_{k+}$) gelijk is aan $Ln$ (het aantal niveaus van B maal het aantal observaties per cel), en (2) de gewogen gemiddelden gelijk zijn aan de corresponderende ongewogen gemiddelden.

De type I kwadratensom van B wordt op een analoge manier berekend als die van A. De formules voor de type I kwadratensom van B zijn dezelfde als die voor de kwadratensom van A, behalve dat er (1) een B komt te staan waar een A staat, (2) een $l$ waar een $k$ staat (en omgekeerd), en (3) een $L$ waar een $K$ staat (en omgekeerd). De formules worden gegeven in de overzichtstabel 3.31.

**De Type II Effectkwadratensom** De type II effectkwadratensom van A wordt berekend met als referentiemodel het model met B. Dit referentiemodel is het volgende:

$$\boxed{y_{kli} = \mu + \beta_l + e_{kli} \quad .}$$

In het gebalanceerde voorbeeldonderzoek is factor A MOBILITEIT en wordt $SS_A$ als volgt berekend:

$$\begin{aligned} SS_A &= 15 \times (10.6 - 9.3)^2 + 15 \times (8 - 9.3)^2 \\ &= 50.7 \quad . \end{aligned}$$

In het ongebalanceerde voorbeeldonderzoek is factor A TRAINING en wordt $SS_A$ als volgt berekend:

$$\begin{aligned} SS_A &= 50 \times (5.28 - 6.28)^2 + 40 \times (7.53 - 6.28)^2 \\ &= 112.5 \quad . \end{aligned}$$

De $\beta_l$ in deze formule is de effectparameter van het l-de niveau van B. Net zoals bij de effectparameters $\alpha_k$, is ook de som van de effectparameters $\beta_l$ gelijk aan 0. Dit model is de tegenhanger van het één-factoriële model met A: de verwachte waarden in de cellen hangen nu alleen af van factor B en niet van factor A. De verwachte waarde in cel $(k, l)$ is $\mu + \beta_l$, hetgeen afhankelijk is van $l$ en onafhankelijk van $k$. Dus, volgens dit model hangen de verwachte waarden in de cellen alleen af van factor B.

Bij de berekening van de type II effectkwadratensom wordt dit één-factoriële model met B vergeleken met het twee-factoriële model met A én B. Dit twee-factoriële model is het volgende:

$$\boxed{y_{kli} = \mu + \alpha_k + \beta_l + e_{kli} \quad .}$$

De verwachte waarde in cel $(k, l)$ is $\mu + \alpha_k + \beta_l$, hetgeen afhangt van $k$ en $l$. Volgens dit model hangen de verwachte waarden in de cellen af van beide factoren.

De type II effectkwadratensom van A wordt berekend door de voorspellingen op basis van het model met A én B, aangeduid met $\widehat{y}_{kli}^{A,B}$, te vergelijken met de voorspellingen op basis van het model met alleen B, aangeduid met $\widehat{y}_{kli}^{B}$. In termen van de ingrediënten van de algemene formule van de effectkwadratensom:

$$\begin{aligned} \widehat{y}_{kli}^{Ref} &= \widehat{y}_{kli}^{B} \qquad \text{en} \\ \widehat{y}_{kli}^{Ref+A} &= \widehat{y}_{kli}^{A,B} \quad . \end{aligned}$$

De voorspellingen $\widehat{y}_{kli}^{B}$ en $\widehat{y}_{kli}^{A,B}$ worden gemaakt op basis van het kleinste-kwadratencriterium. De voorspelling $\widehat{y}_{kli}^{B}$ is gelijk aan $\overline{y}_{\sqcup l}$, het gemiddelde van de $n_{+l}$ observaties in het l-de niveau van B. Bij de andere voorspelling, $\widehat{y}_{kli}^{A,B}$ (gemaakt op basis van het twee-factoriële model met A én B), stuiten we op het probleem dat er geen eenvoudige formule bestaat voor deze kleinste-kwadratenschatter (dit geldt alleen voor een ongebalanceerd design). In de praktijk wordt $\widehat{y}_{kli}^{A,B}$ altijd berekend met de computer. De formule van de type II kwadratensom van A is dan de volgende:

$$\boxed{SS_{A|B} = \sum_{Obs} (\widehat{y}_{kli}^{A,B} - \overline{y}_{\sqcup l})^2 \quad .}$$

Merk op dat $SS_{A|Ref}$ hier geschreven wordt als $SS_{A|B}$ omdat het referentiemodel de factor B bevat. Hierin moet $|B$ gelezen worden als “na voorspelling op basis van B” of “uitgezuiverd voor B”.

Deze formule van de type II kwadratensom $SS_{A|B}$ is zowel geldig in een gebalanceerd als in een ongebalanceerd design. In een gebalanceerd design kan deze formule ook als volgt geschreven worden:

$$SS_{A|B} = \sum_{k=1}^{K} Ln(\overline{y}_{k\bullet} - \overline{y}_{\bullet\bullet})^2 \quad .$$

De afleiding van deze formule wordt gegeven in de appendix op pagina 292. We zien dat, in een gebalanceerd design, de type II kwadratensom $SS_{A|B}$ en de type I kwadratensom $SS_A$, gegeven op pagina 193, gelijk zijn aan elkaar. Dit is een belangrijk resultaat omdat $SS_{A|B}$ en $SS_A$ de teller van de F-statistiek bepalen, en wat er in deze teller staat bepaalt wat er getoetst wordt. Het feit dat, in een gebalanceerd design, $SS_{A|B}$ gelijk is aan $SS_A$ impliceert dat wat er met een F-statistiek getoetst wordt onafhankelijk is van het feit of men uitzuivert voor een andere factor of niet. Deze conclusie is in overeenstemming met het feit dat, in een gebalanceerd design, de gewogen en de ongewogen marginale gemiddelden gelijk zijn. Een F-statistiek met een type I kwadratensom in de teller toetst de nulhypothese van gelijke gewogen marginale verwachte waarden en de F-statistiek met een type II kwadratensom in de teller toetst de nulhypothese van gelijke ongewogen marginale verwachte waarden. Nu, in een gebalanceerd design is er geen verschil tussen deze twee nulhypotheses en dit stemt overeen met het feit dat ook de tellers van de twee F-statistieken gelijk zijn.

In een ongebalanceerd design zijn de type I en de type II kwadratensommen bijna altijd ongelijk. Dit kunnen we illustreren aan de hand van het voorbeeldonderzoek: de type I kwadratensom van factor A, TRAINING, is 112.5 en zijn type II kwadratensom (na uitzuivering voor TYPE LERAAR) is 0.0055. Dus, met de F-statistiek met de type I kwadratensom wordt iets anders getoetst dan met de F-statistiek met de type II kwadratensom: de F-statistiek met de type I kwadratensom toetst de nulhypothese van gelijke gewogen marginale verwachte waarden en de F-statistiek met de type II kwadratensom toetst de nulhypothese van gelijke ongewogen marginale verwachte waarden. In een ongebalanceerd design verschillen deze twee nulhypotheses van elkaar, want de gewogen en de ongewogen marginale verwachte waarden zijn verschillend, en dit stemt overeen met het feit dat ook de tellers van de twee F-statistieken van elkaar verschillen.

De type II kwadratensom van B wordt op een analoge manier berekend als die van A. De formules voor de type II kwadratensom van B zijn dezelfde als die voor de type II kwadratensom van A, behalve dat er (1) een B komt te staan waar een A staat (en omgekeerd), (2) een $l$ waar een $k$ staat (en omgekeerd), en (3) een $L$ waar een $K$ staat (en omgekeerd). De formules worden gegeven in de overzichtstabel 3.31.

### De Errorkwadratensommen

De errorkwadratensom is een maat voor de fouten die we maken als we de observaties voorspellen op basis van één of ander model. De errorkwadratensom staat in de noemer van de F-statistiek en bepaalt hoe gevoelig men toetst met deze F-statistiek. De errorkwadratensom wordt aangeduid met het symbool $SS_{Error|Model}$ en wordt als volgt berekend:

$$\boxed{SS_{Error|Model} = \sum_{Obs} (y_{kli} - \widehat{y}_{kli}^{Model})^2 \quad .}$$

Dus, voor de berekening van $SS_{Error|Model}$ moeten we de observaties $y_{kli}$ vergelijken met de voorspellingen $\widehat{y}_{kli}^{Model}$. Het superscript *Model* in $\widehat{y}_{kli}^{Model}$ geeft aan dat er, afhankelijk van het model, een andere voorspelling gemaakt wordt. Er zijn drie mogelijke voorspellingen waarmee de observaties vergeleken kunnen worden:

1. Voorspellingen op basis van het model met alleen het effect van A, eerder aangeduid met $\widehat{y}_{kli}^{A}$. Deze voorspellingen zijn gelijk aan $\overline{y}_{k\sqcup}$. Vervangen we $\widehat{y}_{kli}^{Model}$ in de algemene formule van de errorkwadratensom door $\overline{y}_{k\sqcup}$, dan krijgen we de formule van $SS_{Error|A}$.

2. Voorspellingen op basis van het model met alleen het effect van B, eerder aangeduid met $\widehat{y}_{kli}^{B}$. Deze voorspellingen zijn gelijk aan $\overline{y}_{\sqcup l}$. Vervangen we $\widehat{y}_{kli}^{Model}$ in de algemene formule van de errorkwadratensom door $\overline{y}_{\sqcup l}$, dan krijgen we de formule van $SS_{Error|B}$.

3. Voorspellingen op basis van het model met het effect van A én het effect van B, eerder aangeduid met $\widehat{y}_{kli}^{A,B}$. Voor een ongebalanceerd design, bestaat er geen eenvoudige formule voor de voorspellingen $\widehat{y}_{kli}^{A,B}$. Voor een gebalanceerd design, is $\widehat{y}_{kli}^{A,B}$ gelijk aan $\overline{y}_{k\bullet} + \overline{y}_{\bullet l} - \overline{y}_{\bullet\bullet}$. (Dit wordt afgeleid in de appendix op pagina 292.) Vervangen we $\widehat{y}_{kli}^{Model}$ in de algemene formule van de errorkwadratensom door deze voorspelling, dan krijgen we de formule van $SS_{Error|A,B}$ voor een gebalanceerd design.

In het bovenste deel van Tabel 3.27 staan de errorkwadratensommen voor het gebalanceerde en het ongebalanceerde voorbeeldonderzoek. We zien dat, voor beide voorbeeldonderzoeken, $SS_{Error|A,B}$ kleiner is dan $SS_{Error|A}$ en $SS_{Error|B}$. Dit geldt algemeen, en dus niet alleen voor deze twee voorbeeldonderzoeken. In de volgende paragraaf komen we hierop terug.

| | $SS_{Totaal}$ | $SS_{Error\|A}$ | $SS_{Error\|B}$ | $SS_{Error\|A,B}$ |
|---|---|---|---|---|
| Gebalanceerd | 232.3 | 181.6 | 160.1 | 109.4 |
| Ongebalanceerd | 1006.056 | 893.556 | 238.16 | 238.1545 |

| | $SS_A$ | $SS_{A\|B}$ | $SS_B$ | $SS_{B\|A}$ |
|---|---|---|---|---|
| Gebalanceerd | 50.7 | 50.7 | 72.2 | 72.2 |
| Ongebalanceerd | 112.5 | 0.0055 | 767.896 | 655.4015 |

Tabel 3.27: Errorkwadratensommen (boven) en effectkwadratensommen (onder) voor het gebalanceerde en het ongebalanceerde voorbeeldonderzoek (zonder rekening te houden met de interactie).

### De Totale Kwadratensom en de Kwadratensom-ontbindingen

We geven nu ook de formule van de totale kwadratensom $SS_{Totaal}$. Deze totale kwadratensom is de errorkwadratensom voor het model zonder effecten. De voorspellingen op basis van het model zonder effecten zijn gelijk aan $\overline{y}_{\sqcup\sqcup}$. Vervangen we $\widehat{y}_{kli}^{Model}$ in de algemene

formule van de errorkwadratensom door $\overline{y}_{\sqcup\sqcup}$, dan krijgen we de formule van $SS_{Totaal}$:

$$SS_{Totaal} = \sum_{Obs} (y_{kli} - \overline{y}_{\sqcup\sqcup})^2 \quad .$$

In het bovenste deel van Tabel 3.27 staan de totale kwadratensommen voor het gebalanceerde en het ongebalanceerde voorbeeldonderzoek. De effectieve berekening van deze kwadratensommen is volledig analoog aan de berekening van een errorkwadratensom; het enige verschil is dat de voorspelling voor elke observatie gelijk is aan $\overline{y}_{\sqcup\sqcup}$.

De totale kwadratensom is op dezelfde manier gerelateerd aan de effect- en de errorkwadratensommen als in de variantieanalyse met één factor. Het volgende geldt namelijk:

$$\begin{aligned} SS_{Totaal} &= SS_A + SS_{Error|A} \\ SS_{Totaal} &= SS_B + SS_{Error|B} \quad . \end{aligned}$$

De eerste van de twee kwadratensom-ontbindingen drukt uit dat de spreiding in de observaties, zoals gemeten door $SS_{Totaal}$, bestaat uit (1) de spreiding die toe te schrijven is aan factor A, zoals gemeten door de effectkwadratensom $SS_A$, en (2) de spreiding die niet toe te schrijven is aan factor A, zoals gemeten door de errorkwadratensom $SS_{Error|A}$. Volledig analoog, de tweede kwadratensom-ontbinding drukt uit dat de spreiding in de observaties, zoals gemeten door $SS_{Totaal}$, bestaat uit (1) de spreiding die toe te schrijven is aan factor B, zoals gemeten door de effectkwadratensom $SS_B$, en (2) de spreiding die niet toe te schrijven is aan factor B, zoals gemeten door de errorkwadratensom $SS_{Error|B}$. Controleer deze kwadratensom-ontbindingen met de gegevens uit de voorbeeldonderzoeken (zie Tabel 3.27).

Het feit dat beide kwadratensom-ontbindingen gelijktijdig waar zijn roept de volgende vragen op:

1. Is een gedeelte van de spreiding die gemeten wordt door $SS_{Error|A}$ toe te schrijven aan factor B?

2. Is een gedeelte van de spreiding die gemeten wordt door $SS_{Error|B}$ toe te schrijven aan factor A?

Beide vragen moeten met "ja" beantwoord worden. Het volgende geldt namelijk:

$$\begin{aligned} SS_{Error|A} &= SS_{B|A} + SS_{Error|A,B} \\ SS_{Error|B} &= SS_{A|B} + SS_{Error|A,B} \quad . \end{aligned}$$

We kunnen deze ontbindingen als volgt begrijpen:

1. $SS_{Error|A}$ is groter dan $SS_{Error|A,B}$ omdat er bij de voorspelling geen rekening is gehouden met factor B. De grootte van het verschil tussen $SS_{Error|A}$ en $SS_{Error|A,B}$ is gelijk aan de hoeveelheid spreiding in de observaties die toe te schrijven is aan factor B bovenop (na uitzuivering voor, gecorrigeerd voor) factor A, zoals gemeten door de type II kwadratensom $SS_{B|A}$.

2. $SS_{Error|B}$ is groter dan $SS_{Error|A,B}$ omdat er bij de voorspelling geen rekening is gehouden met factor A. De grootte van het verschil tussen $SS_{Error|B}$ en $SS_{Error|A,B}$

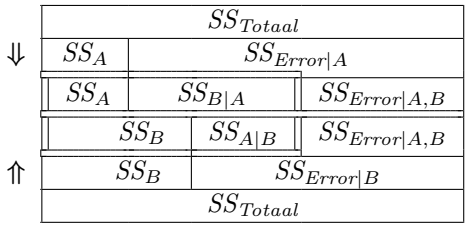
is gelijk aan de hoeveelheid spreiding in de observaties die toe te schrijven is aan factor A bovenop (na uitzuivering voor, gecorrigeerd voor) factor B, zoals gemeten door de type II kwadratensom $SS_{A|B}$.

Controleer deze kwadratensom-ontbindingen met de gegevens uit de voorbeeldonderzoeken (zie Tabel 3.27).

De kwadratensom-ontbindingen voor $SS_{Error|A}$ en $SS_{Error|B}$ kan men substitueren in de kwadratensom-ontbindingen voor $SS_{Totaal}$. Het resultaat is het volgende:

$$SS_{Totaal} = SS_A + SS_{B|A} + SS_{Error|A,B}$$
$$SS_{Totaal} = SS_B + SS_{A|B} + SS_{Error|A,B} \quad .$$

Deze twee kwadratensom-ontbindingen worden in Figuur 3.2 grafisch weergegeven. Van de eerste tot de derde rij van deze figuur wordt die kwadratensom-ontbinding weergegeven waarbij $SS_A$ als eerste afgesplitst wordt. En van de zesde tot de vierde rij (dus, in de omgekeerde richting) wordt die kwadratensom-ontbinding weergegeven waarbij $SS_B$ als eerste afgesplitst wordt. In het gedeelte van de figuur dat met dubbele lijnen omkaderd is staan de twee mogelijke ontbindingen van de effectkwadratensommen: (1) de type I effectkwadratensom van A ($SS_A$) wordt als eerste afgesplitst, gevolgd door de type II effectkwadratensom van B ($SS_{B|A}$), of (2) de type I effectkwadratensom van B ($SS_B$) wordt als eerste afgesplitst, gevolgd door de type II effectkwadratensom van A ($SS_{A|B}$). In een gebalanceerd design zijn deze twee ontbindingen identiek. Dit komt omdat, in een gebalanceerd design, $SS_{B|A}$ gelijk is aan $SS_B$ en $SS_{A|B}$ gelijk is aan $SS_A$.




Figuur 3.2: Grafische voorstelling van de twee verschillende kwadratensom-ontbindingen in een twee-factorieel design (zonder interactie).

De som van de effectkwadratensommen $SS_A$ en $SS_{B|A}$ (of, equivalent hiermee, $SS_B$ en $SS_{A|B}$) wordt ook de *modelkwadratensom* genoemd. Deze modelkwadratensom geeft aan welk gedeelte van de totale kwadratensom toe te schrijven is aan de effecten van de twee factoren. Het resterende gedeelte van de totale kwadratensom is de errorkwadratensom $SS_{Error|A,B}$. De modelkwadratensom wordt in het Engels soms *model sum of squares* genoemd, maar soms ook *corrected model sum of squares*.

Gebruikers van computerprogramma's zoals SPSS zijn vaak verrast door het feit dat, in de ANOVA-tabel, de modelkwadratensom niet gelijk is aan de som van de effectkwadratensommen in deze ANOVA-tabel. Dit is inderdaad het geval in een ongebalanceerd design. Als men de type I kwadratensommen berekent, dan staat er in de ANOVA-tabel $SS_A$ en $SS_B$, en in een ongebalanceerd design is hun som niet gelijk aan de modelkwadratensom [9]. Of men berekent de type II kwadratensom, en dan staat er

[9] Het computerprogramma SPSS verstaat onder de *optie* type I iets anders dan de type I kwadraten-

in de ANOVA-tabel $SS_{A|B}$ en $SS_{B|A}$, en in een ongebalanceerd design is hun som ook niet gelijk aan de modelkwadratensom.

### Nogmaals de F-Statistieken

De kwadratensommen die we beschreven hebben, vormen de ingrediënten van de F-statistieken waarmee het effect van de factoren getoetst wordt. Voor het gemak herhalen we hier de algemene vorm van de F-statistiek:

$$\boxed{F = \frac{MS_{Effect|Ref}}{MS_{Error|Model}} \quad .}$$

In de teller van deze F-statistiek, staat de gemiddelde effectkwadratensom $MS_{Effect|Ref}$, die bepaalt wat men ermee toetst. En in de noemer staat de gemiddelde errorkwadratensom $MS_{Error|Model}$, die bepaalt hoe gevoelig men toetst.

**De Berekening van de Gemiddelde Effectkwadratensom** De gemiddelde effectkwadratensom $MS_{Effect|Ref}$ wordt als volgt berekend:

$$\boxed{MS_{Effect|Ref} = \frac{SS_{Effect|Ref}}{\text{Aantal vrijheidsgraden van } SS_{Effect|Ref}} \quad .}$$

Het aantal vrijheidsgraden van $SS_{Effect|Ref}$ wordt bepaald op basis van de volgende regel:

$$\boxed{\begin{array}{c}\text{aantal vrije parameters in}\\ [\text{referentiemodel} + \text{effect}]\end{array}} \quad - \quad \boxed{\begin{array}{c}\text{aantal vrije parameters in}\\ [\text{referentiemodel}]\end{array}} \quad .$$

We gebruiken de rechte haakjes ([ en ]) om de modelcomponenten "referentiemodel" en "effect" samen te voegen tot één model. Er zijn in totaal vier modellen die een rol spelen bij het bepalen van de vrijheidsgraden: (1) het model zonder effecten, (2) het één-factoriële model met A, (3) het één-factoriële model met B, en (4) het twee-factoriële model met de effecten van A en B. Het aantal vrije parameters in elk van deze vier modellen wordt berekend in Tabel 3.28. Net zoals in de variantieanalyse met één factor, zijn er slechts $K-1$ vríje effectparameters $\alpha_k$. Dit komt omdat de som van deze $K$ effectparameters gelijk is aan 0. Om dezelfde reden zijn er ook slechts $L-1$ vrije effectparameters $\beta_l$.

Gebruikmakend van de aantallen in de laatste kolom van Tabel 3.28 is het eenvoudig om te laten zien dat het aantal vrijheidsgraden van $SS_A$ gelijk is aan $K-1$. Immers, het aantal vrije parameters in [referentiemodel] is gelijk aan 1 en het aantal vrije parameters in [referentiemodel + effect] is gelijk aan $K$. Het aantal vrijheidsgraden van $SS_A$ (i.e., $K-1$) is gelijk aan het aantal vrije effectparameters $\alpha_k$ dat nodig is om het effect van A in het variantieanalysemodel op te nemen. Het aantal vrijheidsgraden van $SS_{A|B}$ is eveneens gelijk aan $K-1$; het aantal vrije effectparameters $\alpha_k$ dat nodig is om het effect van A naast het effect van B in het variantieanalysemodel op te nemen. Op een analoge wijze kan men laten zien dat het aantal vrijheidsgraden van $SS_B$ en $SS_{B|A}$ gelijk is aan $L-1$.

som. Dus, als men binnen SPSS de optie type I aankruist, dan krijgt men níet de type I kwadratensommen. We komen hier later nog op terug, en leggen dan uit wat SPSS dan wél berekent met deze optie.

| Factoren | Vrije Parameters | Aantal vrije parameters |
|---|---|---|
| Geen | $\mu$ | 1 |
| A | $\mu$, $K-1$ vrije $\alpha_k$'s | $K$ |
| B | $\mu$, $L-1$ vrije $\beta_l$'s | $L$ |
| A en B | $\mu$, $K-1$ vrije $\alpha_k$'s, $L-1$ vrije $\beta_l$'s | $K+L-1$ |

Tabel 3.28: Berekening van het aantal vrije parameters in het model zonder effecten, het model met A, het model met B, en het model met de effecten van A en B.

**De Keuze voor een Gemiddelde Effectkwadratensom** De gemiddelde effectkwadratensom $MS_{Effect|Ref}$ bepaalt wat men toetst met de F-statistiek. Men kan voor *Effect* zowel het effect van A als het effect van B invullen. Dit effect moet men nader specificeren door aan te geven wat het referentiemodel is. Indien dit referentiemodel het model zonder effect is, dan is $MS_{Effect|Ref}$ een type I kwadratensom en toetst men met deze F-statistiek een nulhypothese die betrekking heeft op gewogen marginale verwachte waarden. Concreet, bij *Effect*=A, staat in de teller de type I kwadratensom $MS_A$, en wordt de volgende nulhypothese getoetst:

$$\boxed{\mu_{1\sqcup} = \mu_{2\sqcup} = \cdots = \mu_{K\sqcup} \quad ,}$$

de gelijkheid van de gewogen verwachte waarden $\mu_{k\sqcup}$ in de $K$ niveaus van factor A. En bij *Effect*=B, staat in de teller de type I kwadratensom $MS_B$, en wordt een analoge nulhypothese getoetst: de gelijkheid van de gewogen verwachte waarden $\mu_{\sqcup l}$ in de $L$ niveaus van factor B. Dus, in dit geval wordt niet uitgezuiverd voor de andere factor.

Indien het referentiemodel het model is met de andere factor (factor B bij *Effect*=A en factor A bij *Effect*=B), dan is $MS_{Effect|Ref}$ een type II kwadratensom en toetst men met de F-statistiek een nulhypothese die betrekking heeft op ongewogen marginale verwachte waarden. Concreet, bij *Effect*=A, staat in de teller de type II kwadratensom $MS_{A|B}$, en wordt de volgende nulhypothese getoetst:

$$\boxed{\mu_{1\bullet} = \mu_{2\bullet} = \cdots = \mu_{K\bullet} \quad ,}$$

de gelijkheid van de ongewogen verwachte waarden $\mu_{k\bullet}$ in de $K$ niveaus van factor A. En bij *Effect*=B, staat in de teller de type II kwadratensom $MS_{B|A}$, en wordt een analoge nulhypothese getoetst: de gelijkheid van de ongewogen verwachte waarden $\mu_{\bullet l}$ in de $L$ niveaus van factor B. Dus, in dit geval wordt wel uitgezuiverd voor de andere factor. In verreweg de meeste onderzoeken is men geïnteresseerd in de nulhypothese die met de type II kwadratensom getoetst wordt (gelijkheid van de ongewogen verwachte waarden) en niet in de nulhypothese die met de type I kwadratensom getoetst wordt (gelijkheid van de gewogen verwachte waarden).

In een gebalanceerd design zijn de type I gemiddelde kwadratensommen $MS_A$ en $MS_B$ gelijk aan de corresponderende type II gemiddelde kwadratensommen $MS_{A|B}$ en $MS_{B|A}$. Dit stemt overeen met het feit dat, in een gebalanceerd design, de gewogen en de ongewogen marginale verwachte waarden gelijk zijn. Dus, in een gebalanceerd design, kan men voor een bepaalde factor slechts één effect toetsen en niet twee, zoals in een ongebalanceerd design (i.c., het uitgezuiverde en het niet-uitgezuiverde effect).

**De Berekening van de Gemiddelde Errorkwadratensom** De gemiddelde errorkwadratensom $MS_{Error|Model}$ wordt als volgt berekend:

$$\boxed{MS_{Error|Model} = \frac{SS_{Error|Model}}{\text{Aantal vrijheidsgraden van } SS_{Error|Model}} \quad .}$$

Het aantal vrijheidsgraden van $SS_{Error|Model}$ wordt bepaald op basis van de volgende regel:

$$\boxed{\text{aantal observaties}} \quad - \quad \boxed{\text{aantal vrije parameters in [Model]}} \quad .$$

De modellen die in aanmerking komen voor de berekening van een gemiddelde errorkwadratensom zijn (1) het model met A, (2) het model met B en (3) het model met A en B. De aantallen vrije parameters in deze drie modellen worden gegeven in Tabel 3.28.

**De Keuze voor een Gemiddelde Errorkwadratensom** Hoe kleiner de gemiddelde errorkwadratensom in de noemer van de F-statistiek, hoe gevoeliger men toetst. Dit komt omdat, naarmate de gemiddelde errorkwadratensom kleiner is, een kleinere effectkwadratensom voldoende is om een significante F-statistiek te krijgen. Daarom is een kleine gemiddelde errorkwadratensom altijd te verkiezen boven een grote.

Het is niet zo dat men altijd kan kiezen uit meerdere errorkwadratensommen. Concreet, in een F-statistiek met de type II effectkwadratensom $MS_{A|B}$ of $MS_{B|A}$ in de teller, moet $MS_{Error|A,B}$ in de noemer staan. De regel die hier achter zit is dat voor het toetsen van een effect met behulp van de gemiddelde effectkwadratensom $MS_{Effect|Ref}$, zowel *Effect* als *Ref* deel moeten uitmaken van het "Model" in de gemiddelde errorkwadratensom $MS_{Error|Model}$. Er is dus alleen wat te kiezen in een F-statistiek met een type I effectkwadratensom in de teller. In wat volgt, concentreren we ons op $MS_A$, de type I effectkwadratensom waarmee we de gewogen marginale gemiddelden van factor A toetsen. De overwegingen bij de keuze van een gemiddelde errorkwadratensom voor het effect van factor B zijn volledig analoog.

We beginnen met de conclusie: naarmate factor B meer variantie verklaart, wordt het verschil tussen $MS_{Error|A}$ en $MS_{Error|A,B}$ groter, en is $MS_{Error|A,B}$ eerder te verkiezen. We kunnen dit als volgt zien:

1. $SS_{Error|A}$ is altijd groter dan $SS_{Error|A,B}$. Uit de kwadratensom-ontbinding volgt dat het verschil tussen de twee gelijk is aan $SS_{B|A}$. Dus, als we alleen naar de teller van $MS_{Error|A}$ en $MS_{Error|A,B}$ zouden kijken, dan zouden we zeggen dat $MS_{Error|A}$ altijd groter is dan $MS_{Error|A,B}$.

2. Het aantal vrijheidsgraden van $SS_{Error|A}$ $(N-K)$ is altijd groter dan het aantal vrijheidsgraden van $SS_{Error|A,B}$ $(N-K-L+1)$. Het verschil is gelijk aan $L-1$, het aantal vrije $\beta_l$-parameters. Bij de berekening van $MS_{Error|A}$ deelt men dus door een groter getal dan bij $MS_{Error|A,B}$. Dus, als we alleen naar de noemer van $MS_{Error|A}$ en $MS_{Error|A,B}$ zouden kijken, dan zouden we zeggen dat $MS_{Error|A}$ altijd kleiner is dan $MS_{Error|A,B}$.

3. De vraag is nu, waarin de tegengestelde verschillen in teller en noemer resulteren: wint de teller en wordt $MS_{Error|A,B}$ kleiner dan $MS_{Error|A}$, of wint de noemer en

wordt $MS_{Error|A,B}$ groter dan $MS_{Error|A}$? Welnu, naarmate factor B meer variantie verklaart bovenop factor A, is $SS_{B|A}$ groter en dus ook het verschil tussen $SS_{Error|A}$ en $SS_{Error|A,B}$. Hieruit volgt dat, naarmate $SS_{B|A}$ groter is, het eerder het geval is dat de teller wint en $MS_{Error|A,B}$ kleiner is dan $MS_{Error|A}$.

We kunnen dit illustreren aan de hand van de twee voorbeeldonderzoeken. In het gebalanceerde voorbeeldonderzoek is $MS_{Error|A,B}$ kleiner dan $MS_{Error|A}$ en $MS_{Error|B}$. Dus, zowel voor het toetsen van het effect van factor A als voor het toetsen van het effect van factor B, is $MS_{Error|A,B}$ de beste (want de kleinste) gemiddelde errorkwadratensom. In het ongebalanceerde voorbeeldonderzoek is $MS_{Error|A,B}$ kleiner dan $MS_{Error|A}$ maar groter dan $MS_{Error|B}$. Dit laatste komt omdat de uitgezuiverde effectkwadratensom $SS_{A|B}$ (TRAINING uitgezuiverd voor TYPE LERAAR) zo klein is (namelijk, 0.0055). In dit geval wint dus de noemer (het verschil tussen de vrijheidsgraden) en is $MS_{Error|B}$ te verkiezen boven $MS_{Error|A,B}$.

In de praktijk is $MS_{Error|A,B}$ meestal kleiner dan $MS_{Error|A}$ en $MS_{Error|B}$. Men krijgt dus meestal een gevoeliger toets indien men een extra factor opneemt in een onderzoek. Dit komt omdat de effectkwadratensommen $SS_{A|B}$ en $SS_{B|A}$ meestal voldoende groot zijn om het verschil tussen de vrijheidsgraden te compenseren. Ondanks het feit dat $MS_{Error|A,B}$ meestal een betere keuze is dan $MS_{Error|A}$ en $MS_{Error|B}$, wordt er in de praktijk zelden gebruik van gemaakt. De reden hiervoor is dat er nog een andere gemiddelde errorkwadratensom $MS_{Error|Model}$ bestaat die meestal kleiner is dan $MS_{Error|A,B}$. Deze $MS_{Error|Model}$ is gebaseerd op het model met hoofd- én interactie-effecten. Hier komen we nog uitgebreid op terug.

**De F-Toets** De F-toets maakt gebruik van het feit dat, onder een bepaalde nulhypothese, de F-statistiek een $\mathcal{F}$-verdeling heeft met het aantal vrijheidsgraden gelijk aan de vrijheidsgraden van de teller en de noemer van deze F-statistiek. Bijvoorbeeld, de F-statistiek met $MS_A$ in de teller en $MS_{Error|A,B}$ in de noemer, heeft, onder de nulhypothese die betrekking heeft op de gewogen marginale verwachte waarden, een $\mathcal{F}$-verdeling met $K-1$ tellervrijheidsgraden en $N-K-L+1$ noemervrijheidsgraden. De F-statistiek met $MS_{A|B}$ in de teller en $MS_{Error|A,B}$ in de noemer, heeft dezelfde verdeling, maar nu onder de nulhypothese die betrekking heeft op de ongewogen marginale verwachte waarden. Indien men $MS_{Error|A,B}$ in de noemer vervangt door $MS_{Error|A}$, dan stijgt het aantal noemervrijheidsgraden van de $\mathcal{F}$-verdeling van $N-K-L+1$ naar $N-K$.

De F-toets houdt in dat men een p-waarde berekent onder de genoemde $\mathcal{F}$-verdeling. Deze p-waarde is de kans onder deze $\mathcal{F}$-verdeling (en dus ook onder deze nulhypothese) om een F-statistiek te observeren die groter of gelijk is aan de geobserveerde F-statistiek. Deze p-waarde is een maat voor de aannemelijkheid van de nulhypothese. Een kleine p-waarde betekent dat we data geobserveerd hebben die een kleine kans van voorkomen hebben onder de nulhypothese. Als deze p-waarde kleiner is dan 0.05 of 0.01 dan zegt men dat de F-statistiek significant is, en verwerpt men de nulhypothese.

### Overzichtstabellen

We geven nu vier overzichtstabellen met de belangrijkste formules van deze paragraaf (*Kwadratensommen en F-statistieken voor de Situatie Zonder Interactie*).

1. In Tabel 3.29 staan de vier modellen die gebruikt worden voor de berekening van de effect- en de errorkwadratensommen.

2. In Tabel 3.30 staan de F-toetsen voor de gewogen en de ongewogen marginale gemiddelden van de twee factoren.

3. In Tabel 3.31 staan de type I en de type II effectkwadratensommen en hun vrijheidsgraden.

4. In Tabel 3.32 staan de errorkwadratensommen, hun vrijheidsgraden en hun kwadratensomontbindingen.

| Modellen zonder Interactie | | |
|---|---|---|
| | Effecten | Formule |
| (1) | Geen | $y_{kli} = \mu + e_{kli}$ |
| (2) | A | $y_{kli} = \mu + \alpha_k + e_{kli}$ |
| (3) | B | $y_{kli} = \mu + \beta_l + e_{kli}$ |
| (4) | A en B | $y_{kli} = \mu + \alpha_k + \beta_l + e_{kli}$ |

Tabel 3.29: Vier modellen voor een twee-factorieel design, zonder interactie.

| F-toetsen voor Modellen zonder Interactie | | | | |
|---|---|---|---|---|
| $$F = \frac{MS_{Effect\|Ref}}{MS_{Error\|Model}}$$ | | | | |
| | | | $MS_{Error\|Model}$ | |
| Effect | Nulhypothese | $MS_{Effect\|Ref}$ | Meestal klein | Meestal groot |
| A | $\mu_{1\sqcup} = \mu_{2\sqcup} = \ldots = \mu_{K\sqcup}$ | $MS_A$ | $MS_{Error\|A,B}$ | $MS_{Error\|A}$ |
| | $\mu_{1\bullet} = \mu_{2\bullet} = \ldots = \mu_{K\bullet}$ | $MS_{A\|B}$ | $MS_{Error\|A,B}$ | |
| B | $\mu_{\sqcup 1} = \mu_{\sqcup 2} = \ldots = \mu_{\sqcup L}$ | $MS_B$ | $MS_{Error\|A,B}$ | $MS_{Error\|B}$ |
| | $\mu_{\bullet 1} = \mu_{\bullet 2} = \ldots = \mu_{\bullet L}$ | $MS_{B\|A}$ | $MS_{Error\|A,B}$ | |

Tabel 3.30: F-toetsen voor de gewogen en de ongewogen marginale gemiddelden van de factoren A en B, gebruik makend van modellen zonder interactie.

| Type I en Type II Effectkwadratensommen | | | | | | |
|---|---|---|---|---|---|---|
| $SS_{Effect\|Ref} = \sum_{Obs}(\widehat{y}_{kli}^{Ref+Effect} - \widehat{y}_{kli}^{Ref})^2$ | | | | | | |
| Effect | Type | Model-vergelijking | Symbool | Vrijheids-graden | Hoe berekenen? Gebalanceerd | Hoe berekenen? Ongebalanceerd |
| A | I | (2) met (1) | $SS_A$ | $K-1$ | $\sum_{k=1}^{K} Ln(\overline{y}_{k\bullet} - \overline{y}_{\bullet\bullet})^2$ | $\sum_{k=1}^{K} n_{k+}(\overline{y}_{k\sqcup} - \overline{y}_{\sqcup\sqcup})^2$ |
| | II | (4) met (3) | $SS_{A\|B}$ | $K-1$ | als boven | computer |
| B | I | (3) met (1) | $SS_B$ | $L-1$ | $\sum_{l=1}^{L} Kn(\overline{y}_{\bullet l} - \overline{y}_{\bullet\bullet})^2$ | $\sum_{l=1}^{L} n_{+l}(\overline{y}_{\sqcup l} - \overline{y}_{\sqcup\sqcup})^2$ |
| | II | (4) met (2) | $SS_{B\|A}$ | $L-1$ | als boven | computer |

Tabel 3.31: Berekening van de type I en de type II effectkwadratensommen voor de effecten van de factoren A en B.

| Errorkwadratensommen voor Modellen zonder Interactie | | | | | | |
|---|---|---|---|---|---|---|
| $SS_{Error\|Model} = \sum_{Obs}(y_{kli} - \widehat{y}_{kli}^{Model})^2$ | | | | | | |
| | | | Vrijheids- | | Bereken $\widehat{y}_{kli}^{Model}$ | |
| Model | Effecten | Symbool | graden | Ontbinding | Gebalanceerd | Ongebalanceerd |
| (1) | Geen | $SS_{Totaal}$ | $N-1$ | $SS_A + SS_{Error\|A}$<br>$SS_B + SS_{Error\|B}$ | $\overline{y}_{\bullet\bullet}$ | $\overline{y}_{\sqcup\sqcup}$ |
| (2) | A | $SS_{Error\|A}$ | $N-K$ | $SS_{B\|A} + SS_{Error\|A,B}$ | $\overline{y}_{k\bullet}$ | $\overline{y}_{k\sqcup}$ |
| (3) | B | $SS_{Error\|B}$ | $N-L$ | $SS_{A\|B} + SS_{Error\|A,B}$ | $\overline{y}_{\bullet l}$ | $\overline{y}_{\sqcup l}$ |
| (4) | A,B | $SS_{Error\|A,B}$ | $N-K-L+1$ | | $\overline{y}_{k\bullet} + \overline{y}_{\bullet l} - \overline{y}_{\bullet\bullet}$ | computer |

Tabel 3.32: Berekening van de errorkwadratensom op basis van modellen zonder interactie.

*Opgave 9*
Hieronder staan een aantal onderzoeksvragen. Elke onderzoeksvraag betreft het effect van een onafhankelijke variabele binnen een observationeel onderzoek. Noem voor elke onderzoeksvraag minstens één potentiële storende variabele waarvoor men zou moeten uitzuiveren en motiveer dit.

1. Een onderzoeker denkt dat inslaapproblemen verholpen kunnen worden door regelmatige zware lichamelijke inspanning. Hij verdeelde zijn proefpersonen in twee groepen op basis van de mate waarin zij zware lichamelijke inspanning leveren, zoals vastgesteld op basis van deelname aan sport- en fitnessactiviteiten, de aard van iemands werk, en dergelijke. Dan vergelijkt hij de ernst van de inslaapproblemen in deze twee groepen.
2. Een klinisch psycholoog denkt dat ouders die veel met hun kinderen ondernemen zonder hun partner, ook weinig meningsverschillen hebben over de opvoeding. Hij verdeelde de proefgezinnen in twee groepen op basis van het feit of ze veel of weinig met hun kinderen ondernamen zonder hun partner, en vergeleek aantal en ernst van de meningsverschillen in deze twee groepen.
3. Een medewerkster van het Ministerie van Justitie denkt dat zij de kans op herhaling van een misdrijf kan verlagen door een overtreder te confronteren met zijn slachtoffer(s). In een aantal gevangenissen biedt zij de gevangenen de mogelijkheid tot een dergelijke confrontatie. Een aantal jaren later vergelijkt zij het aantal recidives in de groep die aan de confrontatie heeft deelgenomen met het aantal recidives in de groep die dat niet gedaan heeft.

*Opgave 10*
Een onderzoeker is geïnteresseerd in de relatie tussen PERSOONLIJKHEIDSTYPE, STRESS en BLOEDDRUK. Hij heeft de bloeddruk gemeten van 40 personen met persoonlijkheidstype A en 40 met persoonlijkheidstype B. De leefomgeving van elke proefpersoon werd geclassificeerd als stressvol of niet stressvol. In Tabel 3.33 geven we de gemiddelde bloeddruk (vet gedrukt) en het aantal proefpersonen voor elk van de vier cellen van dit onderzoeksdesign.

| | | STRESS | |
|---|---|---|---|
| | | Hoog | Laag |
| PERSOONLIJK-HEIDSTYPE | A | **170**<br>28 | **150**<br>12 |
| | B | **140**<br>12 | **120**<br>28 |

Tabel 3.33: Tabel bij opgave 10.

1. Als je het effect van STRESS wilt scheiden van het effect van PERSOONLIJKHEID-STYPE, welke soort marginale gemiddelden moet je dan gebruiken? Hoe groot zijn deze marginale gemiddelden in dit onderzoek?
2. Is het verschil in gemiddelde bloeddruk tussen de twee persoonlijkheidstypes groter of kleiner wanneer je rekening houdt met (uitzuivert voor) STRESS, dan wanneer je het niet doet? Hou zou je dit inhoudelijk kunnen verklaren? (Voor het antwoord op deze laatste vraag moet je een beroep doen op je eigen redeneervermogen; je kunt het antwoord in dit boek niet vinden.)

*Opgave 11*
Een onderzoeker is geïnteresseerd in het effect van het hebben van een huisdier (hond of kat) op eenzaamheidsgevoelens bij alleenstaanden. Behalve de vraag of men een huisdier bezit (de factor HUISDIER), stelde de onderzoeker ook de vraag in welk type woning men woonde (de factor WONINGTYPE). Hierbij hanteerde hij de volgende categorieën: (1) een woning met een grote tuin (minimaal 300 vierkante meter), (2) een woning met een kleine tuin (minder dan 300 vierkante meter) en (3) een woning zonder tuin. De afhankelijke variabele was de score op een eenzaamheidsvragenlijst (hoe hoger de score, hoe eenzamer).

De gegevens van dit onderzoek worden gegeven in Tabel 3.34. De vetgedrukte getallen zijn celgemiddelden en de andere getallen zijn de aantallen observaties per cel.

| | | WONINGTYPE | | |
|---|---|---|---|---|
| | | Grote Tuin | Kleine Tuin | Geen Tuin |
| HUISDIER | Met Huisdier | **20**<br>10 | **30**<br>30 | **40**<br>60 |
| | Zonder Huisdier | **30**<br>60 | **40**<br>30 | **50**<br>10 |

Tabel 3.34: Tabel bij opgave 11.

1. Als je het effect van HUISDIER wilt scheiden van het effect van WONINGTYPE, welke marginale gemiddelden moet je dan berekenen? Hoe groot zijn deze gemiddelden in dit onderzoek?
2. Verschillen de gemiddelden van de verschillende niveaus van HUISDIER meer van elkaar als je uitzuivert voor het effect van WONINGTYPE dan wanneer je niet uitzuivert voor dit effect? Of verschillen ze juist minder van elkaar?
3. In dit onderzoek is WONINGTYPE een storende variabele. Als men ook voor andere variabelen zou willen uitzuiveren, aan welke voorwaarden moeten die variabelen dan voldoen om storend te zijn?

*Opgave 12*
Geef aan of de onderstaande uitspraken waar of onwaar zijn, en indien onwaar, leg uit waarom.

1. Bij het berekenen van de effectkwadratensom van factor A doet het er niet toe welk model als referentiemodel gebruikt wordt.
2. Als factor X in het referentiemodel zit, dan wordt er voor die factor uitgezuiverd.
3. Een F-statistiek met een type-II effectkwadratensom toetst de nulhypothese dat de gewogen marginale verwachte waarden gelijk zijn.
4. De errorkwadratensom $SS_{Error|A}$ bestaat voor een deel uit variantie die toe te schrijven is aan andere factoren dan factor A.
5. De modelkwadratensom kan slechts op één manier opgesplitst worden in twee effectkwadratensommen.
6. De gemiddelde errorkwadratensom wordt altijd kleiner als men een extra factor in het design opneemt.

*Opgave 13*
Een onderwijspsychologe vraagt zich af of de leesmethode die op de basisschool gebruikt wordt, van invloed is op het niveau van begrijpend lezen aan het einde van groep 8. Ze denkt dat een

methode waarmee kinderen geleerd wordt om tijdens het lezen hardop te denken (methode H) tot betere leesprestaties leidt dan een methode waarmee dat niet gebeurt (methode NH).

De onderwijspsychologe nam van 600 kinderen uit groep 8 een test voor begrijpend lezen af. De scores op deze test zijn de afhankelijke variabele in haar onderzoek. De helft van deze leerlingen heeft onderwijs gevolgd op een school waar methode H gebruikt werd, en de andere helft op een school waar methode NH gebruikt werd. Omdat haar onderzoek observationeel is, is de onderzoekster bang dat het type school (aangeduid als SCHOOLTYPE) waarop de leerling zit een storende variabele is. Zij maakt zich met name zorgen over het verschil tussen reguliere en Montessorischolen. In Tabel 3.35 staan de gemiddelde testscores (vetgedrukt) en het aantal proefpersonen (gewoon gedrukt) voor elk van de vier cellen van dit onderzoeksdesign.

| | | SCHOOLTYPE | |
|---|---|---|---|
| | | Regulier | Montessori |
| METHODE | H | **90**<br>200 | **50**<br>100 |
| | NH | **100**<br>100 | **60**<br>200 |

Tabel 3.35: Tabel bij opgave 13.

1. Leg uit waarom het type school een storende variabele is.
2. Bereken de gewogen en de ongewogen marginale gemiddelden voor de verschillende niveaus van de factor METHODE.
3. Welke methode is het beste als je je antwoord baseert op (1) de gewogen gemiddelden en (2) de ongewogen gemiddelden? Welke gemiddelden zou jij gebruiken om het effect van METHODE te evalueren? Leg uit.
4. Als je wilt weten welke methode het meest effectief is, welk type effectkwadratensom moet je dan gebruiken in je F-toets? Leg uit.

*Opgave 14*
Een onderwijspsychologe is geïnteresseerd in het effect van een nieuwe methode om slechtlezende kinderen te leren lezen. Zij beschikt over de leestestscores van een groep kinderen waarvan de ene helft onderwezen werd met deze nieuwe methode en de andere helft met de oude methode. De toewijzing van de kinderen aan deze twee condities gebeurde op zuiver toevallige wijze. Van elk kind weet zij ook of het, behalve problemen met lezen, ook problemen heeft om zich te concentreren. Omdat de onderzoekster alleen geïnteresseerd is in het effect van de nieuwe versus de oude methode, besluit zij om de gegevens te analyseren met een variantieanalyse met één factor (METHODE). De tweede factor (CONCENTRATIEMOEILIJKHEDEN) laat zij dus gewoon buiten beschouwing. Denk jij dat zij dezelfde resultaten krijgt als met een variantieanalyse met twee factoren? (Ga er hierbij vanuit dat er in elk van de vier cellen evenveel kinderen zitten.) Motiveer je antwoord. Welke methode kan zij volgens jou het beste gebruiken? Waarom?

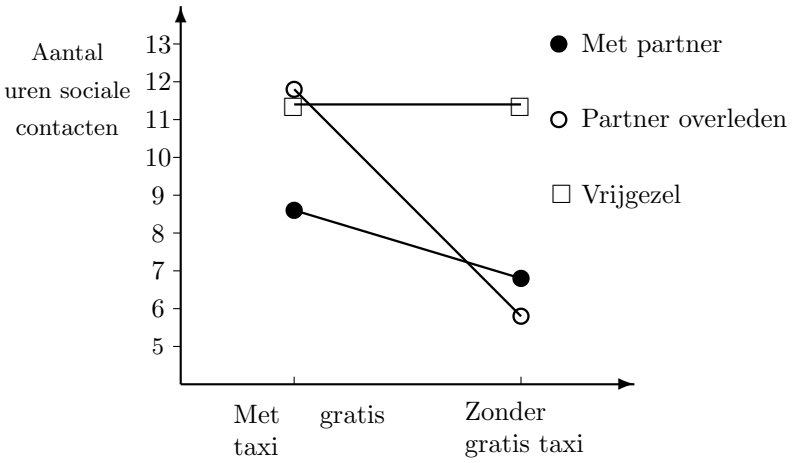
---

### 3.3.9 Interactie tussen Factoren

Om het fenomeen interactie te beschrijven keren we terug naar de gegevens in Tabel 3.14. In het bovenste gedeelte van Figuur 3.3 tonen we een kleine tabel met de marginale

gemiddelden en de celgemiddelden, en in het onderste gedeelte geven we de celgemiddelden weer in een grafiek. We zien dat het effect van MOBILITEIT verschillend is voor de drie types bejaarden: voor bejaarden met partner is er een zwak effect, voor bejaarden waarvan de partner overleden is een sterk effect, en voor bejaarden die vrijgezel zijn is er geen effect.

| | Met partner | Partner verloren | Vrijgezel | Marg Gem. |
|---|---|---|---|---|
| Met gratis taxi | 8.6 | 11.8 | 11.4 | 10.6 |
| Zonder gratis taxi | 6.8 | 5.8 | 11.4 | 8 |
| Marg. Gem. | 7.7 | 8.8 | 11.4 | 9.3 |




Figuur 3.3: Marginale gemiddelden en celgemiddelden van de gegevens uit het voorbeeldonderzoek met grafische voorstelling van de interactie.

Interactie is het fenomeen dat het effect van de ene factor (hier, MOBILITEIT of BURGERLIJKE STAND) verschillend is binnen de verschillende niveaus van de andere factor (respectievelijk, BURGERLIJKE STAND of MOBILITEIT). In de figuur kan men dit zien aan het volgende:

1. Het verschil tussen de groep met en de groep zonder gratis taxi is anders voor elk van de drie types bejaarden (met partner, partner overleden en vrijgezel).

2. De verschillen tussen de drie types bejaarden zijn anders voor de groep met en de groep zonder gratis taxi.

Deze twee patronen komen altijd samen voor. Zij weerspiegelen zich in het feit dat de lijnen voor de drie types niet parallel lopen. Als er geen interactie is, dan hangen de

verschillen tussen twee celgemiddelden niet af van het niveau van de andere onafhankelijke variabele. In dit geval, zeggen we dat de effecten van de onafhankelijke variabelen *additief* zijn.

Interactie is om drie redenen van belang. De eerste en belangrijkste reden is inhoudelijk: als we te maken hebben met interactie, dan kunnen we niet spreken van hét effect van een factor zonder meer. Immers, afhankelijk van het niveau van de andere factor kan dit effect groot of klein zijn. Daarom is het van belang om te onderzoeken of een interactie statistisch significant is.

De tweede reden waarom interactie van belang is, is dat, in een ongebalanceerd design, de effectkwadratensom van een bepaalde factor afhangt van het feit of men uitzuivert voor deze interactie of niet. Dit is analoog aan het feit dat, in een ongebalanceerd design, de type I kwadratensom (geen uitzuivering voor de andere factor) verschilt van de type II kwadratensom (wel uitzuivering voor de andere factor). Het belang van de interactie is dat, in een ongebalanceerd design, de type II kwadratensom verschilt van een derde type effectkwadratensom, de *type III* kwadratensom, die, behalve voor de andere factor, ook uitgezuiverd is voor de interactie tussen de twee factoren.

De derde reden waarom interactie van belang is, is dat de errorkwadratensom kleiner wordt als men bij de voorspelling rekening houdt met de interactie (in vergelijking met de situatie waarin men er geen rekening mee houdt). Dit is des te sterker het geval naarmate de interactie in de gegevens meer uitgesproken is. Voor de gemiddelde errorkwadratensom geldt dat deze meestal kleiner is als men bij de voorspelling rekening houdt met de interactie. Dus, met deze laatste gemiddelde errorkwadratensom is het mogelijk om gevoeliger te toetsen: de gemiddelde effectkwadratensom hoeft niet meer zo groot te zijn om toch te resulteren in een significante toets.

In de volgende twee paragrafen laten we zien hoe, voor de twee voorbeeldonderzoeken, de F-statistieken berekend worden. In de daarna volgende paragrafen geven we de formules die nodig zijn om de F-statistieken ook voor andere gegevens uit te rekenen. We behandelen hier twee types F-statistieken: (1) een F-statistiek waarmee getoetst wordt of een *interactie-effect* significant is, en (2) een F-statistiek waarmee getoetst wordt of het *hoofdeffect* van een factor significant is. Een hoofdeffect heeft te maken met verschillen tussen de niveaus van een factor; verschillen tussen de gewogen of de ongewogen marginale gemiddelden van deze niveaus. En een interactie-effect heeft te maken met verschillen tussen verschillen, en wel op de volgende manier: het interactie-effect is groter naarmate verschillen tussen de celgemiddelden over de niveaus van één factor (A of B) sterker verschillen naargelang het niveau van de andere factor (resp., B of A). Alle F-statistieken zijn gebaseerd op het model met hoofd- en interactie-effecten. Behalve de toetsing van het interactie-effect, stelt dit model ons in staat om, in een ongebalanceerd design, uit te zuiveren voor de interactie. Bovendien levert de voorspelling op basis van dit model een kleinere errorkwadratensom dan $SS_{Error|A,B}$, en wel des te kleiner naarmate de interactie in de gegevens meer uitgesproken is.

### 3.3.10 Het Rekenwerk in een Gebalanceerd Design (Met Interactie)

We laten eerst zien hoe de F-statistiek voor het interactie-effect berekend wordt. Daarna laten we zien hoe de F-statistieken voor de twee hoofdeffecten berekend worden.

#### De Interactie Tussen MOBILITEIT en BURGERLIJKE STAND

Voor het toetsen van de interactie tussen MOBILITEIT en BURGERLIJKE STAND hebben we voorspellingen nodig onder de volgende twee veronderstellingen:

1. Alleen hoofdeffecten van MOBILITEIT en BURGERLIJKE STAND.
2. Hoofdeffecten van MOBILITEIT en BURGERLIJKE STAND én een interactie-effect van MOBILITEIT en BURGERLIJKE STAND.

Onder de eerste veronderstelling wordt er geen rekening gehouden met het interactie-effect, en onder de tweede veronderstelling wel. Onder de eerste veronderstelling, maken we een voorspelling op basis van de volgende regel:

$$\boxed{\text{voorspelling}} = \boxed{\begin{array}{c}\text{globale}\\ \text{gemiddelde}\end{array}} + \boxed{\begin{array}{c}\text{hoofdeffect}\\ \text{van}\\ \text{MOBILITEIT}\end{array}} + \boxed{\begin{array}{c}\text{hoofdeffect van}\\ \text{BURGERLIJKE}\\ \text{STAND}\end{array}} \quad .$$

Van deze regel hebben we reeds eerder gebruik gemaakt (zie paragraaf 3.3.6). De voorspellingen op basis van deze regel staan in de vierde kolom van Tabel 3.36. De getallen in deze kolom zijn precies gelijk aan de getallen in de derde kolom van Tabel 3.22 op pagina 181.

Onder de veronderstelling dat er, behalve de twee hoofdeffecten, ook een interactie-effect is, zijn de voorspellingen gelijk aan de celgemiddelden. Deze celgemiddelden staan in de derde kolom van Tabel 3.36. In deze tabel staan ook de gekwadrateerde verschillen tussen de twee voorspellingen.

We berekenen nu de kwadratensom van het interactie-effect. Dit is een *effect*kwadratensom, net zoals de kwadratensommen van de hoofdeffecten van MOBILITEIT en BURGERLIJKE STAND. De kwadratensom van het interactie-effect is gelijk aan de som over alle observaties van de gekwadrateerde verschillen tussen de twee voorspellingen. Omdat er in elke cel vijf observaties zijn, wordt elk van de zes gekwadrateerde verschillen in de vijfde kolom van Tabel 3.36 vijf keer geteld bij de berekening van deze kwadratensom. Dus,

$$\begin{aligned}\text{Kwadratensom van}\\ \text{het interactie-effect} &= 5 \times [8.6 - 9]^2 + 5 \times [11.8 - 10.1]^2 + 5 \times [11.4 - 12.7]^2 + \\ &\quad 5 \times [6.8 - 6.4]^2 + 5 \times [5.8 - 7.5]^2 + 5 \times [11.4 - 10.1]^2 \\ &= 0.8 + 14.45 + 8.45 + 0.8 + 14.45 + 8.45 \\ &= 47.4 \quad .\end{aligned}$$

De berekening van de errorkwadratensom wordt weergegeven in Tabel 3.37. De gekwadrateerde verschillen tussen de observaties en de voorspellingen in de cel met gratis

| | | Veronderstelling waaronder de voorspelling gemaakt wordt | | Gekwadrateerd verschil tussen de voorspellingen |
|---|---|---|---|---|
| | | Wel interactie | Geen interactie | |
| Met gratis taxi | Met partner | 8.6 | $10.6 + 7.7 - 9.3 = 9$ | $(8.6 - 9)^2$ |
| | Partner overleden | 11.8 | $10.6 + 8.8 - 9.3 = 10.1$ | $(11.8 - 10.1)^2$ |
| | Vrijgezel | 11.4 | $10.6 + 11.4 - 9.3 = 12.7$ | $(11.4 - 12.7)^2$ |
| Zonder gratis taxi | Met partner | 6.8 | $8 + 7.7 - 9.3 = 6.4$ | $(6.8 - 6.4)^2$ |
| | Partner overleden | 5.8 | $8 + 8.8 - 9.3 = 7.5$ | $(5.8 - 7.5)^2$ |
| | Vrijgezel | 11.4 | $8 + 11.4 - 9.3 = 10.1$ | $(11.4 - 10.1)^2$ |

Tabel 3.36: Voorspellingen onder de veronderstelling van wel en geen interactie tussen MOBILITEIT en BURGERLIJKE STAND en hun gekwadrateerde verschillen. Beide voorspellingen worden gemaakt onder de veronderstelling dat er hoofdeffecten zijn van MOBILITEIT en BURGERLIJKE STAND.

taxi en met partner worden aangeduid met $(y_{11i} - \overline{y}_{11})^2$. De gekwadrateerde verschillen voor de andere cellen worden op een analoge manier aangeduid. De errorkwadratensom is 62. Dit is 47.4 minder dan $SS_{Error|A,B}$, de errorkwadratensom onder de veronderstelling dat er alleen hoofdeffecten zijn, berekend in Tabel 3.23 op pagina 182. Dus, door bij de voorspelling ook rekening te houden met de interactie tussen MOBILITEIT en BURGERLIJKE STAND, wordt de errorkwadratensom kleiner. Het verschil tussen de twee errorkwadratensommen is precies gelijk aan de kwadratensom van het interactie-effect. Dit geldt in zijn algemeenheid en niet alleen voor dit voorbeeldonderzoek.

We berekenen nu de gemiddelde kwadratensommen. De gemiddelde kwadratensom van het interactie-effect wordt berekend door deze effectkwadratensom te delen door zijn aantal vrijheidsgraden. Dit aantal vrijheidsgraden wordt als volgt berekend:

$$\boxed{\text{aantal niveaus van MOBILITEIT} - 1} \times \boxed{\text{aantal niveaus van BURGERLIJKE STAND} - 1} \; .$$

Dit product is gelijk aan $(2-1) \times (3-1)$, en dus moeten we delen door 2. De gemiddelde kwadratensom van het interactie-effect is dus gelijk aan $47.4/2 = 23.7$. Analoog, wordt ook de gemiddelde errorkwadratensom berekend door de errorkwadratensom te delen door zijn aantal vrijheidsgraden. Dit aantal vrijheidsgraden wordt als volgt berekend:

$$\boxed{\text{aantal observaties}} - \boxed{\text{aantal niveaus van MOBILITEIT}} \times \boxed{\text{aantal niveaus van BURGERLIJKE STAND}} \; .$$

| | Met partner | | Partner verloren | | Vrijgezel | | Som |
|---|---|---|---|---|---|---|---|
| | $y_{11i}$ | $(y_{11i} - \overline{y}_{11})^2$ | $y_{12i}$ | $(y_{12i} - \overline{y}_{12})^2$ | $y_{13i}$ | $(y_{13i} - \overline{y}_{13})^2$ | |
| Met gratis taxi | 10 | $(10 - 8.6)^2$ | 10 | $(10 - 11.8)^2$ | 12 | $(12 - 11.4)^2$ | |
| | 7 | $(7 - 8.6)^2$ | 13 | $(13 - 11.8)^2$ | 10 | $(10 - 11.4)^2$ | |
| | 9 | $(9 - 8.6)^2$ | 8 | $(8 - 11.8)^2$ | 12 | $(12 - 11.4)^2$ | |
| | 8 | $(8 - 8.6)^2$ | 15 | $(15 - 11.8)^2$ | 13 | $(13 - 11.4)^2$ | |
| | 9 | $(9 - 8.6)^2$ | 13 | $(13 - 11.8)^2$ | 10 | $(10 - 11.4)^2$ | |
| Som | | **5.2** | | **30.8** | | **7.2** | **43.2** |
| | $y_{21i}$ | $(y_{21i} - \overline{y}_{21})^2$ | $y_{22i}$ | $(y_{22i} - \overline{y}_{22})^2$ | $y_{23i}$ | $(y_{23i} - \overline{y}_{23})^2$ | |
| Zonder gratis taxi | 7 | $(7 - 6.8)^2$ | 4 | $(4 - 5.8)^2$ | 11 | $(11 - 11.4)^2$ | |
| | 6 | $(6 - 6.8)^2$ | 8 | $(8 - 5.8)^2$ | 12 | $(12 - 11.4)^2$ | |
| | 7 | $(7 - 6.8)^2$ | 5 | $(5 - 5.8)^2$ | 13 | $(13 - 11.4)^2$ | |
| | 8 | $(8 - 6.8)^2$ | 7 | $(7 - 5.8)^2$ | 11 | $(11 - 11.4)^2$ | |
| | 6 | $(6 - 6.8)^2$ | 5 | $(5 - 5.8)^2$ | 10 | $(10 - 11.4)^2$ | |
| Som | | **2.8** | | **10.8** | | **5.2** | **18.8** |
| | | | | | | | **62** |

Tabel 3.37: Berekening van de errorkwadratensom onder de veronderstelling van een interactie tussen MOBILITEIT en BURGERLIJKE STAND (bovenop de hoofdeffecten van deze factoren).

In dit voorbeeldonderzoek zijn er 30 observaties en 6 ($= 2 \times 3$) cellen, en dus moeten we delen door 24 ($= 30 - 6$):

$$\begin{aligned} \text{Gemiddelde errorkwadratensom} &= \frac{62}{24} \\ &= 2.58 \quad . \qquad \text{(na afronding)} \end{aligned}$$

In de laatste stap wordt de F-statistiek berekend. Dit gebeurt door de gemiddelde kwadratensom van het interactie-effect te delen door de gemiddelde errorkwadratensom:

$$\begin{aligned} F &= \frac{23.7}{2.58} \\ &= 9.19 \quad . \qquad \text{(na afronding)} \end{aligned}$$

Deze F-statistiek geeft een antwoord op de vraag of het interactie-effect dat zichtbaar is in de celgemiddelden (i.c., het feit dat, in de steekproef, het effect van MOBILITEIT verschillend is binnen de verschillende niveaus van BURGERLIJKE STAND) toevallig is. De kans om een F-statistiek te observeren die groter is dan 9.19, gesteld dat het interactie-effect in de steekproef toevallig is, is gelijk aan 0.0011. Deze kans is kleiner dan 0.05 en daarom besluiten we dat het interactie-effect in de steekproef niet toevallig is. Met andere woorden, we besluiten dat, ook in de populatie, het effect van MOBILITEIT verschillend is voor de verschillende niveaus van BURGERLIJKE STAND.

### Het Hoofdeffect van MOBILITEIT

We toetsen het hoofdeffect van MOBILITEIT onder de veronderstelling dat er een hoofdeffect is van BURGERLIJKE STAND en een interactie-effect van MOBILITEIT en BUR-

GERLIJKE STAND. Het verschil met de F-toets met de type II effectkwadratensom is dat, behalve met het hoofdeffect van BURGERLIJKE STAND, nu ook rekening gehouden wordt met het interactie-effect van MOBILITEIT en BURGERLIJKE STAND. De dubbele veronderstelling van een hoofdeffect van BURGERLIJKE STAND en een interactie-effect combineren we met elk van twee andere veronderstellingen: (1) dat er geen hoofdeffect is van MOBILITEIT, en (2) dat er wel een hoofdeffect is van MOBILITEIT. We maken dus voorspellingen onder de volgende twee veronderstellingen:

1. Een hoofdeffect van BURGERLIJKE STAND en een interactie-effect van MOBILITEIT en BURGERLIJKE STAND.

2. Hoofdeffecten van BURGERLIJKE STAND én MOBILITEIT, en een interactie-effect van MOBILITEIT en BURGERLIJKE STAND.

Onder deze tweede veronderstelling doen alle effecten zich voor, de twee hoofdeffecten en het interactie-effect. Onder deze veronderstelling zijn de voorspellingen gelijk aan de celgemiddelden. Deze celgemiddelden staan in de derde kolom van Tabel 3.38.

Voor de voorspelling onder de eerste veronderstelling (alleen een hoofdeffect van BURGERLIJKE STAND en een interactie-effect) maken we gebruik van de volgende regel:

$$\boxed{\text{voorspelling}} = \boxed{\text{globale gemiddelde}} + \boxed{\text{hoofdeffect van BURGERLIJKE STAND}} + \boxed{\text{interactie-effect}} \; .$$

Deze regel kan vereenvoudigd worden. De manier waarop dit gebeurt levert echter weinig inzicht op, en daarom geven we alleen het resultaat:

$$\boxed{\text{voorspelling}} = \boxed{\text{cel-gemiddelde}} - \boxed{\text{effect van MOBILITEIT}} \; .$$

Deze voorspellingen worden gegeven in de vierde kolom van Tabel 3.38. Bijvoorbeeld, voor de cel met gratis taxi en met partner is het celgemiddelde 8.6 en het effect van MOBILITEIT is $10.6 - 9.3$. Het verschil tussen deze twee is $8.6 - 10.6 + 9.3$, het getal in de vierde kolom.

In Tabel 3.38 staan ook de gekwadrateerde verschillen tussen de twee voorspellingen; de voorspelling onder de veronderstelling dat er een hoofdeffect is van MOBILITEIT en die onder de veronderstelling dat er geen hoofdeffect is van MOBILITEIT. Het belangrijke punt hier, is dat deze gekwadrateerde verschillen precies gelijk zijn aan de gekwadrateerde verschillen die we ook tegenkomen in de berekening van de effectkwadratensom op pagina 179: de effectkwadratensom van MOBILITEIT onder de veronderstelling dat er geen hoofdeffect is van de factor BURGERLIJKE STAND, noch een interactie-effect van MOBILITEIT en BURGERLIJKE STAND (i.e., de type I effectkwadratensom). Dus, de effectkwadratensom van MOBILITEIT is hetzelfde als men wel of geen rekening houdt met factor BURGERLIJKE STAND (zijn hoofdeffect en zijn interactie met de andere factor). Verder zullen we zien dat dit alleen het geval is in een gebalanceerd design; in een ongebalanceerd design hangt de effectkwadratensom wél af van het feit of men wel of geen rekening houdt met het effect van de andere factor.

| | | Veronderstelling waaronder de voorspelling gemaakt wordt | | |
|---|---|---|---|---|
| | | Hoofd-effect van MOBILITEIT | Geen hoofd-effect van MOBILITEIT | Gekwadrateerd verschil tussen de voorspellingen |
| Met gratis taxi | Met partner | 8.6 | $8.6 - 10.6 + 9.3$ | $(10.6 - 9.3)^2$ |
| | Partner overleden | 11.8 | $11.8 - 10.6 + 9.3$ | $(10.6 - 9.3)^2$ |
| | Vrijgezel | 11.4 | $11.4 - 10.6 + 9.3$ | $(10.6 - 9.3)^2$ |
| Zonder gratis taxi | Met partner | 6.8 | $6.8 - 8 + 9.3$ | $(8 - 9.3)^2$ |
| | Partner overleden | 5.8 | $5.8 - 8 + 9.3$ | $(8 - 9.3)^2$ |
| | Vrijgezel | 11.4 | $11.4 - 8 + 9.3$ | $(8 - 9.3)^2$ |

Tabel 3.38: Voorspellingen onder de veronderstelling van wel en geen hoofd-effect van MOBILITEIT en hun gekwadrateerde verschillen. Beide voorspellingen worden gemaakt onder de veronderstelling dat er een hoofdeffect is van BURGERLIJKE STAND en een interactie-effect.

Voor de errorkwadratensom is er wél een verschil tussen de situaties waarin men wel of geen rekening houdt met de factor BURGERLIJKE STAND (zijn hoofdeffect en zijn interactie met de andere factor). We bespraken reeds eerder het feit dat wel of geen rekening houden met het hoofdeffect van de andere factor een invloed heeft op de errorkwadratensom: als men geen rekening houdt met het hoofdeffect van BURGERLIJKE STAND, is de errorkwadratensom 181.6 en als men er wel rekening mee houdt daalt de errorkwadratensom naar 109.4. De errorkwadratensom daalt nog verder naar 62 als men bij de voorspelling ook rekening houdt met het interactie-effect van MOBILITEIT en BURGERLIJKE STAND. Het verschil tussen de twee laatste errorkwadratensommen ($109.4 - 62 = 47.4$) is precies gelijk aan de kwadratensom van het interactie-effect. Dus, door rekening te houden met het interactie-effect verkleint de errorkwadratensom met een hoeveelheid gelijk aan de kwadratensom van het interactie-effect.

We berekenen nu de gemiddelde kwadratensommen. De gemiddelde kwadratensom van het hoofdeffect van MOBILITEIT is gelijk aan de gewone kwadratensom (nl., 50.7) omdat het aantal vrijheidsgraden gelijk is aan 1 (het aantal niveaus van de factor MOBILITEIT min 1). De berekening van de gemiddelde errorkwadratensom kwam reeds aan bod in de paragraaf over de F-statistiek voor het interactie-effect. Deze gemiddelde errorkwadratensom is 2.58.

In de laatste stap, wordt opnieuw de F-statistiek berekend door de gemiddelde kwadratensom van het hoofdeffect van MOBILITEIT te delen door de gemiddelde errorkwadratensom:

$$\begin{aligned} F &= \frac{50.7}{2.58} \\ &= 19.65 \quad . \qquad \text{(na afronding)} \end{aligned}$$

Deze F-statistiek geeft een antwoord op de vraag of het verschil tussen de marginale gemiddelden voor de groepen met en zonder gratis taxi toevallig is. Deze vraag wordt

ook beantwoord met de F-statistieken op pagina 179 en 183 die berekend werden zonder rekening te houden met hoofdeffect van BURGERLIJKE STAND en/of het interactie-effect van MOBILITEIT en BURGERLIJKE STAND. Dit is in overeenstemming met het feit dat de tellers van deze drie F-statistieken precies gelijk zijn, want de teller van een F-statistiek geeft aan wat men ermee toetst.

Er is een verschil tussen de drie genoemde F-statistieken (respectievelijk, 7.81, 12.04 en 19.65) omdat hun gemiddelde errorkwadratensommen verschillend zijn. Deze gemiddelde errorkwadratensommen zijn gelijk aan, respectievelijk, 6.49, 4.21 en 2.58. De gemiddelde errorkwadratensom daalt van 6.49 naar 4.21 door bij de voorspelling rekening te houden met het hoofdeffect van BURGERLIJKE STAND, en hij daalt verder naar 2.58 door bij de voorspelling ook rekening te houden met het interactie-effect. De noemer van de F-statistiek geeft aan hoe gevoelig men toetst, en daarom kunnen we stellen dat er gevoeliger getoetst wordt als men bij de voorspelling rekening houdt met de andere factor (zijn hoofdeffect en zijn interactie-effect met de eerste factor). Dit blijkt ook uit de p-waarde van de laatste F-statistiek. Deze is gelijk aan 0.000176, hetgeen kleiner is dan de p-waarden van de andere twee F-statistieken.

### Het Hoofdeffect van BURGERLIJKE STAND

De F-statistiek voor het hoofdeffect van BURGERLIJKE STAND wordt op een analoge manier berekend als die voor het hoofdeffect van MOBILITEIT. We toetsen het hoofdeffect van BURGERLIJKE STAND onder de veronderstelling dat er een hoofdeffect is van MOBILITEIT en een interactie-effect van MOBILITEIT en BURGERLIJKE STAND. Het verschil met de F-toets met de type II effectkwadratensom is dat, behalve met het hoofdeffect van MOBILITEIT, nu ook rekening gehouden wordt met het interactie-effect van MOBILITEIT en BURGERLIJKE STAND. De kwadratensom van het hoofdeffect van BURGERLIJKE STAND is hetzelfde als men bij de voorspelling wel of geen rekening houdt met de factor MOBILITEIT (zijn hoofdeffect en zijn interactie met de andere factor). In alledrie de gevallen vinden we 72.2.

Voor de errorkwadratensom is er wél een verschil tussen de situaties waarin men wel of geen rekening houdt met de factor MOBILITEIT (zijn hoofdeffect en zijn interactie met de andere factor). Zo lieten we in de vorige paragraaf zien dat, door rekening te houden met het interactie-effect, de errorkwadratensom verkleint met een hoeveelheid gelijk aan de kwadratensom van het interactie-effect.

De F-statistiek voor het hoofdeffect van BURGERLIJKE STAND is gelijk aan 13.99, hetgeen groter is dan de F-statistieken voor de situatie waarin men alleen rekening houdt met het hoofdeffect van MOBILITEIT (i.c., 8.57) en de situatie waarin men helemaal geen rekening houdt met de factor MOBILITEIT (i.c., 6.09). Het verschil tussen deze drie F-statistieken komt omdat de gemiddelde errorkwadratensommen verschillend zijn: door bij de voorspelling ook rekening te houden met het interactie-effect, daalt de gemiddelde errorkwadratensom en wordt er gevoeliger getoetst.

Deze F-statistiek geeft een antwoord op de vraag of de verschillen tussen de marginale gemiddelden van de drie niveaus van de factor BURGERLIJKE STAND toevallig zijn. Omdat de p-waarde van deze F-statistiek zo klein is (hij is gelijk aan 0.000094), besluiten we dat deze verschillen niet toevallig zijn.

**Overzichtstabel**

In Tabel 3.39 geven we een overzicht van de kwadratensommen en de F-statistieken voor het toetsen van de hoofd- en interactie-effecten van MOBILITEIT en BURGERLIJKE STAND. Ter vergelijking, geven we in deze tabel ook de resultaten van eerdere analyses waarin geen rekening gehouden werd met het hoofdeffect van de andere factor en het interactie-effect.

### 3.3.11 Het Rekenwerk in een Ongebalanceerd Design (Met Interactie)

We laten eerst zien hoe de F-statistiek voor het interactie-effect berekend wordt. Daarna laten we zien hoe de F-statistieken voor de twee hoofdeffecten berekend worden.

**De Interactie Tussen TRAINING en TYPE LERAAR**

De toetsing van de interactie in een ongebalanceerd design gebeurt in grote lijnen op dezelfde manier als in een gebalanceerd design. Het enige wezenlijke verschil is dat we niet over een eenvoudige formule beschikken om de kwadratensom van het interactie-effect uit te rekenen. Concreet, voor de berekening van de kwadratensom van het interactie-effect hebben we voorspellingen nodig onder de volgende twee veronderstellingen:

1. Alleen hoofdeffecten van TRAINING en TYPE LERAAR.
2. Hoofdeffecten van TRAINING en TYPE LERAAR, én een interactie-effect van TRAINING en TYPE LERAAR.

Onder de eerste veronderstelling, maken we een voorspelling op basis van de volgende regel:

$$\boxed{\text{voorspelling}} = \boxed{\begin{array}{c}\text{globale}\\ \text{gemiddelde}\end{array}} + \boxed{\begin{array}{c}\text{effect van}\\ \text{TRAINING}\end{array}} + \boxed{\begin{array}{c}\text{effect van}\\ \text{TYPE LERAAR}\end{array}} .$$

Anders dan in een gebalanceerd design, beschikken we hier niet over eenvoudige formules voor het effect van TRAINING en het effect van TYPE LERAAR. Deze effecten kunnen we echter wel berekenen met de computer.

Onder de veronderstelling dat er, behalve de twee hoofdeffecten, ook een interactie-effect is, zijn de voorspellingen gelijk aan de celgemiddelden. Hier is dus geen ingewikkelde berekening nodig. Toch is er geen eenvoudige formule voor de kwadratensom van het interactie-effect, want de voorspellingen onder de veronderstelling van alleen hoofdeffecten zijn zo moeilijk te berekenen. We geven deze kwadratensom zonder verdere motivering: hij is gelijk aan 44.1545.

In tegenstelling tot de kwadratensom van het interactie-effect is de errorkwadratensom wél eenvoudig te berekenen: hij is gelijk aan de som van de gekwadrateerde verschillen tussen de individuele observaties en de celgemiddelden (de voorspellingen onder de veronderstelling van hoofd- en interactie-effecten). Maar omdat we voor het ongebalanceerde voorbeeldonderzoek de individuele observaties niet gegeven hebben, geven

| Interactie-effect van MOBILITEIT en BURGERLIJKE STAND |
|---|
| Vergelijking: (MOBILITEIT, BURGERLIJKE STAND, interactie) vs.<br>(MOBILITEIT, BURGERLIJKE STAND)<br>Effectkwadratensom: 47.4<br>Errorkwadratensom: 62<br>F-statistiek: 9.19 |
| Hoofdeffect van MOBILITEIT |
| 1. Zonder rekening te houden met het hoofdeffect van BURGERLIJKE STAND en de interactie |
| Vergelijking: MOBILITEIT vs. niets<br>Effectkwadratensom: 50.7<br>Errorkwadratensom: 181.6<br>F-statistiek: 7.81 |
| 2. Alleen rekening houdend met het hoofdeffect van BURGERLIJKE STAND |
| Vergelijking: (MOBILITEIT, BURGERLIJKE STAND) vs.<br>BURGERLIJKE STAND<br>Effectkwadratensom: 50.7<br>Errorkwadratensom: 109.4<br>F-statistiek: 12.04 |
| 3. Rekening houdend met het hoofdeffect van BURGERLIJKE STAND en de interactie |
| Vergelijking: (MOBILITEIT, BURGERLIJKE STAND, interactie) vs.<br>(BURGERLIJKE STAND, interactie)<br>Effectkwadratensom: 50.7<br>Errorkwadratensom: 62<br>F-statistiek: 19.65 |
| Hoofdeffect van BURGERLIJKE STAND |
| 1. Zonder rekening te houden met het hoofdeffect van MOBILITEIT en de interactie |
| Vergelijking: BURGERLIJKE STAND vs. niets<br>Effectkwadratensom: 72.2<br>Errorkwadratensom: 160.1<br>F-statistiek: 6.09 |
| 2. Alleen rekening houdend met het hoofdeffect van MOBILITEIT |
| Vergelijking: (BURGERLIJKE STAND, MOBILITEIT) vs.<br>MOBILITEIT<br>Effectkwadratensom: 72.2<br>Errorkwadratensom: 109.4<br>F-statistiek: 8.57 |
| 3. Rekening houdend met het hoofdeffect van MOBILITEIT en de interactie |
| Vergelijking: (BURGERLIJKE STAND, MOBILITEIT, interactie) vs.<br>(MOBILITEIT, interactie)<br>Effectkwadratensom: 72.2<br>Errorkwadratensom: 62<br>F-statistiek: 13.99 |

Tabel 3.39: Overzichtstabel van de kwadratensommen en de F-statistieken voor het toetsen van de hoofd- en interactie-effecten van MOBILITEIT en BURGERLIJKE STAND.

we hier de errorkwadratensom zonder verdere motivering: hij is gelijk aan 194. Dit is 44.1545 minder dan $SS_{Error|A,B}$, de errorkwadratensom onder de veronderstelling dat er alleen hoofdeffecten zijn. Dus, door bij de voorspelling ook rekening te houden met de interactie tussen TRAINING en TYPE LERAAR, wordt de errorkwadratensom kleiner. Het verschil tussen de twee errorkwadratensommen is precies gelijk aan de kwadratensom van het interactie-effect.

We berekenen nu de gemiddelde kwadratensommen. De gemiddelde kwadratensom van het interactie-effect wordt berekend door deze effectkwadratensom te delen door zijn aantal vrijheidsgraden. Dit aantal vrijheidsgraden is gelijk aan $(2-1) \times (3-1) = 2$ (het aantal niveaus van TRAINING min 1, maal het aantal niveaus van TYPE LERAAR min 1). De gemiddelde kwadratensom van het interactie-effect is dus gelijk aan $44.1545/2 = 22.08$ (na afronding). Analoog, wordt ook de gemiddelde errorkwadratensom berekend door de errorkwadratensom te delen door zijn aantal vrijheidsgraden. Dit aantal vrijheidsgraden is gelijk aan $90 - 6 = 84$, het aantal observaties min het aantal cellen. De gemiddelde errorkwadratensom is dus gelijk aan $194/84 = 2.31$ (na afronding).

In de laatste stap wordt de F-statistiek berekend door de gemiddelde kwadratensom van het interactie-effect te delen door de gemiddelde errorkwadratensom: $22.08/2.31 = 9.56$ (na afronding). Deze F-statistiek geeft een antwoord op de vraag of het interactie-effect dat zichtbaar is in de celgemiddelden (i.c., het feit dat, in de steekproef, het effect van TRAINING verschillend is binnen de verschillende niveaus van TYPE LERAAR) toevallig is. De kans om een F-statistiek te observeren die groter is dan 9.56, gesteld dat het interactie-effect in de steekproef toevallig is, is slechts 0.000182. Daarom besluiten we dat het interactie-effect in de steekproef niet toevallig is. Met andere woorden, we besluiten dat, ook in de populatie, het effect van TRAINING verschillend is voor de verschillende niveaus van TYPE LERAAR.

### Het Hoofdeffect van TRAINING

De toetsing van een hoofdeffect in een ongebalanceerd design gebeurt in grote lijnen op dezelfde manier als in een gebalanceerd design. Anders dan bij de type II effectkwadratensom, wordt er, behalve voor het hoofdeffect van TYPE LERAAR, nu ook uitgezuiverd voor het interactie-effect van TRAINING en TYPE LERAAR. Het enige wezenlijke verschil met een gebalanceerd design is dat we niet over een eenvoudige formule beschikken om de effectkwadratensom uit te rekenen. Concreet, voor de berekening van de kwadratensom van het hoofdeffect van TRAINING hebben we voorspellingen nodig onder de volgende twee veronderstellingen:

1. Een hoofdeffect van TYPE LERAAR en een interactie-effect van TRAINING en TYPE LERAAR.

2. Hoofdeffecten van TYPE LERAAR én TRAINING, en een interactie-effect van TRAINING en TYPE LERAAR.

Voor de voorspelling onder de eerste veronderstelling (alleen een hoofdeffect van TYPE LERAAR en een interactie-effect) maken we gebruik van de volgende regel:

$$\boxed{\text{voorspelling}} = \boxed{\begin{array}{c}\text{globale}\\ \text{gemiddelde}\end{array}} + \boxed{\begin{array}{c}\text{hoofdeffect van}\\ \text{TYPE LERAAR}\end{array}} + \boxed{\begin{array}{c}\text{interactie-}\\ \text{effect}\end{array}} .$$

Anders dan in een gebalanceerd design, beschikken we hier niet over eenvoudige formules voor het hoofdeffect van TRAINING en het interactie-effect. Deze effecten kunnen we echter wel berekenen met de computer.

Onder de veronderstelling dat er, behalve de twee hoofdeffecten, ook een interactie-effect is, zijn de voorspellingen gelijk aan de celgemiddelden. Hier is dus geen ingewikkelde berekening nodig. Maar omdat de voorspellingen onder de andere veronderstelling (alleen het hoofdeffect van TYPE LERAAR en het interactie-effect) zo moeilijk te berekenen zijn, is er ook geen eenvoudige formule voor de effectkwadratensom van TRAINING. We geven deze effectkwadratensom zonder verdere motivering: hij is precies gelijk aan 0.

Deze effectkwadratensom is een type III kwadratensom. Een type III kwadratensom is een uitgezuiverde kwadratensom: het is de effectkwadratensom van TRAINING uitgezuiverd voor TYPE LERAAR (zowel voor het hoofdeffect van deze factor als voor de interactie met TRAINING). Met deze uitgezuiverde effectkwadratensom toetst men de nulhypothese dat de ongewogen marginale verwachte waarden $\mu_{k\bullet}$ gelijk zijn. Het valt op dat deze effectkwadratensom precies gelijk is aan 0. Dit komt omdat de ongewogen marginale gemiddelden in de steekproef (i.e., $\overline{y}_{1\bullet}$ en $\overline{y}_{2\bullet}$) precies gelijk zijn. Deze uitgezuiverde effectkwadratensom is altijd precies gelijk aan 0 als de ongewogen marginale gemiddelden exact gelijk zijn. Het moet hier opgemerkt worden dat het uiterst zelden voorkomt dat de ongewogen marginale gemiddelden in de steekproef exact gelijk zijn. Toch is dit getallenvoorbeeld nuttig, want het suggereert terecht dat de type III kwadratensom afhangt van het verschil tussen de ongewogen marginale gemiddelden.

Omdat de effectkwadratensom gelijk is aan 0, is ook de F-statistiek gelijk aan 0. De gemiddelde effectkwadratensom (de effectkwadratensom gedeeld door zijn aantal vrijheidsgraden) verschijnt immers in de teller van de F-statistiek. Echter, in de praktijk is de gemiddelde effectkwadratensom vrijwel altijd verschillend van 0, en moet men de gemiddelde errorkwadratensom kennen om de F-statistiek te kunnen berekenen. Deze gemiddelde errorkwadratensom werd ook gebruikt voor de berekening van de F-statistiek van het interactie-effect. Hij is gelijk aan 2.31. Dit is kleiner dan de gemiddelde errorkwadratensom waarin rekening gehouden werd met de hoofdeffecten van TRAINING en TYPE LERAAR maar niet met het interactie-effect (i.e., 2.77).

Met deze F-statistiek toetst men de nulhypothese dat er, in de populatie, geen verschillen zijn tussen de ongewogen marginale gemiddelden voor de twee niveaus van TRAINING (i.e., de $\mu_{k\bullet}$'s). Omdat de F-statistiek gelijk is aan 0, wordt deze nulhypothese niet verworpen.

### Het Hoofdeffect van TYPE LERAAR

Omdat we later de type III effectkwadratensom van TYPE LERAAR nog nodig hebben, berekenen we nu de F-statistiek voor het hoofdeffect van deze factor. Met deze F-

statistiek toetst men het verschil tussen de ongewogen marginale gemiddelden van de verschillende niveaus van TYPE LERAAR (i.e., het voor TRAINING uitgezuiverde verschil tussen de niveaus van TYPE LERAAR). Voor de beantwoording van de vraag of de training een effect heeft, is men niet in dit verschil geïnteresseerd; voor de beantwoording van deze vraag moet men de effectkwadratensom van de onafhankelijke variabele TRAINING uitzuiveren voor de storende variabele TYPE LERAAR, en niet omgekeerd.

De F-statistiek voor het hoofdeffect van TYPE LERAAR wordt op een analoge manier berekend als die voor het hoofdeffect van TRAINING. We beschikken ook hier niet over een eenvoudige formule om de effectkwadratensom uit te rekenen. We geven deze effectkwadratensom zonder verdere motivering: hij is gelijk aan 578.269. Deze effectkwadratensom is 77.1325 kleiner dan de type II effectkwadratensom die alleen uitgezuiverd wordt voor het hoofdeffect van TYPE LERAAR. Dit verschil is toe te schrijven aan het feit dat nu ook uitgezuiverd wordt voor het interactie-effect.

Ook voor de errorkwadratensom is er een verschil tussen de situatie waarin men geen rekening houdt met het interactie-effect en de situatie waarin men dit wel doet. Zoals in één van de vorige paragrafen werd vastgesteld, daalt de errorkwadratensom van 238.1545 naar 194 als bij de voorspelling, behalve met de hoofdeffecten van TRAINING en TYPE LERAAR, ook rekening gehouden wordt met het interactie-effect. Evenzo daalt de gemiddelde errorkwadratensom van 2.77 naar 2.31 door bij de voorspelling rekening te houden met het interactie-effect.

De F-statistiek voor het hoofdeffect van TYPE LERAAR is gelijk aan 125.17. Met deze F-statistiek toetst men de nulhypothese dat er, in de populatie, geen verschillen zijn tussen de ongewogen marginale gemiddelden voor de drie niveaus van TYPE LERAAR (i.e., de $\mu_{\bullet l}$'s). Omdat de p-waarde van deze F-statistiek zo klein is (i.e., kleiner dan 0.00000000001), wordt deze nulhypothese verworpen.

#### Overzichtstabel

In Tabel 3.40 geven we een overzicht van de kwadratensommen en de F-statistieken voor het toetsen van de hoofd- en interactie-effecten van TRAINING en TYPE LERAAR. Ter vergelijking, geven we in deze tabel ook de resultaten van eerdere analyses waarin geen rekening gehouden werd met het hoofdeffect van de andere factor en het interactie-effect.

### 3.3.12 F-statistieken en Kwadratensommen (Met Interactie)

We geven nu de formules die nodig zijn om de F-statistieken ook voor andere gegevens uit te rekenen. Hierbij maken we gebruik van het model met hoofd- en interactie-effecten. We beginnen met een overzicht van de verschillende F-statistieken.

#### De F-statistieken

Alle nulhypotheses worden getoetst met een F-statistiek van de volgende algemene vorm:

$$\boxed{F = \frac{MS_{Effect|Ref}}{MS_{Error|Model}} \quad .}$$

| Interactie-effect van TRAINING en TYPE LERAAR |
|---|
| Vergelijking: (TRAINING, TYPE LERAAR, interactie) vs.<br>(TRAINING, TYPE LERAAR)<br>Effectkwadratensom: 44.1545<br>Errorkwadratensom: 194<br>F-statistiek: 9.56 |
| Hoofdeffect van TRAINING |
| 1. Zonder rekening te houden met het hoofdeffect van TYPE LERAAR<br>en de interactie |
| Vergelijking: TRAINING vs. niets<br>Effectkwadratensom: 112.5<br>Errorkwadratensom: 893.556<br>F-statistiek: 11.08 |
| 2. Alleen rekening houdend met het hoofdeffect van TYPE LERAAR |
| Vergelijking: (TRAINING, TYPE LERAAR) vs.<br>TYPE LERAAR<br>Effectkwadratensom: 0.0055<br>Errorkwadratensom: 238.1545<br>F-statistiek: 0.002 |
| 3. Rekening houdend met het hoofdeffect van TYPE LERAAR en de interactie |
| Vergelijking: (TRAINING, TYPE LERAAR, interactie) vs.<br>(TYPE LERAAR, interactie)<br>Effectkwadratensom: 0<br>Errorkwadratensom: 194<br>F-statistiek: 0 |
| Hoofdeffect van TYPE LERAAR |
| 1. Zonder rekening te houden met het hoofdeffect van TRAINING en de interactie |
| Vergelijking: TYPE LERAAR vs. niets<br>Effectkwadratensom: 767.896<br>Errorkwadratensom: 238.16<br>F-statistiek: 140.256 |
| 2. Alleen rekening houdend met het hoofdeffect van TRAINING |
| Vergelijking: (TYPE LERAAR, TRAINING) vs.<br>TRAINING<br>Effectkwadratensom: 655.4015<br>Errorkwadratensom: 238.1545<br>F-statistiek: 118.304 |
| 3. Rekening houdend met het hoofdeffect van TRAINING en de interactie |
| Vergelijking: (TYPE LERAAR, TRAINING, interactie) vs.<br>(TRAINING, interactie)<br>Effectkwadratensom: 578.269<br>Errorkwadratensom: 194<br>F-statistiek: 125.17 |

Tabel 3.40: Overzichtstabel van de kwadratensommen en de F-statistieken voor het toetsen van de hoofd- en interactie-effecten van TRAINING en TYPE LERAAR.

In deze formule kunnen we voor *Effect* het volgende invullen: (1) het hoofdeffect van A, (2) het hoofdeffect van B, en (3) het interactie-effect van A en B. Dit effect moet men nader specificeren door aan te geven wat het referentiemodel is. Indien men het interactie-effect van A en B wil toetsen, dan kiest men als referentiemodel altijd het model met de hoofdeffecten van A en B; men toetst geen interessante nulhypotheses met andere referentiemodellen dan het model met de hoofdeffecten van A en B. Daarom wordt er in de praktijk alleen gebruik gemaakt van $MS_{A\times B|A,B}$ (*Effect*=$A \times B$ en *Ref*=$A, B$).

Voor de hoofdeffecten van A en B, moet men een onderscheid maken tussen gewogen en ongewogen marginale gemiddelden. Indien men de verschillen tussen de gewogen marginale gemiddelden wil toetsen, dan kiest men het model zonder effecten als referentiemodel. In dit geval, maakt men gebruik van de type I kwadratensommen $SS_A$ en $SS_B$. In verreweg de meeste onderzoeken wil men de verschillen tussen de ongewogen marginale gemiddelden toetsen. In deze onderzoeken kiest men als referentiemodel het model met het hoofdeffect van de andere factor en het interactie-effect. Men maakt dan gebruik van de type III kwadratensommen $SS_{A|B,A\times B}$ (*Effect*=$A$ en *Ref*=$B, A \times B$) en $SS_{B|A,A\times B}$ (*Effect*=$B$ en *Ref*=$A, A \times B$). In een van de vorige paragrafen werd gezegd dat men de verschillen tussen de ongewogen marginale gemiddelden ook kan toetsen met de type II kwadratensommen $SS_{A|B}$ en $SS_{B|A}$, waarbij het referentiemodel alleen het hoofdeffect van de andere factor bevat. Dat klopt, maar alleen als er (in de populatie) geen interactie-effect is. We komen hier nog op terug.

De noemer van de F-statistiek, de gemiddelde errorkwadratensom $MS_{Error|Model}$, bepaalt de gevoeligheid waarmee men toetst. De errorkwadratensom hangt af van het model dat men gebruikt om de observaties te voorspellen. Hier gebruiken we het model met hoofd- en interactie-effecten om de observaties te voorspellen.

### Het Model met Hoofd- en Interactie-effecten

Bij de berekening van de effect- en errorkwadratensommen maken we gebruik van het model met hoofd- en interactie-effecten. Dit model is equivalent met een model waarin elke cel $(k, l)$ een eigen verwachte waarde heeft. Deze verwachte waarde duiden we aan met $\mu_{kl}$. Dit model kunnen we als volgt schrijven:

$$\boxed{y_{kli} = \mu_{kl} + e_{kli} \quad .}$$

De $e_{kli}$'s binnen elk van de $K \times L$ cellen hebben een verwachte waarde gelijk aan 0. Het model met alleen hoofdeffecten is een bijzonder geval van het bovenstaande model. Volgens het model met alleen hoofdeffecten is $\mu_{kl}$ gelijk aan $\mu+\alpha_k+\beta_l$. Om de relatie met het model met alleen hoofdeffecten te verduidelijken, introduceren we nu de *interactie-effectparameters* $\alpha\beta_{kl}$:

$$\boxed{\alpha\beta_{kl} = \mu_{kl} - (\mu + \alpha_k + \beta_l) \quad .}$$

Dus, $\alpha\beta_{kl}$ is het verschil tussen de verwachte waarde volgens het model zonder restricties en de verwachte waarde volgens het model met alleen hoofdeffecten. Hieruit volgt dat $\mu_{kl}$ als volgt geschreven kan worden:

$$\mu_{kl} = \mu + \alpha_k + \beta_l + \alpha\beta_{kl} \quad .$$

Door deze formule te substitueren in het model met de celspecifieke verwachte waarde $\mu_{kl}$ krijgen we het model met hoofd- en interactie-effecten:

$$\boxed{y_{kli} = \mu + \alpha_k + \beta_l + \alpha\beta_{kl} + e_{kli} \quad .}$$

Op de interactie-effectparameters $\alpha\beta_{kl}$ ligt eenzelfde soort restrictie als op de hoofdeffectparameters $\alpha_k$ en $\beta_l$. Voor deze laatste parameters geldt dat hun sommen ($\sum_{k=1}^{K} \alpha_k$ en $\sum_{l=1}^{L} \beta_l$) gelijk zijn aan 0. En voor de $\alpha\beta_{kl}$'s geldt het volgende:

1. Binnen elk niveau van de eerste factor (A) is de som van de interactie-effectparameters over de niveaus van de tweede factor (B) gelijk aan 0. In de vorm van een formule:
$$\sum_{l=1}^{L} \alpha\beta_{kl} = 0, \qquad \text{voor } \ k = 1, \ldots, K.$$

2. Binnen elk niveau van de tweede factor (B) is de som van de interactie-effectparameters over de niveaus van de eerste factor (A) gelijk aan 0. In de vorm van een formule:
$$\sum_{k=1}^{K} \alpha\beta_{kl} = 0, \qquad \text{voor } \ l = 1, \ldots, L.$$

Er liggen dus $K + L$ restricties op de $\alpha\beta_{kl}$'s: binnen elk van de $K$ niveaus van de eerste factor ($k = 1, \ldots, K$) is de som van de $\alpha\beta_{kl}$'s gelijk aan 0, alsook binnen elk van de $L$ niveaus van de tweede factor ($l = 1, \ldots, L$). Het aantal effectieve restricties is echter $K + L$ min 1. Dit komt omdat de genoemde restricties één restrictie gemeenschappelijk hebben, namelijk dat de som van alle $K \times L$ interactie-effectparameters $\alpha\beta_{kl}$ gelijk is aan 0:

$$\sum_{k=1}^{K} \sum_{l=1}^{L} \alpha\beta_{kl} = 0 \quad .$$

Het model met hoofd- en interactie-effecten is equivalent met het model met de celspecifieke verwachte waarde $\mu_{kl}$, want de parameters in het model met hoofd- en interactie-effecten ($\mu$, $\alpha_k$, $\beta_l$, en $\alpha\beta_{kl}$) worden berekend op basis van de $\mu_{kl}$'s. Omgekeerd, kunnen de $\mu_{kl}$'s ook terugberekend worden op basis van deze parameters. Dit illustreren we aan de hand van een getallenvoorbeeld. In Tabel 3.41 tonen we de $\mu_{kl}$'s van een onderzoek waarin A en B elk drie niveaus hebben. Deze $\mu_{kl}$'s staan linksboven in de middelste cellen (niet vet gedrukt). Op basis van deze $\mu_{kl}$'s berekenen we eerst de globale verwachte waarde $\mu$. Deze is gelijk aan het gemiddelde van alle $\mu_{kl}$'s (hier, 8). Daarna berekenen we de hoofdeffectparameters $\alpha_k$. Deze zijn gelijk aan $\mu_{k\bullet} - \mu$, waarbij $\mu_{k\bullet}$ gelijk is aan het gemiddelde van de $\mu_{kl}$'s in de k-de rij. Daarna berekenen we de hoofdeffectparameters $\beta_l$. Deze zijn gelijk aan $\mu_{\bullet l} - \mu$, waarbij $\mu_{\bullet l}$ gelijk is aan het gemiddelde van de $\mu_{kl}$'s in de l-de kolom. Tenslotte berekenen we de interactie-effectparameters $\alpha\beta_{kl}$. Deze zijn gelijk aan $\mu_{kl} - (\mu + \alpha_k + \beta_l)$. Deze $\alpha\beta_{kl}$'s staan rechtsonder vet gedrukt in de middelste cellen. Het is nuttig om te controleren dat je, via de formule $\mu_{kl} = \mu + \alpha_k + \beta_l + \alpha\beta_{kl}$, de oorspronkelijke $\mu_{kl}$'s kunt terugvinden.

| | | B | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | $\mu_{k\bullet}$ | $\alpha_k$ |
| A | 1 | 6<br>**4** | 1<br>**-2** | 5<br>**-2** | 4 | **-4** |
| | 2 | 4<br>**0** | 2<br>**-3** | 12<br>**3** | 6 | **-2** |
| | 3 | 8<br>**-4** | 18<br>**5** | 16<br>**-1** | 14 | **6** |
| | $\mu_{\bullet l}$ | 6 | 7 | 11 | **8** | |
| | $\beta_l$ | **-2** | **-1** | **3** | | |

Tabel 3.41: Matrix met verwachte waarden per cel $\mu_{kl}$ (linksboven in de middelste cellen) en de daarop gebaseerde parameters $\mu$, $\alpha_k$, $\beta_l$ en $\alpha\beta_{kl}$ (vet gedrukt).

### De Kwadratensom van het Interactie-effect

Net zoals alle andere effectkwadratensommen, wordt ook de kwadratensom van het interactie-effect bepaald op basis van de volgende algemene formule:

$$\boxed{SS_{Effect|Ref} = \sum_{Obs} (\widehat{y}_{kli}^{Ref+Effect} - \widehat{y}_{kli}^{Ref})^2 \quad .}$$

Als referentiemodel voor de kwadratensom van het interactie-effect gebuikt men altijd het model met beide hoofdeffecten. Voor de kwadratensom van het interactie-effect krijgen we dan het volgende:

$$\begin{aligned} \widehat{y}_{kli}^{Ref} &= \widehat{y}_{kli}^{A,B} \qquad \text{en} \\ \widehat{y}_{kli}^{Ref+A\times B} &= \widehat{y}_{kli}^{A,B,A\times B} \quad . \end{aligned}$$

In wat volgt, zullen we $\widehat{y}_{kli}^{A,B,A\times B}$ vereenvoudigen door A,B,A×B (de twee hoofdeffecten en het interactie-effect) verkort aan te duiden met [AB].

De voorspellingen $\widehat{y}_{kli}^{[AB]}$ en $\widehat{y}_{kli}^{A,B}$ worden gemaakt op basis van het kleinste-kwadratencriterium. De voorspelling $\widehat{y}_{kli}^{[AB]}$ is gelijk aan $\overline{y}_{kl}$, het gemiddelde in cel $(k, l)$. Dit resultaat is goed te begrijpen omdat het model met hoofd- en interactie-effecten equivalent is met het model met de celspecifieke verwachte waarde $\mu_{kl}$. Voor de andere voorspelling, $\widehat{y}_{kli}^{A,B}$ (gemaakt op basis van het twee-factoriële model met A en B), stuiten we op het probleem dat er geen eenvoudige formule voor bestaat (dit geldt alleen voor een ongebalanceerd design). In de praktijk wordt $\widehat{y}_{kli}^{A,B}$ altijd berekend met de computer. De formule van de kwadratensom van het interactie-effect is dan de volgende:

$$\boxed{SS_{A\times B|A,B} = \sum_{Obs} (\overline{y}_{kl} - \widehat{y}_{kli}^{A,B})^2 \quad .}$$

De bovenstaande formule is zowel geldig voor een gebalanceerd als een ongebalanceerd design. In een gebalanceerd design kan deze formule ook als volgt geschreven

worden:

$$SS_{A\times B|A,B} = \sum_{k=1}^{K}\sum_{l=1}^{L} n[\overline{y}_{kl} - (\overline{y}_{k\bullet} + \overline{y}_{\bullet l} - \overline{y}_{\bullet\bullet})]^2 \quad .$$

Deze formule volgt uit de vorige formule als $\widehat{y}_{kli}^{A,B}$ vervangen wordt door zijn kleinste-kwadratenschatter binnen een gebalanceerd design. Deze kleinste-kwadratenschatter wordt gegeven in de appendix op pagina 292 en volgende.

### De Type III Kwadratensommen van de Hoofdeffecten

De algemene formule voor de kwadratensom van het hoofdeffect van A ziet er als volgt uit:

$$SS_{A|Ref} = \sum_{Obs} (\widehat{y}_{kli}^{Ref+A} - \widehat{y}_{kli}^{Ref})^2 \quad .$$

Als referentiemodel, gebruiken we hier het model met het hoofdeffect van B en het interactie-effect. Voor de kwadratensom van het hoofdeffect van A krijgen we dan de volgende voorspellingen:

$$\begin{aligned} \widehat{y}_{kli}^{Ref} &= \widehat{y}_{kli}^{B,A\times B} \qquad \text{en} \\ \widehat{y}_{kli}^{Ref+A} &= \widehat{y}_{kli}^{A,B,A\times B} = \widehat{y}_{kli}^{[AB]} \quad . \end{aligned}$$

Met deze voorspellingen berekenen we een type III kwadratensom. Het verschil met een type II kwadratensom is dat, behalve voor het hoofdeffect van B, hier ook uitgezuiverd wordt voor het interactie-effect van A en B.

Toepassing van het kleinste-kwaddratencriterium levert op dat $\widehat{y}_{kli}^{[AB]}$ is gelijk aan $\overline{y}_{kl}$, het gemiddelde in cel $(k,l)$. Voor de andere voorspelling, $\widehat{y}_{kli}^{B,A\times B}$ (gemaakt op basis van het model met het hoofdeffect van B en het interactie-effect), bestaat er geen eenvoudige formule, althans niet voor een ongebalanceerd design. In de praktijk wordt $\widehat{y}_{kli}^{B,A\times B}$ altijd berekend met de computer. De formule van de kwadratensom van het hoofdeffect van A is dan de volgende:

$$\boxed{SS_{A|B,A\times B} = \sum_{Obs} (\overline{y}_{kl} - \widehat{y}_{kli}^{B,A\times B})^2 \quad .}$$

Deze formule van de type III kwadratensom is zowel geldig voor een gebalanceerd als een ongebalanceerd design. In een gebalanceerd design kan deze formule ook als volgt geschreven worden:

$$SS_{A|B,A\times B} = \sum_{k=1}^{K} Ln(\overline{y}_{k\bullet} - \overline{y}_{\bullet\bullet})^2 \quad .$$

De afleiding van deze formule wordt gegeven in de appendix op pagina 293. We zien dat, in een gebalanceerd design, de type III kwadratensom $SS_{A|B,A\times B}$ gelijk is aan de type I kwadratensom $SS_A$, gegeven op pagina 193, en de type II kwadratensom $SS_{A|B}$, gegeven op pagina 195. Dit is een belangrijk resultaat omdat, in een F-statistiek, de effectkwadratensom bepaalt wat er getoetst wordt. Het feit dat, in een gebalanceerd

design, de drie types effectkwadratensommen gelijk zijn, impliceert dat wat er met een F-statistiek getoetst wordt onafhankelijk is van het feit of men wel of niet uitzuivert voor een andere factor.

In een ongebalanceerd design is er wél een verschil tussen de gewogen en de ongewogen marginale gemiddelden en toetst een F-statistiek met een type III effectkwadratensom de nulhypothese van gelijke ongewogen verwachte waarden. Deze nulhypothese wordt ook getoetst met een F-statistiek met een type II kwadratensom. Het verschil tussen deze twee F-statistieken is dat een F-statistiek met een type II kwadratensom alleen dan gebruikt mag worden als er geen interactie is (in de populatie). Indien er wel interactie is in de populatie, dan toetst men met deze F-statistiek een andere nulhypothese[10]. Een F-statistiek met een type III kwadratensom, daarentegen, kan altijd gebruikt worden om de nulhypothese van gelijke ongewogen verwachte waarden te toetsen, of er nu sprake is van interactie of niet. Voor een eventuele interactie wordt immers uitgezuiverd in de type III kwadratensom. Omdat we nooit zeker zijn of er wel of geen interactie is in de populatie, is de type III kwadratensom altijd te verkiezen boven de type II kwadratensom.

In een ongebalanceerd design, zijn de type II en de type III kwadratensommen bijna altijd ongelijk. Bijvoorbeeld, in het ongebalanceerde voorbeeldonderzoek is de type II kwadratensom van TRAINING 0.0055 en is de type III kwadratensom precies 0. Een type III kwadratensom is precies gelijk aan 0 als de ongewogen marginale gemiddelden precies gelijk zijn, en dat is hier het geval. Het verschil tussen de type II en de type III kwadratensom is hier bijzonder klein. In de praktijk komt het echter vaak voor dat er wel een substantieel verschil is tussen de type II en de type III kwadratensom. Een voorbeeld hiervan zijn de type II en de type III kwadratensommen van TYPE LERAAR: de type II kwadratensom van TYPE LERAAR is 655.4015 en de type III kwadratensom is 578.296. De grootte van het verschil tussen de type II en de type III kwadratensom hangt af van (1) de grootte van het interactie-effect in de steekproef en (2) de verdeling van het aantal observaties over de cellen.

De type III kwadratensom van B wordt op een analoge manier berekend als die van A. De formules voor de type III kwadratensom van B zijn dezelfde als die voor de type III kwadratensom van A, behalve dat er (1) een B komt te staan waar een A staat (en omgekeerd), (2) een $l$ waar een $k$ staat (en omgekeerd), en (3) een $L$ waar een $K$ staat (en omgekeerd).

### De Errorkwadratensom

De errorkwadratensom wordt berekend op basis van de volgende algemene formule:

$$\boxed{SS_{Error|Model} = \sum_{Obs} (y_{kli} - \widehat{y}_{kli}^{Model})^2 \quad .}$$

We vergelijken hier de observaties met de voorspellingen op basis van het model met hoofd- en interactie-effecten. Deze voorspelling kwam reeds eerder aan bod en is gelijk aan $\overline{y}_{kl}$, het gemiddelde in cel $(k, l)$. Vervangen we $\widehat{y}_{kli}^{Model}$ door $\overline{y}_{kl}$, dan krijgen we de

[10] Deze nulhypothese hangt af van het interactie-effect in de populatie, maar de precieze formulering ervan is verder niet van belang.

volgende formule van de errorkwadratensom:

$$\boxed{SS_{Error|[AB]} = \sum_{Obs} (y_{kli} - \overline{y}_{kl})^2 \quad .}$$

In deze formule is [AB] een verkorte aanduiding voor A,B,A×B, de twee hoofdeffecten en het interactie-effect. Hiermee geven we aan dat we een voorspelling maken op basis van het model met hoofd- en interactie-effecten.

De berekening van $SS_{Error|[AB]}$ voor de gegevens van het gebalanceerde voorbeeldonderzoek wordt gegeven in Tabel 3.37 op pagina 213. De $SS_{Error|[AB]}$ voor deze gegevens is gelijk aan 62. Dit is 47.4 minder dan $SS_{Error|A,B}$, de errorkwadratensom onder de veronderstelling dat er alleen hoofdeffecten zijn. Voor de gegevens van het ongebalanceerde voorbeeldonderzoek, is $SS_{Error|[AB]}$ gelijk aan 194. Dit is 44.1545 minder dan $SS_{Error|A,B}$, de errorkwadratensom onder de veronderstelling dat er alleen hoofdeffecten zijn.

### De Kwadratensom-ontbindingen

De errorkwadratensom $SS_{Error|[AB]}$ is kleiner dan $SS_{Error|A,B}$. Uit de volgende kwadratensom-ontbinding blijkt hoeveel kleiner:

$$SS_{Error|A,B} = SS_{A\times B|A,B} + SS_{Error|[AB]} \quad .$$

$SS_{Error|A,B}$ is groter dan $SS_{Error|[AB]}$ omdat er bij de voorspelling geen rekening is gehouden met de interactie. Het verschil tussen de twee errorkwadratensommen is precies gelijk aan de interactie-effectkwadratensom $SS_{A\times B|A,B}$. Dus, naarmate de interactie belangrijker (groter, sterker) wordt, wordt het verschil tussen $SS_{Error|A,B}$ en $SS_{Error|[AB]}$ groter.

De kwadratensom van het interactie-effect behoort tot de modelkwadratensom, het gedeelte van de totale kwadratensom dat toe te schrijven is aan de effecten van de factoren. Men kan de modelkwadratensom op twee manieren uitrekenen:

$$\begin{aligned} \text{Modelkwadratensom} &= SS_A + SS_{B|A} + SS_{A\times B|A,B} \\ \text{Modelkwadratensom} &= SS_B + SS_{A|B} + SS_{A\times B|A,B} \quad . \end{aligned}$$

Merk op dat er opeenvolgend uitgezuiverd wordt voor niets (in $SS_A$ en $SS_B$), voor het hoofdeffect van de andere factor (in $SS_{B|A}$ en $SS_{A|B}$) en voor de hoofdeffecten van beide factoren (in $SS_{A\times B|A,B}$).

Deze modelkwadratensom wordt meestal afgedrukt in de ANOVA-tabel die men vindt in de uitdraai van een computerprogramma voor variantieanalyse. In een ongebalanceerd design, kan deze modelkwadratensom verschillen van de som van de effectkwadratensommen die ook in deze ANOVA-tabel staan. Immers, als men de type III kwadratensommen berekent, dan staat er in de ANOVA-tabel $SS_{A|B,A\times B}$, $SS_{B|A,A\times B}$ en $SS_{A\times B|A,B}$, en de som van de som van deze drie effectkwadratensommen is niet gelijk aan de modelkwadratensom, althans niet in een ongebalanceerd design.

### Nogmaals de F-Statistieken

We vertrekken opnieuw van de algemene vorm van de F-statistiek:

$$\boxed{F = \frac{MS_{Effect|Ref}}{MS_{Error|Model}} \quad .}$$

In de twee laatste rekenvoorbeelden, werd deze F-statistiek gebruikt voor het toetsen van de volgende nulhypotheses:

1. De nulhypothese dat er, in de populatie, geen interactie-effect is. Voor het toetsen van deze nulhypothese, vervangen we $MS_{Effect|Ref}$ door $MS_{A\times B|A,B}$ en $MS_{Error|Model}$ door $MS_{Error|[AB]}$.
2. De nulhypothese dat er, in de populatie, geen verschil is tussen de ongewogen marginale gemiddelden (voor de niveaus van factor A of die van factor B). Voor het toetsen van deze nulhypothese vervangen we $MS_{Effect|Ref}$ door de type III kwadratensom $MS_{A|B,A\times B}$ ($MS_{B|A,A\times B}$, indien men in factor B geïnteresseerd is) en $MS_{Error|Model}$ door $MS_{Error|[AB]}$.

In wat volgt, geven we de formules van de gemiddelde effectkwadratensommen $MS_{A\times B|A,B}$ en $MS_{A|B,A\times B}$, en de gemiddelde errorkwadratensom $MS_{Error|[AB]}$. Daarna besteden we aandacht aan de F-toetsen van de twee genoemde nulhypotheses.

**De Gemiddelde Kwadratensom van het Interactie-effect** We herhalen eerst de definitie van een gemiddelde effectkwadratensom:

$$\boxed{MS_{Effect|Ref} = \frac{SS_{Effect|Ref}}{\text{Aantal vrijheidsgraden van } SS_{Effect|Ref}} \quad .}$$

Het aantal vrijheidsgraden van $SS_{Effect|Ref}$ wordt bepaald op basis van de volgende regel:

$$\boxed{\begin{array}{c}\text{aantal vrije parameters in}\\ [\text{referentiemodel} + \text{effect}]\end{array}} \quad - \quad \boxed{\begin{array}{c}\text{aantal vrije parameters in}\\ [\text{referentiemodel}]\end{array}} \quad .$$

Er zijn twee modellen die een rol spelen bij het bepalen van de vrijheidsgraden van de kwadratensom van het interactie-effect $SS_{A\times B|A,B}$: (1) het model met de hoofdeffecten van A en B, en (2) het model met de hoofdeffecten van A en B én het interactie-effect $A \times B$. Het aantal vrije parameters in deze modellen wordt berekend in de twee laatste rijen van Tabel 3.42.

Gebruikmakend van de aantallen in de tweede kolom van Tabel 3.42 kan men zien dat het aantal vrijheidsgraden van $SS_{A\times B|A,B}$ gelijk is aan $(K-1)\times(L-1)$. Dit aantal is gelijk aan het aantal vrije effectparameters $\alpha\beta_{kl}$ dat nodig is om het interactie-effect in het variantieanalysemodel op te nemen. Dus,

$$\boxed{MS_{A\times B|A,B} = \frac{SS_{A\times B|A,B}}{(K-1)(L-1)} \quad .}$$

| Factoren | Vrije parameters | Aantal vrije parameters |
|---|---|---|
| A en A×B | $\mu$, $K-1$ vrije $\alpha_k$'s, $(K-1)(L-1)$ vrije $\alpha\beta_{kl}$'s | $K+[(K-1)(L-1)]$ |
| B en A×B | $\mu$, $L-1$ vrije $\beta_l$'s, $(K-1)(L-1)$ vrije $\alpha\beta_{kl}$'s | $L+[(K-1)(L-1)]$ |
| A en B | $\mu$, $K-1$ vrije $\alpha_k$'s, $L-1$ vrije $\beta_l$'s | $K+L-1$ |
| A, B en A×B | $\mu$, $K-1$ vrije $\alpha_k$'s, $L-1$ vrije $\beta_l$'s, $(K-1)(L-1)$ vrije $\alpha\beta_{kl}$'s | $KL$ |

Tabel 3.42: Berekening van het aantal vrije parameters in het model met het hoofdeffect van A en het interactie-effect, het model met het hoofdeffect van B en het interactie-effect, het model met de hoofdeffecten van A en B, en het model met de hoofdeffecten van A en B en het interactie-effect.

De F-statistiek met $MS_{A\times B|A,B}$ in de teller, toetst de nulhypothese dat er, in de populatie, geen interactie-effect is. Deze nulhypothese kan men ook als volgt formuleren:

1. Alle interactie-effectparameters $\alpha\beta_{kl}$ zijn gelijk aan 0.

2. In de populatie, is het effect van de ene factor hetzelfde voor de verschillende niveaus van de andere factor.

**De Gemiddelde Type III Kwadratensommen van de Hoofdeffecten** De gemiddelde type III kwadratensom van het hoofdeffect van A, aangeduid met $MS_{A|B,A\times B}$, is gelijk aan de gewone type III kwadratensom $SS_{A|B,A\times B}$, gedeeld door zijn aantal vrijheidsgraden $K-1$. Dit aantal is gelijk aan het aantal vrije effectparameters $\alpha_k$, het verschil tussen de tweede en de laatste rij van Tabel 3.42.

De F-statistiek met $MS_{A|B,A\times B}$ in de teller, toetst de nulhypothese dat er geen verschil is tussen de ongewogen marginale verwachte waarden voor de verschillende niveaus van factor A. Deze nulhypothese wordt ook getoetst met een F-statistiek met een type II kwadratensom. Een type II kwadratensom mag echter alleen gebruikt worden om deze nulhypothese te toetsen als er geen interactie is (in de populatie). Voor een type III kwadratensom geldt deze beperking niet, want voor een eventuele interactie wordt immers uitgezuiverd in de type III kwadratensom. Omdat we nooit zeker zijn of er interactie is in de populatie, is de type III kwadratensom altijd te verkiezen boven de type II kwadratensom.

De gemiddelde type III kwadratensom van het hoofdeffect van B, $MS_{B|A,A\times B}$, is volledig analoog aan die van A; nu wordt de gewone type III kwadratensom $SS_{B|A,A\times B}$, gedeeld door zijn aantal vrijheidsgraden $L-1$. De F-statistiek met $MS_{B|A,A\times B}$ in de teller, toetst de nulhypothese dat er geen verschil is tussen de ongewogen marginale verwachte waarden voor de verschillende niveaus van factor B.

**De Gemiddelde Errorkwadratensom** Voor het bepalen van $MS_{Error|[AB]}$ vertrekken we van de definitie van een gemiddelde errorkwadratensom:

$$\boxed{MS_{Error|Model} = \frac{SS_{Error|Model}}{\text{Aantal vrijheidsgraden van } SS_{Error|Model}} \quad .}$$

Het aantal vrijheidsgraden van $SS_{Error|Model}$ wordt bepaald op basis van de volgende regel:

$$\boxed{\text{aantal observaties}} \quad - \quad \boxed{\begin{array}{c}\text{aantal vrije parameters in}\\ \text{[Model]}\end{array}} \quad .$$

Voor "Model" nemen we het model met de hoofdeffecten van A en B en het interactie-effect. Dit model heeft $KL$ vrije parameters en dus is het aantal vrijheidsgraden van $SS_{Error|[AB]}$ gelijk aan $N - KL$. Dus,

$$\boxed{MS_{Error|[AB]} = \frac{SS_{Error|[AB]}}{N - KL} \quad .}$$

In de praktijk, is $MS_{Error|[AB]}$ meestal kleiner dan $MS_{Error|A,B}$. Men krijgt dus meestal een gevoeliger toets indien men ook rekening houdt met het interactie-effect bij de voorspelling van de observaties. Dit komt omdat de kwadratensom van het interactie-effect $SS_{A\times B|A,B}$ (i.e., het verschil tussen $SS_{Error|A,B}$ en $SS_{Error|[AB]}$) meestal voldoende groot is om het verschil tussen de vrijheidsgraden (respectievelijk, $N - K - L + 1$ en $N - KL$) te compenseren.

**De F-Toetsen** De F-toets maakt gebruik van het feit dat, onder een bepaalde nulhypothese, de F-statistiek een $\mathcal{F}$-verdeling heeft met het aantal vrijheidsgraden gelijk aan de vrijheidsgraden van de teller en de noemer van deze F-statistiek. We beschouwen hier drie nulhypotheses: (1) de nulhypothese dat er, in de populatie, geen interactie-effect is, (2) de nulhypothese dat er, in de populatie, geen verschil is tussen de ongewogen marginale gemiddelden (voor factor A of factor B), en (3) de nulhypothese dat er, in de populatie, geen verschil is tussen de gewogen marginale gemiddelden.

1. De eerste nulhypothese wordt getoetst met een F-statistiek met $MS_{A\times B|A,B}$ in de teller en $MS_{Error|[AB]}$ in de noemer. Onder de nulhypothese dat er, in de populatie, geen interactie-effect is, heeft deze F-statistiek een $\mathcal{F}$-verdeling met $(K-1)\times(L-1)$ tellervrijheidsgraden en $N - KL$ noemervrijheidsgraden.

2. De tweede nulhypothese wordt getoetst met een F-statistiek met de gemiddelde type III kwadratensom $MS_{A|B,A\times B}$ of $MS_{B|A,A\times B}$ in de teller en $MS_{Error|[AB]}$ in de noemer. Onder de nulhypothese dat er, in de populatie, geen verschil is tussen de ongewogen marginale gemiddelden, heeft deze F-statistiek een $\mathcal{F}$-verdeling met, respectievelijk, $K-1$ (voor factor A) of $L-1$ (voor factor B) tellervrijheidsgraden en $N - KL$ noemervrijheidsgraden.

3. De derde nulhypothese wordt getoetst met een F-statistiek met de gemiddelde type I kwadratensom $MS_A$ of $MS_B$ in de teller en $MS_{Error|[AB]}$ in de noemer. Onder de nulhypothese dat er, in de populatie, geen verschil is tussen de gewogen marginale gemiddelden, heeft deze F-statistiek een $\mathcal{F}$-verdeling met, respectievelijk, $K-1$ (voor factor A) of $L-1$ (voor factor B) tellervrijheidsgraden en $N-KL$ noemervrijheidsgraden.

   Men kan deze nulhypothese ook toetsen met een F-statistiek die een andere noemer heeft: $MS_{Error|A}$ of $MS_{Error|A,B}$ in plaats van $MS_{Error|[AB]}$. Het voordeel van de F-statistiek met $MS_{Error|[AB]}$ in de noemer is dat deze gemiddelde errorkwadratensom meestal het kleinst is, en daarom resulteert in de gevoeligste toets van de genoemde nulhypothese (i.e., de toets die het snelst significant is).

   We bespreken deze laatste F-statistiek alleen voor de volledigheid. In verreweg de meeste onderzoeken is men niet geïnteresseerd in de nulhypothese die met deze F-statistiek getoetst wordt (gelijkheid van de gewogen verwachte waarden). Als het design ongebalanceerd is, dan wil de onderzoeker bijna altijd uitzuiveren voor de aanwezige storende variabelen, en dat kan alleen met een type III kwadratensom.

De F-toets houdt in dat men de kans berekent onder de genoemde $\mathcal{F}$-verdeling (en dus ook onder de genoemde nulhypothese) om een F-statistiek te observeren die groter of gelijk is aan de geobserveerde F-statistiek. Deze p-waarde is een maat voor de aannemelijkheid van de nulhypothese. Een kleine p-waarde betekent dat de geobserveerde data een kleine kans van voorkomen hebben onder de nulhypothese. Als deze p-waarde kleiner is dan 0.05 of 0.01 dan zegt men dat de F-statistiek significant is, en verwerpt men de nulhypothese.

**Het Belang van de p-waarde** Net zoals bij de variantieanalyse met één factor, mag men ook hier het belang van de p-waarde niet overschatten. Soms is deze p-waarde zeer klein ondanks het feit dat de verschillen tussen de gemiddelden waar men in geïnteresseerd is (bijvoorbeeld, de verschillen tussen de ongewogen marginale gemiddelden van factor A) niet bijzonder groot zijn.

Deze situatie kan zich voordoen als het aantal observaties erg groot is. De verklaring voor dit fenomeen is opnieuw dat de teller van de F-statistiek ($MS_{A\times B|A,B}$, $MS_{A|B,A\times B}$, $MS_A$, ...) meestal sterk toeneemt met het aantal observaties en de noemer, $MS_{Error|[AB]}$, helemaal niet.

### De Proportie Verklaarde Variantie

Het is een goede strategie om bij de data-analyse niet alleen te kijken naar de p-waarde van de F-toets. Behalve in de vraag of een bepaald hoofd- of interactie-effect significant is, is men meestal ook geïnteresseerd in de vraag hoe belangrijk dit effect is voor de verklaring van de (variantie in de) afhankelijk variabele. Deze belangrijkheid kunnen we uitdrukken in termen van de proportie verklaarde variantie (PVV).

In de variantieanalyse met één factor, is de PVV de verhouding van de effectkwadratensom $SS_A$ ten opzichte van de totale kwadratensom $SS_{Totaal}$. In de variantieanalyse met twee factoren, hebben we met meer dan één effect te maken, en daarom ook met

meer dan één PVV. Om te beginnen, zijn er de kwadratensommen van de hoofdeffecten van beide factoren, A en B. Voor elke factor, zijn er drie PVV's, elk corresponderend met één van de drie effectkwadratensommen:

1. De type I kwadratensom ($SS_A$ of $SS_B$) gedeeld door $SS_{Totaal}$. Deze PVV drukt uit welk gedeelte van de variantie in de afhankelijke variabele verklaard wordt door een factor.

2. De type II kwadratensom ($SS_{A|B}$ of $SS_{B|A}$) gedeeld door $SS_{Totaal}$. Deze PVV drukt uit welk gedeelte van de variantie in de afhankelijke variabele verklaard wordt door een factor bovenop het hoofdeffect van de andere factor.

3. De type III kwadratensom ($SS_{A|B,A\times B}$ of $SS_{B|A,A\times B}$) gedeeld door $SS_{Totaal}$. Deze PVV drukt uit welk gedeelte van de variantie in de afhankelijke variabele verklaard wordt door een factor bovenop het hoofdeffect van de andere factor en het interactie-effect.

De bijdrage van een factor bovenop een andere factor, wordt ook de *unieke bijdrage* van een factor genoemd. Indien men in deze unieke bijdrage geïnteresseerd is, dan berekent men de PVV met de type-III kwadratensom.

Behalve de drie PVV's voor elk van de twee hoofdeffecten, kan men ook een PVV berekenen voor het interactie-effect. Deze PVV is gelijk aan de kwadratensom van het interactie-effect $SS_{A\times B|A,B}$ gedeeld door $SS_{Totaal}$. Deze PVV drukt uit hoe belangrijk het interactie-effect is voor de verklaring van de variantie in de afhankelijk variabele.

Voor het gebalanceerde voorbeeldonderzoek, is de PVV voor het hoofdeffect van A $50.7/232.2 = 0.22$, die voor het hoofdeffect van B $72.2/232.2 = 0.31$, en die voor het interactie-effect $47.4/232.2 = 0.20$. Voor het ongebalanceerde voorbeeldonderzoek, hebben we meerdere PVV's voor elk hoofdeffect, omdat we hier een onderscheid moeten maken tussen de unieke en de niet-unieke bijdrage.

### Overzichtstabellen

We geven nu vier overzichtstabellen met de belangrijkste formules van deze paragraaf (*Kwadratensommen en F-statistieken voor de Situatie Met Interactie*).

1. In Tabel 3.43 staan de vier modellen die gebruikt worden voor de berekening van de effect- en de errorkwadratensommen.

2. In Tabel 3.44 staan de F-toetsen voor de gewogen en de ongewogen marginale gemiddelden van de twee factoren en voor het interactie-effect.

3. In Tabel 3.45 staan de type III hoofdeffectkwadratensommen, de interactie-effectkwadratensom en hun vrijheidsgraden.

4. In Tabel 3.32 staan de errorkwadratensommen, hun vrijheidsgraden en hun kwadratensomontbindingen.

| Modellen met en zonder Interactie | | |
|---|---|---|
| | Effecten | Formule |
| (4) | A en B | $y_{kli} = \mu + \alpha_k + \beta_l + e_{kli}$ |
| (5) | A en A×B | $y_{kli} = \mu + \alpha_k + \alpha\beta_{kl} + e_{kli}$ |
| (6) | B en A×B | $y_{kli} = \mu + \beta_l + \alpha\beta_{kl} + e_{kli}$ |
| (7) | A, B en A×B | $y_{kli} = \mu + \alpha_k + \beta_l + \alpha\beta_{kl} + e_{kli}$ |

Tabel 3.43: Vier modellen voor een twee-factorieel design, met en zonder interactie.

| F-toetsen voor Modellen met Interactie | | | | |
|---|---|---|---|---|
| $F = \frac{MS_{Effect\|Ref}}{MS_{Error\|Model}}$ | | | | |
| | | | $MS_{Error\|Model}$ | |
| Effect | Nulhypothese | $MS_{Effect\|Ref}$ | Meestal klein | Meestal groot |
| A | $\mu_{1\sqcup} = \mu_{2\sqcup} = \ldots = \mu_{K\sqcup}$ | $MS_A$ | $MS_{Error\|[AB]}$ | $MS_{Error\|A}$, $MS_{Error\|A,B}$ |
| | $\mu_{1\bullet} = \mu_{2\bullet} = \ldots = \mu_{K\bullet}$ | $MS_{A\|B,A\times B}$ | $MS_{Error\|[AB]}$ | |
| B | $\mu_{\sqcup 1} = \mu_{\sqcup 2} = \ldots = \mu_{\sqcup L}$ | $MS_B$ | $MS_{Error\|[AB]}$ | $MS_{Error\|B}$, $MS_{Error\|A,B}$ |
| | $\mu_{\bullet 1} = \mu_{\bullet 2} = \ldots = \mu_{\bullet L}$ | $MS_{B\|A,A\times B}$ | $MS_{Error\|[AB]}$ | |
| A×B | $\alpha\beta_{kl} = 0$ voor alle cellen $(k,l)$ | $MS_{A\times B\|A,B}$ | $MS_{Error\|[AB]}$ | |

Tabel 3.44: F-toetsen voor de gewogen en de ongewogen marginale gemiddelden van de factoren A en B en voor hun interactie.

| Type III Hoofdeffectkwadratensommen en de Interactie-effectkwadratensom | | | | | | |
|---|---|---|---|---|---|---|
| $SS_{Effect\|Ref} = \sum_{Obs}(\widehat{y}_{kli}^{Ref+Effect} - \widehat{y}_{kli}^{Ref})^2$ | | | | | | |
| Effect | Type | Model-vergelijking | Symbool | Vrijheids-graden | Hoe berekenen? Gebalanceerd | Ongebalanceerd |
| A | III | (7) met (6) | $SS_{A\|B,A\times B}$ | $K-1$ | $\sum_{k=1}^{K} Ln(\overline{y}_{k\bullet} - \overline{y}_{\bullet\bullet})^2$ | computer |
| B | III | (7) met (5) | $SS_{B\|A,A\times B}$ | $L-1$ | $\sum_{l=1}^{L} Kn(\overline{y}_{\bullet l} - \overline{y}_{\bullet\bullet})^2$ | computer |
| A×B | | (7) met (4) | $SS_{A\times B\|A,B}$ | $(K-1)(L-1)$ | $\sum_{k=1}^{K}\sum_{l=1}^{L} n[\overline{y}_{kl} - (\overline{y}_{k\bullet} + \overline{y}_{\bullet l} - \overline{y}_{\bullet\bullet})]^2$ | computer |

Tabel 3.45: Berekening van de type III hoofdeffectkwadratensommen voor de factoren A en B en van de interactie-effectkwadratensom.

| Errorkwadratensommen voor een Model met en een Model zonder Interactie | | | | | | |
|---|---|---|---|---|---|---|
| $SS_{Error\|Model} = \sum_{Obs}(y_{kli} - \widehat{y}_{kli}^{Model})^2$ | | | | | | |
| | | | | | Bereken $\widehat{y}_{kli}^{Model}$ | |
| Model | Effecten | Symbool | Vrijheids-graden | Ontbinding | Gebalanceerd | Ongebalanceerd |
| (4) | A,B | $SS_{Error\|A,B}$ | $N-K-L+1$ | $SS_{Error\|A,B} = SS_{A\times B\|A,B}$ $+SS_{Error\|[AB]}$ | $\overline{y}_{k\bullet} + \overline{y}_{\bullet l} - \overline{y}_{\bullet\bullet}$ | computer |
| (7) | A, B, A×B | $SS_{Error\|[AB]}$ | $N-KL$ | | $\overline{y}_{kl}$ | $\overline{y}_{kl}$ |

Tabel 3.46: Berekening van de errorkwadratensom op basis van een model met en een model zonder interactie.

### 3.3.13 Het Nut van het Toevoegen van Factoren

Een onderzoeker is meestal geïnteresseerd in het effect van één welbepaalde onafhankelijke variabele. In de twee voorbeeldonderzoeken zijn de onderzoeksvragen "Is er een effect van MOBILITEIT op de sociale activiteit van bejaarden?" en "Is er een effect van TRAINING op de mate waarin iemand gepest wordt?". Gegeven de interesse in het effect van één welbepaalde onafhankelijke variabele, kan men de vraag stellen wat het nut is van het toevoegen van een extra factor. Voor de twee voorbeeldonderzoeken, komt dit neer op de vraag wat het nut is van het toevoegen van, respectievelijk, de factoren BURGERLIJKE STAND en TYPE LERAAR.

Wij richten onze aandacht op drie punten waarop een extra factor, in principe, van belang kan zijn: (1) datgene wat getoetst wordt, (2) de gevoeligheid waarmee getoetst wordt, en (3) een controle op de algemeenheid van het effect dat getoetst wordt.

1. Datgene wat getoetst wordt met een F-statistiek, wordt bepaald door de effectkwadratensom in de teller van deze F-statistiek. In een design met twee factoren (in vergelijking met een design met slechts één factor) zijn er drie types kwadratensommen in plaats van slechts één: (1) de type I kwadratensom, die niet uitgezuiverd is, en die gelijk is aan de effectkwadratensom in een één-factorieel design, (2) de type II kwadratensom, die uitgezuiverd is voor het hoofdeffect van de andere factor, en (3) de type III kwadratensom, die uitgezuiverd is voor het hoofdeffect van de andere factor en het interactie-effect. In een gebalanceerd design, zijn deze drie types kwadratensommen gelijk. Dus, in een gebalanceerd design maakt het voor datgene wat er getoetst wordt niet uit of men een extra factor in de analyse betrekt of niet, want de effectkwadratensom verandert toch niet als men uitzuivert voor deze extra factor.

   In een ongebalanceerd design verschillen deze drie kwadratensommen wel, en dit stemt overeen met het feit dat men met elke kwadratensom een andere nulhypothese toetst: de nulhypothese van gelijke gewogen (niet-uitgezuiverde) marginale verwachte waarden met de type I kwadratensom en de nulhypothese van gelijke ongewogen (uitgezuiverde) marginale verwachte waarden met de type II en de type III kwadratensom.

2. De gevoeligheid waarmee men toetst met een F-statistiek wordt bepaald door de errorkwadratensom in de noemer van de F-statistiek. In de variantieanalyse met één factor is de noemer van de F-statistiek $MS_{Error|A}$ en in de variantieanalyse met twee factoren is de noemer $MS_{Error|[AB]}$. In bijna alle praktijksituaties is $MS_{Error|[AB]}$ kleiner dan $MS_{Error|A}$. Daarom is een variantieanalyse met twee factoren bijna altijd gevoeliger dan een variantieanalyse met één factor. Met een kleinere gemiddelde errorkwadratensom (de noemer) is een gemiddelde effectkwadratensom (de teller) immers eerder significant.

3. We noemen een effect van een factor A *algemeen* als het even groot is binnen elk van de niveaus van een willekeurige andere factor B. Bijvoorbeeld, het blijkt duidelijk uit Figuur 3.3 dat het effect van MOBILITEIT niet algemeen is: het effect van deze factor is anders voor de bejaarden met partner dan voor de bejaarden

zonder partner. Dus, we zeggen dat het effect van een factor algemeen is, als er geen interactie is tussen de factoren.

We kunnen dus besluiten dat het opnemen van een extra factor (1) het mogelijk maakt om, in een ongebalanceerd design, uit te zuiveren voor een storende variabele, (2) de F-toets meestal gevoeliger maakt, en (3) de mogelijkheid creëert om de algemeenheid van het effect van de eerste factor te onderzoeken.

### 3.3.14 Een Relativering van de Kracht van Statistische Uitzuivering

Onderzoek heeft vaak tot doel om uitspraken te doen over causale verbanden. De ideale methode om dit soort uitspraken te kunnen doen, is het gerandomiseerde experiment met experimentele controle. Randomisering (i.e., random toewijzing aan condities) garandeert dat er geen systematische verschillen zijn tussen de groepen wat betreft storende persoonsvariabelen, en experimentele controle zorgt ervoor dat de groepen niet van elkaar verschillen wat betreft bepaalde (niet noodzakelijk alle!) storende situatievariabelen.

Statistische uitzuivering met behulp van de type III kwadratensom is een uitstekend middel om te corrigeren voor (1) verschillen tussen de groepen op bepaalde storende persoonsvariabelen, als gevolg van het feit dat de toewijzing niet random was, en (2) verschillen tussen de condities op bepaalde storende situatievariabelen, als gevolg van het feit dat experimentele controle van de condities niet mogelijk was. Het is echter geen tovermiddel. De voornaamste beperking van statistische uitzuivering is dat men alleen kan uitzuiveren voor die variabelen waarover men beschikt. Het is best mogelijk dat er, naast deze beschikbare variabelen, nog andere, maar niet-beschikbare, variabelen zijn waarop de groepen van elkaar verschillen. Bijvoorbeeld, in het ongebalanceerde voorbeeldonderzoek kan het best zo zijn dat de groepen met en zonder training niet alleen van elkaar verschillen wat betreft het type leraar, maar ook wat betreft de zogenaamde *locus of control* van de leerlingen. Concreet, het kan best zo zijn dat vooral de leerlingen met een interne locus of control deelnamen aan de training. Gesteld dat leerlingen met een interne locus of control sowieso op eigen kracht (i.e., zonder de training) een einde zouden maken aan het pesten, dan zou ook deze variabele verantwoordelijk zijn voor een verschil in gemiddelde pestscore tussen de groep met en de groep zonder training, zonder dat er sprake is van een causaal verband.

Een tweede beperking van statistische uitzuivering is dat het alleen effectief is als er geen te sterke samenhang is tussen de twee onafhankelijke variabelen. Een voorbeeld van een situatie waarin de samenhang tussen de onafhankelijke variabelen zeer sterk is, wordt getoond in Tabel 3.47. In dit voorbeeld is het zo dat, als men het type leraar kent, men ook bijna zeker weet of de leerling aan de training heeft deelgenomen of niet. De samenhang tussen de twee onafhankelijke variabelen is dus zeer sterk. In een dergelijke situatie zal de type III kwadratensom $SS_{A|B,A\times B}$ niet snel in een significante F-statistiek resulteren, ook al is het verschil tussen de ongewogen marginale gemiddelden groot. Het verschil tussen de ongewogen gemiddelden is immers erg onbetrouwbaar, want het hangt sterk af van slechts een paar observaties. Dit heeft tot gevolg dat de ongewogen marginale gemiddelden zeer sterk van elkaar moeten verschillen opdat de F-statistiek significant is.

| | Actieve Leraar | Reactieve Leraar | Passieve Leraar | Som |
|---|---|---|---|---|
| Met Training | 34 | 1 | 1 | 36 |
| Zonder Training | 1 | 29 | 24 | 54 |
| Som | 35 | 30 | 25 | 90 |

Tabel 3.47: Aantal leerlingen in de zes cellen van het voorbeeldonderzoek voor een situatie waarin de samenhang tussen de twee onafhankelijke variabelen zeer sterk is.

### 3.3.15 Designs met Lege Cellen

In een ongebalanceerd design, kan het gebeuren dat sommige van de $K \times L$ cellen leeg zijn. Zo had het kunnen zijn dat er zich geen leerlingen bevonden in de cel met een actieve leraar en zonder training. Designs met lege cellen vereisen een aparte behandeling omdat (1) het aantal vrijheidsgraden van $SS_{Error|[AB]}$ en $SS_{A\times B|A,B}$ anders is dan in een design zonder lege cellen en (2) de type III kwadratensom $SS_{A|B,A\times B}$ niet uitgerekend kan worden.

We duiden het aantal lege cellen aan met $q$. Het aantal vrijheidsgraden van $SS_{Error|[AB]}$ is dan $N - KL + q$ en het aantal vrijheidsgraden van $SS_{A\times B|A,B}$ is $(K-1)(L-1) - q$.

Omdat de type III kwadratensom niet uitgerekend kan worden, moet men gebruik maken van andere kwadratensommen om de nulhypothese te toetsen dat de ongewogen gemiddelden gelijk zijn. Twee kwadratensommen komen hiervoor in aanmerking: de type II en de *type IV* kwadratensommen, waarvan de laatste hier nog niet aan bod gekomen is. In het verleden werd, in een design met lege cellen, de toets op de ongewogen marginale gemiddelden altijd uitgevoerd met de type II kwadratensom. Strikt genomen, is deze toets alleen geldig als er, in de populatie, geen interactie is tussen A en B. Daarom wordt in de praktijk vaak eerst gekeken naar het resultaat van de F-toets van het interactie-effect. Als de F-statistiek van het interactie-effect niet significant is (en men dus beslist dat er, in de populatie, geen interactie-effect is), toetst men de nulhypothese van gelijke ongewogen marginale verwachte waarden met de type II kwadratensom, en anders niet.

Tegenwoordig kan men in de recente versies van computerprogramma's zoals SPSS ook gebruik maken van de type IV kwadratensom. De type IV kwadratensom probeert de type III kwadratensom zo goed mogelijk te benaderen door voor een zo groot mogelijk gedeelte van het interactie-effect uit te zuiveren. Als er geen lege cellen zijn, dan wordt het interactie-effect volledig uitgezuiverd, en is de type IV gelijk aan de type III kwadratensom. Als er wel lege cellen zijn, dan benadert de type IV kwadratensom (qua mate van uitzuivering voor het interactie-effect) de type III kwadratensom (die zelf niet te berekenen is).

### 3.3.16 De Variantieanalyse met Meer dan Twee Factoren

Variantieanalyse met meer dan twee factoren is niet wezenlijk anders of moeilijker dan variantieanalyse met precies twee factoren. Het nieuwe element is dat er nu meerdere interacties een rol spelen. Bijvoorbeeld, in een design met drie factoren, A, B en C, zijn er drie interacties tussen twee factoren (A×B, A×C en B×C) en één interactie tussen drie factoren (A×B×C). De kwadratensommen van deze interactie-effecten en de bijbehorende F-statistieken, worden op een analoge manier berekend als in een design met slechts twee factoren.

### 3.3.17 Uitvoeren van een Variantieanalyse met de SPSS-procedure GLM-Univariate

Nu volgt een korte beschrijving van de SPSS-procedure GLM-Univariate, die gebruikt wordt om een variantieanalyse uit te voeren. Het is aan te raden om deze beschrijving te lezen vóór een computer waarop SPSS draait. Men kan dan de verschillende vensters bekijken die hier beschreven worden.

Men vindt GLM-Univariate onder *General Linear Model* binnen het menu *Analyze*. In het dialoogvenster van GLM-Univariate kies je (1) de afhankelijke variabele door hem naar het vakje *Dependent Variable* over te brengen en (2) de onafhankelijke variabelen door ze naar het vakje *Fixed Factor(s)* over te brengen. Het vakje *Random Factor(s)* is niet van belang voor de analyses die in dit boek behandeld worden, en het vakje *Covariate(s)* wordt besproken in de paragraaf over covariantieanalyse. Rechtsboven in dit dialoogvenster staan een aantal knoppen. Voor de opgaven in dit boek, zijn alleen de knoppen *Options* en *Model* van belang.

Met de knop *Options* kun je vragen om de ongewogen gemiddelden. SPSS noemt de ongewogen gemiddelden *Estimated Marginal Means*. De factoren en factorcombinaties (door SPSS *Factor Interactions* genoemd) waarvoor je de ongewogen gemiddelden wilt, geef je aan in het veld *Factor(s) and Factor Interactions*.

Met de knop *Model* kun je een keuze maken uit de verschillende types kwadratensommen. Door op deze knop te drukken krijg je een venster waarin je onderaan een keuze kunt maken uit vier types kwadratensommen (I, II, III, en IV). Dit vergt echter enige toelichting, en die wordt hieronder gegeven. Eerst bespreken we kort de andere opties in dit venster. Men kan kiezen uit *Full factorial* en *Custom*. In dit hoofdstuk kiezen wij altijd voor *Full factorial*. Dit betekent dat alle hoofd- en interactie-effecten in de analyse opgenomen worden. (Dit betekent niet dat alle effecten ook voor alle andere effecten uitgezuiverd worden, want dat wordt bepaald door het type kwadratensom dat men kiest; zie verder.) Als men kiest voor *Custom*, dan kan men zelf bepalen welke effecten men in de analyse opneemt. Voor dit hoofdstuk is dit verder niet van belang. Door het vakje bij *Include intercept in model* aan te kruisen, geeft men aan dat men de globale verwachte waarde $\mu$ in het model wil opnemen. Voor alle situaties die in dit boek beschreven worden is het nodig om deze globale verwachte waarde in het model op te nemen.

We geven nu een toelichting bij de keuze uit de vier types kwadratensommen. Het belangrijkste punt hierbij is dat SPSS onder de *optie* type I iets anders verstaat dan een type I *kwadratensom*. SPSS wijkt hier af van de terminologie die men in de handboeken

vindt, en dat is vervelend.

**De Type I Optie** Kiezen voor de type I optie betekent dat er *in een bepaalde volgorde* (hiërarchisch) uitgezuiverd wordt. Stel dat de factoren A en B in deze volgorde staan onder *Factor(s)* in het eerste dialoogvenster. Dan worden met de type I optie de volgende kwadratensommen berekend: $SS_A$ (geen uitzuivering), $SS_{B|A}$ (uitzuivering voor het hoofdeffect van de vorige factor), en $SS_{A\times B|A,B}$ (uitzuivering voor de hoofdeffecten van de twee vorige factoren). In het algemeen (i.e., voor een willekeurig aantal factoren) geldt dat bij de type I optie de kwadratensommen in een bepaalde volgorde berekend worden, en dat er uitgezuiverd wordt voor alle effecten waarvan de kwadratensom reeds berekend werd. Daarom spreekt men ook van *hiërarchische kwadratensommen*. Merk op dat men de type I *optie* van SPSS kan gebruiken om een type I *kwadratensom* te berekenen voor één hoofdeffect. Concreet, met de type I optie, berekent SPSS de type I kwadratensom voor de eerste factor in het lijstje onder *Fixed Factor(s)*; voor de andere factoren wordt geen type I kwadratensom berekend.

Na de hoofdeffecten, berekent SPSS de kwadratensommen van de interactie-effecten. In een twee-factorieel design, is er slechts één interactie-effect en, met de type I optie, wordt zijn kwadratensom uitgezuiverd voor beide hoofdeffecten. Dit wil zeggen, SPSS berekent $SS_{A\times B|A,B}$. In een design met meer dan twee factoren, wordt de kwadratensom van een interactie-effect ook uitgezuiverd voor alle interactie-effecten van dezelfde en een lagere orde.

De gegevens die nodig zijn om de F-toetsen uit te voeren worden meestal gegeven in een ANOVA-tabel. Dit gebeurt ook in de uitvoer van een SPSS-analyse. In Tabel 3.48 staat een deel[11] van de ANOVA-tabel voor de gegevens uit het ongebalanceerde voorbeeldonderzoek. Dit is de ANOVA-tabel die men krijgt met de type I optie en de factoren TRAINING en TYPE LERAAR in deze volgorde in het lijstje onder *Fixed Factor(s)*. Voor de factor TRAINING, wordt de type I kwadratensom $SS_A$ gegeven (zonder uitzuivering), voor de factor TYPE LERAAR, de type II kwadratensom $SS_{B|A}$ (uitzuivering voor het hoofdeffect van A), en voor het interactie-effect de gebruikelijke kwadratensom $SS_{A\times B|A,B}$ (uitzuivering voor de hoofdeffecten van A en B). De modelkwadratensom en de totale kwadratensom $SS_{Totaal}$ staan in de rijtjes met, respectievelijk, *Corrected Model* en *Corrected Total*. De gemiddelde errorkwadratensom in de ANOVA-tabel is $MS_{Error|[AB]}$. Alle F-statistieken in deze tabel werden met deze gemiddelde errorkwadratensom berekend.

**De Type II Optie** Met deze optie, in een twee-factorieel design, berekent SPSS de Type II kwadratensommen, zoals beschreven in dit boek. Voor designs met meer dan twee factoren, moet men de documentatie bij SPSS raadplegen om te achterhalen voor welke andere effecten de effectkwadratensommen uitgezuiverd worden.

**De Type III Optie** Met deze optie berekent SPSS de type III kwadratensommen, zoals beschreven in dit boek. Men zuivert dus uit voor alle andere hoofd- en interactie-

[11] De ANOVA-tabel in de uitvoer van een SPSS-analyse bevat ook de kwadratensom van het zogenaamde *intercept* en een kwadratensom die met *Total* aangeduid wordt. Deze zijn niet van belang voor de analyses die in dit hoofdstuk behandeld worden.

Dependent Variable: PESTSCORE

| Source | Type I Sum of Squares | df | Mean Squares | F | Sig. |
|---|---|---|---|---|---|
| Corrected Model | 812.055 | 5 | 162.411 | 70.308 | .000 |
| TRAINING | 112.500 | 1 | 112.500 | 48.701 | .000 |
| TYPE LERAAR | 655.401 | 2 | 327.701 | 283.723 | .000 |
| TRAINING ∗ TYPE LERAAR | 44.154 | 2 | 22.077 | 9.557 | .000 |
| Error | 194.000 | 84 | 2.310 | | |
| Corrected Total | 1006.055 | 89 | | | |

Tabel 3.48: Deel van de ANOVA-tabel voor de data uit het ongebalanceerde voorbeeldonderzoek, geproduceerd door GLM-Univariate met de type I optie.

effecten. In Tabel 3.49 staat een deel van de ANOVA-tabel die men krijgt met de type III optie voor de gegevens uit het ongebalanceerde voorbeeldonderzoek. Voor de factor TRAINING en TYPE LERAAR worden de type III kwadratensom $SS_{A|B,A\times B}$ en $SS_{B|A,A\times B}$ gegeven, en voor het interactie-effect de gebruikelijke kwadratensom $SS_{A\times B|A,B}$. Alle F-statistieken in de tabel werden met de gemiddelde errorkwadratensom $MS_{Error|[AB]}$ berekend.

Dependent Variable: PESTSCORE

| Source | Type III Sum of Squares | df | Mean Squares | F | Sig. |
|---|---|---|---|---|---|
| Corrected Model | 812.055 | 5 | 162.411 | 70.308 | .000 |
| TRAINING | 0.000 | 1 | 0.000 | 0.000 | 1.000 |
| TYPE LERAAR | 578.269 | 2 | 289.134 | 125.166 | .000 |
| TRAINING ∗ TYPE LERAAR | 44.154 | 2 | 22.077 | 9.557 | .000 |
| Error | 194.000 | 84 | 2.310 | | |
| Corrected Total | 1006.055 | 89 | | | |

Tabel 3.49: Deel van de ANOVA-tabel voor de data uit het ongebalanceerde voorbeeldonderzoek, geproduceerd door GLM-Univariate met de type III optie.

**De Type IV Optie** Met deze optie berekent SPSS de Type IV kwadratensommen. Deze optie is bedoeld voor designs met lege cellen. Als er geen lege cellen zijn, dan is de type IV kwadratensom gelijk aan de type III kwadratensom.

---

**Opgaven Gebalanceerde Designs**

*Opgave 15*
In Tabel 3.50 staan vijf sets van telkens negen $\mu_{kl}$'s uit een design met twee factoren en drie niveaus per factor (vaak aangeduid als een "$3 \times 3$ factorieel design").

1. Bepaal voor elke set de waarden van de $\alpha_k$'s, $\beta_l$'s en $\alpha\beta_{kl}$'s.

2. Geef voor elke set aan welke effecten (A, B of A×B) er zijn in de populatie.
3. Teken een grafiek met de $\mu_{kl}$'s voor de set waarin er een effect is van A en B, maar niet van A×B. Beschrijf het patroon in deze grafiek.

| (a) | | B | | |
|---|---|---|---|---|
| | | 1 | 2 | 3 |
| A | 1 | 5 | 5 | 5 |
| | 2 | 7 | 7 | 7 |
| | 3 | 12 | 12 | 12 |

| (b) | | B | | |
|---|---|---|---|---|
| | | 1 | 2 | 3 |
| A | 1 | 4 | 9 | 14 |
| | 2 | 4 | 9 | 14 |
| | 3 | 4 | 9 | 14 |

| (c) | | B | | |
|---|---|---|---|---|
| | | 1 | 2 | 3 |
| A | 1 | 16 | 12 | 11 |
| | 2 | 13 | 9 | 8 |
| | 3 | 7 | 3 | 2 |

| (d) | | B | | |
|---|---|---|---|---|
| | | 1 | 2 | 3 |
| A | 1 | 18 | 15 | 12 |
| | 2 | 10 | 11 | 15 |
| | 3 | 5 | 7 | 6 |

| (e) | | B | | |
|---|---|---|---|---|
| | | 1 | 2 | 3 |
| A | 1 | 19 | 15 | 14 |
| | 2 | 18 | 10 | 11 |
| | 3 | 8 | 8 | 5 |

Tabel 3.50: Tabel bij opgave 15.

*Opgave 16*
Een onderwijspsycholoog onderzocht de effectiviteit van vier didactische methoden voor het rekenonderwijs in de lagere school[12]. Deze methoden verschilden van elkaar op twee factoren: (1) of de nadruk lag op realistische rekenopgaven of niet (verhaaltjes waar getallen bij te pas komen versus problemen met alleen maar getallen) en (2) of de nadruk lag op *bottom-up* (inductief) leren (vanuit concrete voorbeelden naar het algemene begrip) of op het *top-down* leren (vanuit het algemene begrip naar de concrete voorbeelden). De proefpersonen werden at random aan één van de vier didactische methoden toegewezen. De afhankelijke variabele in dit onderzoek was de score op een rekentoets. In Tabel 3.51 worden, voor elke van de vier cellen, de gemiddelde scores op deze rekentoets gegeven. Er zijn 12 proefpersonen per cel en $MS_{Error|[AB]} = 300$.

1. De onderzoeker beweert dat, in de populatie, het verschil tussen top-down en bottom-up kleiner is als de rekenopgaven realistisch zijn dan wanneer ze dat niet zijn. Klopt dit?
2. Een collega van de onderzoeker beweert dat, in de populatie, het effect van realistische versus niet-realistische rekenopgaven kleiner is in de bottom-up dan in de top-down benadering. Klopt dit?
3. Zijn de effecten van het al dan niet realistisch zijn van de rekenopgaven en het top-down of bottom-up zijn van de benadering additief (in de populatie)?

| | | REALISTISCH | |
|---|---|---|---|
| | | Ja | Nee |
| BENADERING | Top-down | 42 | 27 |
| | Bottom-up | 55 | 48 |

Tabel 3.51: Tabel bij opgave 16.

[12]Voor deze opgave heb ik me laten inspireren door opgave 3 van hoofdstuk 7 (p. 298) van Maxwell en Delaney (1990).

*Opgave 17*
Een onderzoeker wil de effectiviteit evalueren van een aantal therapieën voor de behandeling van depressie. Hij is vooral geïnteresseerd in een vergelijking van een cognitieve therapie (gericht op het veranderen van opvattingen) en een "zuivere" gedragstherapie (gericht op het rechtstreeks veranderen van gedrag). Daarnaast is hij ook geïnteresseerd in het effect van medicatie (anti-depressiva). Hij beschikt over een groep van 60 proefpersonen. Hij twijfelt over twee verschillende designs:

- Een design waarin 30 subjecten at random worden toegewezen aan elk van twee condities: cognitieve therapie en zuivere gedragstherapie.
- Een design waarin 15 subjecten at random worden toegewezen aan elk van vier condities: cognitief, gedrag, cognitief plus medicatie, en gedrag plus medicatie.

1. Welk design (zo er al een is) stelt de onderzoeker in staat om na te gaan of er een interactie is tussen de factor COGNITIEF-VERSUS-GEDRAG en de factor MEDICATIE.
2. De onderzoeker neigt ertoe om het eerste design te gebruiken omdat hij denkt dat het een gevoeliger toets oplevert van het verschil tussen de cognitieve en de zuivere gedragstherapie. Ben je het hiermee eens? Motiveer je antwoord.

*Opgave 18*
In een studie naar de behandeling van hoge bloeddruk werd de effectiviteit van biofeedback en medicinale behandeling onderzocht in een $2 \times 2$ factorieel design. Een steekproef van personen die aan hoge bloeddruk lijden werd verdeeld over de vier cellen die ontstaan door de twee niveaus van de factor BIOFEEDBACK ("wel" en "niet") te combineren met de twee niveaus van de factor MEDICATIE ("wel" en "geen"). Aan het einde van de behandeling werd van al deze personen de bloeddruk gemeten. Deze metingen lagen allemaal tussen 150 en 200 mm kwikdruk.

1. Geef een getallenvoorbeeld met mogelijke celgemiddelden waaruit blijkt dat er, in de steekproef, een effect is van MEDICATIE, maar niet van BIOFEEDBACK en ook geen interactie.
2. Geef een getallenvoorbeeld met mogelijke celgemiddelden waaruit blijkt dat er wel een interactie is in de steekproef.

*Opgave 19*
De effectiviteit van psychotherapie en van het gelijktijdig toedienen van medicijnen werd onderzocht in een $3 \times 2$ factorieel design. De drie niveaus van de factor PSYCHOTHERAPIE zijn "(psycho-)analytisch" (A), "Rogeriaans" (R) en "gedragstherapie" (G) en de twee niveaus van de factor MEDICATIE zijn "wel" (W) en "niet" (N). Een steekproef van patiënten met neurotische klachten werd at random verdeeld over de zes condities van dit design. Een half jaar na het einde van de behandeling werd bij al deze personen een vragenlijst afgenomen waarmee de ernst van de neurotische symptomen vastgesteld werd. De scores op deze schaal lopen van 0 (geen neurotische symptomen) tot 100 (maximale neurotische symptomen).

1. Geef een getallenvoorbeeld met mogelijke celgemiddelden waaruit blijkt dat er, in de steekproef, een effect is van zowel PSYCHOTHERAPIE als van MEDICATIE, maar geen interactie.
2. Geef een getallenvoorbeeld met mogelijke celgemiddelden waaruit blijkt dat er, in de steekproef, wel een interactie is, maar geen hoofdeffect van MEDICATIE.

*Opgave 20*
In Tabel 3.52 worden de celgemiddelden getoond van een onderzoek met een $3 \times 2$ factorieel

design[13]. Het aantal proefpersonen in elk van de cellen was gelijk. De onderzoeker voerde een variantieanalyse uit op deze gegevens en noteerde zijn resultaten in een ANOVA-tabel. In een poging om hem te helpen, heeft zijn dochtertje deze tabel wat bijgekleurd met vingerverf. Hierdoor is een gedeelte van de tabel onleesbaar geworden. Wat er nog van overblijft wordt getoond in Tabel 3.53.

| | | B | |
|---|---|---|---|
| | | 1 | 2 |
| A | 1 | 16 | 18 |
| | 2 | 13 | 17 |
| | 3 | 10 | 16 |

Tabel 3.52: Tabel bij opgave 20.

| Source | Type III Sum of Squares | df | Mean Squares | F | Sig. |
|---|---|---|---|---|---|
| Corrected Model | | | | | |
| A | | | | | |
| B | | | | | |
| A∗B | 44 | | | | |
| Error | 1620 | 60 | | | |
| Corrected Total | 2104 | 65 | | | |

Tabel 3.53: Tabel bij opgave 20.

1. Wat is de waarde van $SS_A$? Tip: in de vrijheidsgraden van $SS_{Error|[AB]}$, die gelijk zijn aan $N - KL$, vind je informatie over het aantal proefpersonen.
2. Wat is de waarde van $SS_B$?
3. Vul de ontbrekende vrijheidsgraden in.
4. Bereken de gemiddelde kwadratensommen voor A, B en A×B.
5. Bereken de F-statistiek voor elk effect. Welke F-statistieken zijn significant op het 0.05-niveau?
6. Bereken de proportie verklaarde variantie door A, B en A×B.
7. De onderzoeker beweert dat het interactie-effect niet significant is. Omdat hij vooral geïnteresseerd is in het hoofdeffect van A, besluit hij om zijn data te heranalyseren als een design met slechts één factor. Hij laat factor B dus gewoon buiten beschouwing. Welke kwadratensommen zal hij vinden? Is het effect van A significant?

*Opgave 21*
Manuck, Kaplan, Adams en Clarckson (1988) deden een onderzoek bij apen naar de invloed van gedragspatronen op atherosclerose, een ziekte van de hartslagader[14]. In een experiment onderzochten zij de effecten van een psychosociale manipulatie (i.c., het regelmatig veranderen

[13]Voor deze opgave heb ik me laten inspireren door opgave 10 van hoofdstuk 7 (pp. 301-302) van Maxwell en Delaney (1990).

[14]Voor deze opgave heb ik me laten inspireren door opgave 14 van hoofdstuk 7 (p. 304) van Maxwell en Delaney (1990).

van de groepssamenstelling) op het onstaan van atherosclerose bij apen met een dominante of een ondergeschikte status. In één van de condities ("instabiel") werden de apen steeds na een periode van een tot drie maanden in een nieuwe sociale groep ondergebracht. In de andere conditie ("stabiel") bleven de apen in dezelfde sociale groep tijdens het gehele onderzoek. In Tabel 3.54 vind je data die vergelijkbaar zijn met die uit het onderzoek van Manuck et al. In de twee onderste rijen van de tabel vind je de resultaten van een aantal berekeningen die je behulpzaam zullen zijn bij de berekening van de kwadratensommen (de tweede onderste rij voor de berekening van $SS_A$, $SS_B$ en $SS_{A\times B|A,B}$, en de onderste rij voor de berekening van $SS_{Error|[AB]}$).

| | SOCIALE STABILITEIT | | | |
|---|---|---|---|---|
| | Stabiel | | Instabiel | |
| | SOCIALE STATUS | | SOCIALE STATUS | |
| | Dominant | Ondergeschikt | Dominant | Ondergeschikt |
| | .23 | .34 | .54 | .39 |
| | .17 | .62 | .68 | .23 |
| | .26 | .54 | .70 | .27 |
| | .32 | .30 | .76 | .49 |
| | .41 | .51 | .58 | .53 |
| | .38 | .44 | .87 | .42 |
| | .49 | .41 | .81 | .34 |
| $\overline{y}_{kl}$ | 0.323 | 0.451 | 0.706 | 0.381 |
| $\sum_{i=1}^{7}(y_{kli}-\overline{y}_{kl})^2$ | 0.075 | 0.077 | 0.085 | 0.072 |

Tabel 3.54: Tabel bij opgave 21.

1. Maak een grafiek, analoog aan de grafiek op pagina 209, met de gemiddelden in de vier cellen. Wat besluit je op basis van deze grafiek?
2. Bereken de F-statistieken voor het hoofdeffect van SOCIALE STABILITEIT, het hoofdeffect van SOCIALE STATUS, en de interactie tussen deze twee factoren. Zijn deze F-statistieken significant op het 0.05-niveau? Presenteer je resultaten in een ANOVA-tabel.

*Opgave 22*

In een onderzoek van Brehm (1981) werd nagegaan in welke mate kinderen hun voorkeur veranderen voor voorwerpen die hen ontzegd worden[15]. Brehm was vooral geïnteresseerd in het verschil tussen jongens en meisjes in de manier waarop ze hun voorkeur veranderen. De ene helft van een groep leerlingen uit de basisschool werd verteld dat zij konden kiezen uit twee voorwerpen. De andere helft werd verteld dat zij één van twee voorwerpen zouden krijgen, maar niet dat zij een keuze konden maken. Elk kind rangordende tien voorwerpen naar voorkeur en kreeg het voorwerp dat het als derde gerangordend had. De kinderen moesten toen opnieuw de voorwerpen rangordenen naar voorkeur. De afhankelijke variabele was de plaats in deze nieuwe rangorde van het vierde voorwerp uit de eerste rangorde (het voorwerp dat zij net niet kregen). In Tabel 3.55 vind je data die vergelijkbaar zijn met die uit het onderzoek van Brehm.

1. Maak een grafiek, analoog aan de grafiek van het voorbeeldonderzoek naar de sociale activiteit van bejaarden, met de gemiddelden in de vier cellen. Wat besluit je op basis van deze grafiek?

[15] Voor deze opgave heb ik me laten inspireren door opgave 15 van hoofdstuk 7 (p. 304) van Maxwell en Delaney (1990).

| GESLACHT | | | |
|---|---|---|---|
| Meisjes | | Jongens | |
| KEUZE | | KEUZE | |
| Ja | Nee | Ja | Nee |
| 4 | 5 | 4 | 5 |
| 3 | 3 | 3 | 4 |
| 6 | 4 | 3 | 4 |
| 4 | 4 | 5 | 6 |
| 7 | 3 | 6 | 3 |
| 5 | 2 | 4 | 7 |
| 4 | 5 | 4 | 6 |
| 5 | 3 | 5 | 5 |
| 5 | 6 | 3 | 6 |

Tabel 3.55: Tabel bij opgave 22.

2. Bereken de F-statistieken voor het hoofdeffect van GESLACHT, het hoofdeffect van KEUZE, en de interactie tussen deze twee factoren. Zijn deze F-statistieken significant op het 0.05-niveau? Presenteer je resultaten in een ANOVA-tabel.

## Opgaven Ongebalanceerde Designs

*Opgave 23*
De personeelsdienst van een grote bank maakt zich zorgen over het grote aantal overuren dat het kaderpersoneel maakt. Men vermoedt dat het overspannen raken van een aantal kaderleden ten minste voor een deel het gevolg is van dit grote aantal overuren. Om het aantal overuren terug te dringen, werd het vaste cursusaanbod van het bedrijf uitgebreid met een cursus time-management. Deze cursus beoogt werknemers te leren efficiënter met hun tijd om te gaan.

Na een jaar, wilde de personeelsdienst het effect van deze cursus evalueren. Daartoe werd aan een zuiver toevallige steekproef van kaderleden gevraagd om gedurende een maand bij te houden hoeveel overuren zij maken. Men vroeg hen ook welke cursus zij het afgelopen jaar gevolgd hebben. (Omdat deze cursussen tijdens werktijden gegeven worden en erg populair zijn, volgt elk kaderlid elk jaar één cursus.) De resultaten van dit onderzoek worden gegeven in Tabel 3.56. In deze tabel wordt een onderscheid gemaakt tussen het wel of niet gevolgd hebben van de time-management cursus (verder aangeduid als CURSUS) en het type kaderfunctie (lager, midden of hoger; verder aangeduid als KADERTYPE). In deze tabel staan het gemiddelde aantal overuren (vetgedrukt) en het aantal proefpersonen (gewoon gedrukt) voor elk van de zes cellen van het onderzoeksdesign.

| | | KADERTYPE | | |
|---|---|---|---|---|
| | | Hoger kader | Midden-kader | Lager kader |
| CURSUS | Time-management cursus | **40**<br>15 | **20**<br>15 | **6**<br>10 |
| | Andere cursus | **60**<br>5 | **30**<br>15 | **15**<br>30 |

Tabel 3.56: Tabel bij opgave 23.

1. Bereken de gewogen en de ongewogen marginale gemiddelden voor de verschillende niveaus van de factor CURSUS.
2. Als je wil weten of de time-management cursus een causaal effect heeft, naar welke gemiddelden moet je dan kijken? Is er een effect? Motiveer je antwoord. (Voor het gemak, vergeet even dat je een effect ook nog statistisch moet toetsen.)
3. Is KADERTYPE een storende variabele? Leg uit waarom wel of waarom niet.
4. Zou je verbaasd zijn als blijkt dat de type I kwadratensom van CURSUS gelijk is aan 0? Leg uit waarom wel of waarom niet.
5. Zou je verbaasd zijn als blijkt dat de type III kwadratensom van CURSUS gelijk is aan 0? Leg uit waarom wel of waarom niet.
6. Waar vind je in Tabel 3.56 informatie over een eventuele interactie tussen CURSUS en KADERTYPE? Leg uit.

*Opgave 24*
Een onderzoekster is geïnteresseerd in de verschillen in de dagdagelijkse activiteiten van mensen met en zonder partner. Zij is in het bijzonder geïnteresseerd in het aantal uur per week dat besteed wordt aan activiteiten in verenigingsverband. Zij benaderde 100 alleenstaanden en 100 mensen met een partner en vroeg hen hoeveel uur zij gemiddeld per week besteden aan activiteiten in verenigingsverband. Omdat het hier gaat om een observationeel onderzoek, maakt de onderzoekster zich zorgen over de mogelijkheid dat een storende variabele (mede) verantwoordelijk is voor een eventueel verschil tussen mensen met en zonder partner. Zij maakt zich met name zorgen over GESLACHT als mogelijke storende variabele. In Tabel 3.57 staat het gemiddeld aantal uren activiteiten in verenigingsverband (vetgedrukt) en het aantal proefpersonen voor elk van de vier cellen van dit onderzoeksdesign.

| | | PARTNER | |
|---|---|---|---|
| | | Ja | Nee |
| GESLACHT | Man | **8**<br>20 | **5**<br>80 |
| | Vrouw | **3**<br>80 | **3**<br>20 |

Tabel 3.57: Tabel bij opgave 24.

1. Bereken de gewogen en de ongewogen marginale gemiddelden voor de verschillende niveaus van de factor PARTNER.
2. Vertrek van Tabel 3.57 en maak een nieuwe tabel waarin er geen samenhang is tussen PARTNER en GESLACHT.
3. Vertrek van Tabel 3.57 en maak een nieuwe tabel waarin er geen samenhang is tussen PARTNER en de afhankelijke variabele.
4. Stel dat de F-toets met de type I kwadratensom van PARTNER significant is. Welke conclusie kan men dan trekken met betrekking tot het verschil tussen alleenstaanden en mensen met een partner? Druk dit uit in termen van de nulhypothese die getoetst wordt.
5. Stel dat de F-toets met de type III kwadratensom van PARTNER significant is. Welke conclusie kan men dan trekken met betrekking tot het verschil tussen alleenstaanden en mensen met een partner? Druk dit uit in termen van de nulhypothese die getoetst wordt.

6. Is het in dit onderzoek te verdedigen dat men, in plaats van gebruik te maken van de type III kwadratensom van PARTNER, gebruik maakt van de type II kwadratensom? Leg uit.

*Opgave 25*
Een psychologe vraagt zich af of de hoeveelheid tijd die een kind op de crèche doorbrengt een effect heeft op zijn of haar cognitieve ontwikkeling. Om dit te onderzoeken nam zij bij 200 kinderen in Groep 2 een intelligentietest af. Deze kinderen werden vervolgens ingedeeld in de groepen "veel crèche" en "weinig crèche", afhankelijk van het aantal uren dat ze als baby en peuter (van 0 tot 3 jaar) in een crèche hebben doorgebracht. De psychologe vermoedt dat het opleidingsniveau van de ouders van het kind een storende variabele kan zijn. Daarom registreerde ze of de ouders van de kinderen hoog of laag opgeleid zijn ("hoog" is een academische of een HBO opleiding en "laag" is de rest). In Tabel 3.58 staat het gemiddelde IQ (vetgedrukt) en het aantal proefpersonen voor elk van de vier cellen van dit onderzoeksdesign.

| | | OPLEIDINGSNIVEAU | |
|---|---|---|---|
| | | Laag | Hoog |
| CRÈCHE | Veel | **98**<br>25 | **110**<br>75 |
| | Weinig | **103**<br>75 | **119**<br>25 |

Tabel 3.58: Tabel bij opgave 25.

1. Leg uit waarom OPLEIDINGSNIVEAU een storende variabele is.
2. Wat is het verschil in gemiddelde IQ tussen kinderen die veel en kinderen die weinig op de crèche hebben gezeten zonder uit te zuiveren voor OPLEIDINGSNIVEAU?
3. Wat is het verschil in gemiddelde IQ tussen kinderen die veel en kinderen die weinig op de crèche hebben gezeten indien men uitzuivert voor OPLEIDINGSNIVEAU?
4. Op basis van de gegevens van dit onderzoek, heeft het veel of weinig naar de crèche gaan een effect op het IQ van het kind? Leg uit.
5. Welk type kwadratensom moet men gebruiken om de significantie van het effect CRÈCHE op IQ te toetsen? Leg uit.
6. Waar of onwaar en leg uit: als de celfrequenties veranderen, dan verandert ook de nulhypothese die men toetst met de type III kwadratensom.

*Opgave 26*
Twee datasets, *a* en *b*, zijn elk afkomstig uit een onderzoek met een ongebalanceerd design met twee factoren, A en B. Beide datasets werden geanalyseerd met GLM-Univariate, en dit op vier verschillende manieren: met de type I optie en de factoren in de volgorde (A,B), met de type I optie en de factoren in de volgorde (B,A), met de type II optie en met de type III optie. De effectkwadratensommen in de uitvoer van deze vier analyses worden gegeven in Tabel 3.59 (in het linkergedeelte, voor dataset *a*, en in het rechtergedeelte voor dataset *b*). De volgorde van de vier kolommen in deze tabellen verschilt van de volgorde waarin de vier analyses hierboven genoemd werden.

1. Vul in, in de onderste regel van de tabellen in het linker- en het rechtergedeelte van Tabel 3.59, van welke van de vier analyses deze effectkwadratensommen afkomstig zijn: type I optie en de factoren in de volgorde (A,B), type I optie en de factoren in de volgorde (B,A), type II optie of de type III optie.

| (a) | Analyse | | | | (b) | Analyse | | | |
|---|---|---|---|---|---|---|---|---|---|
| Effect | 1 | 2 | 3 | 4 | Effect | 1 | 2 | 3 | 4 |
| A | 48.32 | 48.32 | 0.00 | 49.53 | A | 49.53 | 250.00 | 51.33 | 51.33 |
| B | 189.70 | 238.02 | 238.02 | 239.72 | B | 239.72 | 238.02 | 436.69 | 238.02 |
| A×B | 1.98 | 1.98 | 1.98 | 1.98 | A×B | 1.98 | 1.98 | 1.98 | 1.98 |
| Model | 240.00 | 240.00 | 240.00 | 240.00 | Model | 490.00 | 490.00 | 490.00 | 490.00 |
| | | | | | | | | | |

Tabel 3.59: Tabel bij opgave 26.

2. In Tabel 3.60 staan de celgemiddelden (vetgedrukt) en de aantallen observaties per cel voor elk van de twee onderzoeken waarvan sprake hierboven. De factoren A en B hebben elk twee niveaus die aangeduid worden met Niv. 1 en Niv. 2. Met welke tabel van effectkwadratensommen ($a$ of $b$) in Tabel 3.59 corresponderen deze tabellen? Leg uit.

| | | Factor B | | | | Factor B | |
|---|---|---|---|---|---|---|---|
| | | Niv. 1 | Niv. 2 | | | Niv. 1 | Niv. 2 |
| Factor A | Niv. 1 | **4**<br>5 | **10**<br>15 | Factor A | Niv. 1 | **10**<br>5 | **4**<br>15 |
| | Niv. 2 | **2**<br>14 | **7**<br>6 | | Niv. 2 | **7**<br>14 | **2**<br>6 |

Tabel 3.60: Tabel bij opgave 26.

*Opgave 27*
Een onderzoeker had een factor B opgenomen in zijn design omdat hij een andere factor (A) ervoor wilde uitzuiveren. Nadat hij de data verzameld had, stelde hij echter vast dat de type I en de type III kwadratensommen van factor A precies gelijk waren. Kun je dan zeggen dat de onderzoeker factor B net zo goed niet in zijn design had kunnen opnemen? Met andere woorden, zou zijn F-toets precies hetzelfde resultaat opleveren indien hij factor B niet had opgenomen?

*Opgave 28*
Geef aan of de onderstaande uitspraken waar of onwaar zijn. Als een uitspraak onwaar is, verander dan het schuin gedrukte deel van de zin zodat de uitspraak waar wordt. (Het aantal woorden waarmee je een schuin gedrukt deel vervangt hoeft niet gelijk te zijn aan het aantal woorden dát je vervangt.)

1. Als de aantallen observaties per cel gelijk zijn aan elkaar, dan is de Type I kwadratensom gelijk aan *de type III kwadratensom*.
2. *Samenhang tussen de factoren* duidt op interactie.
3. Het gebruik van een F-statistiek met een type III kwadratensom garandeert dat men uitgezuiverd heeft voor *alle mogelijke storende variabelen*.
4. Een F-statistiek met een *type III kwadratensom* in de teller toetst de gelijkheid van de ongewogen gemiddelden onder de aanname dat er geen interactie is tussen de factoren.
5. In designs met lege cellen kan de *type I kwadratensom* niet berekend worden.

*Opgave 29*
In een artikel van Jones (1984) worden gegevens gepresenteerd waaruit blijkt dat een deel van het verschil tussen zwarten en blanken op wiskundetoetsen (blanken doen het hier beter) toe te schrijven is aan een verschil in het aantal vakken wiskunde dat zij in de loop van de middelbare school gevolgd hebben[16]. In Tabel 3.61 staan data die vergelijkbaar zijn met de data die Jones rapporteert. De getallen in deze tabel zijn de scores op een wiskundetoets van een groep zwarte en blanke Amerikanen die 0, 1, 2 of 3 vakken wiskunde gevolgd hebben.

| RAS | | | | | | | |
|---|---|---|---|---|---|---|---|
| Zwart | | | | Blank | | | |
| AANTAL VAKKEN | | | | AANTAL VAKKEN | | | |
| 0 | 1 | 2 | 3 | 0 | 1 | 2 | 3 |
| 45 | 51 | 61 | 71 | 42 | 61 | 63 | 77 |
| 34 | 59 | 73 | 82 | 51 | 48 | 68 | 68 |
| 51 | 53 | 55 | 70 | 39 | 46 | 78 | 79 |
| 54 | 49 | 77 | | 55 | 63 | 60 | 66 |
| 40 | 60 | | | | 55 | 73 | 85 |
| 46 | 65 | | | | | | 80 |
| | 59 | | | | | | |

Tabel 3.61: Tabel bij opgave 29.

1. Wat is de gemiddelde toetsscore voor (afzonderlijk) de zwarte en de blanke studenten, zonder uit te zuiveren voor het aantal vakken dat ze gevolgd hebben? Hoe groot is het verschil tussen deze twee gemiddelden?
2. Wat is de gemiddelde toetsscore voor (afzonderlijk) de zwarte en de blanke studenten, wél uitgezuiverd voor het aantal vakken dat ze gevolgd hebben? Hoe groot is het verschil tussen deze twee gemiddelden?
3. Waarom is het verschil in 2. veel kleiner dan het verschil in 1.?
4. COMPUTEROPDRACHT
   (a) De gegevens in Tabel 3.61 vind je in het SPSS-bestand "witzwart.sav". De variabelen zijn de volgende:
      i. RAS: Zwart (0) of blank (1).
      ii. NVAK: Aantal vakken (0, 1, 2 of 3).
      iii. WISTOETS: Score op de wiskundetoets.
   (b) Toets de significantie van het hoofdeffect van RAS zonder rekening te houden met de storende invloed van het aantal vakken dat men gevolgd heeft.
   (c) Toets de significantie van het hoofdeffect van RAS, uitgezuiverd voor het hoofdeffect van AANTAL VAKKEN.
   (d) Toets de significantie van het hoofdeffect van RAS, uitgezuiverd voor het hoofdeffect van AANTAL VAKKEN en de interactie tussen RAS en AANTAL VAKKEN.

[16] Voor deze opgave heb ik me laten inspireren door opgave 18 van hoofdstuk 7 (pp. 306-307) van Maxwell en Delaney (1990).

## 3.4 Uitzuivering voor een Kwantitatieve Variabele met Covariantieanalyse

### 3.4.1 Voorbeeld: Pesten in de Klas

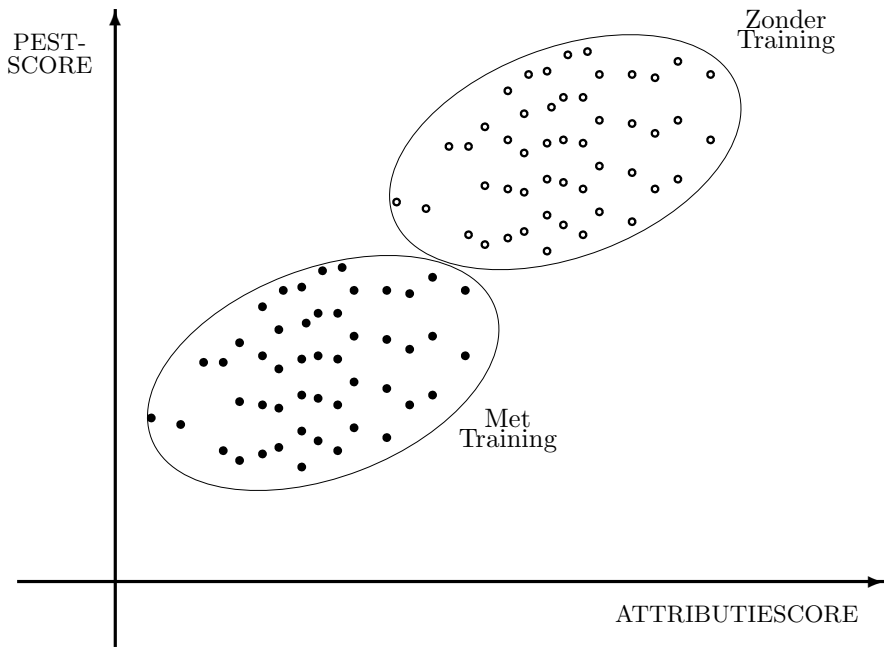
Ter herinnering, het getallenvoorbeeld over pesten in de klas was afkomstig uit een onderzoek naar het effectiviteit van een training in sociale weerbaarheid op de mate waarin leerlingen gepest worden. Een gerandomiseerd experiment is de ideale manier om dit te onderzoeken. Concreet, men wijst de leerlingen op zuiver toevallige wijze toe aan, ofwel de sociale weerbaarheidstraining, ofwel een controleconditie, en vergelijkt de twee groepen (wat betreft de mate waarin de leerlingen gepest worden). In de praktijk is het echter vaak onmogelijk om een gerandomiseerd experiment uit te voeren. Men stelt zich dan tevreden met een observationeel onderzoek en zuivert achteraf uit voor alle storende variabelen waarover men beschikt. In de vorige paragraaf concentreerden we ons op één storende variabele, TYPE LERAAR. We zuiverden uit voor TYPE LERAAR door de ongewogen marginale gemiddelden te berekenen.

Als de storende variabele categorisch is (bijv., TYPE LERAAR), dan komt uitzuiveren neer op het berekenen van ongewogen marginale gemiddelden. Als de storende variabele niet categorisch is maar kwantitatief, dan kan men geen ongewogen marginale gemiddelden berekenen. Uitzuiveren gebeurt met behulp van *uitgezuiverde gemiddelden*. Tussen haakjes, ongewogen marginale gemiddelden zijn ook uitgezuiverde gemiddelden, maar ze zijn van een speciaal soort; de uitgezuiverde gemiddelden die men berekent als de storende variabele kwantitatief is, zijn van een ander soort. Het statistisch toetsen van verschillen tussen uitgezuiverde gemiddelden is mogelijk binnen een speciaal soort variantieanalyse, de covariantieanalyse.

De kwantitatieve storende variabele in het voorbeeldonderzoek is de ATTRIBUTIESCORE. De achtergrond hierbij is dat kinderen van elkaar verschillen wat betreft hun mening over de oorzaak van het pestgedrag. Sommige kinderen leggen de oorzaak vooral bij henzelf; volgens hen ligt het probleem bij het feit dat zij niet adekwaat reageren op het pestgedrag van de anderen. Dit noemen we een *interne* attributie. Andere kinderen leggen de oorzaak vooral bij de anderen; volgens hen ligt het probleem bij pesters en zij zien zichzelf als een willekeurig gekozen slachtoffer. Dit noemen we een *externe* attributie. We zullen later laten zien dat ATTRIBUTIESCORE een storende variabele is. Op dit moment is het voldoende om te weten dat de ATTRIBUTIESCORE bepaald werd bij het begin van het schooljaar. Dit gebeurde met behulp van een vragenlijst. Een hoge score op deze vragenlijst betekent dat de leerling problemen (zoals gepest worden) vooral extern attribueert, en een lage score betekent dat de leerling problemen vooral intern attribueert.

Er was een positief verband tussen ATTRIBUTIESCORE en PESTSCORE: hoe meer de leerlingen hun problemen extern attribueerden, hoe meer ze gepest werden. Er was een negatief verband tussen ATTRIBUTIESCORE en de kans om deel te nemen aan de training: hoe meer de leerlingen hun problemen extern attribueerden, hoe kleiner de kans dat zij aan de training deelnamen. Dit wordt getoond in Figuur 3.4. De observaties in de groepen met en zonder training worden elk omgeven door een ellips. In de groep met training worden observaties aangeduid met dikke punten en in de groep zonder

training met kleine cirkeltjes. Elke ellips toont een positief verband tussen ATTRIBUTIESCORE en PESTSCORE: hoge attributiescores gaan samen met hoge pestscores en lage attributiescores met lage pestscores. De ellips van de groep zonder training staat meer naar rechts dan die van de groep met training. Dit betekent dat leerlingen met een hoge attributiescore een kleinere kans hebben om deel te nemen aan de training dan leerlingen met een lage attributiescore.

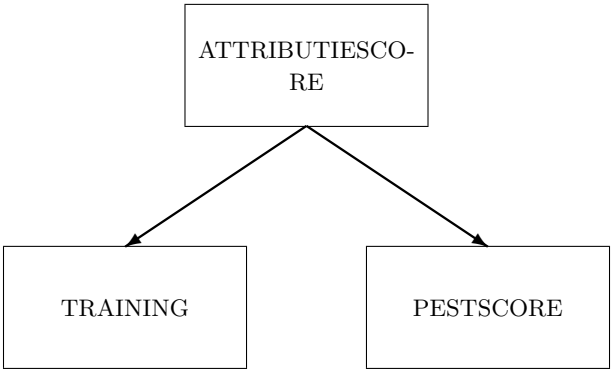



Figuur 3.4: Grafische voorstelling van de attributie- en de pestscores in de groep met en zonder training.

Dit patroon in de data is begrijpelijk als men er vanuit gaat dat naarmate leerlingen hun problemen meer intern attribueren (een lagere attributiescore hebben) zij zich in de omgang met de pesters weerbaarder opstellen. Door deze opstelling worden zij ook minder gepest dan leerlingen die hun problemen meer extern attribueren. Verder zijn het ook vooral de intern attribuerende leerlingen die de training in sociale weerbaarheid zinvol vinden. Omdat zij het pestgedrag (voor een deel) toeschrijven aan de manier waarop zij zichzelf opstellen, is het voor hen zaak om hierin te veranderen. Extern attribuerende leerlingen zien zichzelf als het willekeurige slachtoffer van het pestgedrag, en zien daarom niet in waarom een training in sociale weerbaarheid hier iets aan zou veranderen.

De variabele ATTRIBUTIESCORE wordt een *covariaat* genoemd. Een covariaat is een potentiële storende variabele: door zijn eventuele samenhang met zowel de afhankelijke als de onafhankelijke variabele kan hij verantwoordelijk zijn voor het verband tussen beide. Dit wordt grafisch voorgesteld in Figuur 3.5. In deze figuur wordt de situatie voorgesteld waarin het verband tussen TRAINING en PESTSCORE volledig is toe te

schrijven aan hun beider verband met ATTRIBUTIESCORE. Zoals verder zal blijken, is dit niet het geval in ons voorbeeldonderzoek, want na uitzuivering voor ATTRIBUTIESCORE is er nog steeds een verband tussen TRAINING en PESTSCORE (weliswaar kleiner dan vóór de uitzuivering, maar niet gelijk aan nul).



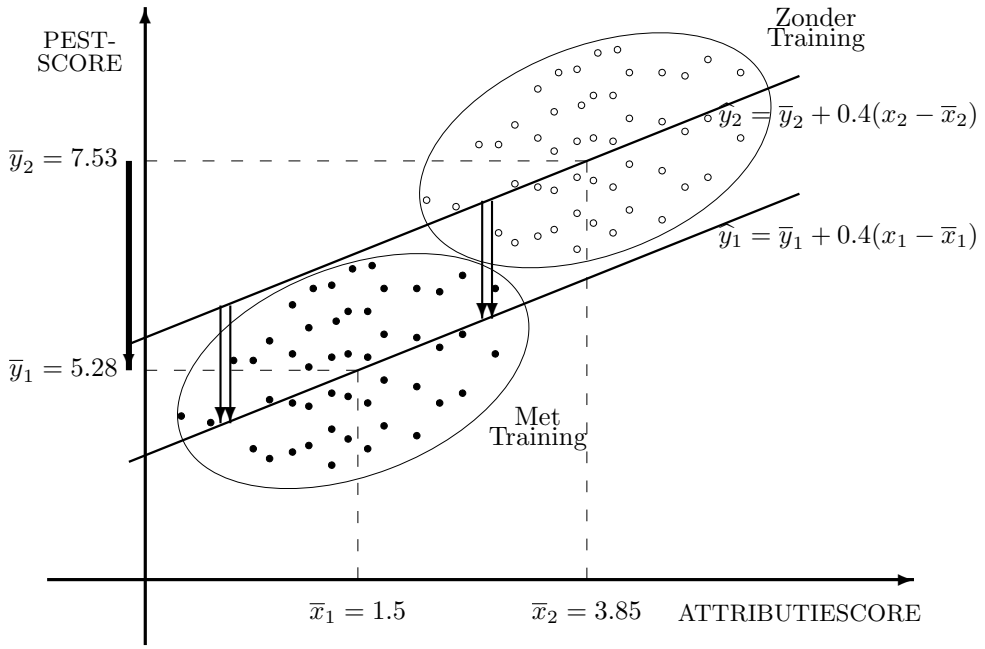

Figuur 3.5: Grafische voorstelling van de invloed van een storende variabele (ATTRIBUTIESCORE) op het verband tussen de afhankelijke variabele (PESTSCORE) en de variabele waarvan men het effect wil onderzoeken (TRAINING).

Het effect van TRAINING kan op twee manieren bepaald worden: met en zonder uitzuivering voor de covariaat. Met uitzuivering houdt in dat men rekening houdt met het verband tussen ATTRIBUTIESCORE en PESTSCORE, en zonder uitzuivering houdt in dat men hier geen rekening mee houdt. Voor het effect van TRAINING zonder uitzuivering voor de covariaat wordt het verschil genomen van de gemiddelde pestscores in de twee groepen. De pestscores in de groep met en de groep zonder training worden aangeduid met, respectievelijk, $y_1$ en $y_2$, en hun gemiddelden met $\overline{y}_1$ en $\overline{y}_2$. In het voorbeeldonderzoek zijn deze gemiddelden gelijk aan, respectievelijk, 5.28 en 7.53, en hun verschil is $\overline{y}_2 - \overline{y}_1 = 7.53 - 5.28 = 2.25$. In Figuur 3.6 wordt dit verschil aangegeven met de dikke pijl.

We bespreken nu het effect van TRAINING mét uitzuivering voor de covariaat. Uitzuiveren voor de covariaat gebeurt met behulp van *regressielijnen*. Deze regressielijnen worden getoond in Figuur 3.6. In deze figuur wordt het verband tussen PESTSCORE en ATTRIBUTIESCORE beschreven met twee regressielijnen, één voor elke groep. De attributiescores in de twee groepen worden aangeduid met $x_1$ en $x_2$, en hun gemiddelden met $\overline{x}_1$ en $\overline{x}_2$. De regressielijn voor de groep met training heeft als predictor $x_1 - \overline{x}_1$, de covariaat voor deze groep min zijn gemiddelde[17]. En de regressielijn voor de groep zonder training heeft als predictor $x_2 - \overline{x}_2$, opnieuw de covariaat voor deze groep min zijn gemiddelde. De voorspellingen in de groep met en in de groep zonder training worden aangeduid met, respectievelijk, $\widehat{y}_1$ en $\widehat{y}_2$. We veronderstellen dat de

[17] In het verschil $x_1 - \overline{x}_1$ staat $x_1$ voor een variabele en $\overline{x}_1$ voor een constante. Dit verschil is een variabele waarvan elk element gelijk is aan het corresponderende element van $x_1$ min het gemiddelde $\overline{x}_1$.

regressiegewichten[18] in de twee groepen gelijk zijn. In het voorbeeldonderzoek zijn ze gelijk aan 0.4. De regressielijn in de eerste groep wordt beschreven door de vergelijking $\widehat{y}_1 = \overline{y}_1 + 0.4(x_1 - \overline{x}_1)$ en die in de tweede groep door $\widehat{y}_2 = \overline{y}_2 + 0.4(x_2 - \overline{x}_2)$. Deze regressielijnen worden getoond in Figuur 3.6. De regressielijn van de eerste groep loopt door het punt $(\overline{x}_1, \overline{y}_1)$. Dit komt omdat de vergelijking $\widehat{y}_1 = \overline{y}_1 + 0.4(x_1 - \overline{x}_1)$ op het punt $x_1 = \overline{x}_1$ de waarde $\widehat{y}_1 = \overline{y}_1$ oplevert. Voor alle $x_1$-waarden kleiner dan $\overline{x}_1$ is $\widehat{y}_1$ kleiner dan $\overline{y}_1$. Dit komt omdat de term $0.4(x_1 - \overline{x}_1)$ dan negatief is. En voor alle $x_1$-waarden groter dan $\overline{x}_1$ is $\widehat{y}_1$ groter dan $\overline{y}_1$. Dit komt omdat $0.4(x_1 - \overline{x}_1)$ dan positief is. De regressielijn van de tweede groep kan op een analoge manier beschreven worden. De twee regressielijnen lopen parallel omdat zij hetzelfde regressiegewicht hebben.




Figuur 3.6: Het verband tussen ATTRIBUTIESCORE en PESTSCORE zoals beschreven met twee regressielijnen, één voor de groep met en één voor de groep zonder training.

Voor het effect van TRAINING mét uitzuivering voor de covariaat wordt het verschil genomen tussen de voorspelde scores voor de twee groepen *bij een vaste (maar willekeurige) waarde van de covariaat*. We nemen dus het verschil tussen de twee regressielijnen. Omdat de regressielijnen parallel zijn is dit verschil hetzelfde voor alle

[18] De regressiegewichten zijn de hellingsgraden van de regressielijnen. Het regressiegewicht is gelijk aan de toename in PESTSCORE die gepaard gaat met een eenheid toename in ATTRIBUTIESCORE.

punten op de X-as. Voor het punt $a$ is dit verschil het volgende:

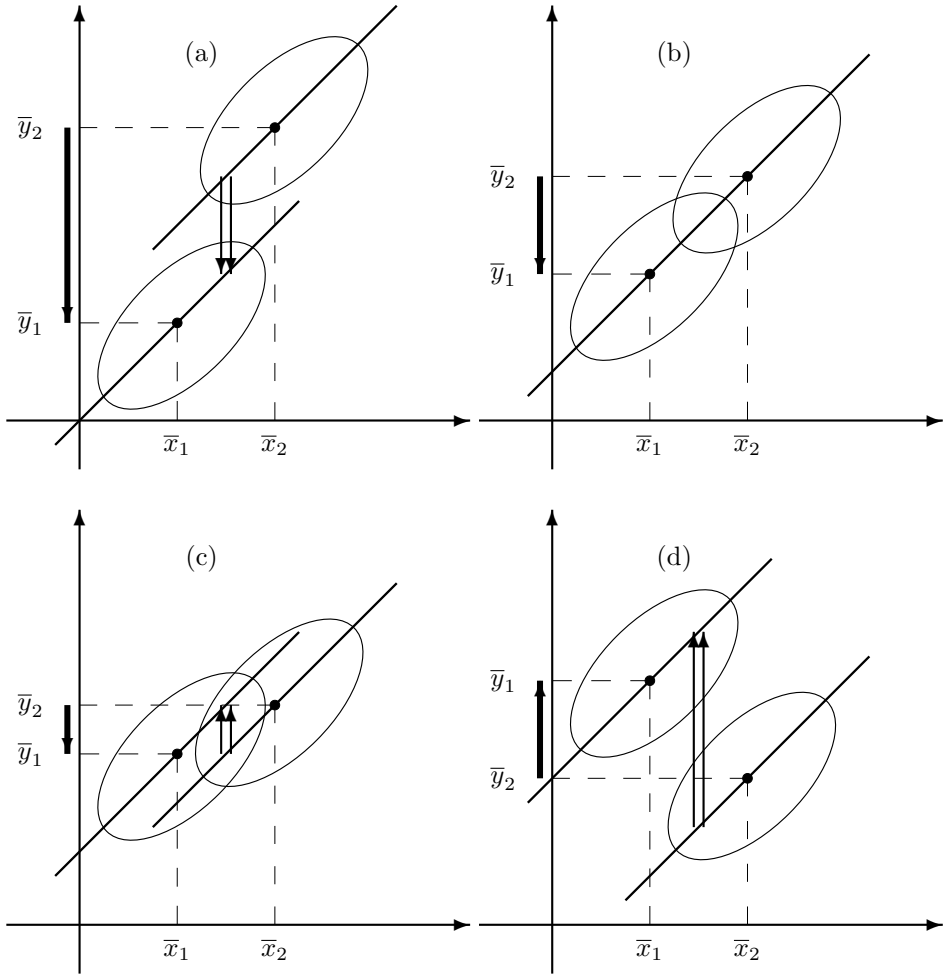
$$\begin{aligned}\widehat{y}_2 - \widehat{y}_1 &= \overline{y}_2 + 0.4(a - \overline{x}_2) - [\overline{y}_1 + 0.4(a - \overline{x}_1)] \\ &= \overline{y}_2 + 0.4(a - \overline{x}_2) - \overline{y}_1 - 0.4(a - \overline{x}_1) \\ &= \overline{y}_2 - \overline{y}_1 + 0.4a - 0.4a - 0.4\overline{x}_2 + 0.4\overline{x}_1 \\ &= \overline{y}_2 - \overline{y}_1 - 0.4(\overline{x}_2 - \overline{x}_1) \quad .\end{aligned}$$

Het is belangrijk om te zien dat het verschil $\widehat{y}_2 - \widehat{y}_1$ niet afhangt van $a$. In Figuur 3.6 wordt dit verschil aangegeven met twee dubbele pijlen. Elke dubbele pijl hoort bij een punt op de X-as. In het voorbeeldonderzoek zijn $\overline{x}_1$ en $\overline{x}_2$ gelijk aan 1.5 en 3.85. Het verschil $\widehat{y}_2 - \widehat{y}_1$ is dan: $7.53 - 5.28 - 0.4(3.85 - 1.5) = 1.31$. Dit is minder dan het gewone verschil tussen de gemiddelden (2.25). Het verschil tussen 2.25 ($\overline{y}_2 - \overline{y}_1$) en 1.31 ($\widehat{y}_2 - \widehat{y}_1$) is toe te schrijven aan de term $-0.4(\overline{x}_2 - \overline{x}_1)$. Deze term zuivert $\overline{y}_2 - \overline{y}_1$ uit voor dat gedeelte dat toe te schrijven is aan het verband tussen $x$ en $y$. We noemen $\overline{y}_2 - \overline{y}_1 - 0.4(\overline{x}_2 - \overline{x}_1)$ dan ook het verschil tussen $\overline{y}_1$ en $\overline{y}_2$ *uitgezuiverd voor* $x$, of kortweg, *het uitgezuiverde verschil*. Men spreekt ook van het (voor $x$) *gecorrigeerde verschil*.

Het verschil tussen de twee regressielijnen correspondeert met het verschil tussen de ongewogen gemiddelden in een ongebalanceerd design. Een verschil tussen ongewogen gemiddelden is gebaseerd op het verschil tussen gemiddelden (voorspellingen) binnen elk niveau van een andere variabele (i.e. $\overline{y}_{21} - \overline{y}_{11}$, $\overline{y}_{22} - \overline{y}_{12}$,... ). Dit correspondeert met de manier waarop we hier het effect van de training bepalen, namelijk als het verschil tussen de voorspellingen bij een vaste waarde van de covariaat.

We willen niet dat onze evaluatie van het effect van de TRAINING afhangt van verschillen tussen de groepen op de covariaat. Daarom is de grootte van het verticale verschil tussen de twee regressielijnen precies wat we nodig hebben om het effect van TRAINING te evalueren. De grootte van dit verschil hangt immers niet af van de verschillen tussen de groepen op de covariaat. Stel dat we de verschillen tussen de groepen op de covariaat veranderen door de twee puntenwolken in Figuur 3.6 in horizontale richting naar elkaar toe zouden schuiven over hun beider regressielijnen (dus, de middelpunten van de twee puntenwolken blijven mooi op de regressielijn liggen). Dit zou geen invloed hebben op het verticale verschil tussen de regressielijnen. Het zou echter wel een invloed hebben op het verschil tussen $\overline{y}_2$ en $\overline{y}_1$. Dit is de reden waarom we het effect van TRAINING evalueren met het verticale verschil tussen de twee regressielijnen.

In dit voorbeeld is het gewone verschil tussen $\overline{y}_1$ en $\overline{y}_2$ groter dan het uitgezuiverde verschil. Dit is niet altijd zo, want soms wordt het verschil kleiner, soms groter en soms klapt de richting van het verschil zelfs om (positief wordt negatief of omgekeerd). Dit wordt geïllustreerd in Figuur 3.7. Van (a) naar (d) wordt de linkerellips telkens een beetje omhoog geschoven en de rechterellips een beetje omlaag. Het gewone verschil tussen $\overline{y}_1$ en $\overline{y}_2$ wordt aangeduid met de dikke pijl, die naar beneden gericht is als $\overline{y}_2$ groter is dan $\overline{y}_1$ en naar boven als $\overline{y}_2$ kleiner is. Het uitgezuiverde verschil wordt aangeduid met de dubbele pijl. Deze pijl is naar beneden gericht als het uitgezuiverde verschil positief is en naar boven als het uitgezuiverde verschil negatief is. In (b) is het uitgezuiverde verschil gelijk aan 0. Dit komt omdat de twee regressielijnen over elkaar liggen. In (c) is $\overline{y}_2 - \overline{y}_1$ positief en het uitgezuiverde verschil negatief. En in (d) is $\overline{y}_2 - \overline{y}_1$ in absolute waarde kleiner dan het uitgezuiverde verschil.




Figuur 3.7: Vier verschillende relaties tussen het gewone verschil $\overline{y}_2 - \overline{y}_1$ en het uitgezuiverde verschil $\overline{y}_2 - \overline{y}_1 - B(\overline{x}_2 - \overline{x}_1)$ (waarin $B$ het regressiegewicht is): (a) verkleining, (b) verkleining tot nul, (c) omkering, en (d) vergroting.

### 3.4.2 Uitgezuiverde Verschillen en Gemiddelden

In het voorbeeldonderzoek hebben we te maken met twee groepen, een groep met en een groep zonder training. In het algemeen, hebben we te maken met een onderzoek met een willekeurig aantal groepen. We introduceren nu wat notatie waarmee we de uitgezuiverde verschillen kunnen uitdrukken voor een willekeurig aantal groepen. Het aantal groepen duiden we aan met $K$, en als index voor deze groepen gebruiken we $k$ ($k = 1, \ldots, K$). De afhankelijke variabele in de k-de groep duiden we aan met $y_k$ en de covariaat met $x_k$. In het voorbeeldonderzoek is PESTSCORE de afhankelijke variabele en ATTRIBUTIESCORE de covariaat. De observaties zijn paren: het eerste lid van elk paar is de afhankelijke variabele en het tweede is de covariaat. Het gemiddelde in de k-de groep duiden we aan met $(\overline{y}_k, \overline{x}_k)$.

De eerste stap in de definitie van het uitgezuiverde verschil is het voorspellen van de observaties op basis van de regressielijnen. Elke groep heeft zijn eigen regressielijn, en die van de k-de groep wordt aangeduid met $\widehat{y}_k$:

$$\boxed{\widehat{y}_k = \overline{y}_k + B(x_k - \overline{x}_k) \quad .}$$

De predictor in de k-de groep is $x_k - \overline{x}_k$, de afwijking van de covariaat ten opzichte van zijn gemiddelde in deze groep. Bij elke waarde van $x_k$ hoort een voorspelling gelijk aan $[\overline{y}_k + B(x_k - \overline{x}_k)]$. Dit is de vergelijking van een rechte lijn met hellingsgraad $B$ door het punt $(\overline{x}_k, \overline{y}_k)$. Dit laatste volgt uit het invullen van $\overline{x}_k$ voor $x_k$ in $[\overline{y}_k + B(x_k - \overline{x}_k)]$. Dus, op de gemiddelde covariaatscore van een groep $(\overline{x}_k)$ voorspellen we op basis van de regressielijn een score die gelijk is aan het gemiddelde van deze groep $(\overline{y}_k)$. Om de plaats te vinden waar de regressielijn de Y-as snijdt, moeten we voor $x_k$ het getal 0 invullen. We krijgen dan $\overline{y}_k - B\overline{x}_k$. Dit is het zogenaamde *intercept* van de regressielijn. In Figuur 3.8 worden drie regressielijnen getoond, elk met een andere waarde voor $\overline{y}_k$ en $\overline{x}_k$.

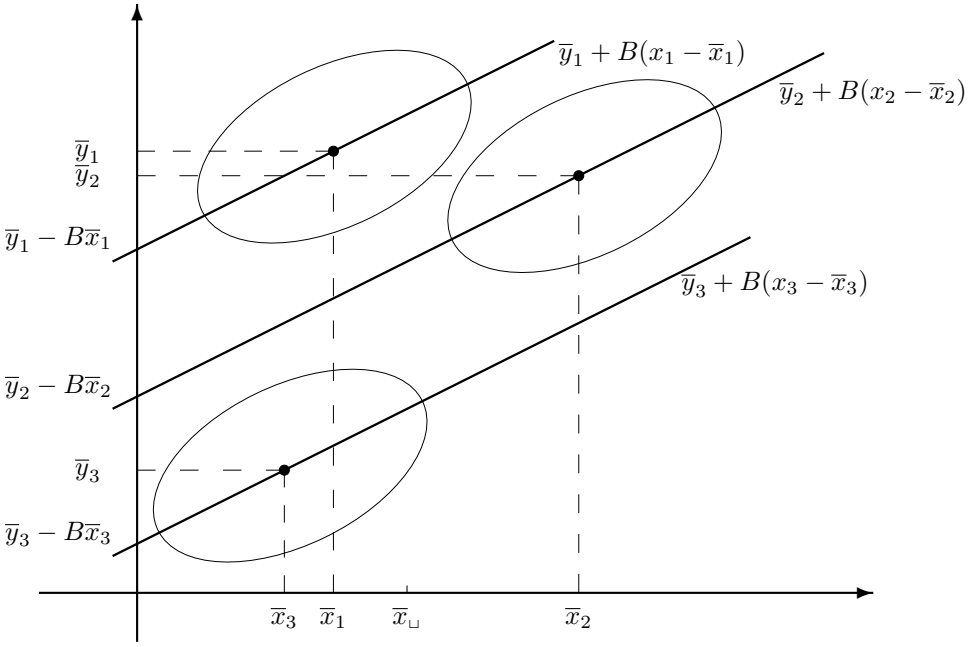
Het uitgezuiverde verschil tussen groep 1 en 2 is de verticale afstand tussen de regressielijnen van deze twee groepen. Omdat deze regressielijnen parallel zijn, krijgt men hetzelfde verschil op elk punt van de X-as. We duiden zo'n punt aan met $a$. Het uitgezuiverde verschil wordt dan als volgt berekend:

$$\begin{aligned}[\overline{y}_2 + B(a - \overline{x}_2)] - [\overline{y}_1 + B(a - \overline{x}_1)] &= \overline{y}_2 - \overline{y}_1 + Ba - Ba - B\overline{x}_2 + B\overline{x}_1 \\ &= \overline{y}_2 - \overline{y}_1 - B(\overline{x}_2 - \overline{x}_1) \quad ,\end{aligned}$$

hetgeen onafhankelijk is van $a$. Op dezelfde manier berekenen we het uitgezuiverde verschil tussen twee andere groepen. Bijvoorbeeld, in een onderzoek met vier groepen is het uitgezuiverde verschil tussen groep 3 en 4 het volgende: $\overline{y}_4 - \overline{y}_3 - B(\overline{x}_4 - \overline{x}_3)$. Met de termen $B(\overline{x}_2 - \overline{x}_1)$ en $B(\overline{x}_4 - \overline{x}_3)$ wordt een gewoon verschil tussen gemiddelden uitgezuiverd voor de covariaat.

Omdat de uitgezuiverde verschillen gelijk zijn voor alle punten op de X-as, kan men ze ook berekenen in het punt 0. De hoogte van een regressielijn op het punt 0 is het intercept van die regressielijn. Het uitgezuiverde verschil tussen twee groepen is dus gelijk aan het verschil tussen de intercepten van die twee groepen.

Als er veel groepen zijn, dan is het aantal uitgezuiverde verschillen nog veel groter: bij 3 groepen zijn er 3 verschillen, bij 4 groepen 6 verschillen, bij 5 groepen 10 verschillen,




Figuur 3.8: Uitgangspunt bij het berekenen van de uitgezuiverde verschillen in een onderzoek met drie groepen: drie parallelle regressielijnen, één voor elk van de drie groepen.

.... Daarom rapporteert men ook vaak *uitgezuiverde (gecorrigeerde) gemiddelden*. Het uitgezuiverde gemiddelde van de k-de groep is de voorspelling op basis van de k-de regressievergelijking voor een bepaald standaardpunt op de X-as. Dit standaardpunt is meestal $\overline{x}_{\sqcup}$, het gemiddelde van alle covariaatscores, en het ligt altijd tussen de covariaatgroepsgemiddelden $\overline{x}_k$ (zie Figuur 3.8). Het uitgezuiverde gemiddelde voor de k-de groep wordt dan als volgt gedefinieerd:

$$\boxed{\text{Uitgezuiverde gemiddelde voor groep } k = \overline{y}_k + B(\overline{x}_{\sqcup} - \overline{x}_k) \quad .}$$

Met deze uitgezuiverde gemiddelden kan men zelf de uitgezuiverde verschillen berekenen. Om de grafiek niet te vol te maken, zijn de uitgezuiverde gemiddelden niet in Figuur 3.8 getekend. Men kan ze zelf tekenen door de snijpunten te bepalen van een verticale lijn boven $\overline{x}_{\sqcup}$ met elk van de drie regressielijnen. De uitgezuiverde gemiddelden zijn gelijk aan de hoogtes van deze snijpunten.

### 3.4.3 De Nulhypothese voor het Uitgezuiverde Populatieverschil

Men is bijna nooit alleen geïnteresseerd in de grootte van de uitgezuiverde verschillen, maar men wil ook weten of ze statistisch significant zijn. Dit is begrijpelijk omdat de uitgezuiverde verschillen in de praktijk bijna altijd verschillen van nul. Maar het is de vraag of dit ook geldt voor hun tegenhangers in de populatie. Voor groep 1 en 2 is het uitgezuiverde populatieverschil het volgende:

$$\mu_2 - \mu_1 - B^{pop}(\overline{x}_2 - \overline{x}_1) \quad .$$

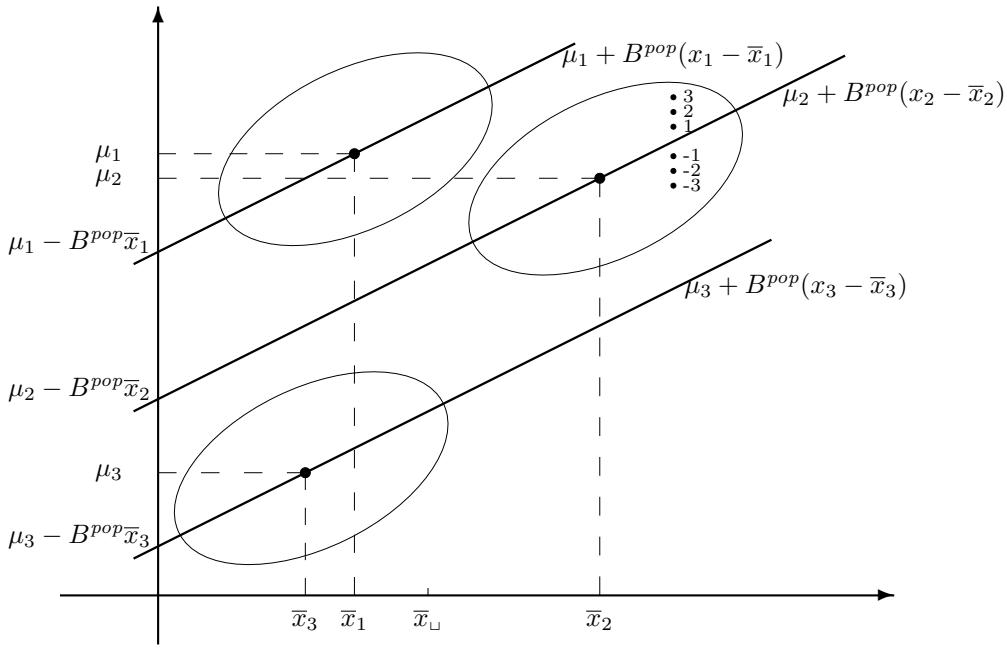
Dit is de populatievariant van $[\overline{y}_2 - \overline{y}_1 - B(\overline{x}_2 - \overline{x}_1)]$, het uitgezuiverde verschil in de steekproef. In de populatievariant van het uitgezuiverde verschil worden de steekproefgrootheden $\overline{y}_1$, $\overline{y}_2$ en $B$ vervangen door de onbekende populatiegrootheden $\mu_1$, $\mu_2$ en $B^{pop}$. De populatiegrootheden $\mu_1$, $\mu_2$ en $B^{pop}$ zijn de $\overline{y}_1$, $\overline{y}_2$ en $B$ die men zou observeren als de steekproef oneindig groot zou zijn. Voor wat volgt, is het handig om het uitgezuiverde populatieverschil ook als volgt te schrijven:

$$(\mu_2 - B^{pop}\overline{x}_2) - (\mu_1 - B^{pop}\overline{x}_1) \quad .$$

We zijn geïnteresseerd in de nulhypothese dat, voor elk paar van groepen, het uitgezuiverde populatieverschil gelijk is aan 0. Deze hypothese komt neer op het feit dat $\mu_k - B^{pop}\overline{x}_k$ hetzelfde is voor alle $K$ groepen:

$$\boxed{\mu_1 - B^{pop}\overline{x}_1 = \mu_2 - B^{pop}\overline{x}_2 = \ldots = \mu_K - B^{pop}\overline{x}_K \quad .}$$

Merk op dat $\mu_k - B^{pop}\overline{x}_k$ het intercept is van de populatieregressielijn $[\mu_k + B^{pop}(x_k - \overline{x}_k])$ (i.e., de plaats waar deze lijn de Y-as snijdt). Dit wordt getoond in Figuur 3.9, de populatietegenhanger van Figuur 3.8. (De cijfers 3, 2, 1, -1, -2, en -3 in Figuur 3.9 zullen later pas uitgelegd worden.) Dus, de nulhypothese kan ook als volgt geformuleerd worden: in de populatie, zijn de intercepten van de $K$ groepsspecifieke regressielijnen gelijk.




Figuur 3.9: Uitgangspunt bij de covariantieanalyse met drie groepen: drie parallelle regressielijnen, één voor elk van drie groepen.

Het is interessant om te kijken wat er met de nulhypothese gebeurt indien de covariaat geen storende variabele is. De covariaat is geen storende variabele indien aan één van de volgende twee voorwaarden is voldaan:

1. Er is geen verband tussen de covariaat en de afhankelijke variabele. Dit betekent dat het populatieregressiegewicht $B^{pop}$ gelijk is aan 0.

2. Er is geen verband tussen de covariaat en de onafhankelijke variabele. Dit betekent dat de gemiddelden van de covariaat $x_k$ in de $K$ groepen, de $K$ $\overline{x}_k$'s, gelijk zijn.

Indien voldaan is aan de eerste voorwaarde ($B^{pop} = 0$), dan is de nulhypothese dezelfde als die van de variantieanalyse met één factor:

$$\mu_1 = \mu_2 = \ldots = \mu_K \quad .$$

Ook indien aan de tweede voorwaarde voldaan is (de $K$ gemiddelden $\overline{x}_k$ zijn gelijk), maar niet aan de eerste ($B^{pop} \neq 0$), dan is de nulhypothese dezelfde als die van de variantieanalyse met één factor. Dit komt omdat, in dit geval, de term $B^{pop}\overline{x}_k$ dezelfde is voor alle groepen. Dus, indien de covariaat geen storende variabele is, dan is de nulhypothese van gelijke intercepten $\mu_k - B^{pop}\overline{x}_k$ dezelfde als de nulhypothese van de variantieanalyse met één factor.

### 3.4.4 De F-statistieken

De nulhypothese dat, in de populatie, de intercepten van de $K$ groepsspecifieke regressielijnen gelijk zijn, wordt getoetst met een F-statistiek van de bekende algemene vorm:

$$\boxed{F = \frac{MS_{Effect|Ref}}{MS_{Error|Model}} \quad .}$$

In deze formule vullen we voor *Effect* de factor A in. Dit is de factor die het verschil tussen de groepen bepaalt. In het voorbeeld over pesten in de klas is factor A de variabele TRAINING. Het referentiemodel is het model met alleen de covariaat X. Dit betekent dat het effect van factor A uitgezuiverd wordt voor deze covariaat. Met deze invulling van de F-statistiek toetsen we de nulhypothese dat, in de populatie, de intercepten van de $K$ groepsspecifieke regressielijnen gelijk zijn.

Met een andere invulling kan de F-statistiek ook gebruikt worden om andere nulhypotheses te toetsen, zoals de nulhypothese dat het populatieregressiegewicht $B^{pop}$ gelijk is aan nul. Deze en andere nulhypotheses zijn niet zo interessant als de nulhypothese dat de intercepten van de $K$ groepsspecifieke regressielijnen gelijk zijn. Hun F-statistieken worden hier toch besproken omdat ze gegeven worden in de uitdraai van computerprogramma's zoals GLM-Univariate.

### 3.4.5 De Effectkwadratensommen

De effectkwadratensom staat in de teller van de F-statistiek en bepaalt wat men toetst. De algemene formule van de effectkwadratensom is als volgt:

$$\boxed{SS_{Effect|Ref} = \sum_{Obs} (\widehat{y}_{ki}^{Ref+Effect} - \widehat{y}_{ki}^{Ref})^2 \quad .}$$

In het rechterlid van deze formule worden de voorspellingen op basis van twee verschillende modellen met elkaar vergeleken: (1) het referentiemodel samen met het effect ($\widehat{y}_{ki}^{Ref+Effect}$) en (2) het referentiemodel zonder het effect ($\widehat{y}_{ki}^{Ref}$). De letter $i$ in de bovenstaande formule is de index voor de proefpersonen binnen de k-de groep.

De nulhypothese die men toetst met een bepaalde effectkwadratensom hangt af van de variabele die men invult voor *Effect* en de variabelen die men opneemt in het referentiemodel (*Ref*). We kunnen voor *Effect* zowel de factor A als de covariaat X invullen, maar in de praktijk zijn we meestal geïnteresseerd in het effect van factor A. Voor *Ref* kunnen we *niks* (het model zonder effecten) invullen, de covariaat X (alleen als *Effect*=A), of de factor A (alleen als *Effect*=X). Door een variabele op te nemen in het referentiemodel zuivert men ervoor uit.

In wat volgt bespreken we de volgende effectkwadratensommen:

1. De kwadratensom van de factor A uitgezuiverd voor de covariaat X, aangeduid met $SS_{A|X}$. Dit is de meest belangrijke effectkwadratensom. De F-statistiek met $SS_{A|X}$ toetst de nulhypothese dat, in de populatie, de intercepten van de $K$ groepsspecifieke regressielijnen gelijk zijn.

2. De kwadratensom van de factor A zonder uitzuivering voor de covariaat, aangeduid met $SS_A$. De F-statistiek met $SS_A$ toetst de nulhypothese dat de $K$ groepsspecifieke verwachte waarden $\mu_k$ gelijk zijn. Deze nulhypothese wordt ook getoetst in de variantieanalyse met één factor.

3. De kwadratensom van de covariaat X uitgezuiverd voor de factor A, aangeduid met $SS_{X|A}$. De F-statistiek met $SS_{X|A}$ toetst de nulhypothese dat, in de populatie, het regressiegewicht van de groepsspecifieke regressielijnen ($B^{pop}$) gelijk is aan nul.

4. De kwadratensom van de covariaat X zonder uitzuivering voor de factor A, aangeduid met $SS_X$. De F-statistiek met $SS_X$ toetst de nulhypothese dat, in de populatie, het regressiegewicht van de regressielijn *over de groepen* gelijk is aan nul. Dezelfde nulhypothese wordt ook getoetst in de enkelvoudige regressieanalyse.

### De Kwadratensom van de Factor Uitgezuiverd voor de Covariaat

Men krijgt de formule van de effectkwadratensom $SS_{A|X}$ als men de voorspellingen op basis van het model met A én X ($\widehat{y}_{ki}^{X,A}$) vergelijkt met de voorspellingen op basis van het model met alleen X ($\widehat{y}_{ki}^{X}$). Het model met A en X wordt ook het *covariantieanalysemodel* genoemd, en het wordt als volgt geschreven:

$$\boxed{y_{ki} = \mu_k + B^{pop}(x_{ki} - \overline{x}_k) + e_{ki} \quad .}$$

De ingredienten van het covariantieanalysemodel worden getoond in Figuur 3.9. In deze figuur staan de drie populatieregressielijnen $[\mu_k + B^{pop}(x_k - \overline{x}_k)]$ (voor $k = 1, \ldots, 3$). Merk op dat de covariaat $x_k$ in de populatieregressielijn $[\mu_k + B^{pop}(x_k - \overline{x}_k)]$ in het covariantieanalysemodel vervangen wordt door $x_{ki}$, de covariaatscore van de i-de proefpersoon in de k-de groep. Het covariantieanalysemodel geeft aan hoe de score $y_{ki}$ van deze proefpersoon opgebouwd is: een systematisch deel, gelijk aan de populatieregressielijn op het punt $x_{ki}$, plus een ruisterm $e_{ki}$.

De ruisterm $e_{ki}$ bepaalt hoeveel de observatie afwijkt van het systematisch deel, $[\mu_k + B^{pop}(x_{ki} - \overline{x}_k)]$, de waarde die we verwachten volgens de populatieregressielijn. In Figuur 3.9 wordt voor zes observaties in de tweede groep (met allemaal dezelfde waarde $x_{ki}$) de waarde van de ruisterm $e_{ki}$ gegeven: voor drie observaties is deze ruis positief (+3, +2, en +1) en voor drie andere observaties is hij negatief (-3, -2, en -1). De verwachte waarde van de ruisterm is gelijk aan 0. Dit betekent dat, voor elke waarde van $x_{ki}$, de ruisterm $e_{ki}$ schommelt rond 0.

Bij de berekening van de effectkwadratensom $SS_{A|X}$ worden de voorspellingen op basis van het covariantieanalysemodel vergeleken met de voorspellingen op basis van het referentiemodel met alleen de covariaat X. Dit referentiemodel is het enkelvoudige regressieanalysemodel. De enkelvoudige regressieanalyse wordt uitgebreid behandeld in het Hoofdstuk *Regressieanalyse*. We gaan er hier slechts kort op in. Het enkelvoudige regressieanalysemodel volgt uit het covariantieanalysemodel als de nulhypothese waar is. Nu, de nulhypothese houdt in dat de intercepten van de groepsspecifieke regressielijnen gelijk zijn. Als deze intercepten gelijk zijn, dan is er slechts één enkele regressielijn waarmee we de data van alle $K$ groepen samen beschrijven.

De voorspellingen op basis van deze enkele regressielijn worden gemaakt op basis van het kleinste-kwadratencriterium. We duiden deze voorspellingen aan met $\widehat{y}^X$. Hierin geeft het superscript $X$ aan dat de voorspelling gemaakt wordt op basis van het enkelvoudige regressieanalysemodel dat alleen de variabele X bevat. Dit model heeft de functie van referentiemodel ($\widehat{y}^{Ref} = \widehat{y}^X$). De voorspelling $\widehat{y}^X$ wordt voorgesteld in Figuur 3.10; het is de lijn die de twee parallelle regressielijnen snijdt. De regressielijn $\widehat{y}^X$ loopt altijd zó dat hij zoveel mogelijk midden door beide puntenwolken loopt.

De voorspellingen op basis van het enkelvoudige regressieanalysemodel ($\widehat{y}^X$) worden vergeleken met de voorspellingen op basis van het covariantieanalysemodel. Deze laatste voorspellingen worden ook gemaakt worden op basis van het kleinste-kwadratencriterium en ze worden aangeduid met $\widehat{y}^{X,A}$. Hierin geeft het superscript $X, A$ aan dat de voorspellingen gemaakt worden op basis van een model dat ook het te toetsen effect bevat ($\widehat{y}^{Ref+Effect} = \widehat{y}^{X,A}$). Om het onderscheid te maken tussen de verschillende groepen, schrijven we in het vervolg $\widehat{y}_k^{X,A}$ (met het subscript $k$) in plaats van $\widehat{y}^{X,A}$. De voorspellingen $\widehat{y}_k^{X,A}$ voor het voorbeeldonderzoek worden voorgesteld in Figuur 3.10. De formule voor deze voorspellingen werd reeds eerder gegeven, namelijk bij de definitie van het uitgezuiverde verschil op pagina 258:

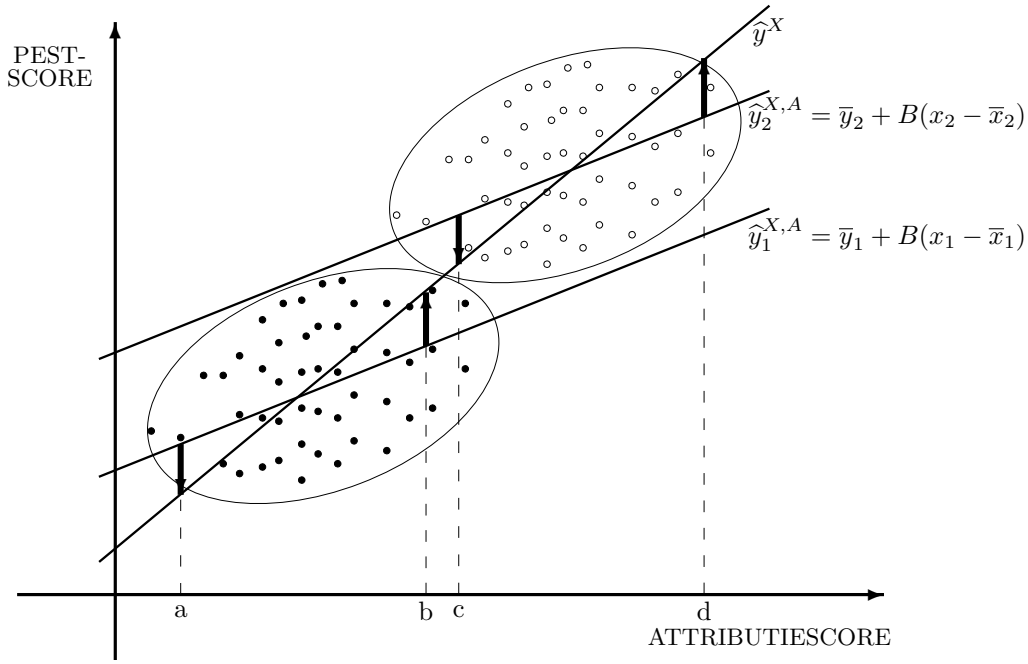
$$\widehat{y}_k^{X,A} = \overline{y}_k + B(x_k - \overline{x}_k) \quad .$$

Het belangrijkste gegeven in Figuur 3.10 is dat de regressielijn van het enkelvoudige regressieanalysemodel ($\widehat{y}^X$) de regressielijnen van het covariantieanalysemodel ($\widehat{y}_k^{X,A}$) snijdt. Naarmate het verticale verschil tussen de twee parallelle regressielijnen ($\widehat{y}_1^{X,A}$ en $\widehat{y}_2^{X,A}$) (het uitgezuiverde verschil) groter wordt, vergroot ook de hoek waaronder $\widehat{y}^X$ de twee parallelle regressielijnen snijdt. Dit komt omdat de regressielijn $\widehat{y}^X$ altijd zoveel mogelijk midden door beide puntenwolken loopt. In het extreme geval waarin het uitgezuiverde verschil gelijk is aan 0 (de twee parallelle regressielijnen liggen over elkaar), is deze snijdingshoek gelijk aan 0. In dit geval, liggen alledrie de regressielijnen ($\widehat{y}^X$, $\widehat{y}_1^{X,A}$, en $\widehat{y}_2^{X,A}$) over elkaar.

We geven nu de formule van de effectkwadratensom $SS_{A|X}$. We krijgen deze formule als we in de algemene formule van de effectkwadratensom $\widehat{y}_{ki}^{Ref+Effect}$ en $\widehat{y}_{ki}^{Ref}$ vervangen door, respectievelijk, $\widehat{y}_{ki}^{X,A}$ en $\widehat{y}_{ki}^{X}$:

$$\begin{aligned} SS_{A|X} &= \sum_{Obs} (\widehat{y}_{ki}^{X,A} - \widehat{y}_{ki}^{X})^2 \\ &= \sum_{k=1}^{K} \sum_{i=1}^{n_k} (\widehat{y}_{ki}^{X,A} - \widehat{y}_{ki}^{X})^2 \quad . \end{aligned}$$

Men krijgt de voorspellingen $\widehat{y}_{ki}^{X,A}$ en $\widehat{y}_{ki}^{X}$ door $x_{ki}$ (de covariaatscore van de i-de proefpersoon in de k-de groep) in te vullen voor $x_k$ in de formules voor de regressielijnen $\widehat{y}_k^{X,A}$ en $\widehat{y}^X$. (De formule voor $\widehat{y}^X$ wordt in dit hoofdstuk niet gegeven, en dat is ook niet nodig. Deze formule wordt wel gegeven in het Hoofdstuk *Regressieanalyse*.) In Figuur 3.10 wordt voor twee covariaatscores in de eerste groep ($a$ en $b$) en twee in de tweede groep ($c$ en $d$) met een dikke pijl aangegeven wat het verschil ($\widehat{y}_{ki}^{X,A} - \widehat{y}_{ki}^{X}$) is. In deze figuur is goed te zien dat $SS_{A|X}$ groter wordt naarmate de verticale afstand tussen de twee parallelle regressielijnen (i.e., het uitgezuiverde verschil) groter wordt.




Figuur 3.10: De drie regressielijnen die betrokken zijn bij de bepaling van de kwadratensom van de factor uitgezuiverd voor de covariaat.

### De Kwadratensom van de Factor Zonder Uitzuivering voor de Covariaat

Het is nuttig om $SS_{A|X}$ te vergelijken met $SS_A$, de kwadratensom van A zonder uitzuivering voor X. Men krijgt de formule van deze effectkwadratensom als men het model zonder effecten als referentiemodel kiest. Deze effectkwadratensom is reeds aan bod gekomen bij de variantieanalyse met één factor. Hij wordt berekend door de volgende twee voorspellingen met elkaar te vergelijken in de formule van de effectkwadratensom:

1. De voorspelling op basis van het referentiemodel ($\widehat{y}_{ki}^{Ref}$), het model zonder effecten. Deze voorspelling is gelijk aan $\overline{y}_{\sqcup}$, het gemiddelde van alle $N$ observaties.

2. De voorspelling op basis van het model met het effect van factor A ($\widehat{y}_{ki}^{Ref+Effect} = \widehat{y}_{ki}^{A}$), het gewone variantieanalysemodel met factor A. Deze voorspelling is gelijk aan $\overline{y}_k$, het gemiddelde van de observaties in het k-de niveau van A.

De formule van $SS_A$ is dan de volgende:

$$\begin{aligned} SS_A &= \sum_{Obs} (\overline{y}_k - \overline{y}_{\sqcup})^2 \\ &= \sum_{k=1}^{K} n_k (\overline{y}_k - \overline{y}_{\sqcup})^2 \quad . \end{aligned}$$

We kennen deze effectkwadratensom van de variantieanalyse met één factor.

Met $SS_A$ kan men de nulhypothese toetsen dat de $K$ groepsspecifieke verwachte waarden $\mu_k$ gelijk zijn. Deze verwachte waarden zijn punten op de Y-as, zoals men kan zien in Figuur 3.9.

Het verschil tussen de nulhypothese die men toetst met $SS_A$ en $SS_{A|X}$ kan men het best beschrijven aan de hand van Figuur 3.6 op pagina 255. Met $SS_A$ toetst men of het verschil tussen de gemiddelden $\overline{y}_1$ en $\overline{y}_2$ significant is. Met andere woorden, met $SS_A$ toetst men of $\mu_1$ gelijk is aan $\mu_2$. En met $SS_{A|X}$ toetst men of het verticale verschil tussen de twee regressielijnen (i.e., $\overline{y}_2 - \overline{y}_1 - B(\overline{x}_2 - \overline{x}_1)$) significant is. Met andere woorden, met $SS_{A|X}$ toetst men of het uitgezuiverde populatieverschil $[\mu_2 - \mu_1 - B^{pop}(\overline{x}_2 - \overline{x}_1)]$ gelijk is aan 0.

### De Kwadratensom van de Covariaat (Met en Zonder Uitzuivering voor het Effect van de Factor)

In de uitdraai van computerprogramma's zoals GLM-Univariate vindt men ook vaak $SS_{X|A}$ of $SS_X$, de kwadratensom van de covariaat met en zonder uitzuivering. Men krijgt de formule van $SS_{X|A}$ als men het model met alleen het effect van A als referentiemodel kiest en de daaruit volgende voorspelling ($\widehat{y}_{ki}^{Ref} = \widehat{y}_{ki}^{A} = \overline{y}_k$) vergelijkt met de voorspelling op basis van het covariantieanalysemodel ($\widehat{y}_{ki}^{Ref+Effect} = \widehat{y}_{ki}^{X,A}$). Met deze kwadratensom kan men de nulhypothese toetsen dat het regressiegewicht bínnen de groepen gelijk is aan nul. Dit betekent dat er bínnen de groepen geen verband is tussen de covariaat en de afhankelijke variabele.

Men krijgt de formule van $SS_X$ als men het model zonder effecten als referentiemodel kiest en de daaruit volgende voorspelling ($\widehat{y}_{ki}^{Ref} = \overline{y}_{\sqcup}$) vergelijkt met de voorspelling op

basis van het enkelvoudige regressieanalysemodel met predictor X ($\widehat{y}_{ki}^{Ref+Effect} = \widehat{y}_{ki}^{X}$). Met deze kwadratensom kan men de nulhypothese toetsen dat het regressiegewicht óver de groepen gelijk is aan nul. Dit betekent dat er óver de groepen geen verband is tussen de covariaat en de afhankelijke variabele.

### 3.4.6 De Errorkwadratensom en de Kwadratensom-ontbinding

De errorkwadratensom verschijnt in de noemer van de F-statistieken waarmee we de hierboven genoemde nulhypotheses kunnen toetsen. Deze errorkwadratensom is $SS_{Error|X,A}$, de errorkwadratensom na voorspelling op basis van het covariantieanalysemodel. De voorspellingen op basis van het covariantieanalysemodel werden eerder aangeduid met $\widehat{y}_{ki}^{X,A}$.

$$SS_{Error|X,A} = \sum_{k=1}^{K}\sum_{i=1}^{n_k}(y_{ki} - \widehat{y}_{ki}^{X,A})^2 \quad .$$

De errorkwadratensom $SS_{Error|X,A}$ is kleiner dan $SS_{Error|A}$, de errorkwadratensom van de variantieanalyse met één factor. Uit de volgende kwadratensom-ontbinding blijkt hoeveel kleiner:

$$SS_{Error|A} = SS_{X|A} + SS_{Error|X,A} \quad .$$

$SS_{Error|A}$ is groter dan $SS_{Error|X,A}$ omdat er bij de voorspelling geen rekening is gehouden met het effect van de covariaat. Het verschil tussen de twee errorkwadratensommen is precies gelijk aan $SS_{X|A}$. Dus, naarmate de covariaat meer variantie verklaart bovenop factor A, is $SS_{Error|A}$ groter dan $SS_{Error|X,A}$.

De kwadratensom van X uitgezuiverd voor A, $SS_{X|A}$, behoort tot de modelkwadratensom. Men kan de modelkwadratensom op de volgende twee manieren uitrekenen:

$$\begin{aligned} \text{Modelkwadratensom} &= SS_{X|A} + SS_A \\ \text{Modelkwadratensom} &= SS_{A|X} + SS_X \quad . \end{aligned}$$

### 3.4.7 Nogmaals De F-Statistieken

Wanneer de kwadratensommen bekend zijn, is de berekening van de F-statistieken precies hetzelfde als in de variantieanalyse. De teller van een F-statistiek, de gemiddelde effectkwadratensom, bepaalt welke nulhypothese getoetst wordt. In wat volgt, komen de vier effectkwadratensommen aan bod die eerder in deze paragraaf geïntroduceerd werden.

De noemer van alle F-statistieken van de covariantieanalyse is $MS_{Error|X,A}$, de gemiddelde errorkwadratensom. Deze gemiddelde errorkwadratensom is gelijk aan $SS_{Error|X,A}$ gedeeld door zijn aantal vrijheidsgraden. Dit aantal is gelijk aan $N - K - 1$, het aantal observaties min het aantal geschatte parameters ($K$ $\mu_k$'s en één $B^{pop}$). Dus,

$$\boxed{MS_{Error|X,A} = \frac{SS_{Error|X,A}}{N - K - 1} \quad .}$$

In de praktijk is $MS_{Error|X,A}$ meestal kleiner dan $MS_{Error|A}$. Men krijgt dus meestal een gevoeliger toets als men bij de voorspelling ook rekening houdt met het effect van

de covariaat. Dit komt omdat $SS_{X|A}$, de voor A uitgezuiverde kwadratensom van X, meestal voldoende groot is om het verschil tussen de vrijheidsgraden (respectievelijk, $N-K-1$ en $N-K$) te compenseren.

### F-toetsen van het Uitgezuiverde en het Niet-Uitgezuiverde Effect van de Factor

Voor de toets van de nulhypothese dat, in de populatie, de intercepten van de $K$ groepsspecifieke regressielijnen gelijk zijn, gebruikt men de F-statistiek met als teller $MS_{A|X}$, de gemiddelde kwadratensom van A uitgezuiverd voor X. Het aantal vrijheidsgraden van $SS_{A|X}$ is $K-1$. Dit aantal is gelijk aan het aantal parameters in het covariantieanalysemodel $(K+1)$ min het aantal parameters in het enkelvoudige regressieanalysemodel (2). Dus,

$$\boxed{MS_{A|X} = \frac{SS_{A|X}}{K-1} \quad .}$$

Onder de nulhypothese dat, in de populatie, de intercepten van de $K$ groepsspecifieke regressielijnen gelijk zijn, heeft deze F-statistiek een $\mathcal{F}$-verdeling met $K-1$ en $N-K-1$ vrijheidsgraden. Als deze F-statistiek significant is, dan besluit men dat, in de populatie, sommige intercepten van de $K$ groepsspecifieke regressielijnen verschillend zijn, en dus dat sommige uitgezuiverde populatieverschillen afwijken van 0.

Voor de toets van de nulhypothese dat de $K$ groepsspecifieke verwachte waarden $\mu_k$ gelijk zijn aan elkaar, gebruikt men de F-statistiek met als teller $MS_A$, de gemiddelde kwadratensom van A zonder uitzuivering voor X. Onder de nulhypothese dat alle $\mu_k$'s gelijk zijn aan elkaar, heeft deze F-statistiek een $\mathcal{F}$-verdeling met $K-1$ tellervrijheidsgraden en $N-K-1$ noemervrijheidsgraden. Als deze F-statistiek significant is, dan besluit men dat, in de populatie, sommige verwachte waarden $\mu_k$ van elkaar verschillen.

Men kan deze nulhypothese ook toetsen met een F-statistiek met als noemer $MS_{Error|A}$ in plaats van $MS_{Error|X,A}$. Het voordeel van de F-statistiek met $MS_{Error|X,A}$ in de noemer is dat deze gemiddelde errorkwadratensom meestal kleiner is dan $MS_{Error|A}$, en daarom resulteert in een meer gevoelige toets.

Het is mogelijk dat de F-statistiek met $MS_A$ (zonder uitzuivering) significant is, maar niet de F-statistiek met $MS_{A|X}$ (met uitzuivering). In Figuur 3.7(b) wordt een situatie getoond die hiermee in overeenstemming is. Ook het omgekeerde is mogelijk: de F-statistiek met $MS_{A|X}$ is significant, maar niet die met $MS_A$. Deze situatie wordt behandeld in een van de opgaven. Tenslotte, is het natuurlijk ook mogelijk dat beide significant zijn of geen van beide.

### F-toetsen van het Uitgezuiverde en het Niet-Uitgezuiverde Effect van de Covariaat

We beschrijven nu twee toetsen die betrekking hebben op het regressiegewicht van de covariaat X. Beide toetsen worden vermeld in de uitdraai van de meeste computerprogramma's, maar men is er meestal niet in geïnteresseerd. De eerste is een toets van de nulhypothese dat er bínnen de groepen geen verband is tussen de covariaat en de

afhankelijke variabele. Deze nulhypothese komt erop neer dat het regressiegewicht $B^{pop}$ in het covariantieanalysemodel gelijk is aan 0. Deze nulhypothese wordt getoetst met de F-statistiek met $MS_{X|A}$ in de teller. Het aantal vrijheidsgraden van $SS_{X|A}$ is 1 (want het covariantieanalysemodel heeft 1 parameter meer dan het variantieanalysemodel) en daarom is $MS_{X|A}$ gelijk aan $SS_{X|A}$. Onder de nulhypothese dat, in de populatie, het regressiegewicht binnen de groepen (i.e., $B^{pop}$ in het covariantieanalysemodel) gelijk is aan 0, heeft deze F-statistiek een $\mathcal{F}$-verdeling met 1 tellervrijheidsgraad en $N - K - 1$ noemervrijheidsgraden.

De tweede toets is een toets van de nulhypothese dat er óver de groepen geen verband is tussen de covariaat en de afhankelijke variabele. Deze nulhypothese kan ook als volgt geformuleerd worden: in de populatie, is het regressiegewicht van de enkelvoudige regressielijn óver de groepen gelijk aan 0. Deze nulhypothese wordt getoetst met de F-statistiek met $MS_X$ in de teller. Het aantal vrijheidsgraden van $SS_X$ is 1 (want het enkelvoudige regressieanalysemodel heeft 1 parameter meer dan het model zonder effecten) en daarom is $MS_X$ gelijk aan $SS_X$. Onder de nulhypothese dat, in de populatie, het regressiegewicht over de groepen gelijk is aan 0, heeft deze F-statistiek een $\mathcal{F}$-verdeling met 1 tellervrijheidsgraad en $N - K - 1$ noemervrijheidsgraden.

### 3.4.8 Overzichtstabellen

We geven nu vier overzichtstabellen met de belangrijkste formules van deze paragraaf.

1. In Tabel 3.62 staan de vier modellen die gebruikt worden voor de berekening van de effect- en de errorkwadratensommen van een covariantieanalyse.

2. In Tabel 3.63 staan de vier F-toetsen die men kan uitvoeren binnen een covariantieanalyse. Voor verreweg de meeste situaties in de praktijk, zijn alleen de F-toetsen van het uitgezuiverde en het niet-uitgezuiverde effect van factor A van belang.

3. In Tabel 3.64 staan de effectkwadratensommen en hun vrijheidsgraden.

4. In Tabel 3.65 staan de errorkwadratensommen, hun vrijheidsgraden en hun kwadratensomontbindingen.

| Modellen gebruikt binnen de Covariantieanalyse | | |
|---|---|---|
| | Effecten | Formule |
| (1) | Geen | $y_{kli} = \mu + e_{ki}$ |
| (2) | A | $y_{kli} = \mu_k + e_{ki}$ |
| (3) | X | zie Hoofdstuk *Regressieanalyse* |
| (4) | X en A | $y_{kli} = \mu_k + B^{pop}(x_{ki} - \overline{x}_k) + e_{ki}$ |

Tabel 3.62: De vier modellen die gebruikt worden binnen een covariantieanalyse met één factor en één covariaat.

| F-toetsen binnen de Covariantieanalyse | | | | |
|---|---|---|---|---|
| $F = \frac{MS_{Effect\|Ref}}{MS_{Error\|Model}}$ | | | | |
| | | | $MS_{Error\|Model}$ | |
| Effect | Nulhypothese | $MS_{Effect\|Ref}$ | Meestal klein | Meestal groot |
| A | $\mu_1 = \ldots = \mu_K$ | $MS_A$ | $MS_{Error\|X,A}$ | $MS_{Error\|A}$ |
| | $\mu_1 - B^{pop}\overline{x}_1 = \ldots = \mu_K - B^{pop}\overline{x}_K$ | $MS_{A\|X}$ | $MS_{Error\|X,A}$ | |
| X | Over-groepen-regressiegewicht = 0 | $MS_X$ | $MS_{Error\|X,A}$ | $MS_{Error\|X}$ |
| | $B^{pop} = 0$ | $MS_{X\|A}$ | $MS_{Error\|X,A}$ | |

Tabel 3.63: F-toetsen binnen de covariantieanalyse met één factor en één covariaat.

| Effectkwadratensommen binnen de Covariantieanalyse | | | | |
|---|---|---|---|---|
| $SS_{Effect\|Ref} = \sum_{Obs}(\widehat{y}_{kli}^{Ref+Effect} - \widehat{y}_{kli}^{Ref})^2$ | | | | |
| Effect | Model-vergelijking | Symbool | Vrijheids-graden | Hoe berekenen? |
| A | (2) met (1) | $SS_A$ | $K-1$ | $\sum_{k=1}^{K} n_k(\overline{y}_k - \overline{y}_{\sqcup})^2$ |
| | (4) met (3) | $SS_{A\|X}$ | $K-1$ | computer |
| X | (3) met (1) | $SS_X$ | 1 | computer |
| | (4) met (2) | $SS_{X\|A}$ | 1 | computer |

Tabel 3.64: Berekening van de effectkwadratensommen van een covariantieanalyse met één factor en één covariaat.

| Errorkwadratensommen binnen de Covariantieanalyse | | | | | |
|---|---|---|---|---|---|
| $SS_{Error\|Model} = \sum_{Obs}(y_{kli} - \widehat{y}_{kli}^{Model})^2$ | | | | | |
| Model | Effecten | Symbool | Vrijheids-graden | Ontbinding | Bereken $\widehat{y}_{kli}^{Model}$ |
| (1) | Geen | $SS_{Totaal}$ | $N-1$ | $SS_A + SS_{Error\|A}$<br>$SS_X + SS_{Error\|X}$ | $\overline{y}_{\sqcup}$ |
| (2) | A | $SS_{Error\|A}$ | $N-K$ | $SS_{X\|A} + SS_{Error\|X,A}$ | computer |
| (3) | X | $SS_{Error\|X}$ | $N-2$ | $SS_{A\|X} + SS_{Error\|X,A}$ | computer |
| (4) | X, A | $SS_{Error\|X,A}$ | $N-K-1$ | | computer |

Tabel 3.65: Berekening van de errorkwadratensommen op basis van één factor en één covariaat.

### 3.4.9 Uitvoeren van een Covariantieanalyse met de SPSS-procedure GLM-Univariate

Als het voor een bepaalde analyse nodig is om uit te zuiveren voor een covariaat, dan moet men die covariaat onder *Covariate(s)* in het dialoogvenster van GLM-Univariate[19] invullen. Met de knop *Options* kan men vragen om de uitgezuiverde gemiddelden. SPSS noemt deze uitgezuiverde gemiddelden *Estimated Marginal Means*. De factoren en factorcombinaties waarvoor men de uitgezuiverde gemiddelden wil, moet men aangeven in het veld *Factor(s) and Factor Interactions*. Als men een covariaat heeft ingevuld in het veld onder *Covariate(s)*, dan zijn de *Estimated Marginal Means* uitgezuiverd voor deze covariaat.

Met de knop *Model* kan men een keuze maken uit de verschillende types effectkwadratensommen. We laten hier de type IV kwadratensom buiten beschouwing. Voor de andere drie types kwadratensommen, beschrijven we de implicaties van de keuze voor (1) een design met één factor en één covariaat en (2) een design met twee factoren en één covariaat. Hierbij gaan we ervan uit dat voor de optie *Full factorial* gekozen is. De implicaties voor een design met één factor en één covariaat zijn de volgende:

1. Met de type I optie worden hiërarchische kwadratensommen berekend. In een design met een covariaat komt de covariaat altijd als eerste in de volgorde. Dit betekent dat de kwadratensom van de covariaat niet uitgezuiverd wordt voor de factor, maar de kwadratensom van de factor wel voor de covariaat. Dus, met deze optie berekent men $SS_X$ en $SS_{A|X}$.

2. Met de type II en type III opties worden $SS_{A|X}$ en $SS_{X|A}$ berekend. Er is geen verschil tussen deze twee opties omdat er maar één factor in het design zit. Er is dus geen interactie waarvoor uitgezuiverd kan worden.

We beschouwen nu een design met twee factoren en één covariaat. We gaan ervan uit dat de factoren A en B in deze volgorde zijn opgenomen in het vakje *Fixed Factors*. De implicaties zijn de volgende:

1. Met de type I optie berekent men $SS_X$, $SS_{A|X}$, $SS_{B|X,A}$ en $SS_{A\times B|X,A,B}$.

2. Met de type II optie berekent men $SS_{X|A,B}$, $SS_{A|X,B}$, $SS_{B|X,A}$ en $SS_{A\times B|X,A,B}$.

3. Met de type III optie berekent men $SS_{X|[AB]}$, $SS_{A|X,B,A\times B}$, $SS_{B|X,A,A\times B}$ en $SS_{A\times B|X,A,B}$.

De gegevens die nodig zijn om de F-toetsen uit te voeren worden meestal gegeven in een ANOVA-tabel. In Tabel 3.66 staat een deel van de ANOVA-tabel die door GLM-Univariate geproduceerd wordt voor de gegevens uit het voorbeeldonderzoek. Het is de ANOVA-tabel die men krijgt met de type I optie. Voor de covariaat ATTRIBUTIESCORE, wordt de niet-uitgezuiverde kwadratensom $SS_X$ gegeven en voor de factor TRAINING, de uitgezuiverde kwadratensom $SS_{A|X}$. Uit de F-statistieken en hun bijbehorende p-waarden kan men besluiten dat er, in de populatie, (1) een verband is tussen de covariaat en de afhankelijke variabele óver de twee groepen (met en zonder training) en (2) een verschil tussen de uitgezuiverde gemiddelden voor de twee groepen.

[19] Deze paragraaf is een vervolg op een eerdere paragraaf over het uitvoeren van een variantieanalyse met GLM-Univariate.

Dependent Variable: PESTSCORE

| Source | Type I Sum of Squares | df | Mean Squares | F | Sig. |
|---|---|---|---|---|---|
| Corrected Model | 203.408 | 2 | 101.704 | 11.024 | .000 |
| ATTRIBUTIESCORE | 156.327 | 1 | 156.327 | 16.944 | .000 |
| TRAINING | 47.081 | 1 | 47.081 | 5.103 | .024 |
| Error | 802.648 | 87 | 9.226 | | |
| Corrected Total | 1006.055 | 89 | | | |

Tabel 3.66: Deel van de ANOVA-tabel voor de data uit het voorbeeldonderzoek bij de covariantieanalyse, geproduceerd door GLM-Univariate met de type I optie.

De ANOVA-tabel die men krijgt met de type III optie verschilt slechts in één regel van Tabel 3.66: in plaats van de niet-uitgezuiverde kwadratensom $SS_X$, krijgt men nu de uitgezuiverde kwadratensom $SS_{X|A}$, die gelijk is aan 90.908. De bijbehorende F-statistiek is significant, en daarom besluiten we dat er, in de populatie, een verband is tussen de covariaat en de afhankelijke variabele bínnen de twee groepen.

### 3.4.10 Twee Redenen om Uit te Zuiveren voor een Covariaat

De belangrijkste reden om uit te zuiveren voor een covariaat, is dat men denkt dat de covariaat een storende variabele is. Een tweede reden om uit te zuiveren voor een covariaat is dat dit ervoor kan zorgen dat de F-toetsen gevoeliger worden. Dit wil zeggen dat de F-toets kleinere verschillen tussen de groepen als significant kan detecteren. We behandelen nu beide redenen meer in detail.

Het grote probleem bij niet-experimenteel onderzoek is dat de groepen van elkaar kunnen verschillen op allerlei storende variabelen, en voor zover die storende variabelen als covariaten geobserveerd kunnen worden, is het belangrijk om ervoor uit te zuiveren. Geformuleerd in termen van de nulhypotheses, omwille van de invloed van de storende variabelen, is men niet geïnteresseerd in de hypothese dat de verschillen tussen de groepsspecifieke verwachte waarden $\mu_k$ gelijk zijn aan 0. Daarentegen, men is wel geïnteresseerd in de hypothese dat de *uitgezuiverde* verschillen $[\mu_k - \mu_m - B^{pop}(\overline{x}_k - \overline{x}_m)]$ gelijk zijn aan 0. (In de formule voor de uitgezuiverde verschillen is $m$ een groepsindex, net zoals $k$: $m = 1, \ldots, K$. We gebruiken deze tweede groepsindex om aan te duiden dat het hier gaat over uitgezuiverde verschillen tussen een willekeurig paar van groepen.).

Vanuit dit gezichtspunt is niet elke covariaat ook een goede covariaat. Een goede covariaat is een storende variabele. Een storende variabele hangt zowel samen met de onafhankelijke als met de afhankelijke variabele. In termen van de ingrediënten van de covariantieanalyse kan men deze dubbele voorwaarde als volgt formuleren:

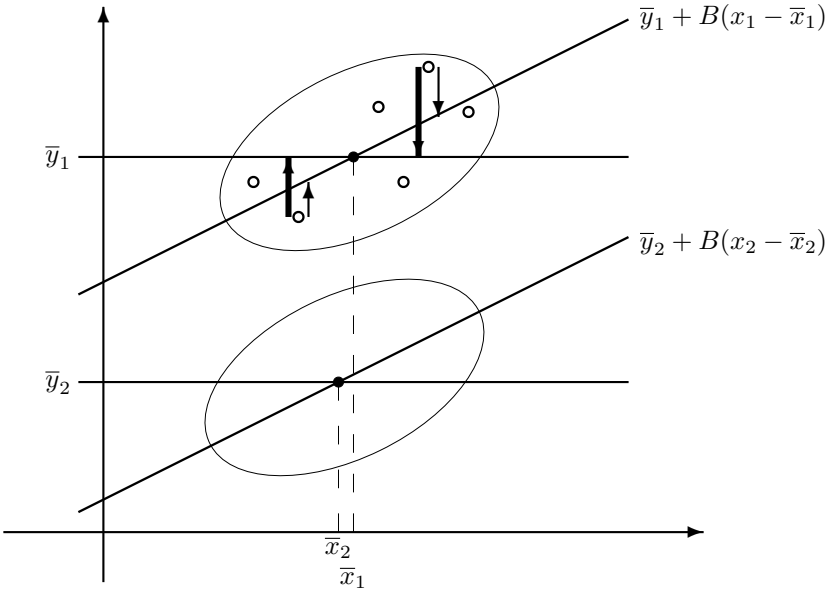
1. De groepen moeten van elkaar verschillen wat betreft hun gemiddelde score op de covariaat. Met andere woorden, de $\overline{x}_k$'s moeten van elkaar verschillen. Als dit niet het geval is, dan zijn de uitgezuiverde verschillen $[\mu_k - \mu_m - B^{pop}(\overline{x}_k - \overline{x}_m)]$ gelijk aan de niet-uitgezuiverde verschillen $\mu_k - \mu_m$, en dan toetst men binnen een covariantieanalyse precies dezelfde nulhypothese als binnen een variantieanalyse.

2. Binnen elk van de groepen moet er een verband zijn tussen de covariaat en de afhankelijke variabele. Met andere woorden, het binnen-groepen regressiegewicht $B^{pop}$ mag niet gelijk zijn aan 0. Als $B^{pop}$ wel gelijk is aan 0, dan zijn de uitgezuiverde verschillen $[\mu_k - \mu_m - B^{pop}(\overline{x}_k - \overline{x}_m)]$ opnieuw gelijk aan de niet-uitgezuiverde verschillen $\mu_k - \mu_m$, en dan toetst men binnen een covariantieanalyse precies dezelfde nulhypothese als binnen een variantieanalyse.

De tweede reden om uit te zuiveren voor een covariaat, is dat dit ervoor kan zorgen dat de F-toets van het effect van factor A gevoeliger wordt. Dit is de reden waarom men ook in een experiment gebruik maakt van covariantieanalyse, ook al kunnen storende variabelen hier in principe uitgesloten worden. Tussen haakjes, het gevoeliger worden van de F-toets doet zich ook voor in een observationeel onderzoek, maar daar is het niet de hoofdreden om uit te zuiveren. Voordat we ingaan op de gevoeligheid van de F-toets, herhalen we kort waarom men met een covariantieanalyse in een experiment dezelfde nulhypothese toetst als met een variantieanalyse. Hierbij gaan we ervan uit dat het een goed experiment is, in de zin dat storende persoonsvariabelen werden uitgesloten door randomisering en storende situatievariabelen door experimentele controle. Dus, in een goed experiment is er, buiten de variabele waarvan men het effect wil onderzoeken, geen variabele waarop de groepen systematisch van elkaar verschillen. Dus, er is ook geen covariaat waarop de groepsgemiddelden $\overline{x}_k$ systematisch van elkaar verschillen. Dit betekent dat de uitgezuiverde verschillen $[\mu_k - \mu_m - B^{pop}(\overline{x}_k - \overline{x}_m)]$, op toevallige schommelingen na, gelijk zijn aan de niet-uitgezuiverde verschillen $\mu_k - \mu_m$. Dus, in een goed experiment is, op toevallige schommelingen na, de nulhypothese die men toetst met een covariantieanalyse dezelfde als de nulhypothese die men toetst met een variantieanalyse.

In de covariantieanalyse wordt het effect van factor A meestal gevoeliger getoetst dan in de variantieanalyse. Dit komt omdat $SS_{Error|A}$, de errorkwadratensom van de variantieanalyse, meestal veel groter is dan $SS_{Error|X,A}$, de errorkwadratensom van de covariantieanalyse. Dit blijkt uit de kwadratensom-ontbinding, die uitdrukt dat, naarmate de covariaat X meer variantie verklaart bovenop factor A (i.e., naarmate $SS_{X|A}$ groter wordt), $SS_{Error|A}$ toeneemt in vergelijking met $SS_{Error|X,A}$. Dit wordt grafisch voorgesteld in Figuur 3.11. In deze figuur worden zes observaties uit de eerste groep aangeduid met kleine cirkeltjes. Voor twee van deze zes observaties wordt hun bijdrage aan $SS_{Error|A}$ en $SS_{Error|X,A}$ aangeduid met, respectievelijk, een dikke en een dunne pijl. De errorkwadratensom van de variantieanalyse ($SS_{Error|A}$) wordt bepaald door de afwijkingen van de punten in de ellipsen (de observaties $y_{ki}$) ten opzichte van de horizontale lijnen met hoogte $\overline{y}_k$. Met name, $SS_{Error|A}$ is gelijk aan de som over alle observaties van de gekwadrateerde verschillen $(y_{ki} - \overline{y}_k)^2$. En de errorkwadratensom van de covariantieanalyse ($SS_{Error|X,A}$) wordt bepaald door de afwijkingen van de punten in de ellipsen ten opzichte van de schuine regressielijnen $\overline{y}_k + B(x_k - \overline{x}_k)$. Met name, $SS_{Error|X,A}$ is gelijk aan de som over alle observaties van de gekwadrateerde verschillen $[y_{ki} - (\overline{y}_k + B(x_{ki} - \overline{x}_k))]^2$. Als er een verband bestaat tussen de afhankelijke variabele en de covariaat (dit wil zeggen, als de ellipsen schuin liggen en $B$ verschilt van 0), dan zijn de afwijkingen van de punten in de ellipsen ten opzichte van de horizontale lijnen gemiddeld genomen groter dan de afwijkingen ten opzichte van de schuine regressielijnen. Hieruit volgt dat $SS_{Error|A}$ groter is dan $SS_{Error|X,A}$. Meestal zal ook de *gemiddelde* errorkwadratensom $MS_{Error|A}$ groter zijn dan $MS_{Error|X,A}$. (Of dit ook

effectief het geval is, is mede afhankelijk van het aantal vrijheidsgraden van deze twee errorkwadratensommen.) Omdat $MS_{Error|A}$ meestal groter is dan $MS_{Error|X,A}$, zal de F-toets uit de covariantieanalyse ook meestal eerder significant zijn dan de F-toets uit de variantieanalyse.




Figuur 3.11: Grafische voorstelling van de errorkwadratensommen van de variantie- en de covariantieanalyse: afwijkingen ten opzichte van de horizontale lijnen voor de variantieanalyse, aangeduid met een dikke pijl, en afwijkingen ten opzichte van de schuine regressielijnen voor de covariantieanalyse, aangeduid met een dunne pijl.

### 3.4.11 Wanneer Stel je de Covariaatscores Vast?

In het bovenstaande werden twee noodzakelijke voorwaarden gegeven opdat men met de covariantieanalyse een andere nulhypothese toetst dan met de variantieanalyse: (1) de groepen moeten van elkaar verschillen wat betreft hun gemiddelde score op de covariaat, en (2) binnen elk van de groepen moet er een verband zijn tussen de covariaat en de afhankelijke variabele. Naast deze twee voorwaarden die ervoor zorgen dát men uitzuivert, is er nog een derde voorwaarde die ervoor zorgt dat men niet teveel uitzuivert, en met name dat men niet ook een deel van het effect van de condities uitzuivert. Deze derde voorwaarde is dat men de scores op de covariaat moet vaststellen voordat de condities een effect kunnen hebben.

Om het belang van deze voorwaarde te zien, stel je een onderzoek voor naar het effect van ontspanningsoefeningen op hoge bloeddruk bij patiënten met een eerste hartinfarct. De afhankelijke variabele in dit denkbeeldige onderzoek was de bloeddruk zoals gemeten acht weken na het infarct. De twee groepen (met en zonder ontspanningsoefeningen)

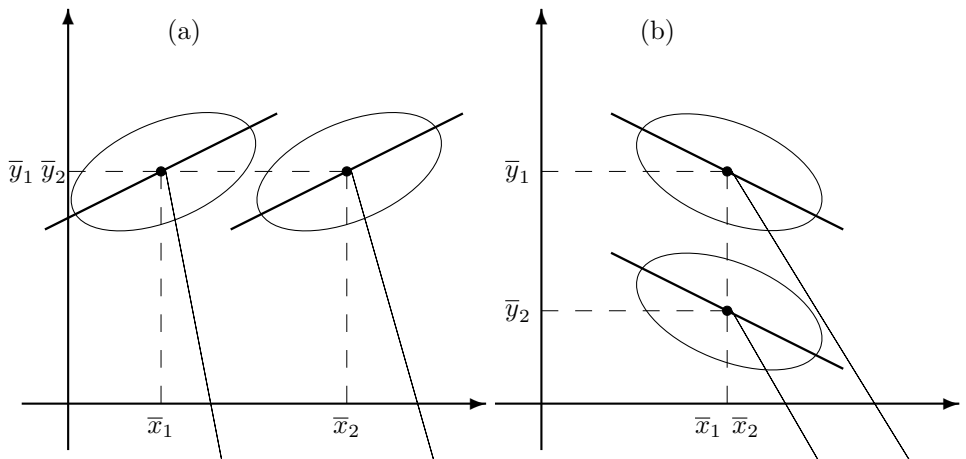
werden niet op zuiver toevallige wijze ingedeeld. De onderzoeker maakte zich daarom zorgen over de mogelijkheid dat de twee groepen al van elkaar verschilden qua gemiddelde bloeddruk voordat ze met de ontspanningsoefeningen waren begonnen. Daarom nam hij als covariaat de bloeddruk zoals gemeten drie weken na het infarct. Nu is het van groot belang of de ontspanningsoefeningen op dat moment reeds begonnen waren of niet. Als ze reeds begonnen waren, en ze hadden op dat moment ook al een effect, dan zou men met het uitzuiveren voor deze covariaat ook (een deel van) het effect van de ontspanningsoefeningen uitzuiveren. Een dergelijke incorrecte toepassing van de covariantieanalyse is uitgesloten als men de scores op de covariaat vaststelt voordat de condities een effect kunnen hebben.

### 3.4.12 Een Relativering van de Kracht van Statistische Uitzuivering

Voor zover men de covariantieanalyse gebruikt om uit te zuiveren voor storende variabelen, past hier dezelfde opmerking als bij de variantieanalyse voor ongebalanceerde designs: statistische uitzuivering is beperkt tot die variabelen (covariaten) waarover men beschikt, en het is best mogelijk dat er, naast deze beschikbare covariaten, nog verschillende andere (maar niet-beschikbare) covariaten zijn waarop de groepen van elkaar verschillen.

---

*Opgave 30*
In Figuur 3.12 worden twee mogelijke resultaten van een observationeel onderzoek gegeven. De ellipsen staan voor de puntenwolken van de scores in de twee groepen. Ook is voor beide groepen de regressielijn van $y$ op $x$ getekend. De variabele $x$ is een persoonsvariabele.




Figuur 3.12: Figuur bij opgave 30.

1. Als de situatie is zoals in deel (a) van Figuur 3.12, welk verschil wordt er getoetst met een variantieanalyse en welk verschil wordt er getoetst met een covariantieanalyse? Leg uit.

2. Dezelfde vraag als in 1., maar nu met betrekking tot deel (b) van Figuur 3.12.
3. Is het in (a) denkbaar dat het observationeel onderzoek in feite een gerandomiseerd experiment is? En in (b)? Leg uit.

*Opgave 31*
De directie van een groot bedrijf vermoedt dat mannelijke werknemers over het algemeen een positievere houding hebben ten opzichte van het op geregelde basis verrichten van overwerk dan vrouwen. Om dit nader te onderzoeken werden uit zowel het vrouwelijke als het mannelijke werknemersbestand at random 60 personen getrokken. Deze personen vulden allen een vragenlijst in betreffende overwerk, waarbij een hoge score duidt op een positieve attitude. De score op deze vragenlijst OVERWERK is de afhankelijke variabele van het onderzoek.

1. Betreft het hier een experiment of een observationeel onderzoek? Leg uit.
2. Mannen bleken gemiddeld hoger te scoren op OVERWERK dan vrouwen. Er werd een variantieanalyse uitgevoerd om dit verschil statistisch te toetsen. De resultaten van deze analyse worden getoond in Tabel 3.67. Welke nulhypothese wordt er getoetst met de F-statistiek in deze tabel en welke conclusie trek je op basis van de uitslag van deze toets? Leg uit.

Dependent Variable: OVERWERK

| | Type I Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Corrected Model | 91.214 | 1 | 91.214 | 9.721 | .002 |
| GESLACHT | 91.214 | 1 | 91.214 | 9.721 | .002 |
| Error | 1107.217 | 118 | 9.383 | | |
| Corrected Total | 1198.431 | 119 | | | |

Tabel 3.67: Tabel bij opgave 31.

Dependent Variable: OVERWERK

| | Type I Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Corrected Model | 533.965 | 2 | 266.982 | 47.012 | .000 |
| SALARIS | 523.839 | 1 | 523.839 | 92.241 | .000 |
| GESLACHT | 10.126 | 1 | 10.126 | 1.783 | .179 |
| Error | 664.466 | 117 | 5.679 | | |
| Corrected Total | 1198.431 | 119 | | | |

Tabel 3.68: Tabel bij opgave 31.

3. Omdat de houding ten opzichte van overwerk mede bepaald kan worden door hoe goed dit betaalt, gaat men in een volgende analyse de salarisschaal van iedere werknemer opnemen als covariaat. De resultaten van een analyse met GLM-Univariate worden getoond in Tabel 3.68. Welke nulhypothese met betrekking tot het effect van GESLACHT wordt hier getoetst? Wat concludeer je op basis van deze resultaten, mede in het licht van de eerder gegeven resultaten van de variantieanalyse? Motiveer je antwoord.

4. Maak een grafiek met SALARIS op de X-as, OVERWERK op de Y-as, en waarin de twee groepen (mannen en vrouwen) als ellipsen worden weergegeven. Zorg ervoor dat deze grafiek in overeenstemming is met de hierboven gegeven resultaten van de variantie- en covariantieanalyse.

*Opgave 32*

Een nieuw soort kattenvoer werd tegelijkertijd in Engeland en in Nederland geïntroduceerd. De reclamecampagne voor de twee landen werd door twee verschillende reclamebureaus verzorgd. Na een jaar werden de twee campagnes geëvalueerd. Hiervoor werd in elk land aan 200 representatieve gezinnen met een kat gevraagd hoeveel blikjes van het nieuwe merk zij gevoerd hebben. Deze variabele wordt aangeduid als BLIKJES. De onafhankelijke variabele is Engeland versus Nederland, en wordt aangeduid als LAND. De gemiddelde aantallen blikjes per land worden gegeven in Tabel 3.69 en de resultaten van de variantieanalyse in Tabel 3.70.

Cell Means

| | BLIKJES | |
|---|---|---|
| LAND | Mean | N |
| Nederland | 2.7 | 200 |
| Engeland | 1.8 | 200 |

Tabel 3.69: Tabel bij opgave 32.

Dependent Variable: BLIKJES

| | Type I Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Corrected Model | 81 | 1 | 81 | 8.955 | 0.003 |
| LAND | 81 | 1 | 81 | 8.955 | 0.003 |
| Error | 3600 | 398 | 9.045 | | |
| Corrected Total | 3681 | 399 | | | |

Tabel 3.70: Tabel bij opgave 32.

1. Wat concludeer je op basis van de resultaten in Tabel 3.69 en 3.70?
2. In dezelfde enquête werd elke familie ook gevraagd wat hun attitude was tegenover blikvoer voor huisdieren (aangeduid als ATTITUDE), aan te geven op een schaal van 0 (totaal overbodig) tot 10 (onmisbaar). Men stelde vast dat (1) in beide landen de correlatie tussen ATTITUDE en BLIKJES positief was, en (2) Nederlanders op deze attitudevraag gemiddeld veel lager scoorden dan Engelsen (zie Tabel 3.71). Stel dat dit de enige informatie is waarover je beschikt (Tabel 3.69 en 3.70 doen even niet mee), in welk land (Nederland of Engeland) verwacht je de grootste consumptie van blikvoer? Motiveer.
3. Stel dat je in een covariantieanalyse ATTITUDE als covariaat opneemt naast de onafhankelijke variabele LAND. Maak een grafiek met ATTITUDE op de X-as, BLIKJES op de Y-as. Geef de twee groepen (Nederlandse en Engelse gezinnen) als ellipsen weer en zorg ervoor dat deze grafiek in overeenstemming is met de hierboven gegeven gemiddelden en de ANOVA-tabel.
4. Wat verwacht je van $SS_{LAND|ATTITUDE}$ in de covariantieanalyse in vergelijking met $SS_{LAND}$ in de variantieanalyse? Leg uit.

Cell Means

| | ATTITUDE | |
|---|---|---|
| LAND | Mean | N |
| Nederland | 2 | 200 |
| Engeland | 8 | 200 |

Tabel 3.71: Tabel bij opgave 32.

5. Wat verwacht je van $SS_{Error|LAND,ATTITUDE}$ in de covariantieanalyse in vergelijking met $SS_{Error|LAND}$ in de variantieanalyse? Leg uit.
6. Welke uitkomst verwacht je van de F-toets van LAND in de covariantieanalyse? Leg uit.
7. Het nieuwe merk blikvoer wordt nu ook in andere landen geïntroduceerd. Welke van de twee reclamebureaus zou je op basis van de hier gerapporteerde data de reclamecampagne laten verzorgen? Leg uit.

*Opgave 33*
Iemand die onderzoek wil doen naar de effectiviteit van verschillende middelbare scholen, raadpleegt een deskundige omtrent de statistische verwerking van zijn gegevens. De deskundige zegt onder andere het volgende: "Met behulp van covariantieanalyse kan men corrigeren voor op voorhand bestaande verschillen tussen de scholen." Wat bedoelt hij hiermee en hoe gaat de covariantieanalyse hier dan in zijn werk?

*Opgave 34*
Een psychogerontologe is geïnteresseerd in het effect van een recent ontwikkelde vriendschapscursus voor bejaarden. Het doel van de cursus is vereenzaamde bejaarden te leren vriendschapsrelaties aan te knopen. De onderzoekster bezocht in totaal 60 bejaarden en stelde hen vragen over hun vriendschapsrelaties vroeger en nu. Op basis van hun antwoorden op de vragen naar hun vroegere vriendschapsrelaties kregen de bejaarden een score van 1 (geen vriendschapsrelaties) tot 10 (zeer veel vriendschapsrelaties). Deze variabele wordt VRIENDVROEGER genoemd. Op het einde van het gesprek bood de onderzoekster de bejaarde aan om deel te nemen aan de vriendschapscursus. Het wel of niet deelnemen aan deze cursus is de onafhankelijke variabele in dit onderzoek, en hij wordt aangeduid als CURSUS.

Een jaar na afloop van de cursus werden alle bejaarden opnieuw bezocht en werd hen gevraagd om gedurende een maand bij te houden hoeveel uren zij elke dag doorbrengen in het gezelschap van een vriend of vriendin. De afhankelijke variabele is het gemiddeld aantal uren per dag dat een bejaarde doorbrengt in het gezelschap van een vriend of vriendin. Deze variabele wordt aangeduid als VRIENDNU. Voor de bejaarden die de vriendschapscursus gevolgd hadden was de gemiddelde VRIENDNU-score 1.4 en voor de bejaarden die de vriendschapscursus niet gevolgd hadden was de gemiddelde VRIENDNU-score 0.6.

1. Met GLM-Univariate werd een variantieanalyse uitgevoerd waarvan de resultaten in Tabel 3.72 getoond worden. Wat concludeer je op basis van deze tabel met betrekking tot de variabele CURSUS? Druk dit uit in termen van de nulhypothese die getoetst wordt.
2. Daarna werd een covariantieanalyse uitgevoerd met als covariaat de variabele VRIENDVROEGER. De resultaten van deze analyse worden getoond in Tabel 3.73. Wat concludeer je op basis van deze tabel met betrekking tot de variabele CURSUS? Druk dit uit in termen van de hypothese die getoetst wordt.
3. Waarom is het zinvol om een covariantieanalyse uit te voeren met VRIENDVROEGER als covariaat en niet met, bijvoorbeeld, het aantal lepeltjes suiker dat men in de koffie doet? (Indien je meer dan één reden weet, geef ze dan allemaal.)

Dependent Variable: VRIENDNU

| | Type I Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Corrected Model | 57.24 | 1 | 57.24 | 3.08 | 0.008 |
| CURSUS | 57.24 | 1 | 57.24 | 3.08 | 0.008 |
| Error | 439.06 | 58 | 7.57 | | |
| Corrected Total | 496.30 | 59 | | | |

Tabel 3.72: Tabel bij opgave 34.

Dependent Variable: VRIENDNU

| | Type I Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Corrected Model | 118.37 | 1 | 118.37 | 17.85 | 0.000 |
| VRIENDVROEGER | 107.13 | 1 | 107.13 | 16.16 | 0.000 |
| CURSUS | 11.24 | 1 | 11.24 | 1.70 | 0.198 |
| Error | 377.93 | 57 | 6.63 | | |
| Corrected Total | 496.30 | 59 | | | |

Tabel 3.73: Tabel bij opgave 34.

4. Maak een grafiek met VRIENDVROEGER op de X-as, VRIENDNU op de Y-as, en waarin de twee groepen (met en zonder vriendschapscursus) als ellipsen worden weergegeven. Zorg ervoor dat deze grafiek in overeenstemming is met de resultaten van de variantie- en de covariantieanalyse.
5. Hoe kan je in Tabel 3.72 en 3.73 zien dat de F-toets van CURSUS in de covariantieanalyse iets anders toetst dan de F-toets van CURSUS in de variantieanalyse.
6. Hoe kan je in Tabel 3.72 en 3.73 zien dat de F-toets van CURSUS in de covariantieanalyse gevoeliger is dan de F-toets van CURSUS in de variantieanalyse.

*Opgave 35*

Als gevolg van een reeks stakingen bij de Nederlandse Spoorwegen, vroeg de spoorwegvakbond een arbeidspsycholoog om te onderzoeken hoe de reizigers tegenover de stakingen staan. De arbeidspsycholoog ontwikkelde een vragenlijst om de attitude ten opzichte van stakingen te meten. Op een willekeurige dag vroeg hij 50 reizigers deze vragenlijst in te vullen. De scores op deze vragenlijst worden aangeduid als ATTITUDE. Daarnaast vroeg hij de reizigers ook naar hun hoogst genoten opleiding en hun inkomen. Op basis van hun antwoord op de eerste vraag verdeelde hij de 50 reizigers in hoog- en laagopgeleiden. Dit vormt de variabele OPLEIDING. De arbeidspsycholoog vraagt zich af of er een verschil is in attitude ten opzichte van staken tussen hoog- en laagopgeleiden.

1. Er werd een variantieanalyse uitgevoerd met GLM-Univariate. De resultaten hiervan worden getoond in Tabel 3.74. Wat concludeer je op basis van deze tabel met betrekking tot de variabele OPLEIDING? Druk dit uit in termen van de nulhypothese die getoetst wordt.
2. De arbeidspsycholoog realiseerde zich dat de attitude ten opzichte van staken ook bepaald kan worden door iemands inkomen. Daarom nam hij in een volgende analyse met GLM-Univariate het inkomen op als covariaat. De resultaten van deze analyse worden getoond

Dependent Variable: ATTITUDE

| | Type III Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Corrected Model | 386.420 | 1 | 386.420 | 302.284 | 0.000 |
| OPLEIDING | 386.420 | 1 | 386.420 | 302.284 | 0.000 |
| Error | 61.360 | 48 | 1.278 | | |
| Corrected Total | 447.780 | 49 | | | |

Tabel 3.74: Tabel bij opgave 35.

Dependent Variable: ATTITUDE

| | Type III Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Corrected Model | 442.305 | 2 | 221.153 | 257.754 | 0.000 |
| INKOMEN | 55.885 | 1 | 55.885 | 65.134 | 0.000 |
| OPLEIDING | 103.728 | 1 | 103.728 | 120.895 | 0.000 |
| Error | 40.341 | 47 | 0.858 | | |
| Corrected Total | 447.780 | 49 | | | |

Tabel 3.75: Tabel bij opgave 35.

in Tabel 3.75. Wat concludeer je op basis van deze tabel met betrekking tot de variabele OPLEIDING? Druk dit uit in termen van de nulhypothese die getoetst wordt.

3. Ga ervan uit dat hoger opgeleiden in vergelijking met lager opgeleiden (1) gemiddeld hoger scoren op de vragenlijst, (2) na uitzuivering voor INKOMEN nog steeds hoger scoren op de vragenlijst, en (3) een hoger inkomen hebben. Maak een grafiek met INKOMEN op de X-as, ATTITUDE op de Y-as, en waarin de twee groepen (hoge en lage opleiding) als ellipsen worden weergegeven. Zorg ervoor dat deze grafiek in overeenstemming is met de resultaten van de variantie- en de covariantieanalyse.
4. Veranderen de kwadratensommen in de rijen OPLEIDING en Error in Tabel 3.75 indien de afstand tussen de binnen-groepen regressielijnen groter wordt maar de afstand van de observaties tot deze regressielijnen gelijk blijft? Zo ja, leg uit hoe ze veranderen.
5. Veranderen de kwadratensommen in de rijen OPLEIDING en Error in Tabel 3.75 indien de afstand van de observaties tot de binnen-groepen regressielijnen groter wordt maar de afstand tussen deze regressielijnen gelijk blijft? Zo ja, leg uit hoe ze veranderen.

*Opgave 36*

In het kader van onderwijsvernieuwing werd beslist om het vak Statistiek 2 in drie didactische varianten aan te bieden. De studenten konden zelf kiezen uit deze drie varianten. Na afloop, werd op de tentamenuitslagen (aangeduid als STAT2) een variantieanalyse uitgevoerd met de door de student gevolgde didactische variant (aangeduid als VORM) als onafhankelijke variabele.

1. Betreft het hier een gerandomiseerd experiment of een observationeel onderzoek? Leg uit.
2. Met GLM-Univariate werd een variantieanalyse uitgevoerd op deze gegevens. De resultaten van deze analyse worden getoond in Tabel 3.76. Welke conclusie trek je op basis van deze analyse?

Dependent Variable: STAT2

| | Type I Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Corrected Model | 169.032 | 2 | 84.516 | 2.886 | .084 |
| VORM | 169.032 | 2 | 84.516 | 2.886 | .084 |
| Error | 2553.105 | 87 | 29.346 | | |
| Corrected Total | 2722.137 | 89 | | | |

Tabel 3.76: Tabel bij opgave 36.

Dependent Variable: STAT2

| | Type I Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Corrected Model | 692.604 | 3 | 230.868 | 9.785 | .000 |
| STAT1 | 524.118 | 1 | 524.118 | 22.215 | .000 |
| VORM | 168.486 | 2 | 84.243 | 3.571 | .031 |
| Error | 2029.533 | 86 | 23.593 | | |
| Corrected Total | 2722.137 | 89 | | | |

Tabel 3.77: Tabel bij opgave 36.

3. Van alle studenten is ook de score bekend van het in het voorafgaande jaar afgelegde tentamen Statistiek 1. Deze tentamenscore (aangeduid als STAT1) wordt als covariaat toegevoegd in een covariantieanalyse. Waarom lijkt STAT1 een veelbelovende covariaat?
4. Met GLM-Univariate werd een covariantieanalyse uitgevoerd met STAT1 als covariaat. De resultaten van deze analyse worden getoond in Tabel 3.77. Welke conclusie trek je op basis van deze analyse met betrekking tot het effect van de verschillende didactische varianten? Vergelijk deze conclusie met de conclusie die je getrokken hebt op basis van de ANOVA-tabel in 3.76. Waar is een eventueel verschil aan toe te schrijven, de teller of de noemer van de F-statistiek?
5. De uitgezuiverde gemiddelden $\overline{y}_k + B(\overline{x}_{\sqcup} - \overline{x}_k)$ werden berekend voor elk van de drie didactische varianten. Deze uitgezuiverde gemiddelden bleken niet of nauwelijks af te wijken van de niet-uitgezuiverde gemiddelden, ook al was er een samenhang tussen STAT1 en STAT2 binnen elk van de drie groepen. Toont dit aan dat de drie groepen gelijkwaardig samengesteld waren wat betreft STAT1-score? Motiveer je antwoord.
6. Maak een grafiek met de STAT1-score op de X-as, de STAT2-score op de Y-as, en waarin de drie groepen als ellipsen worden weergegeven. Zorg ervoor dat deze grafiek in overeenstemming is met (1) de uitgezuiverde en niet-uitgezuiverde gemiddelden en (2) de hierboven gegeven resultaten van de variantie- en de covariantieanalyse.

*Opgave 37*

Een student wil rijlessen nemen. Hij heeft de keuze tussen twee grote rijscholen, A en B. Beide rijscholen hebben 12 instructeurs. Van elke instructeur wordt het slagingspercentage bijgehouden. Dit slagingspercentage wordt berekend als het aantal leerlingen dat bij de eerste examenkans geslaagd was gedeeld door het totale aantal leerlingen dat de instructeur begeleid heeft. Om nu te beoordelen welke rijschool hij het beste kan kiezen, vroeg de student de slagingspercentages van de verschillende instructeurs van de twee rijscholen op. Bij de instruc-

teurs van rijschool A bleken gemiddeld 70 procent leerlingen bij de eerste kans te slagen, tegen gemiddeld 60 procent bij de instructeurs van rijschool B. De student voerde een variantieanalyse uit om dit verschil statistisch te toetsen. De resultaten van deze analyse staan in Tabel 3.78. Hierin wordt de afhankelijke variabele aangeduid als SLAGINGS% en de onafhankelijke als RIJSCHOOL.

1. Welke nulhypothese wordt er getoetst met de F-statistiek in Tabel 3.78? Wat besluit je op basis van deze toets over het effect van RIJSCHOOL op SLAGINGS%? Welke rijschool zou jij kiezen op basis van deze resultaten?

Dependent Variable: SLAGINGS%

| | Type I Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Corrected Model | 600.000 | 2 | 600.000 | 51.563 | .000 |
| RIJSCHOOL | 600.000 | 1 | 51.563 | 51.563 | .000 |
| Error | 256.000 | 22 | 11.636 | | |
| Corrected Total | 856.000 | 23 | | | |

Tabel 3.78: Tabel bij opgave 37.

Dependent Variable: SLAGINGS%

| | Type I Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Corrected Model | 762.000 | 2 | 381.000 | 85.117 | .000 |
| LESSEN | 714.532 | 1 | 714.532 | 159.629 | .000 |
| RIJSCHOOL | 47.468 | 1 | 47.468 | 10.605 | .000 |
| Error | 94.000 | 21 | 4.476 | | |
| Corrected Total | 856.000 | 23 | | | |

Tabel 3.79: Tabel bij opgave 37.

Een vriendin van de student wees hem erop dat het aantal gevolgde rijlessen een storende variabele kan zijn. Daarom vroegen ze aan de rijscholen het gemiddelde aantal rijlessen dat elke instructeur aan zijn of haar leerlingen geeft vóór het eerste examen. Dit gemiddelde aantal rijlessen wordt aangeduid als LESSEN. De twee studenten stelden vast dat de correlatie tussen LESSEN en SLAGINGS% positief is en dat de instructeurs van rijschool A gemiddeld meer rijlessen geven dan die van rijschool B; gemiddeld 50 lessen bij rijschool A en gemiddeld 30 bij rijschool B. De studenten voerden nu een covariantieanalyse uit met LESSEN als covariaat. De resultaten hiervan staan in Tabel 3.79.

2. Welke nulhypothese wordt getoetst met de F-statistiek van RIJSCHOOL in Tabel 3.79? Wat is je conclusie met betrekking tot het effect van de variabele RIJSCHOOL?
3. De uitgezuiverde gemiddelden $\overline{y}_k + B(\overline{x}_{\sqcup} - \overline{x}_k)$ werden berekend voor beide groepen. Het bleek dat dit uitgezuiverde gemiddelde voor rijschool B groter is dan voor rijschool A. Wat kun je hieruit besluiten betreffende de kwaliteit van de instructeurs van de twee rijscholen? Indien je besluit verschilt van je besluit bij de eerste deelopgave, hoe kun je dit verschil verklaren?

4. Maak een grafiek met LESSEN op de X-as, SLAGINGS% op de Y-as, en waarin de drie groepen als ellipsen worden weergegeven. Zorg ervoor dat deze grafiek in overeenstemming is met alle informatie die in deze opgave gegeven is.

*Opgave 38*
In een observationeel onderzoek hebben we een onafhankelijke variabele met twee niveaus. De gemiddelden in de twee groepen worden getoond in Tabel 3.80. Er werd eerst een variantieanalyse uitgevoerd met GLM-Univariate. De resultaten hiervan worden getoond in Tabel 3.81. De afhankelijke variabele wordt aangeduid als AFH en de onafhankelijke als ONAFH.

Cell Means

| | AFH | |
|---|---|---|
| ONAFH | Mean | N |
| Groep 1 | 51.54 | 40 |
| Groep 2 | 48.76 | 40 |

Tabel 3.80: Tabel bij opgave 38.

Dependent Variable: AFH

| | Type I Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Corrected Model | 154.568 | 1 | 154.568 | 1.932 | .385 |
| ONAFH | 154.568 | 1 | 154.568 | 1.932 | .385 |
| Error | 6230.877 | 78 | 79.883 | | |
| Corrected Total | 6385.445 | 79 | | | |

Tabel 3.81: Tabel bij opgave 38.

1. Welke conclusie trek je op basis van deze analyse?

Er was ook een covariaat gemeten (aangeduid als COVAR), en die werd opgenomen in een covariantieanalyse met GLM-Univariate. De uitgezuiverde gemiddelden worden getoond in Tabel 3.82. Deze uitgezuiverde gemiddelden werden berekend op het gemiddelde van alle covariaatscores (over de twee groepen). De resultaten van de covariantieanalyse worden getoond in Tabel 3.83.

| | n | Niet-Uitgezuiverde Gemiddelden | Uitgezuiverde Gemiddelden |
|---|---|---|---|
| Groep 1 | 40 | 51.54 | 45.82 |
| Groep 2 | 40 | 48.76 | 54.48 |

Tabel 3.82: Tabel bij opgave 38.

2. Wat is het verschil tussen de nulhypotheses die in de bovenstaande variantie- en covariantieanalyse getoetst worden?
3. Wat concludeer je nu op basis van beide analyses gezamenlijk? Verduidelijk je interpretatie met een grafiek die in overeenstemming is met de data. Zet de covariaat op de X-as, de afhankelijke variabele op de Y-as, en stel de twee groepen voor met ellipsen.

Dependent Variable: AFH

| | Type I Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Corrected Model | 3792.163 | 2 | 1896.082 | 56.299 | .000 |
| COVAR | 3500.915 | 1 | 3500.915 | 103.949 | .000 |
| ONAFH | 291.248 | 1 | 291.248 | 8.648 | .001 |
| Error | 2593.282 | 77 | 33.679 | | |
| Corrected Total | 6385.445 | 79 | | | |

Tabel 3.83: Tabel bij opgave 38.

*Opgave 39* COMPUTEROPDRACHT

Drie sociaal-psychologen zijn geïnteresseerd in de determinanten van hulpverlenend gedrag. Uit het onderzoek van Latané en Darley (1970) weten zij dat het aantal omstanders een belangrijke determinant is van hulpverlenend gedrag. Het nieuwe element dat zij opnamen in hun onderzoek, is de geboorterang van een persoon binnen zijn of haar gezin van herkomst. Uit een aantal onderzoeken is namelijk gebleken dat eerstgeborenen zich in groepen meer verantwoordelijk gedragen dan latergeborenen. De drie psychologen hebben elk een andere onderzoeksvraag.

1. De hypothese van de eerste psycholoog is dat personen met een lage geboorterang (i.e., personen die zich vooraan bevinden in de geboorterangorde binnen hun gezin van herkomst) meer geneigd zijn om mensen in nood te helpen dan personen met een hoge geboorterang (i.e., personen die zich achteraan bevinden in de geboorterangorde binnen hun gezin van herkomst).
2. De hypothese van de tweede psycholoog is dat personen met een lage geboorterang minder gevoelig zijn voor het aantal omstanders (hun gedrag wordt er minder door bepaald) dan personen met een hoge geboorterang. Tip: het toetsen van deze hypothese heeft niets met uitzuiveren te maken.
3. De derde psycholoog beweert dat, zo er al een verschil is in hulpgedrag tussen personen met een hoge en een lage geboorterang, dit niets te maken heeft met de geboorterang zelf, maar een gevolg is van het feit dat personen met een hoge geboorterang een grotere kans hebben om afkomstig te zijn uit grote gezinnen dan personen met een lage geboorterang. Een eventueel verschil tussen personen met een hoge en een lage geboorterang is dan toe te schrijven aan het feit dat personen die afkomstig zijn uit grote gezinnen beter geleerd hebben om anderen te helpen, door hun ervaringen met broertjes en zusjes.

Om hun hypotheses te toetsen, zetten deze drie psychologen samen een onderzoek op. Dit onderzoek bestond erin dat 135 proefpersonen op zuiver toevallige wijze werden toegewezen aan drie condities: (1) een conditie zonder omstanders, (2) een conditie met één omstander, en (3) een conditie met drie omstanders. De proefpersonen werden ook ingedeeld op basis van hun geboorterang. Hierbij werden de volgende categorieën gehanteerd: eerstgeborene (categorie 1), tweede geborene (categorie 2), en later geborene (categorie 3). Het onderzoeksdesign bevat dus $3 \times 3 = 9$ cellen, waarin zich telkens 15 proefpersonen bevinden. Aan elke proefpersoon werd ook gevraagd wat het aantal kinderen was in zijn of haar gezin van herkomst.

De proefpersonen namen zogezegd deel aan een medisch onderzoek waarin een nieuw type gehoorscherptemeter werd getest. De proefpersoon moest in een wachtkamer plaatsnemen waarin zich de persoon bevond die het hulpgedrag trachtte uit te lokken. In deze wachtkamer bevonden zich ook de eventuele omstanders. De omstanders waren medewerkers van de onderzoeker, en boden geen hulp. Tijdens de wachtperiode begon de hulpvrager symptomen van

maagkrampen te simuleren, en zolang de proefpersoon geen hulp bood werd dit gedrag versterkt (kreunen, over de grond rollen, ...). De afhankelijke variabele is de tijd dat het duurde voordat de proefpersoon opstond en zich bekommerde om de hulpvrager.

De gegevens van dit gefingeerde onderzoek zijn opgeslagen in het SPSS-bestand "hulp.sav". De variabelen zijn de volgende:

1. Proefpersoonnummer, aangeduid als PPNR.
2. Aantal omstanders, aangeduid als OMSTAND.
3. Geboorterang, aangeduid als GEBRANG.
4. Aantal kinderen in het gezin van herkomst, aangeduid als NKIND.
5. Tijd (in seconden) die voorbij ging vooraleer hulp geboden werd, aangeduid als TIJD.

Stel je achtereenvolgens in de plaats van elk van de drie psychologen, en voer die analyse(s) uit waarmee je hun hypotheses zo gevoelig mogelijk kunt toetsen. Alle analyses kunnen uitgevoerd worden met de SPSS-procedure GLM-Univariate.

---

## 3.5 Designs met een Voor- en een Nameting

In de vorige paragrafen hebben we gezien dat men rekening kan houden met de invloed van storende persoonsvariabelen door middel van statistische uitzuivering. Een andere manier om rekening te houden met storende persoonsvariabelen is door gebruik te maken van een design met een voor- en een nameting. Dit houdt in dat men van elke proefpersoon twee waarnemingen heeft: een van vóór de behandeling en een van erna. De idee is nu om de analyse te baseren op de verschilscores "na − voor" (of, equivalent, "voor − na"). Men hoopt dan dat er, door het nemen van verschilscores, gecorrigeerd wordt voor eventuele verschillen tussen de groepen, verschillen die reeds bestaan vóór de behandeling. Zoals we verder zullen zien, kan ook deze methode (net zoals statistische uitzuivering), niet garanderen dat we alleen het effect van de condities onderzoeken, zonder storende invloed van bestaande verschillen tussen de groepen. We moeten ook opmerken dat designs met een voor- en een nameting geen oplossing bieden voor storende situatievariabelen.

We bespreken eerst het design met één groep. Hier kan er natuurlijk geen sprake zijn van een verschil tussen groepen waarvoor er gecorrigeerd moet worden. De reden waarom we dit design behandelen, is dat het soms gebruikt wordt om een uitspraak te doen over een causaal verband tussen de behandeling en de afhankelijke variabele. Zoals verder zal blijken, is dit onder een bepaalde veronderstelling ook mogelijk. Deze veronderstelling is echter niet toetsbaar, en in een aantal situaties is het weinig aannemelijk dat zij opgaat.

### 3.5.1 Het Design met Eén Groep

#### Voorbeeld, Notatie en Toets

We gebruiken hier hetzelfde voorbeeld als in de vorige paragraaf, pesten in de klas. In dit voorbeeld werd er twee keer een vragenlijst afgenomen: een keer in het begin (twee maanden na de eerste schooldag) en een keer aan het einde van het schooljaar. Op basis

van deze vragenlijst werd de pestscore berekend. Voor elke leerling hebben we dus twee pestscores: één van het begin van het schooljaar en één van het einde.

We bekijken nu de pestscores van de leerlingen die aan de training hebben deelgenomen. Hun eerste pestscore duiden we aan met $x$ en hun tweede met $y$. Opnieuw gebruiken we de index $i$ om de leerlingen aan te duiden ($i = 1, \ldots, n$). Dus, $x_i$ staat voor de eerste pestscore van de i-de leerling en $y_i$ staat voor de tweede. Onze uitspraak over het effect van de training is gebaseerd op de verschillen $y_i - x_i$. Deze verschillen duiden we aan met $d_i$. Als deze verschillen overwegend negatief zijn, dan zeggen we dat de training een positief effect heeft (afname van de pestscores). Het gemiddelde verschil duiden we aan met $\overline{d}$. Met dit gemiddelde verschil geven we de grootte van het effect aan. Merk op dat $\overline{d}$ gelijk is aan $\overline{y} - \overline{x}$.

Meestal zijn we niet alleen geïnteresseerd in de grootte van $\overline{d}$, maar willen we ook weten of dit gemiddelde significant afwijkt van 0. Met andere woorden, we willen de nulhypothese toetsen dat de verwachte waarde van het verschil $d = y - x$ (het gemiddelde $\overline{d}$ in de populatie) gelijk is aan 0. Dit kunnen we doen met de *T-toets voor paarsgewijze verschillen* (ook de T-toets voor *gekoppelde paren* genoemd), die in vrijwel alle statistiekhandboeken behandeld wordt.

### Causale Interpretatie

Stel dat $\overline{d}$ duidelijk negatief is en bovendien significant afwijkend van 0. Kunnen we dan besluiten dat er een effect is van TRAINING? Met andere woorden, kunnen we dan besluiten dat er een causaal verband tussen TRAINING en PESTSCORE? Het probleem bij de beantwoording van deze vraag, is dat er slechts één conditie is, namelijk de conditie met training. Hoe kunnen we dan spreken van een effect? Waar moeten we mee vergelijken?

Veronderstel echter dat $x$, de pestscore in het begin van het schooljaar, gelijk is aan de pestscore die men zou observeren op het einde van het schooljaar *indien er geen training geweest zou zijn*. Als dit waar is, dan is $\overline{d}$ ($= \overline{y} - \overline{x}$) precies de gemiddelde toename/afname van de pestscore die toe te schrijven is aan de training. Uit een significante T-toets besluiten we dan dat er een causaal verband is tussen wel of geen training en de pestscore.

Niets garandeert ons echter dat deze veronderstelling waar is. Het kan best zijn dat de pestscore gemiddeld afneemt in de loop van het schooljaar zonder dat daar een training voor nodig is. Leerlingen kunnen bijvoorbeeld op eigen houtje een strategie ontdekken waarmee ze de anderen de lust ontnemen om hen nog langer te pesten. In dit geval betekent een significante T-toets helemaal niet dat er een causaal verband is tussen TRAINING en PESTSCORE. Zoals verder zal blijken, wordt er in het design met twee groepen wel rekening gehouden met de mogelijkheid dat de afhankelijke variabele spontaan (i.e., zonder invloed van de behandeling) kan veranderen in de loop van de tijd.

### 3.5.2 Het Design met Twee Groepen

#### Voorbeeld, Notatie en Toets

We beschouwen hier hetzelfde voorbeeld als bij het design met één groep. We bekijken nu echter niet alleen de scores van de leerlingen die aan de training deelgenomen hebben, maar ook de scores van degenen die er niet aan deelgenomen hebben, maar aan wie de training wel aangeboden werd. De eerste pestscore van de leerlingen die aan de training hebben deelgenomen, duiden we aan met $x_b$ en de tweede met $y_b$. Hierin staat $b$ voor "behandelgroep". De eerste pestscore van de leerlingen die niet aan de training hebben deelgenomen duiden we aan met $x_c$ en de tweede met $y_c$. Hierin staat $c$ voor "controlegroep". Opnieuw gebruiken we de index $i$ om de leerlingen aan te duiden ($i = 1, \ldots, n_b$, voor de behandelgroep, en $i = 1, \ldots, n_c$, voor de controlegroep). Onze uitspraak over het effect van TRAINING is gebaseerd op de verschillen $y_{bi} - x_{bi}$ (voor de behandelgroep) en $y_{ci} - x_{ci}$ (voor de controlegroep). Deze verschillen duiden we aan met, respectievelijk, $d_{bi}$ en $d_{ci}$. De gemiddelde verschillen in de twee groepen duiden we aan met, respectievelijk, $\overline{d}_b$ en $\overline{d}_c$. Deze gemiddelden zijn gelijk aan, respectievelijk, $\overline{y}_b - \overline{x}_b$ en $\overline{y}_c - \overline{x}_c$. Als $\overline{d}_b$ kleiner is dan $\overline{d}_c$, dan is $\overline{d}_b - \overline{d}_c$ negatief, en dan zeggen we dat TRAINING een *positief* effect heeft. Het positieve effect van TRAINING is des te groter (en $\overline{d}_b - \overline{d}_c$ is des te kleiner) naarmate (1) de afname van de pestscores in de behandelgroep groter is dan in de controlegroep of (2) de toename van de pestscores in de behandelgroep kleiner is dan in de controlegroep.

Meestal zijn we niet alleen geïnteresseerd in de grootte van $\overline{d}_b - \overline{d}_c$, maar willen we ook weten of dit verschil significant afwijkt van 0. Met andere woorden, we willen de nulhypothese toetsen dat de verwachte waarde van $d_b$ $(= y_b - x_b)$ gelijk is aan de verwachte waarde van $d_c$ $(= y_c - x_c)$. Dit kunnen we doen met de *T-toets voor onafhankelijke steekproeven*, die beschreven wordt in vrijwel alle statistiekhandboeken. Het enige ongewone aan de T-toets die hier voorgesteld wordt, is dat de gemiddelden die hier vergeleken worden (i.c., $\overline{d}_b$ en $\overline{d}_c$) gemiddelde verschilscores zijn en geen gemiddelden van "gewone" variabelen.

Als er meer dan twee groepen zijn, dan kunnen we geen gebruik maken van de T-toets voor onafhankelijke steekproeven. In dit geval moeten we gebruikmaken van de F-statistiek van de variantieanalyse. Het enige verschil met een gewone toepassing van de variantieanalyse, is dat de afhankelijke variabele nu een verschilscore is (na − voor) in plaats van een gewone variabele. In wat volgt, gaan we ervan uit dat er maar twee groepen zijn.

#### Causale Interpretatie

Stel dat $\overline{d}_b - \overline{d}_c$ duidelijk negatief is en bovendien significant afwijkend van 0. Kunnen we dan besluiten dat er een effect is van de training? Opnieuw kunnen we dit besluit slechts nemen als we bereid zijn om een extra veronderstelling te maken. Deze veronderstelling komt erop neer dat $\overline{d}_c$, de gemiddelde toename/afname van de pestscore in de controlegroep, gelijk is aan de gemiddelde toename/afname die men zou observeren in de behandelgroep *indien de training er niet geweest zou zijn*. Als deze veronderstelling waar is, dan is $\overline{d}_b - \overline{d}_c$ precies de gemiddelde toename/afname van de pestscore die toe

te schrijven is aan de training. Uit een significante T-toets besluiten we dan dat er een causaal verband is tussen wel of geen training en de pestscore.

Niets garandeert ons echter dat deze veronderstelling waar is. Het kan best zijn dat de gemiddelde toename/afname van de pestscore in de controlegroep verschilt van de gemiddelde toename/afname die men zou observeren in de behandelgroep indien de training er niet geweest zou zijn. We kunnen dit concreter maken aan de hand van de zogenaamde *waaierhypothese*. Deze hypothese wordt grafisch voorgesteld in Figuur 3.13. Elke lijn in deze figuur staat voor één persoon. Deze lijnen beschrijven de spontane verandering van de pestscore in de loop van de tijd (i.e., zonder invloed van de training). De vijf onderste lijnen zijn van leerlingen uit de behandelgroep en de vijf bovenste lijnen (de onderbroken lijnen) van leerlingen uit de controlegroep. De zwarte punten in de figuur staan voor de gemiddelde pestscores in de twee groepen, een keer vóór en een keer na de training. De volgende kenmerken van dit patroon zijn van belang:

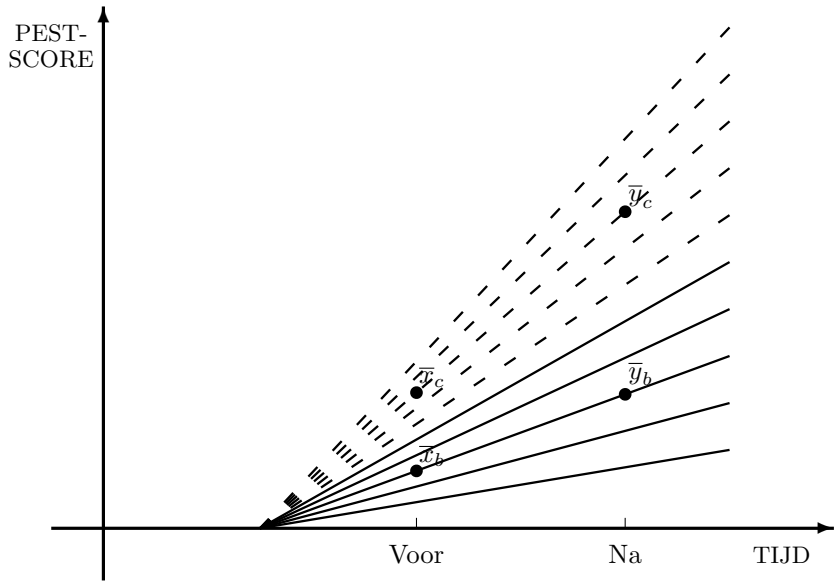
1. Iedereen begint op hetzelfde niveau, maar er zijn grote verschillen in de mate waarin de pestscore toeneemt met de tijd.

2. Leerlingen met lage pestscores nemen deel aan de training en leerlingen met hoge pestscores niet.

Hieruit volgt dat de toename in gemiddelde pestscore veel kleiner is voor de leerlingen in de behandelgroep dan voor de leerlingen in de controlegroep *ook al is er geen effect van de training*. In dit geval betekent een significante T-toets dus niet dat er een causaal verband is tussen TRAINING en PESTSCORE.

De waaierhypothese heeft een belangrijke rol gespeeld in het onderzoek naar de mogelijkheid om kinderen uit achtergestelde milieus cognitief bij te spijkeren zodat ze het op de middelbare school en de universiteit gemiddeld even goed zouden doen als kinderen uit de middenklasse. In de jaren 60 en 70 bestond er vooral in de Verenigde Staten veel interesse voor deze onderzoeksvraag. De afhankelijke variabele in dit type onderzoek zijn schoolresultaten. De onafhankelijke variabele was een training in cognitieve vaardigheden. Deze training werd gevolgd door de kinderen uit de achtergestelde milieus; zij waren de behandelgroep. De kinderen uit de middenklasse volgden deze training niet; zij waren de controlegroep.

Opnieuw duiden we de grootte van het effect aan met $\overline{d}_b - \overline{d}_c$. In tegenstelling tot het voorbeeld over pesten in de klas, spreken we hier van een positief effect als $\overline{d}_b$ groter is dan $\overline{d}_c$. In dit geval is $\overline{d}_b - \overline{d}_c$ positief. Dit geeft immers aan dat de stijging van de schoolresultaten in de behandelgroep groter is dan die in de controlegroep. Nu, indien (1) de waaierhypothese opgaat, (2) de training geen effect heeft, en (3) de kinderen uit de middenklasse gemiddeld betere schoolresultaten hebben vóór de training, dan zal $\overline{d}_b - \overline{d}_c$ negatief zijn. Dit wordt geïllustreerd in Figuur 3.13 (met PESTSCORE vervangen door SCHOOLRESULTATEN). Dit impliceert dat, zelfs al heeft de training een positief effect, $\overline{d}_b - \overline{d}_c$ toch negatief is als dit effect niet groot genoeg is. Dit ziet men in Figuur 3.13 aan het feit dat $\overline{y}_b$ een heel eind in de richting van $\overline{y}_c$ verschoven moet worden om het verschil $\overline{d}_b - \overline{d}_c$ positief te laten worden. Dus, in dit geval is $\overline{d}_b - \overline{d}_c$ ongeschikt om het effect van de training te evalueren.




Figuur 3.13: Grafische voorstelling van de waaierhypothese. De gemiddelde voormetingen van de controle- en de behandelgroep worden aangeduid met, respectievelijk, $\overline{x}_c$ en $\overline{x}_b$. En de gemiddelde nametingen van de controle- en de behandelgroep worden aangeduid met, respectievelijk, $\overline{y}_c$ en $\overline{y}_b$.

### Een Evaluatie vanuit Pragmatisch Oogpunt

Al bij al, hebben we een weinig rooskleurig beeld geschetst van de mogelijkheid om causale verbanden te leggen op basis van gegevens uit observationeel onderzoek en/of onderzoek met een gebrekkige experimentele controle. Ten eerste, uitzuiveren voor storende variabelen in variantie- en covariantieanalyse is alleen maar een oplossing als we die storende variabelen ook kennen (geregistreerd hebben). En ten tweede, de gegevens uit een design met een voor- en een nameting kunnen alleen gebruikt worden voor het leggen van een causaal verband als men bereid is niet-toetsbare veronderstellingen te maken. Omdat psychologisch onderzoek vaak observationeel is en experimentele controle vaak moeilijk, is dit een weinig bevredigende situatie.

De beste houding is "proberen er het beste van te maken". Het volgende kan als richtlijn fungeren:

1. Beantwoord, indien mogelijk, je onderzoeksvraag met een gerandomiseerd experiment en draag zorg voor experimentele controle van de condities.

2. Is dit niet mogelijk, dan is een design met een voor- en een nameting en meerdere groepen een redelijk alternatief. De gegevens die men binnen dit design verzamelt, kan men op twee manieren analyseren. Ten eerste, men kan een T- of een F-toets toepassen op de verschilscores (na − voor). En ten tweede, men kan een covariantieanalyse uitvoeren, met de scores op de voormeting als covariaat. Als men met beide methoden een significant resultaat krijgt, dan mag men enig vertrouwen

hebben in het feit dat de condities een causaal effect hebben. Honderd procent zekerheid heeft men echter niet!

3. Is een design met een voor- en een nameting niet mogelijk, registreer dan zoveel mogelijk variabelen die storend kunnen zijn, en gebruik variantieanalyse voor ongebalanceerde designs en/of covariantieanalyse om ervoor uit te zuiveren. Ook dit geeft geen honderd procent zekerheid over het effect van de condities.

---

*Opgave 40*
Een onderzoekster wilde weten of de onderlinge sociale contacten van bewoners van flatgebouwen toenemen als zij lid worden van een bewonerscommissie. De onderzoekster interviewde alle nieuwe bewoners vijf maanden nadat zij hun nieuwe flat betrokken hebben. Tijdens dit interview bepaalde zij, op basis van de informatie van de bewoner, wat de gemiddelde hoeveelheid tijd per week is die deze bewoner besteedde aan sociale contacten met medebewoners van het flatgebouw. Na drie jaar interviewde zij de bewoners opnieuw. Zij bepaalde weer de gemiddelde hoeveelheid tijd per week die elke bewoner besteedde aan sociale contacten met medebewoners. Zij vroeg ook of de bewoner in de loop van de afgelopen drie jaar lid is geworden van een bewonerscommissie.

1. Stel dat de onderzoekster alleen gebruik maakt van de gegevens van de bewoners die lid zijn geworden van een bewonerscommissie. Hoe kan zij, op basis van deze gegevens, een uitspraak doen over het effect van lidmaatschap van een bewonerscommissie op de hoeveelheid sociale contacten met medebewoners? Welke veronderstelling moet zij maken om deze uitspraak te kunnen doen?
2. Stel dat de onderzoekster nu ook gebruik maakt van de gegevens van de bewoners die geen lid geworden zijn van een bewonerscommissie. Hoe kan zij, op basis van de volledige set gegevens, een uitspraak doen over het effect van lidmaatschap van een bewonerscommissie op de hoeveelheid sociale contacten met medebewoners? Welke veronderstelling moet zij nu maken om deze uitspraak te kunnen doen? Is deze veronderstelling strenger of minder streng dan de veronderstelling bij 1.?

---

## 3.6 Appendix bij Hoofdstuk 3

### Voorspelling op Basis van het Variantieanalysemodel met één Factor

De voorspelling van de observatie $y_{ki}$ op basis van het variantieanalysemodel wordt aangeduid met $\widehat{y}^A_{ki}$. Om een voorspelling te maken op basis van dit model moeten we eerst zijn parameters schatten. Voor de schatting van de parameters blijkt het handiger te zijn als we gebruik maken van het variantieanalysemodel met de parameters $\mu_k$ ($y_{ki} = \mu_k + e_{ki}$) dan van het model met de effectparameters $\alpha_k$ ($y_{ki} = \mu + \alpha_k + e_{ki}$). De schatters van de $\mu_k$-parameters duiden we aan met $\widehat{\mu}_k$. Omdat de $e_{ki}$'s beschouwd worden als onvoorspelbare ruis, zijn de voorspellingen $\widehat{y}^A_{ki}$ gelijk aan $\widehat{\mu}_k$.

We bepalen de schatters $\widehat{\mu}_k$ op basis van het kleinste-kwaddratencriterium. De kleinste-kwadratenschatting van de parameters $\mu_k$ van het variantieanalysemodel houdt

in dat gezocht wordt naar die waarden $\widehat{\mu}_k$ die zorgen voor zo klein mogelijke voorspellingsfouten. Deze voorspellingsfouten zijn de afwijkingen tussen de observaties $y_{ki}$ en hun voorspellingen $\widehat{\mu}_k$. Dus, er wordt gezocht naar die waarden $\widehat{\mu}_k$ die zorgen voor zo klein mogelijke afwijkingen

$$y_{ki} - \widehat{\mu}_k \quad .$$

Letterlijk genomen, zou dit neerkomen op het minimaliseren over alle groepen $k$ en alle proefpersonen $i$ van

$$|y_{ki} - \widehat{\mu}_{ki}| \quad .$$

Wiskundig, blijkt het echter veel handiger te zijn om, in plaats van deze absolute waarde, het gekwadrateerde verschil

$$(y_{ki} - \widehat{\mu}_{ki})^2$$

als de afwijking te definiëren die geminimaliseerd moet worden. Deze overgang is redelijk omdat het zowel bij de absolute waarde als bij het gekwadrateerde verschil om de grootte gaat, afgezien van het teken; de abolute waarde is klein als het kwadraat klein is, en vice versa. Het effect van het kwadraat nemen is dat grotere fouten relatief wat zwaarder worden gewogen.

De keuze voor deze kwadratische afwijking als indicatie voor de fout in de voorspelling leidt tot het volgende criterium:

> **kleinste-kwaddratencriterium:**
> Kies de schatters $\widehat{\mu}_k$ zó dat de som van de gekwadrateerde afwijkingen $(y_{ki} - \widehat{\mu}_k)^2$ zo klein mogelijk wordt. Dus, maak
>
> $$\sum_{Obs} (y_{ki} - \widehat{\mu}_k)^2 \quad \text{minimaal.}$$

De te minimaliseren expressie kunnen we ook als volgt schrijven:

$$\begin{aligned}\sum_{Obs} (y_{ki} - \widehat{\mu}_k)^2 &= \sum_{k=1}^{K} \sum_{i=1}^{n_k} (y_{ki} - \widehat{\mu}_k)^2 \\ &= \sum_{i=1}^{n_1} (y_{1i} - \widehat{\mu}_1)^2 + \sum_{i=1}^{n_2} (y_{2i} - \widehat{\mu}_2)^2 + \ldots + \sum_{i=1}^{n_K} (y_{Ki} - \widehat{\mu}_K)^2 \quad .\end{aligned}$$

Uit deze laatste formule blijkt dat men het minimaliseringsprobleem kan opsplitsen in $K$ analoge deelproblemen, één deelprobleem voor elke groep. Elk deelprobleem is van de volgende vorm:

> Kies de schatter $\widehat{\mu}_k$ zó dat de som van de gekwadrateerde afwijkingen $(y_{ki} - \widehat{\mu}_{ki})^2$ zo klein mogelijk wordt. Dus, maak
>
> $$\sum_{i=1}^{n_k} (y_{ki} - \widehat{\mu}_k)^2 \quad \text{minimaal.}$$

Om dit deelprobleem op te lossen, passen we het trucje toe om tussen de haakjes $-\overline{y}_k+\overline{y}_k$ toe te voegen. Netto wordt dus niets toegevoegd. De te minimaliseren expressie kunnen we dan als volgt uitwerken:

$$\begin{aligned}\sum_{i=1}^{n_k}(y_{ki}-\widehat{\mu}_k)^2 &= \sum_{i=1}^{n_k}(y_{ki}-\overline{y}_k+\overline{y}_k-\widehat{\mu}_k)^2\\ &= \sum_{i=1}^{n_k}[(y_{ki}-\overline{y}_k)+(\overline{y}_k-\widehat{\mu}_k)]^2\\ &= \sum_{i=1}^{n_k}(y_{ki}-\overline{y}_k)^2+\sum_{i=1}^{n_k}(\overline{y}_k-\widehat{\mu}_k)^2+2\sum_{i=1}^{n_k}(y_{ki}-\overline{y}_k)(\overline{y}_k-\widehat{\mu}_k) \quad .\end{aligned}$$

De laatste term in deze formule is gelijk aan 0. Dit zien we als we eerst $(\overline{y}_k-\widehat{\mu}_k)$ voor het somteken brengen en daarna het gedeelte vanaf het somteken als volgt uitwerken:

$$\begin{aligned}\sum_{i=1}^{n_k}(y_{ki}-\overline{y}_k) &= \sum_{i=1}^{n_k}y_{ki}-\sum_{i=1}^{n_k}\overline{y}_k\\ &= n_k\overline{y}_k-n_k\overline{y}_k\\ &= 0 \quad ,\end{aligned}$$

waarin we gebruikmaken van het feit dat $\overline{y}_k$ gelijk is aan $\left(\sum_{i=1}^{n_k}y_{ki}\right)/n_k$.

De te minimaliseren expressie kan nu als volgt geschreven worden:

$$\sum_{i=1}^{n_k}(y_{ki}-\widehat{\mu}_k)^2=\sum_{i=1}^{n_k}(y_{ki}-\overline{y}_k)^2+\sum_{i=1}^{n_k}(\overline{y}_k-\widehat{\mu}_k)^2 \quad .$$

Alleen de laatste term hangt af van $\widehat{\mu}_k$, en hij wordt geminimaliseerd als we $\widehat{\mu}_k$ gelijkstellen aan $\overline{y}_k$, het corresponderende groepsgemiddelde. Dus, de kleinste-kwadraten schatter $\widehat{\mu}_k$ is gelijk aan het corresponderende groepsgemiddelde $\overline{y}_k$. Hiermee is ook aangetoond dat de voorspelling $\widehat{y}^A_{ki}$ gelijk is aan $\overline{y}_k$.

## Afleiding van de Formule van de Type II Kwadratensom voor een Gebalanceerd Design

We vertrekken van de formule van de type II kwadratensom van factor A die zowel geldig is voor een gebalanceerd als een ongebalanceerd design:

$$SS_{A|B}=\sum_{Obs}(\widehat{y}^{A,B}_{kli}-\overline{y}_{\sqcup l})^2 \quad .$$

Deze formule kan herschreven worden voor een gebalanceerd design. In een gebalanceerd design geldt namelijk het volgende:

1. De voorspelling $\widehat{y}^{A,B}_{kli}$ is gelijk aan $\overline{y}_{k\bullet}+\overline{y}_{\bullet l}-\overline{y}_{\bullet\bullet}$. Immers,
   - de voorspelling $\widehat{y}^{A,B}_{kli}$ is gelijk aan de kleinste-kwadratenschatter van de som $\mu+\alpha_k+\beta_l$ in het twee-factoriële model met A en B, en

- in een gebalanceerd design, zijn de kleinste-kwadratenschatters van de parameters $\mu$, $\alpha_k$ en $\beta_l$ de volgende:

$$\begin{aligned} \widehat{\mu} &= \overline{y}_{\bullet\bullet} \\ \widehat{\alpha}_k &= \overline{y}_{k\bullet} - \overline{y}_{\bullet\bullet} \\ B_l &= \overline{y}_{\bullet l} - \overline{y}_{\bullet\bullet} \quad , \end{aligned}$$

- zodat

$$\begin{aligned} \widehat{y}_{kli}^{A,B} &= \widehat{\mu} + \widehat{\alpha}_k + B_l \\ &= \overline{y}_{\bullet\bullet} + (\overline{y}_{k\bullet} - \overline{y}_{\bullet\bullet}) + (\overline{y}_{\bullet l} - \overline{y}_{\bullet\bullet}) \\ &= \overline{y}_{k\bullet} + \overline{y}_{\bullet l} - \overline{y}_{\bullet\bullet} \quad . \end{aligned}$$

2. Het gewogen gemiddelde $\overline{y}_{\sqcup l}$ is gelijk aan het corresponderende ongewogen gemiddelde $\overline{y}_{\bullet l}$.

3. Het aantal observaties per cel, $n_{kl}$, is gelijk voor alle cellen en wordt daarom vervangen door $n$. Het aantal observaties in het k-de niveau van A is daarom gelijk aan het aantal niveaus van B (i.e., $L$) maal $n$ (afgekort, $Ln$).

Voor een gebalanceerd design kan $SS_{A|B}$ daarom als volgt geschreven worden:

$$\begin{aligned} SS_{A|B} &= \sum_{Obs} [(\overline{y}_{k\bullet} + \overline{y}_{\bullet l} - \overline{y}_{\bullet\bullet}) - \overline{y}_{\bullet l}]^2 \\ &= \sum_{k=1}^{K} \sum_{l=1}^{L} \sum_{i=1}^{n} [\overline{y}_{k\bullet} - \overline{y}_{\bullet\bullet}]^2 \quad . \end{aligned}$$

Merk op dat $\overline{y}_{k\bullet} - \overline{y}_{\bullet\bullet}$ (het verschil dat gekwadrateerd wordt in de bovenstaande formule) gelijk is aan de kleinste-kwadratenschatter van de effectparameter $\alpha_k$. Omdat $[\overline{y}_{k\bullet} - \overline{y}_{\bullet\bullet}]^2$ niet afhangt van $l$ en $i$, kan $SS_{A|B}$ als volgt geschreven worden:

$$SS_{A|B} = \sum_{k=1}^{K} Ln(\overline{y}_{k\bullet} - \overline{y}_{\bullet\bullet})^2 \quad .$$

## Afleiding van de Formule van de Type III Kwadratensom voor een Gebalanceerd Design

We vertrekken van de formule van de type III kwadratensom van factor A die zowel geldig is voor een gebalanceerd als een ongebalanceerd design:

$$SS_{A|B,A\times B} = \sum_{Obs} (\overline{y}_{kl} - \widehat{y}_{kli}^{B,A\times B})^2 \quad .$$

Deze formule kan herschreven worden voor een gebalanceerd design. In een gebalanceerd design geldt namelijk het volgende:

1. De voorspelling $\widehat{y}_{kli}^{B,A\times B}$ is gelijk aan $\overline{y}_{kl} - \overline{y}_{k\bullet} + \overline{y}_{\bullet\bullet}$. Immers,

   - de voorspelling $\widehat{y}_{kli}^{B,A\times B}$ is gelijk aan de kleinste-kwadratenschatter van de som $\mu + \beta_l + \alpha\beta_{kl}$ in het referentiemodel met het hoofdeffect van factor B en het interactie-effect en,
   - in een gebalanceerd design, zijn de kleinste-kwadratenschatters van de parameters $\mu$, $\beta_l$ en $\alpha\beta_{kl}$ de volgende:

$$\begin{aligned} \widehat{\mu} &= \overline{y}_{\bullet\bullet} \\ B_l &= \overline{y}_{\bullet l} - \overline{y}_{\bullet\bullet} \\ \widehat{\alpha\beta}_{kl} &= \overline{y}_{kl} - \overline{y}_{k\bullet} - \overline{y}_{\bullet l} + \overline{y}_{\bullet\bullet} \quad , \end{aligned}$$

   - zodat

$$\begin{aligned} \widehat{y}_{kli}^{B,A\times B} &= \widehat{\mu} + B_l + \widehat{\alpha\beta}_{kl} \\ &= \overline{y}_{\bullet\bullet} + (\overline{y}_{\bullet l} - \overline{y}_{\bullet\bullet}) + (\overline{y}_{kl} - \overline{y}_{k\bullet} - \overline{y}_{\bullet l} + \overline{y}_{\bullet\bullet}) \\ &= \overline{y}_{kl} - \overline{y}_{k\bullet} + \overline{y}_{\bullet\bullet} \quad . \end{aligned}$$

2. Het aantal observaties per cel, $n_{kl}$, is gelijk voor alle cellen en wordt daarom vervangen door $n$. Het aantal observaties in het k-de niveau van A is daarom gelijk aan het aantal niveaus van B (i.e., $L$) maal $n$ (afgekort, $Ln$).

Voor een gebalanceerd design kan $SS_{A|B,A\times B}$ daarom als volgt geschreven worden:

$$\begin{aligned} SS_{A|B,A\times B} &= \sum_{Obs} [\overline{y}_{kl} - (\overline{y}_{kl} - \overline{y}_{k\bullet} + \overline{y}_{\bullet\bullet})]^2 \\ &= \sum_{k=1}^{K} \sum_{l=1}^{L} \sum_{i=1}^{n} [\overline{y}_{k\bullet} - \overline{y}_{\bullet\bullet}]^2 \quad . \end{aligned}$$

Merk op dat $\overline{y}_{k\bullet} - \overline{y}_{\bullet\bullet}$ (het verschil dat gekwadrateerd wordt in de bovenstaande formule) gelijk is aan de kleinste-kwadratenschatter van de effectparameter $\alpha_k$. Omdat $[\overline{y}_{k\bullet} - \overline{y}_{\bullet\bullet}]^2$ niet afhangt van $l$ en $i$, kan $SS_{A|B,A\times B}$ als volgt geschreven worden:

$$SS_{A|B,A\times B} = \sum_{k=1}^{K} Ln(\overline{y}_{k\bullet} - \overline{y}_{\bullet\bullet})^2 \quad .$$

# Hoofdstuk 4

# Regressieanalyse

## Inhoudsopgave


**4.1 Inleiding . . . . . 297**

4.1.1 Multivariate Analyse . . . . . 297

4.1.2 Regressieanalyse . . . . . 298

4.1.3 Voorbeeld . . . . . 301

**4.2 Voorspellen Zonder Predictoren . . . . . 302**

4.2.1 Kleinste-kwadratenschatting . . . . . 302

4.2.2 De Gemiddelde Kwadratische Fout . . . . . 303

4.2.3 Voorbeeld . . . . . 305

**4.3 Voorspellen met één Predictor: Enkelvoudige Regressieanalyse . . . . . 306**

4.3.1 Enkelvoudige Regressieanalyse voor Gestandaardiseerde Variabelen . . . . . 306

4.3.2 Enkelvoudige Regressieanalyse voor Ongestandaardiseerde Variabelen . . . . . 308

4.3.3 Verklaarde en Onverklaarde Variantie . . . . . 309

4.3.4 De Proporties Verklaarde en Onverklaarde Variantie . . . . . 310

4.3.5 De Standaardschattingsfout van de Regressie . . . . . 311

4.3.6 Een Zuivere Schatter van de Gemiddelde Kwadratische Fout in de Populatie . . . . . 311

4.3.7 Voorbeeld . . . . . 312

**4.4 Voorspellen met Twee Predictoren: Meervoudige Regressieanalyse . . . . . 318**

4.4.1 Meervoudige Regressieanalyse voor Gestandaardiseerde Variabelen . . . . . 319

4.4.2 Meervoudige Regressieanalyse voor Ongestandaardiseerde Variabelen . . . . . 321

4.4.3 De Meervoudige Correlatiecoëfficiënt . . . . . 322

4.4.4 Voorbeeld . . . . . 325

**4.5 Voorspellen met Willekeurig Veel Predictoren . . . . . 328**
4.5.1 Theorie . . . . . 328
4.5.2 Voorbeeld . . . . . 331
**4.6 De Semi-Partiële of Partcorrelatiecoëfficiënt . . . . . 333**
4.6.1 Theorie . . . . . 333
4.6.2 Voorbeeld . . . . . 336
**4.7 Enige Problemen bij de Interpretatie van Regressiegewichten 341**
4.7.1 Onechte Correlatie . . . . . 341
4.7.2 Suppressie . . . . . 344
4.7.3 Collineariteit . . . . . 348
**4.8 De Statistische Significantie van Regressiegewichten . . . 351**
**4.9 Het Selecteren van Predictoren . . . . . 362**
**4.10 Uitvoeren van een Regressieanalyse met de SPSS-procedure Regression . . . . . 365**
**4.11 Hiërarchische Meervoudige Regressieanalyse . . . . . 371**
4.11.1 HMRA binnen het Onderzoek naar de Rol van Storende Variabelen . . . . . 372
4.11.2 HMRA binnen het Onderzoek naar de Rol van Mediërende Variabelen . . . . . 376
**4.12 Regressieanalyse met Categorische Predictoren . . . . . 378**
4.12.1 Voorbeeld . . . . . 378
4.12.2 De Regressievergelijking . . . . . 379
4.12.3 Statistische Significantie . . . . . 381
4.12.4 Het Algemene Lineaire Model . . . . . 387
4.12.5 Groepsspecifieke Regressiegewichten . . . . . 389
4.12.6 Een Overzicht van de Verschillende Types F-toetsen binnen het Algemene Lineaire Model . . . . . 393
4.12.7 Dummycodering . . . . . 398
4.12.8 Uitvoeren van een Regressieanalyse met de SPSS-procedure GLM-Univariate . . . . . 403
**4.13 Regressieanalyse met Interactie Tussen Twee Kwantitatieve Predictoren . . . . . 411**
4.13.1 Voorbeeld . . . . . 411
4.13.2 De Regressievergelijking met een Interactievariabele . . . . . 415
4.13.3 Inhoudelijke Interpretatie van de Binnen-score Regressielijnen 419
4.13.4 De Regressievergelijking met Niet-gecentreerde Predictoren . 420
4.13.5 Het Belang van de Regressieanalyse met Interactie Tussen Twee Kwantitatieve Predictoren . . . . . 421
**4.14 Logistische Regressieanalyse . . . . . 423**
4.14.1 Kansen in plaats van Scores . . . . . 424
4.14.2 Het Berekenen van de Logistische Regressiecoëfficiënten . . . 427
4.14.3 Eigenschappen van de Logistische Regressiecoëfficiënten . . . 429
4.14.4 De Kwaliteit van de Voorspelling op basis van een Logistische Regressieanalyse . . . . . 429

4.14.5 Het Evalueren van de Effectgrootte . . . . . 433
4.14.6 Statistische Toetsen . . . . . 438
4.14.7 Uitvoeren van een Logistische Regressieanalyse met de SPSS-procedure Logistic Regression . . . . . 440
**4.15 Appendix bij Hoofdstuk 4 . . . . . 452**

# 4.1 Inleiding

## 4.1.1 Multivariate Analyse

Onder de noemer multivariate analyse vallen allerlei technieken voor data-analyse waarbij de scores van personen (of andere onderzoekseenheden) op verscheidene variabelen tegelijkertijd worden bekeken en waar met name wordt bestudeerd hoe deze scores op de verschillende variabelen met elkaar samenhangen. We zullen hier enkele van zulke multivariate analysetechnieken bekijken. Voorlopig concentreren we ons op kwantitatieve variabelen, alhoewel we later ook enige aandacht zullen besteden aan categorische variabelen. De onderlinge samenhang van kwantitatieve variabelen wordt meestal uitgedrukt in de Pearson- of productmoment-correlatiecoëfficiënt.

Gegeven een geconstateerde samenhang tussen de scores op verschillende variabelen, weergegeven in een matrix van onderlinge correlaties, kunnen we de volgende doelen hebben:

1. We kunnen proberen deze samenhang te benutten, in de zin dat we willen onderzoeken hoe de score van een willekeurige persoon op een van de variabelen zo goed mogelijk te voorspellen valt uit de scores van diezelfde persoon op de andere variabelen. Dit aspect komt aan de orde in de (meervoudige) regressieanalyse, het onderwerp van dit hoofdstuk.

   Zó geformuleerd, lijkt het erop dat men met de regressieanalyse een louter praktisch doel nastreeft, het zo goed mogelijk voorspellen van een variabele los van een inhoudelijke vraag die men hiermee probeert te beantwoorden. Dit klopt niet, want vaak heeft men bij het toepassen van regressieanalyse een inhoudelijk doel voor ogen, namelijk om voor een bepaald psychologisch construct te achterhalen door welke van een aantal andere constructen dit wordt gedetermineerd. Dit onderzoekt men dan door de theoretische concepten te operationaliseren in observeerbare variabelen en die vervolgens in een regressieanalyse op te nemen. Met deze regressieanalyse onderzoekt men dan hoe de operationalisering van één psychologisch construct voorspeld kan worden door de operationaliseringen van de andere constructen.

2. We kunnen een bepaald aspect van deze samenhang isoleren en eventueel statistisch toetsen. Het meest voorkomend is de vraag of een bepaalde variabele een bijdrage levert aan de voorspelling van een andere variabele *bovenop een aantal andere variabelen*. Van deze laatste groep variabelen zegt men ook dat men ervoor *uitzuivert*. Deze vraag is de kern van veel onderzoeksvragen die men in de literatuur aantreft

(o.a. in wetenschappelijke tijdschriften, proefschriften en afstudeerscripties). Het beantwoorden van deze vraag behoort ook tot de regressieanalyse.

3. We kunnen proberen de geconstateerde samenhang te verklaren door te veronderstellen dat al de geobserveerde variabelen op een bepaalde manier zijn opgebouwd uit een veel kleiner aantal onderliggende, niet-observeerbare variabelen, de zogenaamde *factoren*. De waargenomen correlaties worden dan verklaard uit de veronderstelling dat alle variabelen deze zelfde factoren gemeenschappelijk hebben. Deze benadering wordt verder uitgewerkt in een variant van de *factoranalyse*, de factoranalyse met ongecorreleerde residuen, die in het volgende hoofdstuk behandeld wordt.

4. In plaats van de correlaties tussen de variabelen, kunnen we ook proberen zoveel mogelijk variantie van deze variabelen te verklaren door een veel kleiner aantal factoren. Met andere woorden, in plaats van te zoeken naar factoren die zo goed mogelijk de correlaties verklaren, zoeken we nu naar factoren die zoveel mogelijk variantie van deze variabelen verklaren. Deze benadering wordt verder uitgewerkt in een andere variant van de factoranalyse, de *hoofdcomponentenanalyse*, die ook in het volgende hoofdstuk behandeld wordt.

### 4.1.2 Regressieanalyse

In zijn zuivere vorm, houdt de regressieanalyse zich bezig met een probleem van zeer praktische aard: hoe zijn scores op één variabele het beste te voorspellen op basis van scores op andere variabelen? We zullen nu eerst heel kort dit probleem wat concreter formuleren en daarbij alvast wat notatie en terminologie introduceren. In volgende paragrafen wordt dit alles meer in detail uitgewerkt, waarbij we met het allereenvoudigste geval zullen beginnen (voorspellen zónder predictoren) en slechts stap voor stap (voorspellen met één en met twee predictoren) tot de behandeling van het algemene probleem komen (voorspellen met willekeurig veel predictoren).

De ene variabele, waarvan we met behulp van een regressieanalyse de scores willen voorspellen wordt de *criterium*- of ook wel *afhankelijke* variabele genoemd. Deze wordt hier algemeen aangeduid met de letter $Y$. De variabelen die gebruikt worden om het criterium te voorspellen heten de *predictor*- of ook wel *onafhankelijke* variabelen[1]. Zij worden hier algemeen aangeduid met de letter $X$ als er slechts één predictor is, en genummerde $X_1, X_2, \ldots, X_m$ als er verscheidene, hier dus $m$, predictoren in het spel zijn.

Het doel van de regressieanalyse is de score op het criterium $Y$ te voorspellen voor personen waarvoor deze score onbekend is, maar waarvan wel scores op de predictoren $X_1, X_2, \ldots, X_m$ gegeven zijn. Om zo'n voorspelling ergens op te baseren is het nodig

[1] De termen afhankelijke en onafhankelijke variabele zijn bekend uit het effectonderzoek. Merk op dat het gebruik van deze termen hier echter ruimer (eigenlijk, minder precies) is. In een effectonderzoek ligt het vast welke de onafhankelijke (idealiter, door de experimentator gemanipuleerde) variabele is en welke de afhankelijke (de door de respons van proefpersonen bepaalde). In de regressiecontext is de rolverdeling in onafhankelijke en afhankelijke variabelen echter vrij te kiezen en kan deze ook wisselen: de afhankelijke variabele in de ene analyse kan een onafhankelijke variabele zijn in een andere analyse en omgekeerd. Wij houden in het vervolg de meer directe termen criterium en predictor aan.

dat we over een zekere steekproef van, zeg $n$, personen beschikken waarvan we wél scores op zowel de criteriumvariabele als op de predictorvariabelen hebben waargenomen. Met name zijn dan in zo'n steekproef de volgende overzichtsstatistieken gegeven (of te berekenen):

1. gemiddelden $\overline{Y}$, $\overline{X}_1$, $\overline{X}_2$, ... van alle variabelen.
2. standaarddeviaties $\mathrm{SD}(Y)$, $\mathrm{SD}(X_1)$, $\mathrm{SD}(X_2)$, ... van alle variabelen.
3. correlaties $r_{y1}$, $r_{y2}$, $r_{y3}$, ... tussen criterium- en predictorvariabelen.
4. correlaties $r_{12}$, $r_{13}$, $r_{23}$, $r_{14}$, ... tussen de predictorvariabelen onderling.

Om de notatie zo eenvoudig mogelijk te houden schrijven we $r_{y1}$ in plaats van het meer volledige $r_{yx_1}$ voor de correlatie tussen $Y$ en $X_1$, enz. Evenzo schrijven we $r_{12}$ in plaats van het meer volledige $r_{x_1x_2}$ voor de correlatie tussen $X_1$ en $X_2$. Kortweg, in subscripten staan cijfers voor de corresponderende $X$-variabelen. In Tabel 4.1 wordt een voorbeeld gegeven van dergelijke data, met bijbehorende overzichtsstatistieken. Dit voorbeeld wordt aan het eind van deze paragraaf nader geïntroduceerd.

Op basis van deze gegevens moeten we een regel bepalen om voor personen buiten de steekproef, waarvan alleen scores op de predictoren $X_1$ t/m

$X_m$ bekend zijn, zo goed mogelijk een score op $Y$ te voorspellen. Als we zo'n voorspelde $Y$-score aanduiden met $\widehat{Y}$, dus met een $\widehat{\ }$ boven de variabelennaam, dan komt dit er dus op neer dat $\widehat{Y}$ bepaald moet worden als een functie van de $X$-scores:

$$\widehat{Y} = f(X_1, \ldots, X_m) \quad .$$

Zo'n $\widehat{Y}$ heet dan de regressie van $Y$ op $X_1$ t/m $X_m$. De algemene functie $f$ kan hier in principe een willekeurig ingewikkelde vorm hebben, maar deze vrijheid gaan we meteen drastisch inperken: we zullen alleen lineaire functies $f$ beschouwen. Concreet wil dit zeggen dat de algemene voorspellingsregel de volgende vorm heeft:

$$\boxed{\widehat{Y} = A + B_1\, X_1 + B_2\, X_2 + \cdots + B_m\, X_m \quad .}$$

Een zo bepaalde $\widehat{Y}$ heet dan ook de *lineaire* regressie van $Y$ op $X_1$ t/m $X_m$. Hierin blijft dus als enige vrijheid over het kiezen van de coëfficiënten $A$, $B_1, B_2, \ldots, B_m$. Aangezien de gegevens in de steekproef onze enige basis vormen voor deze voorspelling, kiezen we ervoor deze coëfficiënten zó te kiezen dat de voorspelling optimaal is voor de personen in de steekproef (waarvoor we de voorspelling immers kunnen vergelijken met werkelijk geobserveerde waarden). De optimale bepaling van $A$ en $B_1$ t/m $B_m$ in termen van de steekproefgegevens is dan ook de centrale vraag in het vervolg. Direct hieraan gekoppeld wordt echter ook het antwoord gezocht op een tweede vraag, namelijk hoe goed, met een optimale keuze, de voorspelling het dan wel doet in de steekproef.

Deze twee vragen, naar de optimale voorspelling en de kwaliteit ervan, zullen we echter niet meteen behandelen voor het algemene geval met verschillende, zeg $m$, predictoren dat we hierboven hebben geschetst. In plaats daarvan gaan we stapsgewijs van het meest eenvoudige geval naar dit algemene geval toewerken. We beginnen in de

| pp | VLG | EM | ZB | AC |
|---|---|---|---|---|
| 1 | 9 | 41 | 20 | 40 |
| 2 | 4 | 55 | 22 | 35 |
| 3 | 11 | 42 | 20 | 44 |
| 4 | 10 | 48 | 20 | 34 |
| 5 | 6 | 36 | 21 | 39 |
| 6 | 10 | 44 | 22 | 42 |
| 7 | 12 | 44 | 21 | 36 |
| 8 | 16 | 32 | 26 | 39 |
| 9 | 12 | 30 | 22 | 45 |
| 10 | 12 | 38 | 21 | 34 |
| 11 | 13 | 46 | 21 | 36 |
| 12 | 7 | 46 | 19 | 37 |
| 13 | 14 | 39 | 22 | 39 |
| 14 | 9 | 54 | 22 | 39 |
| 15 | 16 | 43 | 26 | 34 |
| 16 | 12 | 51 | 22 | 43 |
| 17 | 15 | 41 | 22 | 41 |
| 18 | 10 | 48 | 20 | 41 |
| 19 | 6 | 50 | 22 | 33 |
| 20 | 4 | 57 | 21 | 32 |
| 21 | 13 | 41 | 18 | 33 |
| 22 | 11 | 36 | 22 | 35 |
| 23 | 13 | 43 | 22 | 38 |
| 24 | 8 | 40 | 25 | 32 |
| 25 | 8 | 41 | 20 | 42 |
| 26 | 8 | 46 | 20 | 33 |
| 27 | 11 | 39 | 22 | 31 |
| 28 | 12 | 40 | 21 | 41 |
| 29 | 11 | 38 | 19 | 35 |
| 30 | 15 | 34 | 23 | 40 |
| 31 | 10 | 52 | 22 | 40 |
| 32 | 9 | 46 | 19 | 39 |
| Gem | 10.53 | 43.16 | 21.41 | 37.56 |
| Var | 9.62 | 41.51 | 3.24 | 14.37 |
| SD | 3.10 | 6.44 | 1.80 | 3.79 |

| Corr. | VLG | EM | ZB |
|---|---|---|---|
| EM | -0.55 | | |
| ZB | 0.35 | -0.20 | |
| AC | 0.24 | -0.20 | -0.04 |

Tabel 4.1: Voorbeeld van een personen × variabelen matrix van steekproefgegevens, met gemiddelden, varianties, standaarddeviaties en correlaties. Zie tekst.

volgende paragraaf zelfs met het triviaal lijkende geval van nul predictoren: welke $Y$-score voorspellen we zonder enige informatie over scores op andere variabelen? Daarna beschouwen we achtereenvolgens het voorspellingsprobleem gegeven één predictor $X$, gegeven twee predictoren $X_1$ en $X_2$, en tenslotte gegeven een willekeurig aantal van $m$ predictoren $X_1, X_2, \ldots, X_m$. Deze opbouw zal duidelijk maken dat we in al deze situaties dezelfde principes toepassen. Bij het uitwerken van deze principes treden er met het toenemen van het aantal predictoren een aantal complicaties op. Deze zullen in het volgende dan ook een voor een aan de orde komen.

### 4.1.3 Voorbeeld

De behandelde theorie zal in de volgende paragrafen geïllustreerd worden aan de hand van data van een onderzoek van Stalmeier, Unic, Verhoef en van Daal (1999). Deze data worden getoond in Figuur 4.1. Het betreft hier de scores van 32 proefpersonen op vier vragenlijsten. De proefpersonen zijn mannen bij wie een tumor aan de stembanden is geconstateerd en die de keuze hebben uit twee behandelingen: een ingrijpende maar veilige behandeling (laryngectomie-operatie), of een niet-ingrijpende behandeling met blijvend risico (radiotherapie). Om tot een gefundeerde keuze te komen doorliepen deze patiënten een programma waarin zij enerzijds voorlichting kregen over de alternatieven en anderzijds voorkeursoordelen gaven over allerlei kwesties waarin kwaliteit en kwantiteit van resterende levensjaren tegen elkaar moesten worden afgewogen. Op basis van deze gegevens stelden medisch-besliskundigen dan per patiënt een keuzeadvies op. De vier vragenlijsten in Figuur 4.1 hebben betrekking op een evaluatieonderzoek van dit programma, uitgevoerd na afloop van het programma, maar vóór het advies is uitgebracht. De vragenlijst VLG vraagt in hoeverre men geneigd zal zijn in zijn uiteindelijke beslissing het te verstrekken advies te volgen (maximumscore van 20 punten). Dit nemen we als de criteriumvariabele. Als mogelijke predictoren hiervoor beschouwen we de vragenlijst EM, die vraagt naar de emotionele belasting tengevolge van het doorlopen van het programma (maximumscore van 63); de vragenlijst ZB, die evalueert in welke mate men vindt dat men via het doorlopen programma eigenlijk zelf beslist (maximumscore van 28); en de vragenlijst AC, die meet in hoeverre men de doorlopen procedure als een acceptabele of zelfs tevredenstemmende manier beschouwt om tot een beslissing te komen (maximumscore van 56). In de correlatiematrix onderaan Figuur 4.1 zien we dat de samenhang tussen deze variabelen de richting heeft die we verwachten: men is des te meer geneigd het te verstrekken advies te volgen naarmate men de doorlopen procedure emotioneel als minder belastend heeft ervaren ($r_{\mathrm{VLG,EM}} = -0.548$), naarmate men deze meer ziet als het bevragen van de eigen mening waardoor men eigenlijk zelf beslist ($r_{\mathrm{VLG,ZB}} = 0.353$), en naarmate men deze meer acceptabel vindt ($r_{\mathrm{VLG,AC}} = 0.243$). Verder correleert de mate van emotionele belasting (licht) negatief met ZB ($r_{\mathrm{EM,ZB}} = -0.197$) en AC ($r_{\mathrm{EM,AC}} = -0.204$), terwijl deze laatste twee variabelen vrijwel ongecorreleerd zijn ($r_{\mathrm{ZB,AC}} = -0.043$).

## 4.2 Voorspellen Zonder Predictoren

We beginnen met het meest eenvoudige geval, waarin we $Y$-scores moeten voorspellen als er helemaal geen predictorscores voorhanden zijn. De voorspellingsregel $\widehat{Y}$ kan dus geen gebruik maken van enige persoonspecifieke informatie. Dan is het duidelijk dat we voor iedere willekeurige persoon $i$ dezelfde score zullen voorspellen. Dit wil zeggen, de voorspelling is van de vorm

$$\widehat{Y}_i = A$$

voor een vast getal $A$. Dit getal moeten we bepalen aan de hand van de verdeling van $Y$-scores in de steekproef, de enige informatie waarover we beschikken. De twee vragen die we hier dus moeten beantwoorden zijn: (1) welke keuze voor het getal $A$ is de beste en (2) hoe goed wordt de voorspelling met zo'n beste keuze voor $A$?

### 4.2.1 Kleinste-kwadratenschatting

Om te kunnen zeggen welke keuze voor $A$ de beste is moeten we over een criterium[2] beschikken op basis waarvan we de kwaliteit van voorspellingen heel concreet kunnen vergelijken. We gaan hier gebruikmaken van het *kleinste-kwadratencriterium*. Dit criterium hebben we in hoofdstuk 3 gebruikt om voorspellingen te maken voor de verschillende variantie- en covariantieanalysemodellen.

Om iets te zeggen over de kwaliteit van voorspellingen, ligt voor de hand te kijken naar de zogenaamde *residuen* van de voorspelling in de steekproef. Dit zijn de verschillen tussen de geobserveerde en de voorspelde scores van de personen in de steekproef. Dus, voor persoon $i$:

$$\text{Residu}_i = Y_i - \widehat{Y}_i \quad .$$

Een voorspellingsregel $\widehat{Y}$ is goed wanneer deze verschillen over het geheel genomen klein zijn, en slecht wanneer deze verschillen groot zijn. We willen dus met onze voorspellingsregel de gemiddelde grootte van de residuen zo klein mogelijk maken. Letterlijk genomen, zou dit neerkomen op het minimaliseren over alle proefpersonen $i$ van

$$|Y_i - \widehat{Y}_i| \quad .$$

Wiskundig, blijkt het echter veel handiger te zijn om, in plaats van deze absolute waarde, het gekwadrateerde verschil

$$(Y_i - \widehat{Y}_i)^2$$

als de afwijking te definiëren die we moeten minimaliseren. Deze overgang is redelijk omdat het zowel bij de absolute waarde als bij het kwadraat van het residu om de grootte gaat, afgezien van het teken; de abolute waarde is klein als het kwadraat klein is, en vice versa. Het effect van het kwadraat nemen is dat grotere fouten relatief wat zwaarder worden gewogen. De keuze voor deze kwadratische afwijking als indicatie voor de fout in de voorspelling leidt tot het volgende zeer algemene criterium voor een beste voorspelling:

[2] Het woord "criterium" wordt hier gebruikt in de betekenis van een formele voorwaarde waaraan het getal $A$ moet voldoen, en niet in de betekenis van een variabele die voorspeld wordt op basis van een aantal andere variabelen.

**Kleinste-Kwadratencriterium:**
Beschouw voor elke voorspellingsregel $\widehat{Y}$ de residuen $Y - \widehat{Y}$ in de steekproef en kies dan $\widehat{Y}$ zó dat het kwadraat van deze residuen gemiddeld zo klein mogelijk wordt. Dus, maak

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 \qquad \text{minimaal} \quad .$$

Dit kleinste-kwadratencriterium zullen we in het vervolg meerdere keren gebruiken om een optimale voorspellingsregel $\widehat{Y}$ te definiëren. Laten we nu bekijken wat het in het zeer eenvoudige geval zonder predictoren oplevert. We hebben al gezien dat hier de voorspelling de vorm $\widehat{Y}_i = A$ heeft, waarin $A$ gewoon een getal is dat we volgens bovenstaand criterium zo moeten bepalen dat

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - A)^2 \qquad \text{minimaal}$$

wordt. Als we deze voorwaarde uitwerken blijkt de uitkomst te zijn: kies

$$A = \overline{Y} \quad ,$$

het gemiddelde van de $Y$-scores. (De afleiding van dit resultaat wordt gegeven in de appendix op pagina 452.) Dus, bij gebrek aan predictoren is de beste strategie om voor een willekeurige persoon de gemiddelde $Y$-score te voorspellen:

$$\widehat{Y}_i = \overline{Y} \quad .$$

### 4.2.2 De Gemiddelde Kwadratische Fout

De tweede vraag die we nu nog moeten beantwoorden betreft de kwaliteit van deze beste oplossing. Een voor de hand liggende maat hiervoor is de grootte van de gemiddelde kwadratische fout die we via ons criterium hebben geminimaliseerd. Welnu, als we als voorspelling $\widehat{Y} = \overline{Y}$ invullen, komt deze gemiddelde kwadratische fout uit op

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \quad ,$$

de definitie van de variantie van de $Y$-scores, verder aangeduid als $\text{VAR}(Y)$. Dus, naarmate de $Y_i$'s korter bij het gemiddelde $\overline{Y}$ liggen, is $\text{VAR}(Y)$ kleiner.

Een nadeel van de variantie is dat hij de voorspellingsfouten niet op dezelfde schaal uitdrukt als de observaties $Y_i$ zelf. Dit blijkt uit het feit dat, indien de observaties met een constante $c$ vermenigvuldigd worden, de variantie $c^2$ keer groter wordt. (De afleiding van dit resultaat wordt gegeven in de appendix op pagina 453.) Daarom, willen we deze index voor de kwaliteit van de voorspelling uitdrukken op dezelfde schaal als die van de

$Y$-variabele, dan moeten we het kwadrateren weer wegwerken door de wortel te trekken. Het resultaat is dan dat we als maat voor de gemiddelde fout bij de voorspelling zonder predictor vinden:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2} \quad ,$$

de standaarddeviatie van de $Y$-scores, verder aangeduid als $\mathrm{SD}(Y)$.

De variantie $\mathrm{VAR}(Y)$ heeft nog een tweede nadeel: hij is een *onzuivere schatter* van de populatievariantie. Dit vraagt om enige toelichting. Het uitgangspunt bij het begrip zuiverheid is de mogelijkheid om de $n$ observaties $Y_i$ op te vatten als een steekproef uit een (veel) grotere populatie. De variantie in deze populatie duiden we aan met $\sigma^2$ (de Griekse letter $\sigma$ spreekt men uit als "sigma"). Men hecht meer waarde aan $\sigma^2$ dan aan $\mathrm{VAR}(Y)$ omdat $\sigma^2$ een maat is voor de kwaliteit van de voorspelling in de gehele populatie, terwijl $\mathrm{VAR}(Y)$ alleen de kwaliteit van de voorspelling in de steekproef uitdrukt. Daarom ligt het voor de hand om $\mathrm{VAR}(Y)$ te evalueren als schatter van $\sigma^2$: we vinden $\mathrm{VAR}(Y)$ een goede schatter als hij meestal kort bij $\sigma^2$ ligt en een slechte schatter als hij er meestal ver vanaf ligt. Nu, een nadeel van $\mathrm{VAR}(Y)$ is dat hij $\sigma^2$ systematisch ónderschat. Dit wil zeggen, bij herhaalde trekkingen van een steekproef van omvang $n$ uit de populatie, is $\mathrm{VAR}(Y)$ gemiddeld genomen (over deze herhaalde trekkingen) kleíner dan $\sigma^2$. Een schatter die gemiddeld genomen gelíjk is aan $\sigma^2$ is zuiver; een schatter die gemiddeld genomen kleiner of groter is dan $\sigma^2$ is ónzuiver.

We staan nu even stil bij de reden waarom $\mathrm{VAR}(Y)$ een onzuivere schatter is. Stel dat we de verwachte waarde[3] van de variabele $Y$ kennen en aanduiden met $\mu$ (spreek uit "mu"). Men kan aantonen dat de volgende schatter zuiver is:

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu)^2 \quad .$$

Deze schatter verschilt van $\mathrm{VAR}(Y)$ doordat het steekproefgemiddelde $\overline{Y}$ is vervangen door de verwachte waarde $\mu$ (ook het populatiegemiddelde genoemd). Nu, van alle mogelijke waarden, ligt het steekproefgemiddelde $\overline{Y}$ het dichtst bij de $n$ observaties $Y_i$, zoals uitgedrukt in het kleinste-kwadratencriterium. Dit impliceert dat, indien we in de bovenstaande zuivere schatter de verwachte waarde $\mu$ vervangen door het steekproefgemiddelde $\overline{Y}$, en dus de formule van $\mathrm{VAR}(Y)$ krijgen, we een kleiner getal krijgen dan eerst. Dit verklaart het feit dat $\mathrm{VAR}(Y)$ de onbekende $\sigma^2$ systematisch ónderschat.

Er bestaat een zeer eenvoudige manier om $\mathrm{VAR}(Y)$ te corrigeren voor deze systematische onderschatting: in plaats van de som $\sum_{i=1}^{n}(Y_i - \overline{Y})^2$ te delen door de steekproefomvang $n$ (of te vermenigvuldigen met $\frac{1}{n}$), delen we ze door[4] $n-1$. Het resultaat,

$$\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 \quad ,$$

[3]De verwachte waarde van $Y$ is het gemiddelde van $Y$ berekend over alle personen (onderzoeksobjecten) in de populatie. Daarom wordt deze verwachte waarde ook vaak het *populatiegemiddelde* genoemd.

[4]Uitleggen waarom het precies $n-1$ moet zijn, en niet $n-0.5$ of $n-3.7$, zou ons te ver voeren.

is een zuivere schatter van $\sigma^2$. Toch blijkt het handig te zijn om voorlopig verder te werken met de $\text{VAR}(Y)$ die berekend wordt met de steekproefomvang $n$ in de noemer. Echter, wanneer we het hebben over het statistisch toetsen van regressiegewichten, dan is het van belang om de kwaliteit van de voorspelling uit te drukken in een formule waarin niet gedeeld wordt door $n$ maar door een kleiner getal. We komen hier nog op terug.

Samenvattend vinden we dus bij het voorspellen zonder predictoren de bekende grootheden gemiddelde, variantie en standaarddeviatie terug. Meer theoretisch is het belang van bovenstaande exercitie dat we deze resultaten vinden door toepassing van een algemeen principe, het kleinste-kwadratencriterium, dat in het vervolg zal terugkeren. Meer praktisch gezien is van belang dat we hiermee expliciet een vergelijkingspunt hebben waartegen we de resultaten van voorspellingen op basis van een of meerdere predictoren kunnen afzetten. Met name zullen we het effect van predictoren op de kwaliteit van de voorspelling in het vervolg bepalen door de gemiddelde kwadratische fout van de voorspelling mét predictoren te vergelijken met de variantie van $Y$, die we nu immers kunnen interpreteren als de gemiddelde kwadratische fout die we maken zonder de hulp van enige predictor. Evenzo zal voor de voorspelling mét predictoren de voorspellingsfout, uitgedrukt op de schaal van de criteriumvariabele $Y$, worden vergeleken met de standaarddeviatie van $Y$, die, zoals we gezien hebben, deze fout meet bij de voorspelling zonder predictor.

### 4.2.3 Voorbeeld

In ons voorbeeld van Tabel 4.1 komt de in deze paragraaf beschreven situatie erop neer dat we voor een nieuwe patiënt, waarvoor we geen scores op de variabelen EM, ZB, of AC hebben, de score op VLG dienen te voorspellen. In Tabel 4.1 is dan alleen de in de eerste kolom gegeven verdeling van de VLG-scores in de steekproef van belang. Volgens het kleinste-kwadratencriterium is de beste voorspelling voor onze nieuwe patient een score van 10.53 punten op VLG, aangezien het steekproefgemiddelde $\overline{\text{VLG}}$ gelijk is aan 10.53. De kwadratische fout $(\text{VLG} - \widehat{\text{VLG}})^2$ die we met de voorspelling $\widehat{\text{VLG}} = 10.53$ punten mogen verwachten is gemiddeld gelijk aan $\text{VAR}(\text{VLG}) = 9.62$ punten in het kwadraat. Deze gemiddelde kwadratische fout is minimaal. Dit wil zeggen, als we voor een willekeurige patiënt iets anders voorspellen dan $\overline{\text{VLG}} = 10.53$ punten, dan moeten we een gemiddelde kwadratische fout verwachten die hoger uitvalt dan 9.62 punten in het kwadraat.

De rare eenheid van deze kwadratische fout, punten in het kwadraat, illustreert dat dit geen goede beschrijvende maat is voor de fout die we maken. Om die te verkrijgen moeten we de wortel uit dit getal nemen: $\sqrt{\text{VAR}(\text{VLG})} = \text{SD}(\text{VLG}) = 3.10$ punten. Dus, we mogen verwachten dat we er met deze voorspelling zo'n ruim 3 punten naast te zitten, aangezien in de steekproef $\text{SD}(\text{VLG})$ gelijk is aan 3.10 punten. Met andere woorden, hoewel een score van 10 à 11 punten onze beste gok is, zullen we helemaal niet verbaasd zijn als deze patiënt later een score zo laag als 7 of zo hoog als 14 punten blijkt te behalen op de variabele VLG.

## 4.3 Voorspellen met één Predictor: Enkelvoudige Regressieanalyse

Nu gaan we de meer interessante situatie bekijken waarin we wél over predictoren beschikken. We beginnen met het meest eenvoudige geval van een enkele predictor $X$ voor de criteriumvariabele $Y$. Nu willen we dus de voorspelling voor een persoon $i$, $\widehat{Y}_i$, laten afhangen van $X_i$, de score van persoon $i$ op de predictorvariabele. Deze afhankelijkheid, dit wil zeggen de manier waarop we $\widehat{Y}$ als functie van $X$ schrijven, kan in principe een willekeurig ingewikkelde vorm hebben, maar, zoals we al in de inleiding hebben gesteld, we gaan ons hier beperken tot de zeer eenvoudige vorm van een lineaire functie. Concreet komt dit erop neer dat de voorspelling $\widehat{Y}$, voor bepaalde vaste getallen $A$ en $B$, er als volgt uit ziet:

$$\boxed{\widehat{Y} = A + B\,X \quad .}$$

De coëfficiënt $A$ wordt ook vaak de *regressieconstante* of het *intercept* genoemd, en $B$ het *regressiegewicht voor* $X$. Zo'n voorspellingsregel heet de *(lineaire) regressievergelijking* van $Y$ op $X$.

Van de coëfficiënten $A$ en $B$, is het regressiegewicht $B$ het belangrijkst. Zoals we verder zullen zien, beschrijft $B$ de richting van het verband tussen $X$ en $Y$ (stijgend, dalend, of vlak). Onderzoeksvragen worden vaak geformuleerd in termen van de richting van het verband tussen twee variabelen, en daarom is het regressiegewicht $B$ een interessante grootheid. Dit regressiegewicht is op te vatten als de toename in de voorspelling $\widehat{Y}$ als de predictor $X$ met één eenheid toeneemt (i.e., van $X$ naar $X + 1$). De regressieconstante $A$ op haar beurt, is op te vatten als de voorspelling $\widehat{Y}$ die men maakt voor iemand die score van 0 heeft op de predictor $X$.

Met betrekking tot de bovenstaande regressievergelijking gaan we weer de bekende vragen stellen: welke keuze voor de getallen $A$ en $B$ is de beste en hoe goed is dan de voorspelling? Voor de keuze van de beste $A$ en $B$ passen we weer het kleinste-kwadratencriterium toe op de residuen $Y - \widehat{Y}$: we bepalen $A$ en $B$ zó dat, over de steekproef genomen,

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 \qquad \text{minimaal}$$

wordt. Invullen van de regressievergelijking voor $\widehat{Y}$ levert dan meer concreet op dat we moeten zorgen dat

$$\frac{1}{n}\sum_{i=1}^{n}[Y_i - (A + B\,X_i)]^2 \qquad \text{minimaal}$$

wordt door een geschikte keuze van de coëfficiënten $A$ en $B$.

### 4.3.1 Enkelvoudige Regressieanalyse voor Gestandaardiseerde Variabelen

De oplossing die voldoet aan het kleinste-kwadratencriterium blijkt eenvoudiger te zijn als zowel $X$ als $Y$ uitgedrukt worden in *standaardscores* $\mathbf{x}$ en $\mathbf{y}$. Standaardscores worden

berekend door (1) van de originele scores (hier, $X$ en $Y$) het gemiddelde (respectievelijk, $\overline{X}$ en $\overline{Y}$) af te trekken en (2) het resulterende verschil te delen door hun standaarddeviatie (respectievelijk, $\mathrm{SD}(X)$ en $\mathrm{SD}(Y)$):

$$\mathbf{x} = \frac{X - \overline{X}}{\mathrm{SD}(X)} \quad , \qquad \mathbf{y} = \frac{Y - \overline{Y}}{\mathrm{SD}(Y)} \quad .$$

Standaardscores op een variabele worden dus aangeduid met de corresponderende vetgedrukte kleine letter. Het resultaat dat we vinden voor standaardscores zullen we dan door substitutie van de bovenstaande definitie van de standaardscore weer terugvertalen naar een uitdrukking voor de oorspronkelijke variabelen $X$ en $Y$.

De gezochte regressievergelijking voor de gestandaardiseerde variabelen $\mathbf{x}$ en $\mathbf{y}$ schrijven we als

$$\boxed{\widehat{\mathbf{y}} = a + b\,\mathbf{x} \quad ,}$$

waarin de coëfficiënten worden aangeduid met kleine letters $a$ en $b$. Deze zijn nu zo te bepalen, dat

$$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \widehat{\mathbf{y}}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}[\mathbf{y}_i - (a + b\,\mathbf{x}_i)]^2$$

zo klein mogelijk wordt. Het uitwerken van deze voorwaarde leidt tot het volgende resultaat:

$$\boxed{\begin{aligned} a &= 0 \\ b &= r_{xy} \quad , \end{aligned}}$$

met $r_{xy}$ als de correlatie tussen predictor $X$ en criterium $Y$. Dit resultaat wordt afgeleid in de appendix op pagina 453. De gestandaardiseerde regressievergelijking krijgt dus de volgende zeer eenvoudige vorm:

$$\widehat{\mathbf{y}} = r_{xy}\mathbf{x} \quad .$$

Volgens deze regel wordt een standaardscore in $Y$ voorspeld door $r_{xy}$ maal de gegeven standaardscore in $X$.

Merk op dat $\widehat{\mathbf{y}}$ géén gestandaardiseerde variabele is. Dit komt omdat de variantie van een gestandaardiseerde variabele gelijk is aan 1 en de variantie van $\widehat{\mathbf{y}}$ bijna altijd kleiner is dan 1. Dit volgt uit het feit dat de variantie van een variabele (hier, de variabele $\mathbf{x}$) maal een constante (hier, $r_{xy}$) gelijk is aan het kwadraat van die constante (dus, $r_{xy}^2$) maal de variantie van de eerstgenoemde variabele (hier, $\mathrm{VAR}(\mathbf{x}) = 1$), hetgeen wordt afgeleid in de appendix op pagina 453. Dus,

$$\begin{aligned} \mathrm{VAR}(\widehat{\mathbf{y}}) &= \mathrm{VAR}(r_{xy}\mathbf{x}) \\ &= r_{xy}^2\mathrm{VAR}(\mathbf{x}) \\ &= r_{xy}^2 \quad . \end{aligned}$$

Omdat de correlatie $r_{xy}$ in absolute waarde altijd kleiner of gelijk is aan 1, geldt hetzelfde voor de variantie van $\widehat{\mathbf{y}}$.

Om te bepalen hoe goed deze voorspelling voor standaardscores is bekijken we weer haar gemiddelde kwadratische fout. Het blijkt dat in dit geval

$$\boxed{\frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \widehat{\mathbf{y}}_i)^2 = 1 - r_{xy}^2 \quad .}$$

Dit resultaat wordt afgeleid in de appendix op pagina 453. Dus, de gemiddelde kwadratische fout, onze maat voor de kwaliteit van de voorspelling, wordt kleiner (en de voorspelling dus beter) naarmate de correlatie $r_{xy}$ korter bij $+1$ of $-1$ ligt.

### 4.3.2 Enkelvoudige Regressieanalyse voor Ongestandaardiseerde Variabelen

De oplossing voor de gestandaardiseerde regressievergelijking $\widehat{\mathbf{y}} = r_{xy}\mathbf{x}$ en de grootte van de bijbehorende gemiddelde kwadratische fout kunnen we nu terugvertalen naar de originele scores $X$ en $Y$, eenvoudigweg door overal $\mathbf{x}$ en $\widehat{\mathbf{y}}$ te herschrijven als een functie van de corresponderende originele scores. Het resultaat van deze herschrijfoperatie zijn de volgende regressiecoëfficiënten:

$$\boxed{\begin{aligned} B &= r_{xy}\frac{\mathrm{SD}(Y)}{\mathrm{SD}(X)} \\ A &= \overline{Y} - B\,\overline{X} \quad . \end{aligned}}$$

Dit resultaat wordt afgeleid in de appendix op pagina 455. Dus, het ongestandaardiseerde regressiegewicht $B$ is gelijk aan het gestandaardiseerde regressiegewicht $b = r_{xy}$ maal de factor $\mathrm{SD}(Y)/\mathrm{SD}(X)$. En de ongestandaardiseerde regressieconstante $A$ is gelijk aan het criteriumgemiddelde $\overline{Y}$ min het product van het predictorgemiddelde $\overline{X}$ en het regressiegewicht $B$.

Ook de gemiddelde kwadratische fout voor de originele scores $Y$ is gemakkelijk te halen uit het resultaat voor de standaardscores:

$$\boxed{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 = (1 - r_{xy}^2)\mathrm{VAR}(Y) \quad .}$$

Ook dit resultaat wordt afgeleid in de appendix op pagina 455. We zien dat de gemiddelde kwadratische fout voor ongestandaardiseerde variabelen gelijk is aan die voor gestandaardiseerde variabelen (i.e., $1 - r_{xy}^2$) *maal* $\mathrm{VAR}(Y)$.

We zien dat in de voorspelling van $Y$ op basis van een predictor $X$ de correlatie $r_{xy}$ tussen $X$ en $Y$ op twee manieren een belangrijke rol speelt. Ten eerste bij het bepalen van de beste regressievergelijking (dit is het duidelijkste zichtbaar in de vergelijking voor standaardscores) en ten tweede bij het bepalen van de kwadratische fout die bij deze regressie gemaakt wordt. In de bovenstaande uitdrukking voor deze fout is het bijvoorbeeld duidelijk dat ze kleiner wordt naarmate $r_{xy}$ dichter bij $+1$ of $-1$ (en daarmee $r_{xy}^2$ dichter bij 1) ligt. Met andere woorden, de voorspelling wordt beter naarmate er

een sterkere (positieve óf negatieve) samenhang bestaat tussen criterium en predictor. Dit is natuurlijk wat we vooraf konden verwachten, maar we hebben hier de precieze uitdrukking gevonden voor hoe de kwaliteit van de voorspelling afhangt van de hoogte van de correlatiecoëfficiënt.

### 4.3.3 Verklaarde en Onverklaarde Variantie

In de paragraaf over het voorspellen zonder predictor hebben we gezien dat daarbij de gemiddelde kwadratische fout gegeven werd door $\mathrm{VAR}(Y)$. Vergelijken we dit met de bovenstaande uitdrukking voor de gemiddelde kwadratische fout met één predictor, dan zien we dat het opnemen van de predictor $X$ de gemiddelde kwadratische fout terugbrengt van $\mathrm{VAR}(Y)$ tot $(1 - r_{xy}^2)\mathrm{VAR}(Y)$. Dit verschil

$$\mathrm{VAR}(Y) - (1 - r_{xy}^2)\mathrm{VAR}(Y) = r_{xy}^2\mathrm{VAR}(Y)$$

is op te vatten als het gedeelte van $\mathrm{VAR}(Y)$ dat door $X$ verklaard of voorspeld wordt en daarom de (door $X$) *verklaarde variantie* van $Y$ genoemd wordt. Het resterende deel $(1 - r_{xy}^2)\mathrm{VAR}(Y)$ is dan op te vatten als het gedeelte van $\mathrm{VAR}(Y)$ dat niet door $X$ verklaard wordt. Daarom wordt het de (door $X$) *onverklaarde variantie* van $Y$ genoemd.

Echter, niet alleen de variantie van $Y$ kan opgesplitst worden in een verklaard en een onverklaard gedeelte, ook de variabele $Y$ zelf kan opgesplitst worden in een verklaard en een onverklaard gedeelte. Dit brengt ons bij een tweede manier om naar de regressieanalyse te kijken. Het uitgangspunt hierbij is de opsplitsing van $Y$ in een gedeelte $\widehat{Y}$ dat door $X$ wordt verklaard en het residudeel $Y - \widehat{Y}$ dat niet met $X$ samenhangt:

$$Y = \widehat{Y} + (Y - \widehat{Y}) \quad .$$

De tweede manier om naar de regressieanalyse te kijken bestaat erin dat we zoeken naar een voorspellingsregel $\widehat{Y}$ (een lineaire functie van $X$) waarvoor geldt dat $\widehat{Y}$ en $Y - \widehat{Y}$ niet met elkaar samenhangen. Wat preciezer, we willen dat de correlatie tussen $\widehat{Y}$ en $Y - \widehat{Y}$ gelijk is aan 0. Welnu, de eis van een nulcorrelatie tussen $\widehat{Y}$ en $Y - \widehat{Y}$ is equivalent met het kleinste-kwadratencriterium (de eis dat de gemiddelde kwadratische fout minimaal moet zijn); als aan het ene criterium voldaan is, is automatisch aan het andere voldaan. Intuïtief kan men dit als volgt zien: als men wil dat de gemiddelde kwadratische fout minimaal is, dan moet de voorspellingsformule $\widehat{Y}$ al het nuttige uit $X$ halen zodat het overblijvende deel ($Y - \widehat{Y}$, de voorspellingsfout) alleen nog maar bestaat uit datgene wat níet samenhangt met de predictor $X$, en dus ook niet met $\widehat{Y}$, die een functie is van $X$ ($\widehat{Y} = A + BX$). Om deze equivalentie ook precies aan te tonen, komt er wat formulewerk bij kijken, en dat laten we hier achterwege.

De nulcorrelatie tussen $\widehat{Y}$ en $Y - \widehat{Y}$ leidt tot een handig resultaat: de variantie van $Y$ is gelijk aan de som van de varianties van zijn samenstellende delen, $\widehat{Y}$ en $Y - \widehat{Y}$. In de vorm van een formule:

$$\mathrm{VAR}(Y) = \mathrm{VAR}(\widehat{Y}) + \mathrm{VAR}(Y - \widehat{Y}) \quad .$$

Deze formule is een bijzonder geval van een meer algemene formule voor de variantie van een som van variabelen die afgeleid wordt in de appendix op pagina 457. De eerste term

in het rechterlid, $\text{VAR}(\widehat{Y})$, blijkt gelijk te zijn aan wat we eerder de door $X$ verklaarde variantie van $Y$ genoemd hebben. En de tweede term, $\text{VAR}(Y - \widehat{Y})$, blijkt gelijk te zijn aan wat we eerder de door $X$ onverklaarde variantie van $Y$ genoemd hebben. Eerder stelden we immers de onverklaarde variantie gelijk aan de gemiddelde kwadratische fout en deze laatste is gelijk aan $\text{VAR}(Y - \widehat{Y})$, zoals aangetoond wordt in de appendix op pagina 458. Dus,

$$\boxed{\text{onverklaarde variantie} = \text{VAR}(Y - \widehat{Y}) = (1 - r_{xy}^2)\text{VAR}(Y) \quad .}$$

De resterende term, $\text{VAR}(\widehat{Y})$, is gelijk aan

$$\begin{aligned}\text{VAR}(Y) - \text{VAR}(Y - \widehat{Y}) &= \text{VAR}(Y) - (1 - r_{xy}^2)\text{VAR}(Y) \\ &= r_{xy}^2\text{VAR}(Y) \quad ,\end{aligned}$$

hetgeen we eerder de verklaarde variantie genoemd hebben. Dus,

$$\boxed{\text{verklaarde variantie} = \text{VAR}(\widehat{Y}) = r_{xy}^2\text{VAR}(Y) \quad .}$$

### 4.3.4 De Proporties Verklaarde en Onverklaarde Variantie

De varianties $\text{VAR}(\widehat{Y})$ en $\text{VAR}(Y - \widehat{Y})$ geven aan hoe belangrijk $\widehat{Y}$ en $Y - \widehat{Y}$ zijn als onderdeel van $Y$: hoe groter $\text{VAR}(\widehat{Y})$ in vergelijking met $\text{VAR}(Y - \widehat{Y})$, hoe meer variantie er verklaard wordt. Een voor de hand liggende maat voor de kwaliteit van de voorspelling is dan ook de verhouding van $\text{VAR}(\widehat{Y})$ ten opzichte van $\text{VAR}(Y)$. Deze verhouding wordt de *proportie verklaarde variantie* genoemd. Zij geeft aan welk deel van $\text{VAR}(Y)$ verklaard wordt door de voorspelling op basis van $X$. Dezelfde informatie wordt gegeven door de verhouding van $\text{VAR}(Y - \widehat{Y})$ ten opzichte van $\text{VAR}(Y)$. Deze verhouding wordt de *proportie onverklaarde variantie* genoemd.

De formule voor de proportie verklaarde variantie is de volgende:

$$\boxed{\text{proportie verklaarde variantie} = \frac{\text{VAR}(\widehat{Y})}{\text{VAR}(Y)} = r_{xy}^2 \quad .}$$

En de formule voor de proportie onverklaarde variantie is de volgende:

$$\boxed{\text{proportie onverklaarde variantie} = \frac{\text{VAR}(Y - \widehat{Y})}{\text{VAR}(Y)} = 1 - r_{xy}^2 \quad .}$$

Met behulp van deze begrippen kunnen we de kwaliteit van de voorspelling $\widehat{Y}$ bondig weergeven: de op de predictor $X$ gebaseerde voorspelling $\widehat{Y}$ verklaart een proportie $r_{xy}^2$ van de variantie van $Y$, terwijl een proportie $(1-r_{xy}^2)$ van de variantie van $Y$ onverklaard blijft.

### 4.3.5 De Standaardschattingsfout van de Regressie

De gemiddelde kwadratische fout wordt gebruikt om de grootte van de voorspellingsfouten uit te drukken. In het bovenstaande hebben we laten zien dat die gemiddelde kwadratische fout hetzelfde is als $\mathrm{VAR}(Y - \widehat{Y})$. De gemiddelde kwadratische fout is gebaseerd op gekwadrateerde verschillen. Daarom hebben we te maken met een grootheid die op een andere schaal wordt uitgedrukt dan de criteriumvariabele. Om een grootheid te verkrijgen die op dezelfde schaal uitgedrukt wordt als de criteriumvariabele, nemen we de vierkantswortel uit $\mathrm{VAR}(Y - \widehat{Y})$. Zo krijgen we als maat voor de voorspellingsfout de standaarddeviatie van de residuen $Y - \widehat{Y}$. Deze is via de correlatiecoëfficiënt $r_{xy}$ eenvoudig uit te drukken als een proportie van de standaarddeviatie van de $Y$-scores. Uit

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 = \mathrm{VAR}(Y - \widehat{Y}) = (1 - r_{xy}^2)\,\mathrm{VAR}(Y)$$

volgt immers via worteltrekken dat

$$\boxed{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2} = \mathrm{SD}(Y - \widehat{Y}) = \sqrt{(1 - r_{xy}^2)}\,\mathrm{SD}(Y) \quad .}$$

Aangezien $\mathrm{SD}(Y)$ de gemiddelde fout is bij het voorspellen zonder predictoren (zie paragraaf 4.2.2), zien we dat het benutten van de predictorscores $X$ deze fout een factor $\sqrt{1 - r_{xy}^2}$ kleiner maakt.

De zo gedefinieerde voorspellingsfout van de regressie $\mathrm{SD}(Y - \widehat{Y})$ wordt in het Nederlands de *standaardschattingsfout van de regressie* genoemd. In het Engels heet ze *standard error of estimate* of ook wel *R.M.S. error for regression*, waarin R.M.S. de afkorting is van *Root-Mean-Square*, de vierkantswortel (*Root*) van het gemiddelde (*Mean*) van het kwadraat (*Square*) van de residuen.

### 4.3.6 Een Zuivere Schatter van de Gemiddelde Kwadratische Fout in de Populatie

In de paragraaf over voorspellen zonder predictoren merkten we op dat de gemiddelde kwadratische fout, $\mathrm{VAR}(Y)$ voor die situatie, een onzuivere schatter is van gemiddelde kwadratische fout in de populatie. Hetzelfde geldt voor de huidige situatie, voorspellen met één predictor: $\mathrm{VAR}(Y - \widehat{Y})$ is een systematische onderschatting van de variantie van de voorspellingsfouten $Y - \widehat{Y}$ in de populatie, verder aangeduid met $\sigma^2$. Dit is niet vreemd, want de voorspellingen $\widehat{Y}$ zijn zó gekozen dat de som $\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$ minimaal is, hetgeen equivalent is met het criterium dat $\mathrm{VAR}(Y - \widehat{Y})$ (de genoemde som gedeeld door $n$) minimaal is. Door deze keuze krijgen we een te rooskleurig beeld van de variantie van de voorspellingsfouten in de populatie.

Opnieuw bestaat er echter een zeer eenvoudige manier om $\mathrm{VAR}(Y - \widehat{Y})$ te corrigeren voor deze systematische onderschatting: in plaats van de som $\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$ te delen

door de steekproefomvang $n$ (of te vermenigvuldigen met $\frac{1}{n}$), delen we ze door $n-2$. Het resultaat,

$$\frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 \quad ,$$

is een zuivere schatter van $\sigma^2$. Het belangrijke punt hier, is dat nu gedeeld wordt door $n-2$ in plaats van door $n-1$, zoals in de situatie zonder predictoren. Dit heeft te maken met het feit dat nu twéé onbekenden geschat worden ($A$ en $B$) in plaats van één in de situatie zonder predictoren (alleen $A$). Wanneer we de statistische toetsing van regressiegewichten bespreken, komen we hier op terug.

### 4.3.7 Voorbeeld

In Tabel 4.2 staan de resultaten van de regressie van de criteriumvariabele VLG op de predictor EM, die van de drie predictoren in Tabel 4.1 in absolute waarde de hoogste correlatie met de variabele VLG heeft. Alle berekeningen zijn uitgeschreven voor de eerste drie en de laatste drie proefpersonen uit de steekproef. Indien gewenst kunnen ze, op basis van de gegevens in Tabel 4.1, voor de gehele steekproef worden nagerekend. In Figuur 4.1 zijn de scores op VLG en EM weergegeven in een puntenwolk, waarin de proefpersonen uit Tabel 4.2 apart zijn aangegeven.

| pp | VLG | EM | **vlg** | **em** | $\widehat{\mathbf{vlg}}$ | $\widehat{\text{VLG}}$ | DEV | DEV$^2$ | RES | RES$^2$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 9 | 41 | -0.49 | -0.33 | 0.18 | 11.10 | 0.57 | 0.32 | -2.10 | 4.41 |
| 2 | 4 | 55 | -2.11 | 1.84 | -1.01 | 7.40 | -3.13 | 9.78 | -3.40 | 11.59 |
| 3 | 11 | 42 | 0.15 | -0.18 | 0.10 | 10.84 | 0.31 | 0.09 | 0.16 | 0.03 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 30 | 15 | 34 | 1.44 | -1.42 | 0.78 | 12.95 | 2.42 | 5.84 | 2.05 | 4.21 |
| 31 | 10 | 52 | -0.17 | 1.37 | -0.75 | 8.20 | -2.33 | 5.45 | 1.80 | 3.25 |
| 32 | 9 | 46 | -0.49 | 0.44 | -0.24 | 9.78 | -0.75 | 0.56 | -0.78 | 0.61 |
| Gem | 10.53 | 43.16 | 0.00 | 0.00 | 0.00 | 10.53 | 0.00 | 2.89 | 0.00 | 6.73 |
| Var | 9.62 | 41.51 | 1.00 | 1.00 | 0.30 | 2.89 | ⟵ | ↩ | 6.73 | ↩ |
| SD | 3.10 | 6.44 | 1.00 | 1.00 | 0.55 | 1.70 | | | 2.59 | |

| Cor. | VLG |
|---|---|
| EM | -0.55 |

$r^2 = 0.30$

$21.92 = A$

$b = -0.55 \quad -0.26 = B$

Tabel 4.2: De regressie van VLG op EM. Zie tekst.

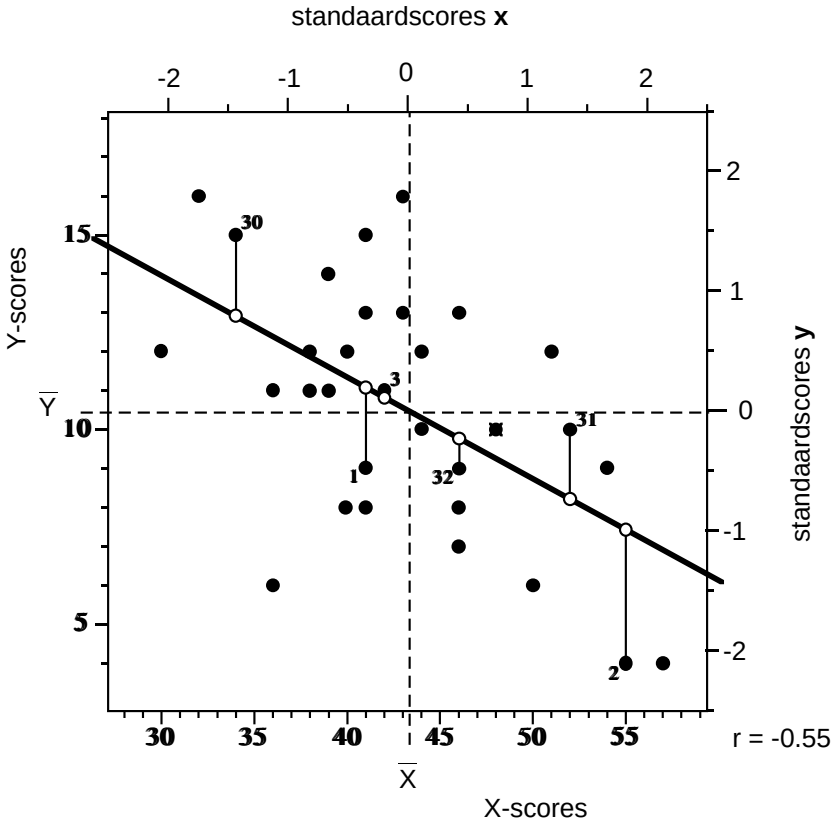
In de kolommen **vlg** en **em** van Tabel 4.2 zijn de variabelen VLG en EM uitgedrukt in standaardscores. (Ook hier noteren we de variabelen in hun standaardscorevorm met kleine vetgedrukte letters.) Zo is, in de eerste rij, $\mathbf{vlg}_1 = -0.49$ berekend als

$$\mathbf{vlg}_1 = \frac{\text{VLG}_1 - \overline{\text{VLG}}}{\text{SD(VLG)}} = \frac{9 - 10.53}{3.10} = -0.49 \quad ,$$

en analoog in de tweede rij $\mathbf{em}_2 = 1.84$ als

$$\mathbf{em}_2 = \frac{55 - 43.16}{6.44} = 1.84 \quad .$$

De standaardscores **vlg** en **em** hebben uiteraard beide een gemiddelde van 0 en een variantie van 1. Terwijl in Figuur 4.1 voor de linker- en de benedenas de originele score-eenheden gebruikt zijn, zijn er rechts en boven assen in standaardeenheden toegevoegd. De kolommen **vlg** en **em** van Tabel 4.2 geven de coördinaten van de datapunten met betrekking tot deze assen in standaardeenheden.




Figuur 4.1: Spreidingsdiagram van de variabelen VLG ($= Y$) en EM ($= X$) met de regressielijn van VLG op EM. Gesloten cirkels zijn datapunten, open cirkels op de lijn zijn voorspelde VLG-scores van de proefpersonen 1, 2, 3, 30, 31 en 32 (vergelijk met Tabel 4.1, van wie de residuen zijn ingetekend als lijnstukjes). De assen links en onder geven de variabelen in hun originele score-eenheden, de assen rechts en boven in standaardeenheden.

In de volgende kolom $\widehat{\mathbf{vlg}}$ staan de op basis van EM voorspelde standaardscores van de criteriumvariabele. Zoals we in deze paragraaf gezien hebben worden deze berekend

met de formule

$$\widehat{\mathbf{vlg}}_i = r_{\mathrm{VLG,EM}}\mathbf{em}_i \quad .$$

Volgens Tabel 4.1 (en nog herhaald onderaan Tabel 4.2) is de correlatie tussen criterium en predictor gelijk aan $r_{\mathrm{VLG,EM}} = -0.55$, en dus wordt $\widehat{\mathbf{vlg}}_1$ in Tabel 4.2 berekend als

$$\widehat{\mathbf{vlg}}_1 = (-0.55) \times (-0.33) = 0.18 \quad ,$$

en

$$\widehat{\mathbf{vlg}}_2 = (-0.55) \times 1.84 = -1.01 \quad ,$$

enzovoort. In termen van standaardscores (dus in Figuur 4.1 met betrekking tot de assen rechts en boven) liggen al deze voorspellingen op de rechte lijn door de oorsprong met richtingscoëfficiënt gelijk aan $r_{\mathrm{VLG,EM}} = -0.55$, die in Figuur 4.1 is getekend.

Voor de voorspelling van de ongestandaardiseerde variabele VLG moeten de coëfficiënten $A$ en $B$ van de regressielijn worden bepaald. Volgens de resultaten van deze paragraaf krijgen we hiervoor:

$$B = r_{\mathrm{VLG,EM}} \frac{\mathrm{SD(VLG)}}{\mathrm{SD(EM)}} = (-0.55) \times \frac{3.10}{6.44} = -0.26 \quad ,$$

en

$$A = \overline{\mathrm{VLG}} \ - \ B\,\overline{\mathrm{EM}} = 10.53 \ - \ (-0.26) \cdot 43.16 = 21.92 \quad .$$

In originele scores wordt de regressievergelijking van VLG op EM dus:

$$\widehat{\mathrm{VLG}} = 21.92 \ - \ 0.26\,\mathrm{EM} \quad .$$

Deze voorspellingen zijn verzameld in de kolom $\widehat{\mathrm{VLG}}$ van Tabel 4.2. Bijvoorbeeld, de voorspelling van 11.10 punten op VLG voor de eerste proefpersoon resulteert als de uitkomst van $21.92 - 0.26 \cdot 41$, enzovoort. Bovenstaande vergelijking voor de ongestandaardiseerde scores is de vergelijking van de in Figuur 4.1 getekende regressielijn, maar nu met betrekking tot de assen links en onder.

In de kolom $\widehat{\mathrm{VLG}}$ blijkt ook dat de variantie van de voorspelde VLG-scores gelijk is aan 2.89. Om expliciet te maken hoe deze variantie is berekend zijn in de kolom DEV de voorspellingen $\widehat{\mathrm{VLG}}$ omgezet in deviatiescores door er overal $\overline{\mathrm{VLG}} = 10.53$ van af te trekken (het gemiddelde van VLG is namelijk ook het gemiddelde van $\widehat{\mathrm{VLG}}$). Voor de eerste rij, bijvoorbeeld, is $0.57 = 11.10 - 10.53$. In de volgende kolom zijn deze deviatiescores gekwadrateerd ($0.32 = 0.57^2$, enzovoort) en het gemiddelde van deze kolom met kwadraten van deviatiescores is dan per definitie gelijk aan de variantie van de variabele $\widehat{\mathrm{VLG}}$.

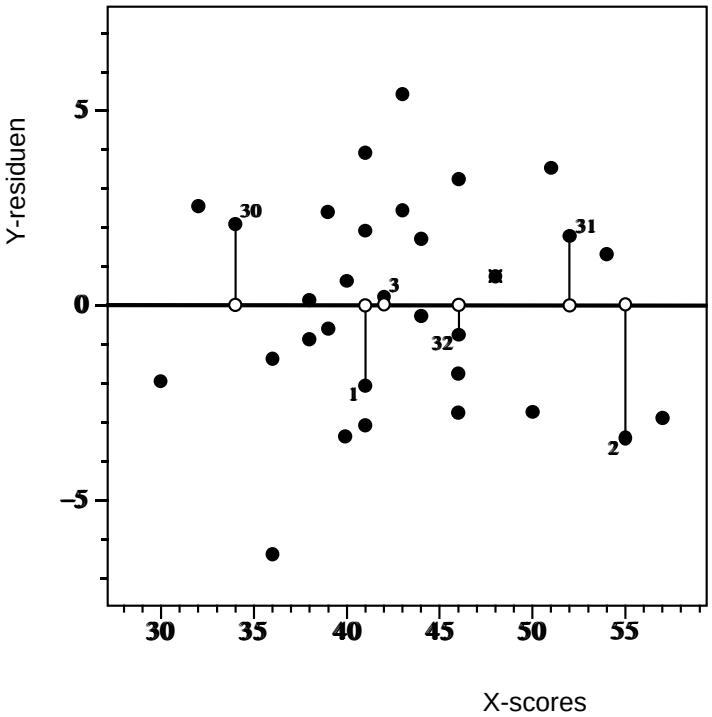
Om de kwaliteit van de voorspelling via deze regressie te onderzoeken worden de residuen bepaald. Deze staan weergegeven in de kolom RES van Tabel 4.2. Zo is $\mathrm{RES}_1 = -2.10$ berekend als

$$\mathrm{RES}_1 = \mathrm{VLG}_1 \ - \ \widehat{\mathrm{VLG}}_1 = 9 \ - \ 11.10 = -2.10 \quad .$$

Het gemiddelde van deze residuen is gelijk aan 0 en hun variantie blijkt gelijk te zijn aan 6.73. Ook dit is nog eens expliciet gemaakt door in de laatste kolom van Tabel 4.2

de kwadraten van de residuen te berekenen ($4.41 = (-2.10)^2$, enzovoort) en hierover te middelen. Dit levert een gemiddelde kwadratische fout van 6.73 op, die, aangezien de residuen gemiddelde 0 hebben, ook meteen de onverklaarde variantie is.

In Figuur 4.1 corresponderen deze residuen met de verticale afstand van de datapunten tot de regressielijn. Voor de zes proefpersonen uit Tabel 4.2 is de grootte van het residu hier weergegeven door een lijnstukje. De residuen zijn nog eens apart uitgezet in Figuur 4.2. De idee dat deze residuscores (verticaal uitgezet) het gedeelte van VLG vormen dat niet met EM (horizontaal uitgezet) samenhangt uit zich hier in het ontbreken van enige trend of structuur in deze figuur. De punten liggen hier kriskras om het nulniveau, en zo hoort het dus ook in een residuendiagram. Het enige interessante is hier de gemiddelde afstand van de residuen tot de nullijn. Volgens Tabel 4.2 moet deze ongeveer gelijk zijn aan $\mathrm{SD}(\mathrm{VLG} - \widehat{\mathrm{VLG}}) = \sqrt{6.73} = 2.59$, welke waarde goed in overeenstemming lijkt met Figuur 4.2.




Figuur 4.2: Spreidingsdiagram van de residuen van de regressie van VLG ($= Y$) op EM ($= X$) (zie Figuur 4.1). Van dezelfde proefpersonen als in Figuur 4.1 is de grootte van het residu door een lijnstukje weergegeven.

We bekijken in dit voorbeeld nog de opsplitsing van de variantie van VLG in verklaarde en onverklaarde variantie. Zoals we eerder zagen, vallen deze twee delen van de variantie samen met, respectievelijk, de variantie van de voorspelde scores en de variantie

van de residuen. Inderdaad zien we in Tabel 4.2 dat de opsplitsing

$$\mathrm{VAR}(\mathrm{VLG}) = \mathrm{VAR}(\widehat{\mathrm{VLG}}) \ + \ \mathrm{VAR}(\mathrm{VLG} - \widehat{\mathrm{VLG}})$$

hier opgaat, en wel in de vorm

$$9.62 = 2.89 \ + \ 6.73 \quad .$$

De verhouding

$$\frac{\mathrm{VAR}(\widehat{\mathrm{VLG}})}{\mathrm{VAR}(\mathrm{VLG})} = \frac{2.89}{9.62} = 0.30$$

is dus de proportie verklaarde variantie, welke inderdaad gelijk is aan

$$r^2_{\mathrm{VLG,EM}} = (-0.55)^2 = 0.30 \quad ,$$

zoals eerder in deze paragraaf werd afgeleid. De andere verhouding

$$\frac{\mathrm{VAR}(\mathrm{VLG} - \widehat{\mathrm{VLG}})}{\mathrm{VAR}(\mathrm{VLG})} = \frac{6.73}{9.62} = 0.70$$

geeft de proportie onverklaarde variantie, die korter berekend had kunnen worden als

$$1 - r^2_{\mathrm{VLG,EM}} = 1 - (-0.55)^2 = 1 - 0.30 = 0.70 \quad .$$

Via de regressie op EM wordt dus 30% van de variantie van VLG verklaard, waarmee nog 70% onverklaarde variantie overblijft.

Vergelijken we tenslotte de voorspellingsfout die we mogen verwachten gebruikmakend van de regressie op EM met de fout die we mogen verwachten zonder predictor. Volgens deze en de vorige paragraaf komt dit neer op een vergelijking van de standaardschattingsfout $\mathrm{SD}(\mathrm{VLG} - \widehat{\mathrm{VLG}}) = 2.59$ met de standaarddeviatie van het criterium $\mathrm{SD}(\mathrm{VLG}) = 3.10$. Dan zien we dat de fout mét regressie gelijk is aan de fractie

$$\frac{2.59}{3.10} = 0.84$$

van de fout zonder regressie. Deze verkleiningsfactor is inderdaad gelijk aan

$$\sqrt{1 - r^2_{\mathrm{VLG,EM}}} = \sqrt{0.70} = 0.84 \quad ,$$

zoals eerder in deze paragraaf werd gesteld. Het gebruiken van een predictor met een toch behoorlijke correlatie van $-0.55$ met het criterium knabbelt dus slechts $100\% - 84\% = 16\%$ af van de fout zonder predictor. Dit illustreert dat de reductie van de voorspellingsfout via de factor $\sqrt{1 - r^2}$ nogal langzaam gaat. Bijvoorbeeld, er is een correlatie extremer dan $\pm 0.85$ vereist om deze fout met behulp van één predictor te halveren.

---

*Opgave 1*

Met welke van de onderstaande uitspraken ben je het eens en met welke ben je het oneens? Als je het oneens bent met een uitspraak, leg dan uit waarom.

1. De ongestandaardiseerde regressieconstante $A$ is gelijk aan de gestandaardiseerde regressieconstante $a$ als de gemiddelden van $Y$ en $X$ gelijk zijn aan elkaar.
2. Het ongestandaardiseerde regressiegewicht $B$ is gelijk aan het gestandaardiseerde regressiegewicht $b$ als de standaarddeviaties van $Y$ en $X$ gelijk zijn aan elkaar.
3. De standaarddeviatie van een variabele zegt ons iets over hoe goed we deze variabele kunnen voorspellen zonder dat we iets meer weten van de personen (meer dan alleen maar de scores op die ene variabele).
4. Naarmate de variantie van de afhankelijke variabele $Y$ groter is, zal ook de variantie van de voorspellingsfouten $Y - \widehat{Y}$ groter zijn (de correlatie $r_{xy}$ blijft gelijk).

*Opgave 2*
Een onderzoeker in de motivatiepsychologie wil de wet van Yerkes-Dodson onderzoeken. Die wet zegt dat er een omgekeerd U-vormig verband is tussen motivatie en performantie. De onderzoeker verzamelt gegevens bij een groep rechtenstudenten. Motivatie werd geoperationaliseerd door een vragenlijst, en performantie door het gemiddelde examencijfer op het einde van het jaar. Om het verband tussen motivatie en performantie te onderzoeken wordt een enkelvoudige lineaire regressieanalyse uitgevoerd met de einduitslag als criterium en de motivatiescore als predictor. Wat denk je van dit onderzoek?

*Opgave 3*
Een onderzoekster was geïnteresseerd in de voorspelling van het aantal sexuele relaties dat iemand heeft tijdens zijn of haar studententijd. Daartoe maakte zij gebruik van de volgende predictor: het aantal seconden dat men in een gestandaardiseerd groepsgesprek het woord richt tot een persoon van het geslacht waarvoor hij of zij een voorkeur heeft. De onderzoekster beweerde dat het regressiegewicht $B$ vrijwel gelijk is aan 0, maar dat de proportie verklaarde variantie toch behoorlijk is (i.c., 0.51). Is dit mogelijk? Leg uit waarom wel of waarom niet.

*Opgave 4*
Leg uit, aan de hand van de formules in de tekst, dat de onverklaarde variantie altijd kleiner is dan de variantie van $Y$, ook al heeft de predictor (bijvoorbeeld, lichaamslengte) helemaal niets te maken met de afhankelijke variabele (bijvoorbeeld, inkomen). Ga er hierbij vanuit dat een correlatie die precies gelijk is aan 0 in de praktijk vrijwel nooit voorkomt.

*Opgave 5*
De directie van een groot bedrijf in de informatietechnologie ontwerpt een nieuwe selectieprocedure voor haar programmeurs. Eén van de selectie-instrumenten is een kleine programmeeropdracht. De score op deze opdracht is de tijd (in minuten) die de kandidaat nodig heeft om deze opdracht uit te voeren. Deze variabele wordt verder aangeduid als TIJD. Om de voorspellende waarde van deze programmeeropdracht te evalueren, heeft men een groot aantal kandidaten deze programmeeropdracht laten uitvoeren. Na een periode van twee jaar vraagt men aan hun directe chef een beoordeling van de kwaliteit van hun werk (op een schaal van 0 tot 10), verder aangeduid als OORDEEL. Stel, men doet nu drie regressieanalyses:

1. OORDEEL wordt voorspeld met TIJD.
2. OORDEEL wordt voorspeld met TIJD, maar TIJD wordt nu niet in minuten maar in uren uitgedrukt.
3. OORDEEL wordt voorspeld met TIJD, maar OORDEEL wordt nu niet op een schaal van 0 tot 10 uitgedrukt, maar op een schaal van 0 tot 100 (door de oorspronkelijke cijfers met 10 te vermenigvuldigen).

Geef aan, voor elk van de grootheden in de rijen van Tabel 4.3, of zij groter worden, kleiner worden, of gelijk blijven bij de stap van 1. naar 2. (tweede kolom) en de stap van 1. naar 3. (derde kolom). Ga er vanuit dat $A$, $B$ en $b$ positief zijn.

| | TIJD: minuut $\rightarrow$ uur | OORDEEL: 0-10 $\rightarrow$ 0-100 |
|---|---|---|
| VAR($X$) | | |
| VAR($Y$) | | |
| $r_{xy}^2$ | | |
| VAR($\widehat{Y}$) | | |
| SD($Y - \widehat{Y}$) | | |
| $B$ | | |
| $A$ | | |
| $b$ | | |

Tabel 4.3: Tabel bij opgave 5.

*Opgave 6*
Stel dat we de variabele VLG uit het voorbeeld voorspellen op basis van de variabele ZB.

1. Wat is de voorspelde standaardscore van persoon 1 en persoon 2?
2. Bereken $A$ en $B$.
3. Wat is de voorspelde ongestandaardiseerde score van persoon 1 en persoon 2?
4. Hoe groot is de verklaarde en de onverklaarde variantie?
5. Hoe groot is de proportie verklaarde variantie?
6. Wat is de standaardschattingsfout?

---

## 4.4 Voorspellen met Twee Predictoren: Meervoudige Regressieanalyse

Nu bekijken we de situatie waarin we de voorspelling van de criteriumvariabele $Y$ niet baseren op één predictor $X$, maar op twee predictoren die we $X_1$ en $X_2$ zullen noemen. Op basis van steekproefgegevens waarin we scores op $X_1$, $X_2$ en $Y$ hebben waargenomen willen we nu de lineaire regressievergelijking bepalen van $Y$ op de predictoren $X_1$ en $X_2$. Dit wil zeggen, nu willen we in de vergelijking

$$\boxed{\widehat{Y} = A + B_1\, X_1 + B_2\, X_2}$$

de coëfficiënten $A$, $B_1$ en $B_2$ zó kiezen dat de residuen $Y - \widehat{Y}$ weer geminimaliseerd worden volgens het kleinste-kwadratencriterium. De coëfficiënt $B_1$ heet hier het regressiegewicht van $X_1$ en $B_2$ is het regressiegewicht van $X_2$.

### 4.4.1 Meervoudige Regressieanalyse voor Gestandaardiseerde Variabelen

#### De Gestandaardiseerde Regressiecoëfficiënten

Net zoals bij enkelvoudige regressie, is de oplossing die voldoet aan het kleinste-kwadratencriterium eenvoudiger voor de regressievergelijking

$$\boxed{\widehat{\mathbf{y}} = a + b_1\,\mathbf{x}_1 + b_2\,\mathbf{x}_2}$$

voor de gestandaardiseerde variabelen $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{y}$. Als we hiervoor $a$, $b_1$ en $b_2$ bepalen volgens het kleinste-kwadratencriterium vinden we als oplossing:

$$\boxed{\begin{array}{rcl} a & = & 0 \\ b_1 & = & \dfrac{r_{1y} - r_{12}r_{2y}}{1 - r_{12}^2} \\ b_2 & = & \dfrac{r_{2y} - r_{12}r_{1y}}{1 - r_{12}^2} \end{array}} \quad .$$

Met deze keuze van $b_1$ en $b_2$ wordt de vergelijking voor de gestandaardiseerde variabelen dus

$$\widehat{\mathbf{y}} = b_1\,\mathbf{x}_1 + b_2\,\mathbf{x}_2 \quad .$$

De $b_1$ en $b_2$ zijn de gestandaardiseerde regressiegewichten, die, vanwege historische redenen, ook vaak *bètagewichten*[5] worden genoemd.

We zien dat met twee predictoren de situatie een stuk complexer is geworden dan met één predictor. Terwijl in de situatie met één predictor het ene bètagewicht eenvoudig gelijk was aan de correlatiecoëfficiënt tussen predictor en criterium ($b = r_{xy}$), is in de situatie met twee predictoren elk bètagewicht een vrij ingewikkelde functie van alle correlatiecoëfficiënten $r_{1y}$, $r_{2y}$ en $r_{12}$. Zo speelt bijvoorbeeld de correlatie tussen $Y$ en $X_1$ een rol bij de bepaling van het bètagewicht voor $X_2$, en speelt ook $r_{12}$, de correlatie tussen de predictoren onderling, mee. Als we hier wat preciezer naar kijken, zien we dat in het speciale geval dat de beide predictoren ongecorreleerd zouden zijn, dus $r_{12} = 0$, de formules voor $b_1$ en $b_2$ plotsklaps weer zeer eenvoudig worden en volkomen analoog aan het geval met slechts één predictor. Substitutie van $r_{12} = 0$ in bovenstaande formules voor $b_1$ en $b_2$ resulteert namelijk in $b_1 = r_{1y}$ en $b_2 = r_{2y}$. Alleen in dit uitzonderlijke geval van twee volledig ongecorreleerde predictoren valt het gestandaardiseerde regressiegewicht van een predictor samen met de correlatie van deze predictor met het criterium.

#### De Bijdrage van een Predictor aan de Voorspelde Criteriumscores

De formules voor de regressiegewichten bij twee predictoren zijn ingewikkelder dan bij één predictor. Om uit te leggen hoe dit komt, concentreren we ons op $b_2$, het regressiegewicht voor $\mathbf{x}_2$. De uitleg voor $b_1$, het regressiegewicht voor $\mathbf{x}_1$, is volledig

[5] Ze worden in de literatuur dan ook vaak met de Griekse letter $\beta$ genoteerd. Dit is echter verwarrend, omdat Griekse letters meestal worden gebruikt voor populatiegrootheden (parameters), terwijl deze bètagewichten zuivere steekproefgrootheden zijn. Daarom worden ze hier, hoewel wel bètagewicht genoemd, genoteerd met de Latijnse kleine letter $b$.

analoog. De formule voor $b_2$ is zo ingewikkeld omdat, in de regressievergelijking met twee predictoren, het regressiegewicht $b_2$ weegt hoe zwaar de score $\mathbf{x}_2$ meetelt in de voorspelling $\widehat{\mathbf{y}}$, *gegeven de bijdrage van* $\mathbf{x}_1$ *aan de voorspelling*. Terwijl $\mathbf{x}_2$ als enige predictor een gewicht van $r_{2y}$ zou krijgen, moet dit gewicht nu gecorrigeerd worden omdat een deel van de door $\mathbf{x}_2$ voorspelde score ook al door $\mathbf{x}_1$ wordt voorspeld. De grootte van deze correctie hangt af van zowel $r_{1y}$, die meet hoeveel $\mathbf{x}_1$ van de $\mathbf{y}$-score voorspelt, alsook van $r_{12}$, die meet welk deel van wat $\mathbf{x}_1$ voorspelt samenvalt met wat $\mathbf{x}_2$ voorspelt. In feite representeert het product $r_{12}\, r_{1y}$ in de teller van $b_2$ het gedeelte van de samenhang tussen $\mathbf{x}_2$ en $\mathbf{y}$ dat vía $\mathbf{x}_1$ loopt. Met andere woorden, dit is de samenhang die tussen $\mathbf{x}_2$ en $\mathbf{y}$ bestaat louter als gevolg van de samenhang tussen $\mathbf{x}_1$ en $\mathbf{x}_2$ (met een sterkte gelijk aan $r_{12}$) en $\mathbf{x}_1$ en $\mathbf{y}$ (met een sterkte gelijk aan $r_{1y}$). Aangezien dit product $r_{12}r_{1y}$ in de teller van $b_2$ wordt afgetrokken van $r_{2y}$, representeert deze teller juist het gedeelte van de samenhang tussen $\mathbf{x}_2$ en $\mathbf{y}$ dat níet via $\mathbf{x}_1$ loopt, maar dat $\mathbf{x}_2$ juist buiten $\mathbf{x}_1$ om aan de voorspelling toevoegt. De noemer van $b_2$ is als een normering te beschouwen (het verschil $r_{2y} - r_{12}r_{1y}$ wordt uitgedrukt als het aantal keer dat het in $1 - r_{12}^2$ zit), die niet differentieert tussen de predictoren.

Het is níet zo dat eerst de bijdrage van $\mathbf{x}_1$ vastgelegd wordt (door de keuze van $b_1$) en daarna de bijdrage van $\mathbf{x}_2$ (door de keuze van $b_2$) gegeven de eerder vastgelegde bijdrage van $\mathbf{x}_1$: $b_1$ en $b_2$ worden sámen bepaald, op zo'n manier dat zij samen zorgen voor een zo klein mogelijke gemiddelde kwadratische fout. Als we zeggen dat $b_2$ weegt hoe zwaar de score $\mathbf{x}_2$ meetelt in de voorspelling $\widehat{\mathbf{y}}$ gegeven de bijdrage van $\mathbf{x}_1$ aan de voorspelling, dan bedoelen we daarmee dat $b_1$ samen met $b_2$ bepaald wordt. Dit hele verhaal met betrekking tot $b_2$ geldt natuurlijk ook voor $b_1$, met de rol van $\mathbf{x}_1$ en $\mathbf{x}_2$ verwisseld.

Een concrete consequentie van deze beschouwing is dat het regressiegewicht $b_2$ in ieder geval níet geïnterpreteerd kan worden als hét gewicht van de variabele $\mathbf{x}_2$ in de voorspelling $\widehat{\mathbf{y}}$. Als we de eerste predictor $\mathbf{x}_1$ namelijk vervangen door een andere variabele, dan veranderen daarmee de waardes van $r_{1y}$ en $r_{12}$, en dus ook de waarde van het gewicht $b_2$. De waarde van $b_2$ kan dus slechts relatief geïnterpreteerd worden: het is het gewicht van predictor $\mathbf{x}_2$ in de voorspelling $\widehat{\mathbf{y}}$, gegeven een welbepaalde andere predictor $\mathbf{x}_1$ in de regressievergelijking. Ook hier natuurlijk hetzelfde verhaal voor $b_1$.

### De Bijdrage van een Predictor aan de Proportie Verklaarde Variantie

Wanneer in het bovenstaande gesproken werd over "de bijdrage van een predictor aan de voorspelling", dan werd dit opgevat als de bijdrage van deze predictor aan $\widehat{\mathbf{y}}$, de voorspelde criteriumscores. Nu, de bijdrage van een predictor aan $\widehat{\mathbf{y}}$ wordt uitgedrukt in het regressiegewicht van deze predictor: hoe groter dit regressiegewicht (in absolute waarde), hoe groter de verandering in $\widehat{\mathbf{y}}$ als gevolg van een verandering in de predictorscore (bijvoorbeeld, een toename met 1 eenheid). Echter, "de bijdrage van een predictor aan de voorspelling" kan nog op een andere manier opgevat worden, namelijk als de bijdrage van deze predictor aan de proportie verklaarde variantie. Het uitgangspunt is dat, door te voorspellen op basis van twee predictoren in plaats van één, de gemiddelde kwadratische fout daalt en de proportie verklaarde variantie dus stijgt. Met twee predictoren kan men immers beter voorspellen dan met één. Nu, de bijdrage van predictor $\mathbf{x}_2$ aan de proportie verklaarde variantie is gelijk aan het verschil tussen de volgende twee proporties verklaarde variantie: (1) de proportie verklaarde variantie op basis van een meervoudige

regressievergelijking met $\mathbf{x}_1$ én $\mathbf{x}_2$, en (2) de proportie verklaarde variantie op basis van een enkelvoudige regressievergelijking met alleen $\mathbf{x}_1$. Zoals verder meer in detail besproken zal worden, is het verschil tussen deze twee proporties verklaarde variantie uit te rekenen als het kwadraat van een zogenaamde *semi-partiële* of *part*correlatiecoëfficiënt. Dus, waar de bijdrage van een predictor aan $\widehat{\mathbf{y}}$ uitgedrukt wordt in het regressiegewicht van deze predictor, wordt de bijdrage van een predictor aan de proportie verklaarde variantie uitgedrukt in een gekwadrateerde partcorrelatiecoëfficiënt.

Het belangrijke punt is niet alleen dat "de bijdrage van een predictor aan de voorspelling" op twee manieren opgevat kan worden, maar ook dat deze twee opvattingen tot zeer verschillende conclusies kunnen leiden. Bijvoorbeeld, het is mogelijk dat $\mathbf{x}_2$ een belangrijke bijdrage levert aan $\widehat{\mathbf{y}}$ (het regressiegewicht $b_2$ is groot) maar dat $\mathbf{x}_2$ toch een zeer kleine bijdrage levert aan de proportie verklaarde variantie (de gekwadrateerde partcorrelatiecoëfficiënt is klein). Zoals verder zal blijken, doet dit fenomeen zich voor als er een zeer sterke correlatie is tussen $\mathbf{x}_1$ en $\mathbf{x}_2$.

We hebben gezien dat een regressiegewicht in een meervoudige regressievergelijking slechts relatief geïnterpreteerd kan worden: het is het gewicht van een predictor in de voorspelling van $\mathbf{y}$, gegeven een welbepaalde andere predictor in de regressievergelijking. Ook de gekwadrateerde partcorrelatiecoëfficiënt, onze kwantificatie van de bijdrage van een predictor aan de proportie verklaarde variantie, kan slechts relatief geïnterpreteerd worden. Men kan immers laten zien dat ook de partcorrelatiecoëfficiënt van een predictor $\mathbf{x}_2$ afhangt van de andere predictor $\mathbf{x}_1$ die men samen met $\mathbf{x}_2$ gebruikt om $\mathbf{y}$ te voorspellen. Dus, we kunnen besluiten dat, hoe ook de bijdrage van een predictor aan de voorspelling opgevat wordt (als bijdrage aan $\widehat{\mathbf{y}}$ of als bijdrage aan de proportie verklaarde variantie), deze bijdrage altijd afhangt van de andere predictor die men ook in de regressievergelijking opneemt.

### 4.4.2 Meervoudige Regressieanalyse voor Ongestandaardiseerde Variabelen

De uiteenzetting was tot hiertoe vooral gebaseerd op de gestandaardiseerde regressievergelijking, omdat daar de formules nog het eenvoudigst zijn en de onderliggende logica nog het duidelijkst. Voor praktische voorspellingstoepassingen werken we echter graag met de ongestandaardiseerde regressievergelijking, waarin de oorspronkelijk geobserveerde scores $X_1$ en $X_2$ direct ingevuld kunnen worden. Aan de ongestandaardiseerde gewichten zien we hoe zwaar deze oorspronkelijke, direct interpreteerbare scores $X_1$ en $X_2$ meewegen in de voorspelling van het criterium $Y$.

Op dezelfde manier als bij één predictor volgt deze ongestandaardiseerde vergelijking echter gemakkelijk uit de gestandaardiseerde: we hoeven hierin slechts de definities van $\mathbf{x}_1$, $\mathbf{x}_2$ en $\mathbf{y}$ in termen van de oorspronkelijke scores te substitueren. Dit leidt dan tot de ongestandaardiseerde regressievergelijking

$$\widehat{Y} = A + B_1\, X_1 + B_2\, X_2 \quad ,$$

met als coëfficiënten

$$\boxed{\begin{aligned} B_1 &= b_1 \frac{\mathrm{SD}(Y)}{\mathrm{SD}(X_1)} \\ B_2 &= b_2 \frac{\mathrm{SD}(Y)}{\mathrm{SD}(X_2)} \\ A &= \overline{Y} - B_1 \overline{X}_1 - B_2 \overline{X}_2 \quad . \end{aligned}}$$

Hier zijn dus de ongestandaardiseerde $B$-gewichten uitgedrukt in de boven gegeven gestandaardiseerde bètagewichten. Merk op dat deze omzetting van $b$- naar $B$-gewichten dezelfde is als bij één predictor. Ook daar hebben we $B = b\,\mathrm{SD}(Y)/\mathrm{SD}(X)$, met $b = r_{xy}$. Ook de definitie van $A$ is hier duidelijk analoog aan die bij één predictor.

### 4.4.3 De Meervoudige Correlatiecoëfficiënt

Nu gaan we kijken naar de kwaliteit van de voorspelling met twee predictoren. Dit bespreken we eerst in termen van de gestandaardiseerde vergelijking. De omzetting naar de ongestandaardiseerde vergelijking is daarna weer heel eenvoudig. In paragraaf 4.3 hebben we gezien dat in een voorspelling met één predictor de gemiddelde kwadratische fout van de gestandaardiseerde regressievergelijking op een eenvoudige manier afhangt van de correlatie tussen criterium en predictor:

$$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \widehat{\mathbf{y}}_i)^2 = 1 - r_{xy}^2 \quad .$$

Bij de voorspelling met twee predictoren blijkt er een soortgelijke eenvoudige uitdrukking voor de gemiddelde kwadratische fout te bestaan in termen van een correlatie tussen het criterium enerzijds en de twee predictoren anderzijds.

#### Veralgemening van de Gewone Correlatiecoëfficiënt

Tot nog toe verwees het begrip correlatie naar een relatie tussen twee variabelen. Nu veralgemenen we dit begrip naar de relatie tussen, aan de ene kant, één variabele ($\mathbf{y}$) en, aan de andere kant, een stelstel van variabelen ($\mathbf{x}_1$ en $\mathbf{x}_2$). Dit nieuwe begrip wordt een *meervoudige correlatiecoëfficiënt* genoemd. We bedoelen ermee de gewone correlatie tussen $\mathbf{y}$ en díe lineaire combinatie van $\mathbf{x}_1$ en $\mathbf{x}_2$ die $\mathbf{y}$ het beste voorspelt. Die combinatie is precies $b_1\,\mathbf{x}_1 + b_2\,\mathbf{x}_2$, oftewel $\widehat{\mathbf{y}}$, zoals bepaald in de regressieanalyse. Dus de meervoudige correlatie tussen $\mathbf{y}$ enerzijds en $\mathbf{x}_1$ en $\mathbf{x}_2$ anderzijds, die we aanduiden met $R_{\mathbf{y}\cdot\mathbf{x}_1\mathbf{x}_2}$ of korter $R_{\mathbf{y}\cdot 12}$, is gelijk aan de gewone correlatie $r_{\mathbf{y}\widehat{\mathbf{y}}}$ tussen de variabelen $\mathbf{y}$ en $\widehat{\mathbf{y}}$:

$$\boxed{R_{\mathbf{y}\cdot 12} = r_{\mathbf{y}\widehat{\mathbf{y}}} \quad .}$$

Wat betreft de notatie merken we nog op dat we in de subscripten van bovenstaande correlaties $\mathbf{y}$, $\widehat{\mathbf{y}}$, $\mathbf{x}_1$ en $\mathbf{x}_2$ kunnen vervangen door hun ongestandaardiseerde versies, aangezien correlaties tussen gestandaardiseerde variabelen gelijk zijn aan de correlaties tussen de corresponderende ongestandaardiseerde variabelen. We kunnen dus $R_{y\cdot 12}$ schrijven voor $R_{\mathbf{y}\cdot 12}$ en $r_{y\widehat{y}}$ voor $r_{\mathbf{y}\widehat{\mathbf{y}}}$.

### Uitrekenen van de Meervoudige Correlatiecoëfficiënt

Via bovenstaande definitie is de meervoudige correlatiecoëfficiënt enerzijds praktisch uit te rekenen als de correlatie tussen de geobserveerde en de voorspelde $Y$-scores, en anderzijds ook theoretisch uit te drukken in termen van de correlaties $r_{1y}$, $r_{2y}$ en $r_{12}$. Dit laatste is mogelijk door de formule voor $r_{\mathbf{y}\widehat{\mathbf{y}}}$ uit te werken met de substitutie $\widehat{\mathbf{y}} = b_1\,\mathbf{x}_1 + b_2\,\mathbf{x}_2$. Nogal uitgebreid rekenwerk, dat we hier niet laten zien, leidt dan tot de volgende berekeningsformule:

$$R_{y\cdot 12} = \sqrt{b_1\,r_{1y} + b_2\,r_{2y}} \quad ,$$

en dus

$$\boxed{R^2_{y\cdot 12} = b_1\,r_{1y} + b_2\,r_{2y} \quad .}$$

Hierin kunnen nog de definities van de regressiegewichten $b_1$ en $b_2$ gesubstitueerd worden om een uitdrukking alleen in de correlatiecoëfficiënten $r_{1y}$, $r_{2y}$ en $r_{12}$ te verkrijgen. Aangezien de formule er niet inzichtelijker op wordt, laten we dit hier achterwege.

Behalve de bovenstaande formule bestaat er nog een tweede manier om de meervoudige correlatiecoëfficiënt te berekenen. Deze tweede berekeningsmethode maakt gebruik van de partcorrelatiecoëfficiënt, die in een van de volgende paragrafen besproken zal worden. Voor berekeningen met de hand, is deze tweede berekeningsmethode meestal eenvoudiger.

In het vervolg zal meestal uit de context duidelijk zijn op welk criterium en welke predictoren de meervoudige correlatiecoëfficiënt betrekking heeft, en dan schrijven we eenvoudigweg $R$, zonder subscripten. In de rest van deze paragraaf, schrijven we steeds $R$ voor $R_{y\cdot 12}$.

### Gemiddelde Kwadratische Fout, Verklaarde en Onverklaarde Variantie

Het belang van de meervoudige correlatiecoëfficiënt $R$ is dat deze, bij meer dan één predictor, de rol overneemt die de correlatie $r_{xy}$ speelt bij één predictor. Zo valt af te leiden dat de gemiddelde kwadratische fout van de gestandaardiseerde regressievergelijking gelijk is aan $1 - R^2$:

$$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \widehat{\mathbf{y}}_i)^2 = 1 - R^2 \quad .$$

Voor de ongestandaardiseerde vergelijking moeten we dit weer vermenigvuldigen met $\mathrm{VAR}(Y)$:

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 = (1 - R^2)\mathrm{VAR}(Y) \quad .$$

Net zoals bij één predictor, geldt ook hier dat $\widehat{Y}$ en $Y - \widehat{Y}$ ongecorreleerd zijn. Hieruit volgt de gekende opsplitsing van $\mathrm{VAR}(Y)$ in verklaarde en onverklaarde variantie:

$$\mathrm{VAR}(Y) = \mathrm{VAR}(\widehat{Y}) \;+\; \mathrm{VAR}(Y - \widehat{Y}) \quad .$$

Omdat $\text{VAR}(Y-\widehat{Y})$ gelijk is aan de gemiddelde kwadratische fout, kan men deze variantie berekenen als $(1-R^2)\text{VAR}(Y)$. Hieruit volgt dat $\text{VAR}(\widehat{Y})$ gelijk is aan $R^2\text{VAR}(Y)$. We krijgen de proportie verklaarde variantie als we $\text{VAR}(\widehat{Y})$ delen door $\text{VAR}(Y)$:

$$\boxed{\text{proportie verklaarde variantie} = \frac{\text{VAR}(\widehat{Y})}{\text{VAR}(Y)} = R^2 \quad .}$$

En we krijgen de proportie onverklaarde variantie als we $\text{VAR}(Y-\widehat{Y})$ delen door $\text{VAR}(Y)$:

$$\boxed{\text{proportie onverklaarde variantie} = \frac{\text{VAR}(Y-\widehat{Y})}{\text{VAR}(Y)} = 1-R^2 \quad .}$$

Deze overeenkomst in formules met het geval van één predictor wekt weinig verbazing als we de definitie van meervoudige correlatie ook in dat geval gewoon toepassen. Aangezien bij één predictor de voorspelling $\widehat{Y}$ een lineaire functie is van $X$ (met een negatieve coëfficiënt als $r_{xy} < 0$), is de correlatie tussen $Y$ en $\widehat{Y}$ ook gelijk aan de correlatie tussen $Y$ en $X$ (de absolute waarde hiervan als ze negatief is). De correlatie is immers ongevoelig voor lineaire transformaties van de variabelen. Dus, bij één predictor valt de meervoudige correlatie $R$, gedefinieerd als $R = r_{y\hat{y}}$, samen met de gewone correlatie $r_{xy}$ (om precies te zijn, met $|r_{xy}|$), zodat ook geldt dat $r_{xy}^2 = R^2$.

Op deze manier verschijnt $R^2$, het kwadraat van de meervoudige correlatie, als een algemene, natuurlijke maat voor hoe goed de voorspelling het doet: een hoge $R^2$ (dicht bij 1) betekent dat de spreiding in de geobserveerde $Y$-scores, zoals gemeten door $\text{VAR}(Y)$, grotendeels te voorspellen valt vanuit de scores op de predictoren $X_1$ en $X_2$, en daarmee ook weerspiegeld wordt in een corresponderende spreiding van de voorspellingen $\widehat{Y}$, zoals gemeten door $\text{VAR}(\widehat{Y})$. In zo'n geval zijn de verschillen tussen geobserveerde en voorspelde scores, zoals gemeten door $\text{VAR}(Y-\widehat{Y})$, relatief klein. Een lage $R^2$ (veel kleiner dan 1) duidt op precies de omgekeerde situatie. Nu valt de spreiding in geobserveerde $Y$-scores slechts in geringe mate te voorspellen vanuit de $X_1$- en $X_2$-scores, zodat de hierop gebaseerde voorspellingen $\widehat{Y}$ ook veel minder uiteen zullen lopen. In dit geval, zullen de voorspelde $Y$-scores dicht bij elkaar liggen, en daarmee liggen ze ook alle dicht bij het gemiddelde $\overline{Y}$, dat we immers voorspellen als $X_1$ en $X_2$ helemaal geen informatie verschaffen betreffende $Y$. In zo'n geval is $\text{VAR}(\widehat{Y})$ relatief klein en blijven de verschillen $Y-\widehat{Y}$ over het geheel genomen groot, wat resulteert in een relatief grote $\text{VAR}(Y-\widehat{Y})$.

### De Standaardschattingsfout van de Regressie

Evenals bij één predictor kunnen we de boven besproken gemiddelde kwadratische fout door worteltrekking weer omzetten in de standaardschattingsfout van de regressie. Deze standaardschattingsfout is gelijk aan de standaarddeviatie van de residuen $Y-\widehat{Y}$. Net zoals bij de formules voor de verklaarde en de onverklaarde variantie, neemt ook hier de meervoudige correlatiecoëfficiënt $R$ de rol van de gewone correlatiecoëfficiënt $r_{xy}$ over. Dit wil zeggen, uit

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i-\widehat{Y}_i)^2 = \text{VAR}(Y-\widehat{Y}) = (1-R^2)\,\text{VAR}(Y)$$

volgt nu

$$\boxed{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i-\widehat{Y}_i)^2} = \mathrm{SD}(Y-\widehat{Y}) = \sqrt{(1-R^2)}\,\mathrm{SD}(Y)}$$

als uitdrukking voor de standaardschattingsfout en dus is nu $\sqrt{1-R^2}$ de factor waarmee de voorspellingsfout afneemt door gebruik te maken van de twee predictoren $X_1$ en $X_2$.

**Een Zuivere Schatter van de Gemiddelde Kwadratische Fout in de Populatie**

Net zoals voor de situatie met één predictor, is ook hier $\mathrm{VAR}(Y-\widehat{Y})$ een systematische onderschatting van de variantie van de voorspellingsfouten $Y-\widehat{Y}$ in de populatie, aangeduid met $\sigma^2$. Opnieuw bestaat er echter een zeer eenvoudige manier om $\mathrm{VAR}(Y-\widehat{Y})$ te corrigeren voor deze systematische onderschatting: in plaats van de som $\sum_{i=1}^{n}(Y_i-\widehat{Y}_i)^2$ te delen door de steekproefomvang $n$ (of te vermenigvuldigen met $\frac{1}{n}$), delen we ze door $n-3$. Het resultaat,

$$\frac{1}{n-3}\sum_{i=1}^{n}(Y_i-\widehat{Y}_i)^2 \quad ,$$

is een zuivere schatter van $\sigma^2$. Het belangrijke punt is dat nu gedeeld wordt door $n-3$ in plaats van door $n-2$, zoals in de situatie met één predictor. Dit heeft te maken met het feit dat nu dríe onbekenden geschat worden ($A$, $B_1$ en $B_2$) in plaats van twee in de situatie met één predictor ($A$ en $B$).

### 4.4.4 Voorbeeld

In Tabel 4.4 staat de regressie van VLG uitgewerkt met, naast EM, nu de variabele ZB als tweede predictor toegevoegd. De scores op deze variabele zijn in de kolom **zb** ook in standaardeenheden gegeven ($-0.78 = \frac{20-21.41}{1.80}$, enzovoort). Met de onderaan Tabel 4.4 herhaalde onderlinge correlaties kunnen volgens de formules op pagina 319 de bètagewichten worden berekend. Zo krijgen we:

$$b_{\mathrm{EM}} = \frac{r_{\mathrm{VLG,EM}} - r_{\mathrm{EM,ZB}}\, r_{\mathrm{VLG,ZB}}}{1-r^2_{\mathrm{EM,ZB}}} = \frac{-0.55-(-0.20)\times 0.35}{1-(-0.20)^2} = -0.50 \quad ,$$

$$b_{\mathrm{ZB}} = \frac{r_{\mathrm{VLG,ZB}} - r_{\mathrm{EM,ZB}}\, r_{\mathrm{VLG,EM}}}{1-r^2_{\mathrm{EM,ZB}}} = \frac{0.35-(-0.20)\times(-0.55)}{1-(-0.20)^2} = 0.26 \quad .$$

Merk op dat deze bètagewichten inderdaad verschillen van de corresponderende correlatiecoëfficiënten, zoals eerder werd aangekondigd. De verschillen zijn hier echter niet al te groot, wat toe te schrijven valt aan het feit dat, alhoewel de twee predictoren onderling gecorreleerd zijn, deze correlatie met een waarde van $-0.20$ vrij laag is.

Met deze bètagewichten kunnen dan de voorspelde standaardscores $\widehat{\mathbf{vlg}}$ worden berekend, gebruikmakend van de gestandaardiseerde regressievergelijking

$$\widehat{\mathbf{vlg}} = b_{\mathrm{EM}}\,\mathbf{em} \;+\; b_{\mathrm{ZB}}\,\mathbf{zb} \quad .$$

| pp | VLG | EM | ZB | **vlg** | **em** | **zb** | $\widehat{\mathbf{vlg}}$ | $\widehat{\text{VLG}}$ | RES |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 9 | 41 | 20 | -0.49 | -0.33 | -0.78 | -0.03 | 10.43 | -1.43 |
| 2 | 4 | 55 | 22 | -2.11 | 1.84 | 0.33 | -0.83 | 7.95 | -3.95 |
| 3 | 11 | 42 | 20 | 0.15 | -0.18 | -0.78 | -0.11 | 10.19 | 0.81 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 30 | 15 | 34 | 23 | 1.44 | -1.42 | 0.89 | 0.93 | 13.43 | 1.57 |
| 31 | 10 | 52 | 22 | -0.17 | 1.37 | 0.33 | -0.60 | 8.67 | 1.33 |
| 32 | 9 | 46 | 19 | -0.49 | 0.44 | -1.34 | -0.56 | 8.79 | 0.21 |
| Gem | 10.53 | 43.16 | 21.41 | 0.00 | 0.00 | 0.00 | 0.00 | 10.53 | 0.00 |
| Var | 9.62 | 41.51 | 3.24 | 1.00 | 1.00 | 1.00 | 0.36 | 3.49 | 6.13 |
| SD | 3.10 | 6.44 | 1.80 | 1.00 | 1.00 | 1.00 | 0.60 | 1.87 | 2.48 |

| Cor. | VLG | EM |
|---|---|---|
| EM | -0.55 | |
| ZB | 0.35 | -0.20 |

$R^2 = 0.36$

$11.48 = A$

$b_{\text{EM}} = -0.50 \quad -0.24 = B_{\text{EM}}$

$b_{\text{ZB}} = 0.26 \quad 0.44 = B_{\text{ZB}}$

Tabel 4.4: De regressie van VLG op de predictoren EM en ZB. Zie tekst.

De resultaten hiervan staan in de kolom $\widehat{\mathbf{vlg}}$, met bijvoorbeeld in de tweede rij $-0.83 = (-0.50) \times 1.84 + 0.26 \times 0.33$. Voor praktische doeleinden hebben we echter liever de voorspellingen voor de originele, ongestandaardiseerde scores. Daartoe zetten we de gestandaardiseerde $b$-gewichten om in de ongestandaardiseerde $B$-gewichten volgens de op pagina 322 gegeven formules. Dus,

$$B_{\text{EM}} = b_{\text{EM}} \frac{\text{SD(VLG)}}{\text{SD(EM)}} = -0.50 \frac{3.10}{6.44} = -0.24 \quad ,$$

$$B_{\text{ZB}} = b_{\text{ZB}} \frac{\text{SD(VLG)}}{\text{SD(ZB)}} = 0.26 \frac{3.10}{1.80} = 0.44 \quad ,$$

$$A = 10.53 - (-0.24) \times 43.16 - 0.44 \times 21.41 = 11.48 \quad .$$

We verkrijgen zo de ongestandaardiseerde regressievergelijking

$$\widehat{\text{VLG}} = 11.48 \; - \; 0.24\,\text{EM} \; + \; 0.44\,\text{ZB} \quad ,$$

op basis waarvan de voorspellingen in de kolom $\widehat{\text{VLG}}$ zijn berekend. Zo is daar in de eerste rij de voorspelling $\widehat{\text{VLG}}_1 = 10.43$ verkregen als

$$\widehat{\text{VLG}}_1 = 11.48 \; - \; 0.24 \times 41 \; + \; 0.44 \times 20 = 10.43$$

en in de tweede rij

$$\widehat{\text{VLG}}_2 = 11.48 \; - \; 0.24 \times 55 \; + \; 0.44 \times 22 = 7.95 \quad ,$$

enzovoort.

Terwijl bij één predictor al deze voorspellingen mooi op een lijn liggen, de regressielijn, is de situatie bij twee predictoren meetkundig iets moeilijker voor te stellen. Nu liggen de voorspellingen alle in een vlak binnen een driedimensionale ruimte. Wanneer we de lengte- en de breedteas van deze ruimte associëren met de twee predictoren EM en ZB, en de hoogteas met het criterium VLG, dan vormen de scores van de proefpersonen op deze drie variabelen 32 datapunten in deze ruimte. Het regressievlak ligt in deze ruimte dan zo dat, gemiddeld genomen, de verticale afstand van deze punten tot het vlak zo klein mogelijk is gemaakt. (Vergelijk met Figuur 4.1, waar dit alles met één dimensie minder wordt weergegeven.)

In deze figuur van een vlak in een driedimensionale ruimte corresponderen de zo klein mogelijk gemaakte verticale afstanden met de residuen van de regressie op de twee predictoren EM en ZB. Deze zijn algebraïsch berekend in kolom RES van Tabel 4.4. Dus $-1.43 = 9 - 10.43$, en $-3.95 = 4 - 7.95$, enzovoort. Als maat voor de gemiddelde grootte van deze residuen vinden we voor deze regressie op twee predictoren een standaardschattingsfout van $\mathrm{SD}(\mathrm{VLG} - \widehat{\mathrm{VLG}}) = 2.48$. Dit is beter dan de met slechts één predictor bereikte $\mathrm{SD}(\mathrm{VLG} - \widehat{\mathrm{VLG}}) = 2.59$ (zie Tabel 4.2), maar de verbetering is bepaald niet dramatisch te noemen.

We krijgen hetzelfde beeld wanneer we gaan kijken naar de gemiddelde kwadratische fout en de proportie verklaarde variantie. Deze laatste is gelijk aan het kwadraat van de meervoudige correlatiecoëfficiënt die, volgens de formule op pagina 323, berekend kan worden als

$$R^2 = b_{\mathrm{EM}}\, r_{\mathrm{VLG,EM}} \;+\; b_{\mathrm{ZB}}\, r_{\mathrm{VLG,ZB}} = (-0.50) \times (-0.55) \;+\; 0.26 \times 0.35 = 0.36 \quad .$$

De twee predictoren EM en ZB verklaren samen dus 36% van de variantie van VLG. Vergelijkend met de 30% die door EM alleen werd verklaard (zie Tabel 4.2) moeten we concluderen dat het toevoegen van ZB als predictor voor slechts 6% extra variantie verklaart. Dit is een niet te verwaarlozen, maar toch bescheiden bijdrage.

Met $R^2 = 0.36$ komt de meervoudige correlatiecoëfficiënt zelf dus uit op

$$R = \sqrt{0.36} = 0.60 \quad .$$

Volgens de theorie van deze paragraaf is dit de correlatie tussen geobserveerde en voorspelde VLG-scores:

$$r_{\mathrm{VLG},\widehat{\mathrm{VLG}}} = R = 0.60 \quad ,$$

We zouden deze waarde ook vinden als we op de normale manier de correlatie zouden berekenen voor de scores in de kolommen VLG en $\widehat{\mathrm{VLG}}$ van Tabel 4.4.

Ook de opsplitsing van de variantie van VLG in verklaarde en onverklaarde variantie is weer rechtstreeks uit Tabel 4.4 te halen. Ook hier blijkt weer dat

$$\mathrm{VAR}(\mathrm{VLG}) = \mathrm{VAR}(\widehat{\mathrm{VLG}}) \;+\; \mathrm{VAR}(\mathrm{VLG} - \widehat{\mathrm{VLG}}) \quad ,$$

want

$$9.62 = 3.49 \;+\; 6.13 \quad .$$

In overeenstemming met de resultaten van deze paragraaf, correspondeert dit met een proportie

$$\frac{\mathrm{VAR}(\widehat{\mathrm{VLG}})}{\mathrm{VAR}(\mathrm{VLG})} = \frac{3.49}{9.62} = 0.36 = R^2$$

verklaarde variantie en een proportie

$$\frac{\mathrm{VAR}(\mathrm{VLG} - \widehat{\mathrm{VLG}})}{\mathrm{VAR}(\mathrm{VLG})} = \frac{6.13}{9.62} = 0.64 = 1 - R^2$$

onverklaarde variantie.

De verhouding tussen de standaardschattingsfout van deze regressie op twee predictoren en de voorspellingsfout zonder predictoren, hangt ook af van de meervoudige correlatiecoëfficiënt. Volgens Tabel 4.4 wordt deze verhouding gegeven door

$$\frac{\mathrm{SD}(\mathrm{VLG} - \widehat{\mathrm{VLG}})}{\mathrm{SD}(\mathrm{VLG})} = \frac{2.48}{3.10} = 0.80 \quad ,$$

wat inderdaad gelijk is aan $\sqrt{1 - R^2}$, aangezien

$$\sqrt{1 - R^2} = \sqrt{1 - 0.36} = \sqrt{0.64} = 0.80 \quad .$$

De twee predictoren EM en ZB brengen de voorspellingsfout dus terug tot 80% van de fout zonder predictor, oftewel een reductie van 20%.

## 4.5 Voorspellen met Willekeurig Veel Predictoren

### 4.5.1 Theorie

Met meer predictoren wordt de situatie gelukkig niet wezenlijk anders dan met twee. Bijvoorbeeld, voor drie predictoren hebben we een ongestandaardiseerde regressievergelijking van de vorm

$$\widehat{Y} = A + B_1\, X_1 + B_2\, X_2 + B_3\, X_3 \quad ,$$

met bijbehorende gestandaardiseerde versie van de vorm

$$\widehat{\mathbf{y}} = b_1\, \mathbf{x}_1 + b_2\, \mathbf{x}_2 + b_3\, \mathbf{x}_3 \quad .$$

Hierin ontbreekt de regressieconstante $a$ want voor gestandaardiseerde regressievergelijkingen blijkt altijd $a = 0$ te gelden.

Het algemene geval met $m$ predictoren kunnen we schrijven als

$$\widehat{Y} = A + B_1\, X_1 + B_2\, X_2 + \ldots + B_m\, X_m$$

voor de oorspronkelijke variabelen $Y$, $X_1$, $X_2$, $\ldots$ $X_m$ en

$$\widehat{\mathbf{y}} = b_1\, \mathbf{x}_1 + b_2\, \mathbf{x}_2 + \ldots + b_m\, \mathbf{x}_m$$

voor de gestandaardiseerde variabelen. Vanaf $m = 3$ worden de formules voor de gestandaardiseerde regressiegewichten $b_1$, $b_2$, $\ldots$, $b_m$ al snel zeer ingewikkeld, zodat het ook geen enkele zin heeft ze hier te vertonen. Het principe blijft echter hetzelfde als bij twee predictoren. Het regressiegewicht $b_j$ van predictor $\mathbf{x}_j$ hangt, behalve van de correlatie $r_{jy}$ tussen predictor $X_j$ en $Y$ ook af van de correlaties die $Y$ heeft met de

andere predictoren en van de correlaties tussen alle predictoren onderling. Alleen in het uitzonderlijke geval dat álle predictoren $X_1$ t/m $X_m$ onderling ongecorreleerd zijn worden de formules voor de $b_j$ weer heel eenvoudig en vallen de gewichten weer samen met de desbetreffende correlatiecoëfficiënt. Dus, alleen in dat geval geldt $b_j = r_{jy}$. De formules voor het algemene geval – dit is, wanneer er correlaties tussen de predictoren bestaan – zijn ingewikkeld omdat het gewicht $b_j$ weegt hoeveel de predictor $\mathbf{x}_j$ bijdraagt aan de voorspelling $\widehat{\mathbf{y}}$, gegeven de bijdragen van de andere predictoren aan de voorspelling. Opnieuw kan de bètacoëfficiënt $b_j$ niet geïnterpreteerd worden als hét gewicht van de variabele $\mathbf{x}_j$ in de voorspelling van $\widehat{\mathbf{y}}$ zonder meer, maar slechts als het gewicht van $\mathbf{x}_j$ in de voorspelling van $\widehat{\mathbf{y}}$, gegeven de welbepaalde collectie overige predictoren in de regressievergelijking. Iedere wijziging in deze verzameling van overige predictoren doet in principe de waarde van $b_j$ veranderen.

Net zoals bij regressieanalyse met twee predictoren past ook hier de opmerking dat de "bijdrage van een predictor aan de voorspelling" ook opgevat kan worden als de bijdrage van deze predictor aan de proportie verklaarde variantie. Zoals in de volgende paragraaf meer in detail besproken wordt, kan deze bijdrage uitgedrukt worden in een gekwadrateerde partcorrelatiecoëfficiënt.

Gegeven de gestandaardiseerde regressiegewichten vinden we de gewichten van de ongestandaardiseerde vergelijking $\widehat{Y} = A + B_1\,X_1 + B_2\,X_2 + \ldots + B_m\,X_m$ op dezelfde manier als in paragraaf 4.4 bij twee predictoren. Voor elke $j = 1, \ldots, m$ wordt het ongestandaardiseerde $B$-gewicht van predictor $X_j$ gegeven door

$$B_j = b_j\,\frac{\mathrm{SD}(Y)}{\mathrm{SD}(X_j)} \quad ,$$

terwijl de constante $A$ de waarde krijgt

$$A = \overline{Y} - B_1\,\overline{X}_1 - B_2\,\overline{X}_2 - \ldots - B_m\,\overline{X}_m \qquad .$$

Deze gewichten worden dus gebruikt om de originele $Y$-scores te voorspellen gegeven geobserveerde scores op de predictoren $X_1$ t/m $X_m$.

De kwaliteit van de voorspelling wordt weer uitgedrukt in de meervoudige correlatiecoëfficiënt $R$, en met name zijn kwadraat $R^2$. Analoog aan de situatie met twee predictoren is deze $R$ gedefinieerd als de correlatie tussen de geobserveerde $Y$ en de volgens de regressievergelijking voorspelde $\widehat{Y}$. Als deze vergelijking is gebaseerd op $m$ predictoren $X_1$ t/m $X_m$ is de officiële notatie voor deze meervoudige correlatie $R_{y\cdot x_1 x_2 \ldots x_m}$ ofwel, iets simpeler, $R_{y\cdot 12\ldots m}$. Voor de praktijk, hebben we alleen de gekwadrateerde meervoudige correlatie $R^2_{y\cdot 12\ldots m}$ nodig. Deze kan als volgt berekend worden:

$$R^2_{y\cdot 12\ldots m} = b_1\,r_{1y} + b_2\,r_{2y} + \ldots + b_m\,r_{my} \quad .$$

Hierin zijn de $b_1, b_2, \ldots, b_m$ de ingewikkelde bètacoëfficiënten waarvoor we geen algemene formule hebben gegeven. Dus, in deze vorm heeft bovenstaande formule geen direct praktisch nut. Computerprogramma's die een meervoudige regressieanalyse uitvoeren hebben echter geen moeite de juiste waarde van $R$ en $R^2$ te berekenen. Behalve met de bovenstaande formule, kan men de meervoudige correlatie ook berekenen met behulp van partcorrelatiecoëfficiënten. Deze berekeningsmethode wordt in de volgende paragraaf besproken.

De meervoudige correlatie $R$ is altijd een getal tussen 0 en 1. Net zoals bij twee predictoren, is het zó dat via de regressievergelijking een proportie van $R^2$ van de variantie van $Y$ wordt verklaard, dit wil zeggen, terug te vinden is als de variantie van de voorspelling $\widehat{Y}$. De rest, de proportie $1 - R^2$ van de variantie van $Y$, blijft onverklaard en is terug te vinden als de variantie van de residuen $Y - \widehat{Y}$. Hoe hoger $R^2$ (hoe dichter bij 1), des te beter is de voorspelling. Ook met betrekking tot de standaardschattingsfout van de regressie verandert er niets wezenlijks in vergelijking met de situatie met twee predictoren. In het algemeen, is deze vierkantswortel uit de gemiddelde kwadratische fout gelijk aan de standaarddeviatie van de residuen, die kan worden berekend als $\mathrm{SD}(Y - \widehat{Y}) = \sqrt{(1 - R^2)}\,\mathrm{SD}(Y)$, waarbij dus in het algemeen de voorspellingsfout van de regressie op de predictoren $X_1$ t/m $X_k$ gelijk is aan de fractie $\sqrt{1 - R^2}$ van de voorspellingsfout zonder predictoren.

Net zoals voor de situatie met één en twee predictoren, is ook hier $\mathrm{VAR}(Y - \widehat{Y})$ een systematische onderschatting van de variantie van de voorspellingsfouten $Y - \widehat{Y}$ in de populatie ($\sigma^2$). Men kan $\mathrm{VAR}(Y - \widehat{Y})$ als volgt corrigeren voor deze systematische onderschatting: in plaats van de som $\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$ te delen door de steekproefomvang $n$, delen we ze door $n - m - 1$. Het resultaat,

$$\frac{1}{n - m - 1}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 \quad ,$$

is een zuivere schatter van $\sigma^2$. We delen dus door het aantal observaties ($n$) min het aantal onbekenden dat we geschat hebben (hier, $m + 1$, één $A$ en $m$ $B_k$'s).

Niet alleen de ongecorrigeerde variantie $\mathrm{VAR}(Y - \widehat{Y})$ is een onzuivere schatter van zijn tegenhanger in de populatie; hetzelfde geldt voor $R^2$. Net zoals $\mathrm{VAR}(Y - \widehat{Y})$ een te rooskleurig beeld geeft van de variantie van de voorspellingsfouten in de populatie, geeft ook de $R^2$ in de steekproef waarop de regressievergelijking is berekend een te rooskleurig beeld van de $R^2$ in de populatie. Deze populatie-$R^2$ is op te vatten als de $R^2$ berekend op alle personen in de populatie. De steekproef-$R^2$ is een systematische overschatting van deze populatie-$R^2$. Het eenvoudigste voorbeeld hiervan is de $R^2$ van een enkelvoudige regressieanalyse (die gelijk is aan $r_{xy}^2$, de gekwadrateerde gewone correlatie) in het geval dat de populatie-$R^2$ (hier dus de populatie-$r_{xy}^2$) gelijk is aan 0. Nu, de steekproef-$r_{xy}^2$ is altijd groter of gelijk aan 0, en dus hebben we te maken met een systematische óverschatting.

Er bestaan formules waarmee men $R^2$ probeert te corrigeren voor deze systematische overschatting. De meest bekende, die ook gegeven wordt door SPSS, wordt aangeduid als $R^2_{Adjusted}$ (het Engelse *adjusted* staat voor *gecorrigeerd*):

$$R^2_{Adjusted} = R^2 \;-\; \frac{m}{n - m - 1}\,(1 - R^2) \quad .$$

Hierin staat $m$ voor het aantal predictoren in de vergelijking en $n$ voor het aantal personen in de steekproef. We kunnen in de formule zien dat de correctie groter wordt naarmate er meer predictoren (grote $m$) zijn en naarmate de steekproef kleiner (kleine $n$) is. Dit is in overeenstemming met het feit dat de mate van overschatting groter is naarmate er meer predictoren zijn en naarmate de steekproef kleiner is.

Een onvolkomenheid van deze correctieformule is dat zij geldig is als de populatie-$R^2$ gelijk is aan nul (i.e., in de populatie levert geen enkele predictor een bijdrage aan de voorspelling van het criterium). In dit geval, is de steekproef-$R^2_{Adjusted}$ een zuivere schatter van de populatie-$R^2$. Indien de populatie-$R^2$ echter níet gelijk is aan nul, dan hebben we die zekerheid niet. Toch is het in de praktijk altijd beter om $R^2_{Adjusted}$ te gebruiken in plaats van $R^2$, ook al is de populatie-$R^2$ verschillend van 0.

### 4.5.2 Voorbeeld

Ter illustratie zijn in Tabel 4.5 de resultaten van de regressie van de variabele VLG op alle drie gepresenteerde predictoren gegeven. De bespreking hiervan kan kort zijn, aangezien het veelal een herhaling betreft van eerdere tabellen. Via ingewikkelde formules, die hierboven niet gegeven zijn, werden de bètagewichten $b_{\mathrm{EM}} = -0.46$, $b_{\mathrm{ZB}} = 0.27$ en $b_{\mathrm{AC}} = 0.16$ berekend. Op basis hiervan zijn de voorspelde standaardscores in de kolom $\widehat{\mathbf{vlg}}$ berekend uit de standaardscores op de predictorvariabelen (die ditmaal niet in de tabel zijn opgenomen). Voor de voorspelling in termen van de originele scores worden de bètagewichten op de inmiddels bekende wijze omgezet in ongestandaardiseerde $B$-gewichten, bijvoorbeeld nu voor de predictor AC:

$$B_{\mathrm{AC}} = b_{\mathrm{AC}} \frac{\mathrm{SD(VLG)}}{\mathrm{SD(AC)}} = 0.16 \times \frac{3.10}{3.79} = 0.13 \quad ,$$

en voor $A$:

$$A = 10.53 - (-0.22) \times 43.16 - 0.46 \times 21.41 - 0.13 \times 37.56 = 5.31 \quad .$$

Met behulp van deze gewichten volgen dan weer de voorspellingen in kolom $\widehat{\mathrm{VLG}}$, zoals in de eerste rij:

$$10.68 = 5.31 - 0.22 \times 41 + 0.46 \times 20 + 0.13 \times 40 \quad .$$

Zo'n regressievergelijking met drie predictoren, hier dus

$$\widehat{\mathrm{VLG}} = 5.31 \ - \ 0.22\,\mathrm{EM} \ + \ 0.46\,\mathrm{ZB} \ + \ 0.13\,\mathrm{AC} \quad ,$$

valt nauwelijks meer meetkundig voor te stellen. In totaal hebben we dan immers vier onderling loodrechte assen nodig, een voor elk van de variabelen VLG, EM, ZB en AC. Theoretisch gesproken bepaalt bovenstaande regressievergelijking dan een driedimensionale deelruimte in een vierdimensionale hyperruimte, maar hierbij kunnen we ons niet veel meer voorstellen.

Uit de bètagewichten valt ook te berekenen dat de proportie verklaarde variantie nu gelijk is aan

$$R^2 = (-0.46) \times (-0.55) \ + \ 0.27 \times 0.35 \ + \ 0.16 \times 0.24 = 0.39 \quad .$$

In overeenstemming hiermee, is in de opsplitsing

$$\begin{aligned} \mathrm{VAR(VLG)} &= \mathrm{VAR}(\widehat{\mathrm{VLG}}) \ + \ \mathrm{VAR}(\mathrm{VLG} - \widehat{\mathrm{VLG}}) \\ 9.62 &= 3.73 \ + \ 5.90 \quad , \end{aligned}$$

| pp | VLG | EM | ZB | AC | $\widehat{\mathbf{vlg}}$ | $\widehat{\text{VLG}}$ | RES |
|---|---|---|---|---|---|---|---|
| 1 | 9 | 41 | 20 | 40 | 0.05 | 10.68 | -1.68 |
| 2 | 4 | 55 | 22 | 35 | -0.87 | 7.83 | -3.83 |
| 3 | 11 | 42 | 20 | 44 | 0.14 | 10.98 | 0.02 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 30 | 15 | 34 | 23 | 40 | 1.00 | 13.63 | 1.37 |
| 31 | 10 | 52 | 22 | 40 | -0.44 | 9.15 | 0.85 |
| 32 | 9 | 46 | 19 | 39 | -0.50 | 8.97 | 0.03 |
| Gem | 10.53 | 43.16 | 21.41 | 37.56 | 0.00 | 10.53 | 0.00 |
| Var | 9.62 | 41.51 | 3.24 | 14.37 | 0.39 | 3.73 | 5.90 |
| SD | 3.10 | 6.44 | 1.80 | 3.79 | 0.62 | 1.93 | 2.43 |

| Cor. | VLG | EM | ZB |
|---|---|---|---|
| EM | -0.55 | | |
| ZB | 0.35 | -0.20 | -0.04 |
| AC | 0.24 | -0.20 | -0.04 |

$R^2 = 0.39$

$5.31 = A$

$b_{\text{EM}} = -0.46$ $\quad -0.22 = B_{\text{EM}}$

$b_{\text{ZB}} = 0.27$ $\quad 0.46 = B_{\text{ZB}}$

$b_{\text{AC}} = 0.16$ $\quad 0.13 = B_{\text{AC}}$

Tabel 4.5: De regressie van VLG op alle drie predictoren EM en ZB en AC. Zie tekst.

nu 39% (3.73/9.62) van de variantie van VLG via de voorspelling $\widehat{\text{VLG}}$ verklaard door de predictoren EM, ZB, en AC, terwijl nog 61% (5.90/9.62) van deze variantie onverklaard blijft. Vergeleken met de situatie met twee predictoren (zie Tabel 4.4) wordt er nu 3% meer variantie verklaard door AC als derde predictor toe te voegen.

De geringe bijdrage van de derde predictor blijkt ook uit de geringe afname van de standaardschattingsfout. Bij twee predictoren bedroeg deze 2.48 (zie Tabel 4.4), en dit blijkt nu in Tabel 4.5 slechts gedaald te zijn tot $\text{SD}(\text{VLG} - \widehat{\text{VLG}}) = 2.43$. Wat dit betreft levert het toevoegen van de derde predictor dus nauwelijks waarneembare winst. De waarde van 2.43 kan natuurlijk weer gemakkelijk gevonden worden met behulp van de meervoudige correlatiecoëfficiënt en de standaarddeviatie van VLG:

$$\text{SD}(\text{VLG} - \widehat{\text{VLG}}) = \sqrt{(1 - R^2)}\,\text{SD}(\text{VLG}) = \sqrt{0.61} \times 3.10 = 0.78 \times 3.10 = 2.43 \quad ,$$

waaruit meteen blijkt dat met het benutten van deze drie predictoren in de voorspelling nog steeds 78% van de oorspronkelijke voorspellingsfout aanwezig blijft.

Samenvattend concluderen we met betrekking tot dit voorbeeld, dat in hoeverre patiënten die het medisch-besliskundige programma doorlopen het te verstrekken advies ook zullen opvolgen (VLG) slechts in bescheiden mate te voorspellen valt op basis van de volgende predictoren: (1) hoe emotioneel belastend dit programma wordt gevonden (EM), (2) in hoeverre het wordt ervaren als een bevragen van de eigen mening (ZB) en (3) in hoeverre het als een acceptabele beslissingsprocedure wordt ervaren (AC). Informatie betreffende deze drie variabelen reduceert de gemiddelde kwadratische fout met 39%, wat neerkomt op een vermindering van de voorspellingsfout met 22%. Hierbij wordt

verreweg het grootste aandeel geleverd door de ervaren emotionele belasting. Het als acceptabel beoordeeld worden verschaft nauwelijks nog extra informatie.

## 4.6 De Semi-Partiële of Partcorrelatiecoëfficiënt

### 4.6.1 Theorie

In het voorafgaande is benadrukt dat, bij meerdere predictoren, "de bijdrage van een predictor aan de voorspelling" op twee manieren opgevat kan worden: (1) als bijdrage van deze predictor aan de voorspelde criteriumscores $\widehat{\mathbf{y}}$ en (2) als bijdrage van deze predictor aan de proportie verklaarde variantie. In deze paragraaf werken we dit tweede punt wat preciezer uit met behulp van de semi-partiële correlatiecoëfficiënt. Deze wordt in het Engels, en daardoor ook in de uitvoer van computerprogramma's voor data-analyse, meestal partcorrelatiecoëfficiënt genoemd, welke term we dan ook hier zullen aanhouden.

Om de idee van de partcorrelatie uit te leggen, vertrekken we van een regressievergelijking met twee predictoren $X_1$ en $X_2$. In de gestandaardiseerde versie:

$$\widehat{\mathbf{y}} = b_1\,\mathbf{x}_1 + b_2\,\mathbf{x}_2 \quad .$$

We willen onderzoeken welke bijdrage elk van de predictoren levert aan de proportie verklaarde variantie bovenop de proportie verklaarde variantie op basis van de andere predictor alleen. Bijvoorbeeld, wat is de bijdrage van $X_2$ bovenop $X_1$? Deze vraag kan beantwoord worden in drie stappen:

1. De predictorvariabele $X_2$ voegt slechts iets toe voorzover de score op $X_2$ nieuwe informatie bevat, onafhankelijk van de al in de score op $X_1$ aanwezige informatie. Andersom uitgedrukt, voorzover de score op $X_2$ al te voorspellen valt uit de score op $X_1$ voegt ze niets toe aan de proportie verklaarde variantie.

2. De score op $X_2$ wordt opgesplitst in een deel dat uit $X_1$ te voorspellen valt en in een residu-deel dat dan niet meer met $X_1$ samenhangt. Dit laatste deel wordt enigszins beeldend ook wel het *voor $X_1$ uitgezuiverde deel van $X_2$* genoemd en alleen dit deel van de $X_2$-score levert een extra bijdrage aan de proportie verklaarde variantie, gegeven dat $X_1$ al als predictor meedoet. (In plaats van "uitzuiveren" wordt ook vaak de term "uitpartialiseren" gebruikt.)

   Het uitzuiveren gebeurt door middel van een enkelvoudige regressieanalyse waarin $X_2$ voorspeld wordt op basis van $X_1$. De voorspelling van $X_2$ duiden we aan met $\widehat{X}_2$ en de voorspellingsfout met $X_2 - \widehat{X}_2$. Omdat $X_2 - \widehat{X}_2$ niet correleert met $\widehat{X}_2$ (net zoals $Y - \widehat{Y}$ niet correleert met $\widehat{Y}$), correleert $X_2 - \widehat{X}_2$ ook niet met $X_1$. De voorspelling $\widehat{X}_2$ is immers een lineaire functie van $X_1$. Daarom noemen we $X_2 - \widehat{X}_2$ het voor $X_1$ uitgezuiverde deel van $X_2$.

3. De grootte van de extra bijdrage van $X_2$ is direct afhankelijk van de correlatie van $Y$ met het voor $X_1$ uitgezuiverde deel van $X_2$. Deze correlatie heet de semi-partiële correlatie of, korter, de partcorrelatie van $Y$ met $X_2$, uitgezuiverd voor

$X_1$. Deze coëfficient wordt genoteerd met het symbool $r_{y(x_2 \cdot x_1)}$ of, korter, met $r_{y(2\cdot 1)}$. De partcorrelatie kan via de volgende formule worden uitgedrukt in de gewone correlatiecoëfficiënten $r_{2y}$, $r_{12}$ en $r_{1y}$:

$$\boxed{r_{y(2\cdot 1)} = \frac{r_{2y} - r_{12}\, r_{1y}}{\sqrt{1 - r_{12}^2}} \quad .}$$

Merk op dat dit veel lijkt op de formule voor het bètagewicht $b_2$ van $X_2$, gegeven op pagina 319. Het verschil is dat er nu een vierkantswortel verschijnt in de noemer. Een belangrijk verschil tussen de partcorrelatiecoëfficiënt en het bètagewicht is dat de partcorrelatiecoëfficiënt, zoals elke correlatiecoëfficiënt, tussen $-1$ en $+1$ ligt. Het bètagewicht $b_2$, daarentegen, is niet begrensd; het kan waarden aannemen tussen min en plus oneindig. Dit komt omdat het bètagewicht $b_2$ gelijk is aan de partcorrelatiecoëfficiënt gedeeld door de standaarddeviatie van het voor $\mathbf{x}_1$ uitgezuiverde deel van $\mathbf{x}_2$, die gelijk is aan $\sqrt{1 - r_{12}^2}$. Deze standaarddeviatie ligt tussen 0 en 1, en dit verklaart het feit dat het bètagewicht $b_2$ onbegrensd is. Partcorrelatiecoëfficiënt en bètagewicht hebben altijd hetzelfde teken en als de ene gelijk is aan nul dan is de andere ook gelijk aan nul.

Het blijkt nu dat we met behulp van deze partcorrelatie de proportie door $X_1$ en $X_2$ gezamenlijk verklaarde variantie van $Y$, die wordt gegeven door $R^2_{y\cdot 12}$, eenvoudig kunnen opsplitsen in een deel dat door $X_1$ alleen al wordt verklaard en in een extra deel dat wordt verklaard door $X_2$ als tweede predictor toe te voegen. Er geldt namelijk de volgende opsplitsing:

$$\boxed{R^2_{y\cdot 12} = r^2_{y1} \;+\; r^2_{y(2\cdot 1)} \quad ,}$$

waarin $r^2_{y1}$ de proportie door $X_1$ alleen verklaarde variantie van $Y$ weergeeft en waarin als nieuw element $r^2_{y(2\cdot 1)}$ verschijnt als de toename in proportie verklaarde variantie die te danken is aan het toevoegen van $X_2$ als tweede predictor. Wat betreft de notatie, moeten we hier opmerken dat de punt in het subscript van $R^2_{y\cdot 12}$ iets anders betekent dan de punt in het subscript van $r^2_{y(2\cdot 1)}$: in het eerste geval betekent het "voorspeld op basis van" en in het tweede geval "uitgezuiverd voor".

In het bovenstaande verhaal kunnen we de rol van $X_1$ en $X_2$ uiteraard verwisselen. Dit wil zeggen, we kunnen onderzoeken wat de extra bijdrage is van $X_1$ aan de proportie verklaarde variantie buiten de voorspelling die al op basis van $X_2$ wordt gedaan. Dit leidt dan tot het beschouwen van $r_{y(1\cdot 2)}$, de partcorrelatie van $Y$ met $X_1$ met daaruit $X_2$ uitgezuiverd. De formule hiervoor is natuurlijk dezelfde als die voor $r_{y(2\cdot 1)}$, met alleen de rol van de subscripten 1 en 2 verwisseld. Ook geldt de bovenstaande opsplitsing van $R^2_{y\cdot 12}$ met de rol van $X_1$ en $X_2$ verwisseld:

$$R^2_{y\cdot 12} = r^2_{y2} \;+\; r^2_{y(1\cdot 2)} \quad ,$$

waarin nu $r^2_{y2}$ de proportie verklaarde variantie door $X_2$ alleen weergeeft, en $r^2_{y(1\cdot 2)}$ de toename in proportie verklaarde variantie die te danken is aan het toevoegen van $X_1$ als tweede predictor.

**Twee Opsplitsingen**
Aangezien dezelfde $R^2$ op twee manieren opgesplitst kan worden, moet het eindresultaat van de twee opsplitsingen blijkbaar altijd identiek zijn. Met andere woorden, er moet altijd gelden dat

$$r_{y1}^2 \ + \ r_{y(2\cdot1)}^2 = r_{y2}^2 \ + \ r_{y(1\cdot2)}^2 \quad ,$$

wat zo op het eerste oog, ook na onmiddellijke substitutie van de formules voor $r_{y(2\cdot1)}$ en $r_{y(1\cdot2)}$, helemaal niet duidelijk is. Het vergt enig handig en/of vlijtig rekenwerk (voor de liefhebber!) om na te gaan of dit inderdaad altijd klopt. (Het handige rekenwerk begint met bovenstaande gelijkheid om te zetten in de equivalente gelijkheid $r_{y1}^2 - r_{y2}^2 = r_{y(1\cdot2)}^2 - r_{y(2\cdot1)}^2$, en dan rechts, na substitutie van de formules, in de teller de gelijkheid $P^2 - Q^2 = (P+Q)(P-Q)$ toe te passen.)

Bekijken we nu nog het speciale geval waarin de twee predictoren $X_1$ en $X_2$ ongecorreleerd zijn. We herinneren ons nog uit paragraaf 4.4 dat in deze uitzonderlijke situatie de gestandaardiseerde regressiegewichten $b_1$ en $b_2$ heel eenvoudig worden: ze worden gelijk aan de correlatie van $Y$ met de corresponderende predictor. Het blijkt dat hier ook de partcorrelaties heel eenvoudig worden: ze worden gelijk aan de gewone (niet-uitgezuiverde) correlaties. Als $X_1$ en $X_2$ ongecorreleerd zijn, valt er namelijk vanuit de score van een persoon op $X_1$ niets te voorspellen over zijn score op $X_2$ en vice versa. Dit impliceert dat het voor $X_1$ uitgezuiverde deel van $X_2$ gelijk is aan $X_2$ zelf (er valt niets uit te zuiveren) en daarmee is natuurlijk ook de partcorrelatie van $Y$ met dit uitgezuiverde deel van $X_2$ gewoon gelijk aan de correlatie van $Y$ met $X_2$. Op dezelfde manier vallen met betrekking tot $X_1$ partcorrelatie en gewone correlatie samen. Dit blijkt ook meteen uit de formule voor de partcorrelatie. Het ongecorreleerd zijn van $X_1$ en $X_2$ betekent dat $r_{12} = 0$ en wanneer we dit invullen in de hierboven gegeven formule voor $r_{y(2\cdot1)}$ vereenvoudigt deze inderdaad tot $r_{y(2\cdot1)} = r_{y2}$. Evenzo volgt, met verwisseling van $X_1$ en $X_2$, dat $r_{y(1\cdot2)} = r_{y1}$.

Als gevolg van deze gelijkheden wordt in het geval van ongecorreleerde predictoren ook de opsplitsing van $R^2$ in twee afzonderlijke bijdragen zeer eenvoudig. De twee mogelijke opsplitsingen die we boven hebben beschouwd (extra bijdrage van $X_2$ ten opzichte van $X_1$ en omgekeerd) vallen nu immers samen:

$$R_{y\cdot12}^2 = r_{y1}^2 \ + \ r_{y2}^2 \quad .$$

Zo blijken in dit geval de bijdragen van $X_1$ en $X_2$ aan de voorspelling ook volledig onafhankelijk (zonder overlap) te zijn: de proportie door $X_1$ en $X_2$ gezamenlijk verklaarde variantie is gewoon de proportie $r_{y1}^2$ die via een enkelvoudige regressie van $Y$ op $X_1$ zou worden verklaard, plus de proportie $r_{y2}^2$ die via een enkelvoudige regressie van $Y$ op $X_2$ zou worden verklaard. In het algemeen, met gecorreleerde predictoren, geldt dit dus niet, omdat dan partcorrelaties niet samenvallen met de corresponderende gewone correlaties.

De idee van partcorrelaties valt gemakkelijk uit te breiden naar de situatie met meer dan twee predictoren. Met een derde predictor $X_3$ erbij kunnen we ons afvragen wat de extra bijdrage is van deze $X_3$ aan de voorspelling van $Y$, buiten de voorspelling die al op basis van $X_1$ en $X_2$ kan worden gedaan. Via een redenering als hierboven komen we dan

tot de conclusie dat we hiervoor het deel van de $X_3$-score moeten bepalen dat niet door $X_1$ en $X_2$ samen al wordt voorspeld. De correlatie van $Y$ met dit voor $X_1$ en $X_2$ uitgezuiverde deel van $X_3$ zal de extra bijdrage van $X_3$ bepalen. Deze partcorrelatie wordt genoteerd als $r_{y(x_3 \cdot x_1 x_2)}$ of, korter, $r_{y(3 \cdot 12)}$. Via een formule, die te ingewikkeld en te weinig inzichtelijk is om hier te vertonen, kan deze partcorrelatie uitgedrukt worden in de gewone correlaties tussen alle betrokken variabelen. Tussen $r_{y(3 \cdot 12)}$ en het bètagewicht $b_3$ blijkt een soortgelijke betrekking te bestaan als die tussen $r_{y(2 \cdot 1)}$ en $b_2$ bij het uitzuiveren van één andere predictor: nu wordt $b_3$ verkregen door de partcorrelatie van $\mathbf{y}$ met het voor $\mathbf{x}_1$ en $\mathbf{x}_2$ uitgezuiverde deel van $\mathbf{x}_3$ te delen door de standaarddeviatie van dit uitgezuiverde deel van $\mathbf{x}_3$. De partcorrelatie ligt tussen -1 en +1, maar het bètagewicht $b_3$ is onbegrensd. Bètagewicht en partcorrelatie hebben altijd hetzelfde teken en het bètagewicht is dán gelijk aan nul als de desbetreffende partcorrelatie gelijk is aan nul.

De partcorrelatie met twee variabelen uitgezuiverd draagt op de verwachte manier bij aan een formule voor $R^2$: de verhoging van $R^2$ als gevolg van het toevoegen van $X_3$ is gelijk aan $r^2_{y(3 \cdot 12)}$. Voor drie predictoren kunnen we $R^2$ bijvoorbeeld opsplitsen als

$$\boxed{R^2_{y \cdot 123} = r^2_{y1} \ + \ r^2_{y(2 \cdot 1)} \ + \ r^2_{y(3 \cdot 12)} \quad ,}$$

waarbij we eerst de bijdrage van $X_1$ bekijken, dan de extra bijdrage van $X_2$ bovenop $X_1$ en tenslotte de extra bijdrage van $X_3$ ten opzichte van $X_1$ en $X_2$ gezamenlijk. Maar we kunnen ook volgens een willekeurige andere volgorde opsplitsen, zoals bijvoorbeeld

$$R^2_{y \cdot 123} = r^2_{y2} \ + \ r^2_{y(3 \cdot 2)} \ + \ r^2_{y(1 \cdot 23)} \quad ,$$

met eerst de bijdrage van $X_2$, dan de extra bijdrage van $X_3$ bovenop $X_2$ en tenslotte de extra bijdrage van $X_1$ bovenop $X_3$ en $X_2$ gezamenlijk. Er zijn natuurlijk nog andere opsplitsingen mogelijk, die alle optellen tot dezelfde $R^2_{y \cdot 123}$, de totale proportie door $X_1$, $X_2$ en $X_3$ gezamenlijk verklaarde variantie. In het speciale geval van drie onderling ongecorreleerde predictoren (dus $r_{12} = 0$, $r_{13} = 0$ en $r_{23} = 0$) vallen, om dezelfde reden als bij twee predictoren, alle partcorrelaties weer samen met de corresponderende gewone correlaties. De totale proportie verklaarde variantie is dan gelijk aan de som van de proporties die door de afzonderlijke predictoren in een enkelvoudige regressieanalyse zouden worden verklaard:

$$R^2_{y \cdot 123} = r^2_{y1} \ + \ r^2_{y2} \ + \ r^2_{y3} \quad .$$

Er is dan geen enkele overlap in de bijdragen aan de voorspelling.

### 4.6.2 Voorbeeld

In de bespreking van Tabel 4.4 zagen we dat het toevoegen van ZB als tweede predictor de proportie verklaarde variantie verhoogde van 0.30 tot 0.36, dus met 0.06. Als we nu de partcorrelatie van VLG met ZB uitgezuiverd voor EM berekenen, vinden we

$$r_{\mathrm{VLG(ZB \cdot EM)}} = \frac{r_{\mathrm{VLG,ZB}} - r_{\mathrm{EM,ZB}}\, r_{\mathrm{VLG,EM}}}{\sqrt{1 - r^2_{\mathrm{EM,ZB}}}} = \frac{0.35 - (-0.20) \times (-0.55)}{\sqrt{1 - (-0.20)^2}} = 0.25 \quad ,$$

zodat inderdaad

$$r^2_{\mathrm{VLG(ZB \cdot EM)}} = 0.25^2 = 0.06 \quad .$$

Dit illustreert de in deze paragraaf gegeven opsplitsing van de proportie door twee predictoren verklaarde variantie:

$$\begin{aligned} R^2_{\mathrm{VLG \cdot EM,ZB}} &= r^2_{\mathrm{VLG,EM}} \ + \ r^2_{\mathrm{VLG(ZB \cdot EM)}} \\ 0.36 &= 0.30 \ + \ 0.06 \quad . \end{aligned}$$

In een enkelvoudige regressieanalyse met ZB als enige predictor zou deze variabele een proportie $r^2_{\mathrm{VLG,ZB}} = 0.35^2 = 0.12$ van de variantie van VLG verklaren. Als dan EM als tweede predictor wordt toegevoegd, komt de partcorrelatie van VLG met EM uitgezuiverd voor ZB uit op

$$r_{\mathrm{VLG(EM \cdot ZB)}} = \frac{r_{\mathrm{VLG,EM}} - r_{\mathrm{EM,ZB}}\, r_{\mathrm{VLG,ZB}}}{\sqrt{1 - r^2_{\mathrm{EM,ZB}}}} = \frac{-0.55 - (-0.20) \times 0.35}{\sqrt{1 - (-0.20)^2}} = -0.49 \quad ,$$

zodat met

$$r^2_{\mathrm{VLG(EM \cdot ZB)}} = (-0.49)^2 = 0.24$$

de 36% verklaarde variantie nu opgesplitst kan worden als

$$\begin{aligned} R^2_{\mathrm{VLG \cdot EM,ZB}} &= r^2_{\mathrm{VLG,ZB}} \ + \ r^2_{\mathrm{VLG(EM \cdot ZB)}} \\ 0.36 &= 0.12 \ + \ 0.24 \quad . \end{aligned}$$

Als nu AC als derde predictor wordt toegevoegd, kan via een ingewikkelde formule (die hierboven niet gegeven werd) de partcorrelatie van VLG met AC uitgezuiverd voor EM en ZB worden berekend als

$$r_{\mathrm{VLG(AC \cdot EM,ZB)}} = 0.17 \quad ,$$

zodat

$$r^2_{\mathrm{VLG(AC \cdot EM,ZB)}} = 0.17^2 = 0.03 \quad .$$

Hiermee komt, zoals we in de bespreking van Tabel 4.5 hebben gezien, de totale proportie verklaarde variantie inderdaad op 0.39:

$$\begin{aligned} R^2_{\mathrm{VLG \cdot EM,ZB,AC}} &= R^2_{\mathrm{VLG \cdot EM,ZB}} \ + \ r^2_{\mathrm{VLG(AC \cdot EM,ZB)}} \\ 0.39 &= 0.36 \ + \ 0.03 \quad . \end{aligned}$$

---

*Opgave 7*

We voeren een regressieanalyse uit op de gegevens uit ons voorbeeld, maar nu met EM en AC als predictoren, in plaats van EM en ZB.

1. Wat zijn de (gestandaardiseerde) bètagewichten $b_{\mathrm{EM}}$ en $b_{\mathrm{AC}}$?
2. Wat zijn de (ongestandaardiseerde) $B$-gewichten $B_{\mathrm{EM}}$ en $B_{\mathrm{AC}}$ en het intercept $A$?

3. Wat zijn de voorspelde VLG-waarden voor persoon 1 en persoon 2?
4. Wat is de proportie verklaarde variantie?
5. Wat is de verklaarde en de onverklaarde variantie?
6. Hoe groot is de reductie van de voorspellingsfout door gebruik te maken van EM en AC?

*Opgave 8*
Stel dat je over drie variabelen beschikt: $X$, $Y$ en $Z$. De correlaties tussen deze drie variabelen worden gegeven in Tabel 4.6.

| | $Y$ | $X$ |
|---|---|---|
| $X$ | 0.70 | |
| $Z$ | 0.45 | 0.30 |

Tabel 4.6: Tabel bij opgave 8.

1. Je wilt $Y$ voorspellen op basis van $X$. Hoe groot is de proportie verklaarde variantie?
2. Je wilt $Y$ voorspellen op basis van $X$ en $Z$. Hoe groot is de proportie verklaarde variantie?

*Opgave 9*
In Tabel 4.7 staan de correlaties tussen een criteriumvariabele ($Y$) en vier predictorvariabelen ($X_1$, $X_2$, $X_3$ en $X_4$).

| | $Y$ | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|---|
| $X_1$ | 0.70 | | | |
| $X_2$ | 0.70 | 0.99 | | |
| $X_3$ | 0.28 | 0.40 | 0.40 | |
| $X_4$ | 0.00 | 0.36 | 0.36 | 0.00 |

Tabel 4.7: Tabel bij opgave 9.

1. Zou je $X_1$ opnemen in een enkelvoudige regressievergelijking? Motiveer.
2. Zou je $X_4$ opnemen in een enkelvoudige regressievergelijking? Motiveer.
3. Zou je $X_2$ naast $X_1$ opnemen in een meervoudige regressievergelijking met twee predictoren? Motiveer.
4. Zou je $X_3$ naast $X_1$ opnemen in een meervoudige regressievergelijking met twee predictoren? Motiveer.
5. Zou je $X_4$ naast $X_1$ opnemen in een meervoudige regressievergelijking met twee predictoren? Motiveer.

*Opgave 10*
Een ontwikkelingspsycholoog wilde een voorspelling maken van het leesniveau waarop LOM-kinderen uit Groep 4 zich over één jaar zouden bevinden. Als predictor gebruikte hij het huidige leesniveau, zoals vastgesteld met een test voor technisch lezen (LEZENGR4). Het criterium was eveneens de score op een test voor technisch lezen, maar een met een wat hogere moeilijkheidsgraad, afgenomen in Groep 5 (LEZENGR5). De variabelen LEZENGR4 en LEZENGR5 bleken 0.70 te correleren.

1. Hoe groot is de proportie variantie in LEZENGR5 die verklaard wordt door LEZENGR4?

2. Een collega raadde hem aan om, naast LEZENGR4, ook het gemiddeld aantal boeken dat een kind per maand leest (BOEKEN) als predictor te gebruiken. De proportie verklaarde variantie door beide predictoren samen was 0.85. De correlatie tussen BOEKEN en LEZENGR4 was 0.0. Wat is dan de correlatie tussen BOEKEN en het criterium, LEZENGR5? Laat zien hoe je aan je antwoord komt.

*Opgave 11*
In Tabel 4.8 worden de correlaties gegeven tussen een criterium $Y$ en zes predictoren $X_1$ t/m $X_6$. Stel dat $X_1$ al is opgenomen als predictor en dat er nog gezocht wordt naar een geschikte tweede predictor. Bespreek voor de andere mogelijke predictoren ($X_2$ t/m $X_6$) waarom je deze wel of niet als tweede predictor op zou nemen.

| | $Y$ | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ |
|---|---|---|---|---|---|---|
| $X_1$ | 0.80 | | | | | |
| $X_2$ | 0.40 | 0.50 | | | | |
| $X_3$ | 0.80 | 1.00 | 0.50 | | | |
| $X_4$ | 0.20 | 0.10 | 0.10 | 0.10 | | |
| $X_5$ | 0.00 | 0.00 | 0.10 | 0.00 | 0.30 | |
| $X_6$ | -0.30 | -0.10 | 0.10 | -0.10 | 0.00 | -0.20 |

Tabel 4.8: Tabel bij opgave 11.

*Opgave 12*
We beschouwen de regressie van een criterium $Y$ op twee predictoren $X_1$ en $X_2$. Waar of niet waar, en leg uit:

1. Als $Y$ en $X_2$ ongecorreleerd zijn, dan is het gewicht van $X_2$ in de regressievergelijking gelijk aan 0.
2. Als $X_2$ in de regressievergelijking een gewicht gelijk aan 0 heeft, dan zijn $Y$ en $X_2$ ongecorreleerd.

*Opgave 13*
Een klinisch psycholoog wilde een voorspelling maken van de opnameduur bij patiënten met een neurose. Hij maakte hierbij gebruik van regressieanalyse. Zijn criterium was het aantal dagen dat de patiënt opgenomen is (DAGENOPNAME). Als predictor maakt hij gebruik van een zelfbeoordelingsvragenlijst waarin gepeild wordt naar de mate waarin de patiënt zelf last ervaart van zijn symptomen (de zogenaamde subjectieve-lastschaal). De scores op deze schaal werden aangeduid als SUBJLAST. De variabelen DAGENOPNAME en SUBJLAST correleerden 0.50.

1. Hoe groot is de proportie variantie in DAGENOPNAME die verklaard wordt door SUBJLAST?
2. Een collega raadde hem aan om, naast SUBJLAST, ook het oordeel van een psychiater over de ernst van problematiek (OPSYCHIA) als predictor te gebruiken. Dit oordeel werd uitgedrukt op een schaal van 1 (zeer milde stoornis) tot 10 (zeer ernstige stoornis). De proportie verklaarde variantie door beide predictoren samen was 0.61. De correlatie tussen OPSYCHIA en SUBJLAST is gelijk aan 0.0. Wat is dan de correlatie tussen OPSYCHIA en DAGENOPNAME? Laat zien hoe je aan je antwoord komt.
3. Als je weet dat er geen correlatie is tussen SUBJLAST en OPSYCHIA, is het dan nog zinvol om de partcorrelatie te berekenen tussen DAGENOPNAME en OPSYCHIA uitgezuiverd voor SUBJLAST? Motiveer je antwoord.

*Opgave 14*
Een ontwikkelingspsycholoog wilde een voorspelling maken van het gemak waarmee kinderen leren lezen in het begin van het basisonderwijs. Haar criteriumvariabele was de score op een leestest (LEESTEST). Als predictor verwachtte zij veel van een zogenaamde benoemingstest. De items van een benoemingstest zijn plaatjes met voorwerpen die kinderen moeten benoemen. Deze test kan afgenomen worden op jonge leeftijd, voordat met het leesonderwijs begonnen is. De score op de benoemingstest werd aangeduid als BENOEMTEST. Haar verwachting bleek te kloppen: BENOEMTEST en LEESTEST correleerden 0.60.

1. Hoe groot is de proportie variantie in LEESTEST die verklaard wordt door BENOEMTEST?
2. Een collega raadde haar aan om, naast BENOEMTEST, ook een zogenaamde schoolrijpheidstest als predictor te gebruiken, en dan een méérvoudige regressieanalyse uit te voeren. De score op deze schoolrijpheidstest werd aangeduid als SCHRIJPTEST. De variabelen SCHRIJPTEST en LEESTEST correleerden 0.50. De proportie verklaarde variantie door beide predictoren samen was 0.44. De psychologe was verbaasd dat dit minder was dan de som van de proporties verklaarde variantie door de twee predictoren afzonderlijk (binnen enkelvoudige regressieanalyses). Hoe kun je dit verklaren?
3. De psychologe overwoog om, naast BENOEMTEST (SCHRIJPTEST werd nu niet gebruikt als predictor), ook een non-verbale intelligentietest als predictor op te nemen. De score op deze test werd aangeduid als NVINTTEST. De variabelen NVINTTEST en LEESTEST correleerden niet met elkaar. De nulcorrelatie met het criterium gaf de psychologe weinig hoop op een bijdrage van de predictor aan de voorspelling. Was deze verwachting terecht? Motiveer je antwoord.

*Opgave 15*
Een RIAGG heeft een behandelprogramma opgezet voor werkloze ex-psychiatrisch patiënten. De doelstelling van het programma is het klaarstomen van de deelnemers voor een herintrede op de arbeidsmarkt. De verantwoordelijke psychologe wilde een voorspelling maken van het aantal weken dat elk van de deelnemers nodig had om een betaalde baan te vinden (ZOEKTIJD). Als predictor verwachtte de psychologe veel van de score op een schaal voor communicatieve vaardigheden (COMVAARD). Deze verwachting bleek te kloppen want COMVAARD correleerde -0.60 met ZOEKTIJD.

1. Omdat COMVAARD negatief correleerde met ZOEKTIJD verwachtte een collega van de psychologe dat de proportie verklaarde variantie negatief zou zijn. Ben je het hiermee eens? Motiveer je antwoord.
2. De psychologe vroeg zich af of het gebruik van een tweede predictor (naast COMVAARD) de voorspelling zou verbeteren. Concreet, zij was van plan om een meervoudige regressieanalyse uit te voeren met als predictoren COMVAARD en de score op een extraversievragenlijst (EXTRAV). De variabele EXTRAV correleerde -0.60 met ZOEKTIJD en 0.50 met COMVAARD. Hoeveel steeg de proportie verklaarde variantie door het toevoegen van EXTRAV als predictor naast COMVAARD?
3. De psychologe overwoog ook of een test voor schriftelijke taalvaardigheid (TAALVAARD) de voorspelling zou verbeteren. De variabelen TAALVAARD en ZOEKTIJD correleerden niet met elkaar. Mag je hieruit besluiten dat TAALVAARD de proportie verklaarde variantie niet zou doen stijgen als men haar zou opnemen als predictor naast COMVAARD (de variabele EXTRAV werd niet gebruikt als predictor)? Motiveer je antwoord.

*Opgave 16*
Een onderzoekster beschikt over 10 predictoren waarmee zij een bepaalde criteriumvariabele wil voorspellen. Met een meervoudige regressieanalyse met 5 van deze 10 predictoren behaalt

zij een proportie verklaarde variantie van 0.98. Wat kan er gebeuren als de overblijvende 5 predictoren aan de regressievergelijking toegevoegd worden:

1. Verandert de proportie verklaarde variantie en waar hangt dit eventueel van af?
2. Veranderen de regressiegewichten van de eerste vijf predictoren en waar hangt dit eventueel van af?

---

## 4.7 Enige Problemen bij de Interpretatie van Regressiegewichten

Gewapend met het begrip partcorrelatie kunnen we nu wat preciezer een aantal gevallen beschrijven waarin een naïeve interpretatie van de verkregen regressiegewichten zeer misleidend zou zijn. We zullen enkele typische voorbeelden geven, waaruit zal blijken dat er geen algemeen, inzichtelijk verband is tussen bètagewichten en gewone correlaties of tussen partcorrelaties en gewone correlaties.

### 4.7.1 Onechte Correlatie

#### Voorbeeld en Omschrijving

Als eerste voorbeeld, stel dat bij leerlingen van een lagere school een intelligentietest is afgenomen. De scores op deze intelligentietest zijn het criterium en ze worden aangeduid als IQ. In de correlatiematrix in Tabel 4.9 staan de correlaties tussen IQ en twee predictoren, de leeftijd en het gewicht van de kinderen, aangeduid als LEEFTIJD ($X_1$) en GEWICHT ($X_2$). Als we nu met deze gegevens IQ willen voorspellen met behulp van een regressievergelijking met LEEFTIJD als enige predictor, dan zien we dat LEEFTIJD een bètagewicht van $r_{1y} = 0.60$ zou krijgen:

$$\widehat{\mathbf{y}} = 0.60\,\mathbf{x}_1 \quad ,$$

en de predictor LEEFTIJD alleen verklaart daarmee 36% van de variantie van IQ:

$$R^2_{y\cdot 1} = r^2_{1y} = 0.36 \quad .$$

In een regressievergelijking met GEWICHT als enige predictor krijgt deze predictor een bètagewicht van $r_{2y} = 0.48$. Dus,

$$\widehat{\mathbf{y}} = 0.48\,\mathbf{x}_2 \quad ,$$

waarmee de predictor GEWICHT alleen 23% van de variantie van IQ verklaart:

$$R^2_{y\cdot 2} = r^2_{2y} = 0.23 \quad .$$

Gezien de substantiële bijdrage van GEWICHT, kunnen we alleen maar ontevreden zijn met het resultaat van de regressie van IQ op beide predictoren (LEEFTIJD en GEWICHT) gezamenlijk. Voor de regressievergelijking

$$\widehat{\mathbf{y}} = b_1\,\mathbf{x}_1 + b_2\,\mathbf{x}_2$$

| | | $Y$ | $X_1$ | |
|---|---|---|---|---|
| IQ | $Y$ | 1 | | $R^2_{y\cdot 12} = 0.36$ |
| LEEFTIJD | $X_1$ | 0.60 | 1 | $R^2_{y\cdot 1} = 0.36$ |
| GEWICHT | $X_2$ | 0.48 | 0.80 | $R^2_{y\cdot 2} = 0.23$ |

bètagewichten :

$$b_1 = \frac{0.60 - 0.80 \times 0.48}{1 - (0.80)^2} = \frac{0.216}{0.36} = 0.60$$

$$b_2 = \frac{0.48 - 0.80 \times 0.60}{1 - (0.80)^2} = \frac{0}{0.36} = 0$$

partcorrelatiecoëfficiënt :

$$r_{y(2\cdot 1)} = \frac{0.48 - 0.80 \times 0.60}{\sqrt{1 - (0.80)^2}} = 0$$

Tabel 4.9: Voorbeeld van onechte correlatie die niet bijdraagt in de regressievergelijking.

berekenen we met behulp van de in paragraaf 4.4 gegeven formules voor $b_1$ en $b_2$ (zie Tabel 4.9):

$$b_1 = 0.60$$
$$b_2 = 0 \quad .$$

De predictor LEEFTIJD heeft hier hetzelfde gewicht als bij de enkelvoudige regressie, maar de predictor GEWICHT draagt nu met een bètagewicht $b_2 = 0$ helemaal niets bij:

$$R^2_{y\cdot 12} = b_1\, r_{1y} + b_2\, r_{2y} = 0.60 \times 0.60 \ + \ 0 \times 0.48 = 0.36 \quad .$$

Terwijl GEWICHT met $r_{2y} = 0.48$ toch een behoorlijke correlatie heeft met IQ, blijkt deze variabele niets bij te dragen aan een voorspelling van IQ als de variabele LEEFTIJD al als predictor is opgenomen.

De verklaring hiervoor is dat in zo'n geval niet de correlatie van IQ met GEWICHT van belang is, maar de partcorrelatie van IQ met GEWICHT uitgezuiverd voor LEEFTIJD, dus met dat deel van de GEWICHT-scores dat niet voorspeld kan worden uit LEEFTIJD. Dan blijkt inderdaad $r_{y(2\cdot 1)} = 0$ (zie Tabel 4.9). De inhoudelijke interpretatie hiervan is dat de geconstateerde samenhang tussen IQ en GEWICHT geheel via de variabele LEEFTIJD loopt: zwaardere leerlingen behalen gemiddeld hogere IQ-scores alleen maar omdat zwaardere leerlingen gemiddeld ouder zijn, en deze behalen gemiddeld hogere IQ-scores dan jongere leerlingen. Anders gezegd, $r_{y(2\cdot 1)} = 0$ drukt uit dat, bínnen een groep leerlingen van een bepaalde leeftijd (bijvoorbeeld, 8 jaar), er geen verband bestaat tussen GEWICHT en IQ. Algemeen geformuleerd, gegeven een bepaalde $X_1$-score is er geen samenhang tussen $Y$ en $X_2$. Daarom, als de $X_1$-score (hier, LEEFTIJD) reeds als predictor is opgenomen, geeft een $X_2$-score (hier, GEWICHT) geen extra informatie over de te verwachten $Y$-score (hier, IQ).

### Definitie

In het beschreven voorbeeld noemt men de geobserveerde correlatie tussen IQ en GEWICHT wel eens *onecht* (in het Engels, *spurious*). Dit is eigenlijk geen goede term, want de correlatie bestaat, is dus echt; men wil er echter mee aangeven dat zij geen enkel causaal verband tussen de twee variabelen weergeeft, maar volledig verklaard kan worden door wederzijdse verbanden met een derde variabele (hier dus LEEFTIJD). Deze gedachte wordt ook verwoord in de definitie van een onechte correlatie. Deze definitie houdt het volgende in:

1. De criteriumvariable (in het voorbeeld, IQ) correleert substantieel (positief of negatief) met elk van de twee predictoren (in het voorbeeld, LEEFTIJD en GEWICHT).

2. Eén van de twee predictoren (LEEFTIJD) correleert sterker met de criteriumvariabele dan de andere predictor (GEWICHT).

3. De partcorrelatie van de criteriumvariabele met de zwak correlerende predictor uitgezuiverd voor de sterk correlerende predictor, is gelijk aan 0.

Deze definitie geeft goed de gedachte weer die achter het begrip onechte correlatie zit. Echter, deze definitie is in de praktijk vaak moeilijk toepasbaar. Daar zijn twee redenen voor:

1. Het is onduidelijk wanneer een correlatie substantieel is. Is 0.20 genoeg, 0.30, 0.40 ...?

2. Een partcorrelatie is in de praktijk vrijwel nooit precies gelijk aan nul.

Om deze definitie in de praktijk te kunnen toepassen heeft men twee vuistregels nodig die ons zeggen (1) wanneer een correlatie substantieel is (bijvoorbeeld, groter dan 0.20), en (2) wanneer een partcorrelatie klein genoeg is als nul beschouwd te kunnen worden (bijvoorbeeld, kleiner dan 0.05). Deze vuistregels zijn arbitrair en daarom ook vatbaar voor kritiek. De gedachte achter het begrip onechte correlatie hangt echter niet af van deze arbitraire vuistregels.

Het voorbeeld dat we hebben beschouwd is duidelijk een geconstrueerd voorbeeld: enerzijds met een bètagewicht en partcorrelatie die precies op nul uitkomen, anderzijds met variabelen die, inhoudelijk gezien, al suggereren wat er aan de hand is. Dit is gedaan om het behandelde verschijnsel zo duidelijk en zuiver mogelijk te illustreren. In de praktijk worden bètagewichten en partcorrelatie natuurlijk nooit precies nul, maar dat is ook niet essentieel. Het gaat erom dat een bepaalde variabele, die behoorlijk hoog met het criterium correleert, gegeven één of meerdere andere predictoren een onbeduidend bètagewicht krijgt in de regressievergelijking en een onbeduidende partcorrelatie. Blijkbaar is de samenhang tussen deze variabele en het criterium vrijwel geheel terug te voeren op die andere predictor(en). In de praktijk, anders dan in een geconstrueerd voorbeeld, volgt deze conclusie alleen uit de data-analyse (i.c., de berekening van het bètagewichten en de partcorrelatie), zonder dat het al vooraf duidelijk hoeft te zijn vanuit een inhoudelijke beschouwing van de variabelen. De data-analyse voegt dan iets toe

aan de inhoudelijke interpretatie van de relatie tussen verschillende variabelen. Soortgelijke opmerkingen over de gekunsteldheid van de voorbeelden en de functie daarvan gelden ook voor wat nog volgt.

### 4.7.2 Suppressie

**Voorbeeld en Omschrijving**

Nu volgt een tweede voorbeeld waarin bètagewicht en correlatie tot schijnbaar tegenstrijdige conclusies leiden betreffende de samenhang tussen twee variabelen. Terwijl hierboven werd geïllustreerd dat soms een behoorlijke correlatie kan samengaan met een bètagewicht van nul in de regressievergelijking, gaan we nu zien dat het omgekeerde ook mogelijk is. Onder sommige omstandigheden kan een variabele die geheel niet correleert met de criteriumvariabele toch met een interessant bètagewicht in de regressievergelijking verschijnen, en daarmee wezenlijk bijdragen aan de voorspelling van dit criterium. Het voorbeeld is duidelijk weer geconstrueerd en behelst drie fysieke variabelen, die psychologisch niet erg interessant zijn, maar die het wel mogelijk maken een inhoudelijke verklaring te geven van wat er aan de hand is. In het midden van de bespreking zullen we aangeven dat er ook psychologisch interessante invullingen van deze variabelen denkbaar zijn waarbij hetzelfde verschijnsel optreedt.

In Tabel 4.10 staan de correlaties tussen LENGTE, GEWICHT en BUIKOMVANG in een steekproef van volwassen mannen. Opvallend is hier natuurlijk de nulcorrelatie tussen LENGTE en BUIKOMVANG. Grofweg geeft dit weer dat dik zijn en dun zijn in dezelfde verhouding voorkomt bij korte mensen en bij lange mensen, wat best een redelijke benadering van de werkelijkheid zou kunnen zijn. De precieze waarde van nul is uiteraard weer gekozen voor illustratieve doeleinden. We stellen ons nu als taak om op basis van de gegevens de variabele LENGTE te voorspellen met de variabelen GEWICHT en BUIKOMVANG als mogelijke predictoren. Uit de correlaties is direct duidelijk hoe de regressievergelijkingen van LENGTE op de predictoren afzonderlijk eruit zien. Als enkelvoudige predictor krijgt GEWICHT een gestandaardiseerd regressiegewicht gelijk aan $r_{1y} = 0.70$, waarmee 49% ($r_{1y}^2 = 0.49$) van de variantie van LENGTE wordt verklaard. Als enkelvoudige predictor presteert BUIKOMVANG uiteraard niets, omdat hij ongecorreleerd is met LENGTE: $r_{2y} = 0$, dus het bètagewicht wordt nul en 0% van de variantie van LENGTE wordt verklaard. Naïef verwacht men dan dat GEWICHT en BUIKOMVANG het samen niet beter zullen doen dan GEWICHT alleen, maar dit blijkt heel anders uit te pakken. Bij een regressieanalyse op de beide predictoren verkrijgen we als gestandaardiseerde regressiegewichten (zie Tabel 4.10):

$$\begin{aligned} b_1 &= 1.09 \\ b_2 &= -0.66 \quad , \end{aligned}$$

die leiden tot

$$R^2_{y\cdot 12} = b_1\, r_{1y} + b_2\, r_{2y} = 1.09 \times 0.70 \;+\; (-0.66) \times 0 = 0.77 \quad .$$

Nu krijgt de niet met het criterium gecorreleerde predictor dus een bètagewicht dat duidelijk van nul afwijkt. Het resultaat is dat het toevoegen van een predictor die op

zich 0% van de variantie van het criterium verklaart de proportie verklaarde variantie dus heeft verhoogd van 49% naar 77%, een stijging met zo'n 28%.

| | | $Y$ | $X_1$ | |
|---|---|---|---|---|
| LENGTE | $Y$ | 1 | | $R^2_{y\cdot 12} = 0.77$ |
| GEWICHT | $X_1$ | 0.70 | 1 | $R^2_{y\cdot 1} = 0.49$ |
| BUIKOMVANG | $X_2$ | 0.00 | 0.60 | $R^2_{y\cdot 2}$=0 |

bètagewichten :

$$b_1 = \frac{0.70 - 0.60 \times 0}{1 - (0.60)^2} = \frac{0.70}{0.64} = 1.09$$

$$b_2 = \frac{0 - 0.60 \times 0.70}{1 - (0.60)^2} = \frac{-0.42}{0.64} = -0.66$$

partcorrelatiecoëfficiënt :

$$r_{y(2\cdot 1)} = \frac{0 - 0.60 \times 0.70}{\sqrt{1 - (0.60)^2}} = -0.525$$

Tabel 4.10: Voorbeeld van suppressie, waarin een niet met het criterium correlerende variabele toch wezenlijk kan bijdragen aan de voorspelling.

Dit verschijnsel wordt *suppressie* genoemd en in deze situatie heet de predictor $X_2$ (BUIKOMVANG), die niet met het criterium maar wel met een andere predictor $X_1$ (GEWICHT) correleert, een *suppressorvariabele* voor $X_1$. De idee achter deze benaming is, dat de functie van $X_2$ in de regressie erin bestaat een bepaalde voor de voorspelling irrelevante component van de andere predictor $X_1$ te onderdrukken. We zouden ook kunnen zeggen dat $X_2$ corrigeert voor deze component van $X_1$. In dit geval wordt LENGTE voorspeld door GEWICHT, en tussen deze variabelen bestaat inderdaad een samenhang. Maar anderzijds wordt de predictor GEWICHT ook gedeeltelijk bepaald door hoe dik iemand is. Het is dus nuttig voor deze component te corrigeren. Kort gezegd, het gewicht van een persoon moet uit de lengte of uit de breedte komen en om te weten in hoeverre het uit de lengte komt (voor de voorspelling van LENGTE) is het erg nuttig te weten in hoeverre het uit de breedte komt (via BUIKOMVANG). Anders geformuleerd, binnen de groep van personen met een bepaalde $X_1$-score (hier GEWICHT) is er wel degelijk een sterke (hier negatieve) samenhang tussen $Y$ (hier LENGTE) en de suppressorvariabele $X_2$ (hier BUIKOMVANG). Dit feit komt precies tot uiting in de partcorrelatie van LENGTE met BUIKOMVANG uitgezuiverd voor GEWICHT: $r_{y(2\cdot 1)} = -0.525$ (zie Tabel 4.10). De door de toevoeging van $X_2$ veroorzaakte stijging van de proportie verklaarde variantie kan met de op pagina 334 gegeven opsplitsingsformule

voor $R^2$ ook precies aan deze partcorrelatie worden toegeschreven:

$$\begin{aligned} R^2_{y\cdot 12} &= r^2_{y1} \ + \ r^2_{y(2\cdot 1)} \\ &= (0.70)^2 \ + \ (-0.525)^2 \\ &= 0.49 \ + \ 0.28 = 0.77 \quad . \end{aligned}$$

Het voorbeeld met deze fysieke variabelen is zo gekozen om het fenomeen suppressie zo duidelijk mogelijk naar voren te laten komen. Het is niet moeilijk om ook puur psychologische variabelen te bedenken waarbij dit fenomeen optreedt. Bijvoorbeeld, vervang in het bovenstaande LENGTE door IQ, GEWICHT door SCHOOLSUCCES en BUIKOMVANG door MOTIVATIE, dan zou in principe hetzelfde verhaal op kunnen gaan. Bij een variabele als SCHOOLSUCCES kan het ook, hetzij uit de lengte (IQ), hetzij uit de breedte (MOTIVATIE) komen. Daarom, ook al correleren over alle personen IQ en MOTIVATIE nauwelijks met elkaar, mogen we weer wel een (negatieve) samenhang verwachten tussen deze twee variabelen binnen een groep personen met een gelijke mate van SCHOOLSUCCES. Kortom, we zouden hetzelfde patroon kunnen vinden als in Tabel 4.10.

### Definitie

Het kunstmatige voorbeeld met LENGTE, GEWICHT en BUIKOMVANG is gekozen om het begrip suppressie zo duidelijk mogelijk te formuleren. In de praktijk is het meestal niet zo duidelijk of er in een bepaalde situatie wel of geen sprake is van suppressie. Dit roept de vraag op hoe suppressie precies gedefinieerd is.

Het begrip suppressie omvat twee ingrediënten: (1) een bepaald patroon van correlaties tussen de criteriumvariabele en de twee predictorvariabelen, en (2) een bepaalde inhoudelijke verklaring van dit patroon. Het tweede ingrediënt, de inhoudelijke verklaring, is vooral van historisch belang. We beginnen met het eerste ingrediënt, het patroon in de correlaties. Voor wat betreft dit patroon, moeten we een onderscheid maken tussen een strikte en een ruime definitie. De strikte definitie houdt het volgende in:

1. Een substantiële correlatie (positief of negatief) tussen de criteriumvariabele en één van de twee predictoren (in het voorbeeld, LENGTE en GEWICHT).

2. Een substantiële correlatie (positief of negatief) tussen de twee predictoren (in het voorbeeld, GEWICHT en BUIKOMVANG).

3. (Ondanks 1. en 2.) Een nulcorrelatie tussen de criteriumvariabele en de tweede predictor (in het voorbeeld, LENGTE en BUIKOMVANG).

Net zoals de definitie van een onechte correlatie, is ook deze definitie in de praktijk vaak moeilijk toe te passen. Daar zijn opnieuw twee redenen voor:

1. Het is onduidelijk wanneer een correlatie substantieel is. Is 0.20 genoeg, 0.30, 0.40 ...?

2. Een correlatie is in de praktijk vrijwel nooit precies gelijk aan nul.

De ruime definitie van suppressie is veel beter toe te passen in de praktijk. Dit komt omdat de ruime definitie (1) geen gewag maakt van een "substantiële correlatie", en (2) niet vereist dat de suppressorvariabele een nulcorrelatie heeft met het criterium. Deze ruime definitie vertrekt van de gedachte dat een suppressorvariabele $X_2$, die zwak correleert met het criterium, ervoor zorgt dat een andere predictor ($X_1$), die (in absolute waarde) sterker correleert met het criterium, beter gaat voorspellen. Met name, de bijdrage van $X_1$ (in ons voorbeeld, GEWICHT) aan de proportie verklaarde variantie is groter als de suppressorvariabele $X_2$ (in ons voorbeeld, BUIKOMVANG) mee opgenomen wordt in de regressievergelijking dan wanneer zij alleen in de regressievergelijking zit. Dit kan ook uitgedrukt worden in termen van een gekwadrateerde partcorrelatie die vergeleken wordt met een gekwadrateerde gewone correlatie. De gekwadrateerde partcorrelatie is die tussen de criteriumvariabele en de sterk correlerende andere predictor uitgezuiverd voor de zwak correlerende suppressorvariabele ($r^2_{y(1\cdot 2)}$). Deze gekwadrateerde partcorrelatie is gelijk aan de toename in de proportie verklaarde variantie als die andere predictor $X_1$ opgenomen wordt in de regressievergelijking bovenop de suppressorvariabele $X_2$. En de gekwadrateerde gewone correlatie is die tussen de criteriumvariabele en de andere predictor ($r^2_{1y}$). Dit is de proportie verklaarde variantie op basis van een regressievergelijking met alleen $X_1$ als predictor. Nu, de ruime definitie van suppressie houdt in dat $r^2_{y(1\cdot 2)}$ groter moet zijn dan $r^2_{1y}$. In dit geval is de bijdrage aan de proportie verklaarde variantie van predictor $X_1$ groter als ook de suppressorvariabele $X_2$ in de regressievergelijking zit dan wanneer dit niet het geval is.

We kunnen deze ruime definitie van suppressie illustreren aan de hand van een getallenvoorbeeld. De correlaties tussen de drie variabele $Y$, $X_1$, en $X_2$, zijn de volgende:

$$\begin{aligned} r_{1y} &= 0.6 \\ r_{2y} &= 0.2 \\ r_{12} &= 0.8 \quad . \end{aligned}$$

Omdat de correlatie tussen de criteriumvariabele en de potentiële suppressorvariabele, $r_{2y}$, niet gelijk is aan 0, is er geen sprake van suppressie volgens de strikte definitie. Volgens de ruime definitie is dat echter wel het geval. De partcorrelatie $r_{y(1\cdot 2)}$ is de volgende:

$$r_{y(1\cdot 2)} = \frac{0.6 - 0.2 \times 0.8}{\sqrt{1 - 0.80^2}} = 0.73 \quad .$$

Dit is groter dan $r_{1y}$ ($= 0.6$) en daarom geldt hetzelfde voor de corresponderende gekwadrateerde correlaties. Dus, de suppressorvariabele $X_2$ zorgt ervoor dat de bijdrage van $X_1$ aan de voorspelling groter wordt, ook al correleert $X_2$ zelf ook een beetje met $Y$.

We bekijken nu kort het tweede ingrediënt van het begrip suppressie, de inhoudelijke verklaring van het gevonden patroon van correlaties. In het voorbeeld met LENGTE, GEWICHT en BUIKOMVANG, komt deze verklaring erop neer dat GEWICHT bestaat uit twee componenten waarvan er één (BUIKOMVANG) níet gecorreleerd is met de criteriumvariabele LENGTE en de ander *in sterke mate* gecorreleerd is. Door het inbrengen van een zuivere meting van de ongecorreleerde component (een rechtstreekse meting van de BUIKOMVANG) kan deze component onderdrukt worden waardoor alleen die

component overblijft die sterk met LENGTE correleert. De naam "suppressie", wat "onderdrukking" betekent, verwijst naar dit type verklaring. Het is, historisch gezien, de eerste inhoudelijke verklaring van het bovengenoemde patroon van correlaties.

Het is niet altijd mogelijk om een patroon van correlaties dat duidt op suppressie ook inhoudelijk te verklaren in termen van een component die (1) ongecorreleerd is met het criterium en (2) onderdrukt wordt door een andere variabele (i.c., de suppressorvariabele). Hiervan worden voorbeelden gegeven in Opgaven 18, 19 en 20. Toch spreekt men in zo'n situatie ook vaak van suppressie. Dus, in de praktijk wordt ook van suppressie gesproken als het niet duidelijk is wat nu precies onderdrukt wordt. In dit geval, hangt het gebruik van de term suppressie alleen af van het bovengenoemde patroon van correlaties.

### 4.7.3 Collineariteit

#### Voorbeeld en Omschrijving

In het volgende voorbeeld wordt de idee van gecorreleerde predictoren tot in het extreme doorgetrokken. In Tabel 4.11 staan de gegevens die nodig zijn om een voorspelling te maken van de score op een bepaalde test, afgenomen bij leerlingen van een basisschool. Deze criteriumvariabele wordt aangeduid als TEST ($Y$). De predictoren zijn de leeftijd van de leerlingen, aangeduid als LEEFTIJD ($X_1$), en de klas (tegenwoordig "groep" genoemd, en lopend van 1 t/m 8) waarin de leerling zit, aangeduid als GROEP ($X_2$). Deze twee predictoren hangen natuurlijk heel sterk samen en in Tabel 4.11 is aangenomen dat de desbetreffende correlatie 0.99 bedraagt. Vanwege deze nauwe samenhang correleert TEST uiteraard vrijwel even hoog met de twee predictoren, hier 0.58 met LEEFTIJD en 0.57 met GROEP. Een direct gevolg van deze bijna-gelijkheid is dat beide predictoren het even goed doen als enkelvoudige predictor: LEEFTIJD alleen voorspelt een proportie van $0.58^2 = 0.336$ van de variantie van TEST, terwijl voor GROEP alleen deze proportie verklaarde variantie op $0.57^2 = 0.325$ uitkomt. Gezien deze vrijwel equivalente rol van $X_1$ en $X_2$ ten opzichte van het criterium gebeurt er iets onverwachts wanneer beide predictoren in de voorspelling worden opgenomen. Dan blijken (zie Tabel 4.11) de bètagewichten voor $X_1$ en $X_2$ gelijk te worden aan, respectievelijk, $b_1 = 0.79$ en $b_2 = -0.21$. Ze wijken dus duidelijk van elkaar af en hebben zelfs een tegengesteld teken. Een naïeve interpretatie van de bètagewichten als indicatoren voor de samenhang zou hier dus tot de conclusie leiden dat de testscores positief samenhangen met LEEFTIJD, maar negatief met GROEP. Dit is natuurlijk onzin. Het onzinnige van zo'n conclusie blijkt ook als we bedenken dat het verschil tussen de in Tabel 4.11 gerapporteerde $r_{1y}$ en $r_{2y}$ dermate klein is, dat het met een minieme wijziging in de steekproefgegevens (bijvoorbeeld, één persoon toevoegen of weglaten) net zo goed andersom had kunnen uitvallen: een correlatie van 0.57 tussen $Y$ en $X_1$ en een van 0.58 tussen $Y$ en $X_2$. Met een verwisseling van de $r$-waarden verwisselen dan natuurlijk ook de $b$-waarden. De conclusie zou dan precies omgekeerd uitvallen. Het hele verschil in $b$-waarden kan zelfs te wijten zijn aan afrondingsfouten. Als $r_{1y}$ en $r_{2y}$ precies gelijk zijn aan elkaar:

$$r_{y1} = r_{y2} = 0.575 \quad ,$$

dan worden plotseling wel gelijke bètagewichten gevonden, in dit geval

$$b_1 = b_2 = 0.29 \quad .$$

| | | $Y$ | $X_1$ | |
|---|---|---|---|---|
| TEST | Y | 1 | | $R^2_{y\cdot 12} = 0.337$ |
| LEEFTIJD | $X_1$ | 0.58 | 1 | $R^2_{y\cdot 1} = 0.336$ |
| GROEP | $X_2$ | 0.57 | 0.99 | $R^2_{y\cdot 2} = 0.325$ |

bètagewichten :

$$b_1 = \frac{0.58 - 0.99 \times 0.57}{1 - (0.99)^2} = \frac{0.0157}{0.0199} = 0.79$$

$$b_2 = \frac{0.57 - 0.99 \times 0.58}{1 - (0.99)^2} = \frac{-0.0042}{0.0199} = -0.21$$

partcorrelatiecoëfficiënten :

$$r_{y(1\cdot 2)} = \frac{0.58 - 0.99 \times 0.57}{\sqrt{1 - (0.99)^2}} = \frac{0.0157}{0.1411} = 0.11$$

$$r_{y(2\cdot 1)} = \frac{0.57 - 0.99 \times 0.58}{\sqrt{1 - (0.99)^2}} = \frac{-0.0042}{0.1411} = -0.03$$

Tabel 4.11: Voorbeeld van collineariteit: bijna perfecte correlatie tussen predictoren.

### Definitie

De situatie met vrijwel perfect met elkaar correlerende predictoren (hetzij positief, hetzij negatief) wordt gevat onder de noemer *collineariteit*. Dit woord verwijst naar het feit dat in een dergelijke situatie de ene predictor vrijwel een lineaire transformatie is van de andere. In het gegeven voorbeeld is dit al duidelijk zonder dat er iets uitgerekend moet worden: omdat kinderen naar Groep 1 gaan als ze vier jaar worden, wordt het verband tussen LEEFTIJD en GROEP beschreven met de formule "LEEFTIJD = 3.5 + GROEP".

Anders dan onechte correlatie en suppressie, is collineariteit te definiëren in termen van één enkele correlatie: collineariteit is een sterke correlatie tussen de predictoren. Maar net zoals de definities van een onechte correlatie en van suppressie, is ook deze definitie in de praktijk vaak moeilijk toe te passen. De reden is de onduidelijkheid over wat nu een sterke correlatie is. In plaats van een arbitraire vuistregel te hanteren (bijvoorbeeld, een correlatie groter dan 0.95 noemen we sterk), ziet men collineariteit beter als een continuüm: naarmate de predictoren sterker met elkaar correleren, is de collineariteit meer uitgesproken en wordt het in deze paragraaf beschreven patroon beter benaderd.

### Instabiele Bètagewichten

Uit de bovenstaande bespreking van het voorbeeld moet duidelijk zijn geworden dat, als er sprake is van collineariteit, het onzinnig is om aan de verkregen bètagewichten ook maar enige betekenis toe te kennen. Het teken van het bètagewicht hoeft bijvoorbeeld niet gelijk te zijn aan het teken van de corresponderende gewone correlatie tussen criterium en predictor.

De bètagewichten zijn echter niet alleen oninterpreteerbaar, maar ook nog zeer instabiel: over verschillende steekproeven (uit dezelfde populatie!) berekend, zullen dezelfde bètagewichten zeer verschillende uitkomsten geven. Binnen één steekproef uit zich deze instabiliteit in grote standaardfouten[6] (SE's) voor deze gewichten.

Instabiele bètagewichten wekken ten onrechte het vermoeden dat de voorspellingen die op basis van deze bètagewichten gemaakt worden, ook instabiel zijn en om die reden weinig variantie kunnen verklaren. In de appendix op pagina 458 wordt, aan de hand van een getallenvoorbeeld, geïllustreerd dat instabiele bètagewichten wel degelijk gepaard kunnen gaan met een hoge proportie verklaarde variantie.

In geval van collineariteit is er een simpele manier om van deze instabiele bètagewichten af te komen. Aangezien $X_1$ en $X_2$ vrijwel dezelfde informatie bevatten betreffende $Y$, zorgt toevoeging van de ene predictor aan de andere nauwelijks voor een stijging in $R^2$; er wordt dus niets extra verklaard. In het voorbeeld blijkt dat de $R^2$ met twee predictoren ($R^2_{y\cdot 12} = 0.337$) nauwelijks uitkomt boven de $R^2$ met één predictor ($R^2_{y\cdot 1} = 0.336$ en $R^2_{y\cdot 2} = 0.325$); zie Tabel 4.11. Het opnemen van beide predictoren heeft dus geen nut: van twee (vrijwel) equivalente predictoren hoeft er slechts één in de regressievergelijking te verschijnen. Dit blijkt ook uit de gekwadrateerde partcorrelaties $r^2_{y(1\cdot 2)} = 0.11^2 = 0.012$ en $r^2_{y(2\cdot 1)} = (-0.03)^2 = 0.001$ (zie Tabel 4.11). Deze gekwadrateerde partcorrelaties zijn precies gelijk aan de verschillen $R^2_{y\cdot 12} - R^2_{y\cdot 2} = 0.337 - 0.325 = 0.012$, respectievelijk, $R^2_{y\cdot 12} - R^2_{y\cdot 1} = 0.337 - 0.336 = 0.001$. Merk op dat, alhoewel de formules van de bètagewichten en de partcorrelaties erg op elkaar lijken (alleen het wortelteken in de noemer verschilt), de coëfficiënten zelf erg kunnen verschillen. Het verschil tussen bètagewichten en partcorrelaties neemt toe naarmate de predictoren sterker met elkaar correleren.

Tot slot van deze paragraaf herhalen we dat de voorbeelden hier geconstrueerd zijn om zo typisch en zuiver mogelijk een bepaald fenomeen te illustreren dat zich bij meervoudige regressie kan voordoen en waar een naïeve interpretatie van regressiegewichten tot misleidende conclusies zou leiden. In de praktijk komen we dergelijke gevallen zelden in zo'n zuivere vorm tegen, maar het is goed ons te realiseren dat, gegeven correlaties tussen de predictoren onderling, een afgezwakte mengvorm van deze verschijnselen zich in principe altijd voordoet. In elke situatie met gecorreleerde predictoren is er sprake van een zekere mate van collineariteit, zorgen predictoren die laag met het criterium correleren ervoor dat andere predictoren een grotere bijdrage leveren aan de voorspelling (suppressie) en corrigeren predictoren die hoog met het criterium correleren voor onechte gedeelten van de correlaties tussen andere predictoren en het criterium (onechte corre-

[6] Net zoals in de context van het toetsen van verschillen tussen gemiddelden (met behulp van een T-toets), kunnen we ook in de regressieanalyse uit de steekproefgegevens voor ieder regressiegewicht een bijbehorende standaardfout schatten als indicatie voor de spreiding van dit regressiegewicht over verschillende steekproeven. Hierop komen we in de volgende paragraaf nog terug.

latie). Alleen in het theoretische geval van allemaal onderling ongecorreleerde predictoren doet iedere predictor zogezegd alleen maar zijn eigen werk, onafhankelijk van wat de andere predictoren doen.

## 4.8 De Statistische Significantie van Regressiegewichten

Tot hiertoe hebben we ons alleen bezig gehouden met het beschrijven van het verband tussen een criteriumvariabele en een set van predictorvariabelen. Dit verband werd uitgedrukt als een lineaire combinatie van predictoren die het criterium zo goed mogelijk voorspelt. Hierbij drukken de regressiegewichten uit hoe groot de bijdrage is van de individuele predictoren aan deze voorspelling. Deze regressiegewichten zijn vrijwel nooit precies gelijk aan 0. Daarom kunnen we voor geen enkele predictor zeggen dat hij helemaal niets bijdraagt aan de voorspelling.

Op een bepaalde manier kunnen de regressiegewichten ons echter bedriegen: doordat ze verschillen van 0, wekken ze de indruk dat de bijbehorende predictor ook in de populatie bijdraagt aan de voorspelling, terwijl hij dat niet doet. Om hier een idee van te krijgen, keren we even terug naar de enkelvoudige regressie, waar hetzelfde principe (bedrieglijke voorspellende waarde) werkzaam is. Stel dat we een criteriumvariabele willen voorspellen op basis van toevalsgetallen. Bijvoorbeeld, voor elke persoon uit een populatie gooien we één keer met een dobbelsteen, en het aantal ogen dat boven ligt gebruiken we als predictor om de criteriumvariabele te voorspellen. Stel dat de populatie oneindig groot is (bijvoorbeeld, iedereen die nu leeft en in de toekomst nog zal leven). Dan is de correlatie in de populatie precies gelijk aan 0, en daarmee ook het regressiegewicht, eenvoudig omdat toevalsgetallen nergens mee samenhangen. Echter, deze populatie krijgen we nooit in zijn geheel te zien; we krijgen slechts te maken met een steekproef uit deze populatie. In deze steekproef zal de correlatie tussen het criterium en de toevalsgetallen vrijwel nooit precies gelijk zijn aan 0, en daarmee zal ook het regressiegewicht nooit precies gelijk zijn aan 0. Het regressiegewicht zal wel schómmelen rond 0 als we meerdere steekproeven trekken uit de populatie (elke steekproef is een nieuwe groep personen), en op elke steekproef een enkelvoudige regressieanalyse uitvoeren: in sommige steekproeven zullen hoge waarden op de criteriumvariabele eerder samengaan met hoge toevalsgetallen (resulterend in positieve regressiegewichten), en in andere steekproeven eerder met lage toevalsgetallen (resulterend in negatieve regressiegewichten).

In tegenstelling tot dit hypothetische voorbeeld, waarin we ons meerdere steekproeven kunnen voorstellen, de ene keer met een positief en de andere keer met een negatief regressiegewicht, hebben we in de praktijk bijna altijd met slechts één steekproef te maken. Het is dan de vraag hoe we op basis van deze ene steekproef kunnen bepalen of het aannemelijk is dat het regressiegewicht netjes rond 0 zal schommelen als we telkens een nieuwe steekproef zouden trekken. Dit kan door middel van statistische toetsing. Men toetst dan de nulhypothese dat het regressiegewicht *in de populatie* gelijk is aan 0. Hoe deze toetsing precies in zijn werk gaat, wordt verderop uitgelegd.

In de meervoudige regressieanalyse is de situatie in principe dezelfde. Het enige verschil is dat nu een andere nulhypothese getoetst wordt: we toetsen nu de nulhypothese dat, binnen een regressievergelijking met meerdere predictoren, een bepaalde predictor

in de populatie een regressiegewicht heeft gelijk aan 0. Het is belangrijk om op te merken dat, voor een bepaalde predictor, de nulhypothese afhangt van wélke andere predictoren mee in de regressievergelijking opgenomen worden. Dit geldt niet voor de enkelvoudige regressie, want daar is er maar één predictor. Dit komt overeen met het eerder benadrukte feit dat, in de meervoudige regressie, de waarde van een regressiegewicht sterk afhankelijk kan zijn van de andere predictoren, en met name van de correlaties met deze andere predictoren. Dit fenomeen op het niveau van de steekproef heeft dus zijn tegenhanger op het niveau van de populatie.

De nulhypothese dat een populatieregressiegewicht gelijk is aan 0, kan nog op twee andere manieren geformuleerd worden:

1. De predictor levert, in de populatie, geen extra bijdrage aan de proportie verklaarde variantie bovenop de andere predictoren.

2. In de populatie, is de partcorrelatie van het criterium met de predictor uitgezuiverd voor de andere predictoren, gelijk aan 0.

Het is makkelijk te zien dat deze twee formuleringen equivalent zijn; de toename in de proportie verklaarde variantie is immers gelijk aan het kwadraat van de partcorrelatie. Het is ook makkelijk te zien dat de partcorrelatie alleen gelijk is aan 0 als het corresponderende regressiegewicht gelijk is aan 0, en omgekeerd; de tellers van de formules voor de partcorrelatie en het regressiegewicht zijn immers gelijk.

Het toetsen van de nulhypothese dat het populatieregressiegewicht gelijk is aan nul is mogelijk omdat we uit de data in de steekproef voor ieder regressiegewicht ook een bijbehorende *standaardfout* (SE voor het Engelse *standard error*) kunnen bepalen. Dit gebeurt volledig analoog aan de manier waarop men de standaardfout van een steekproefgemiddelde bepaalt. Men mag deze SE niet verwarren met de standaardschattingsfout van de regressie. Deze laatste is de standaarddeviatie van de voorspellingsfouten $Y - \widehat{Y}$, terwijl het hier gaat over de standaarddeviatie van een regressiegewicht. De standaardfout van een regressiegewicht is (een schatting van) de standaarddeviatie van dit regressiegewicht, berekend over een oneindig aantal steekproeven van dezelfde omvang uit dezelfde populatie. De verdeling van het regressiegewicht over dit oneindig aantal steekproeven wordt ook de steekproevenverdeling van dit regressiegewicht genoemd. De standaardfout van het j-de regressiegewicht $B_j$ wordt aangeduid met $\mathrm{SE}(B_j)$. Er bestaat een formule waarmee men, gebruikmakend van de data van één enkele steekproef, een schatting kan maken van $\mathrm{SE}(B_j)$, de standaarddeviatie van de steekproevenverdeling van $B_j$. Deze formule is één van de belangrijkste resultaten uit de statistiek.

De grootheid waarmee we de nulhypothese over het populatieregressiegewicht toetsen verkrijgen we door het geobserveerde regressiegewicht $B_j$ te delen door zijn standaardfout. Deze grootheid wordt de T-statistiek genoemd:

$$\boxed{T = \frac{B_j}{\mathrm{SE}(B_j)} \quad .}$$

Deze T-statistiek is volledig analoog aan de T-statistiek van de paarsgewijze T-toets en de T-toets voor onafhankelijke steekproeven. Onder de nulhypothese heeft de bovenstaande T-statistiek een $\mathcal{T}$-verdeling met $n - m - 1$ vrijheidsgraden (met $n$ het aantal

personen in de steekproef en $m$ het aantal predictoren in de regressievergelijking)[7]. We kunnen dus een bijbehorende tweezijdige p-waarde bepalen. Deze p-waarde is de kans onder de $\mathcal{T}$-verdeling om een T-statistiek te observeren die in absolute waarde groter is dan de geobserveerde T-statistiek (in *absolute* waarde omdat we het hier over een tweezijdige[8] toets hebben). In de uitvoer van SPSS wordt deze p-waarde gegeven onder het kopje "Sig." of "Sig. T". We kunnen dan bekijken of deze p-waarde onder een gekozen significantieniveau van, bijvoorbeeld, $\alpha = 0.05$ komt. Is de p-waarde van zo'n toets inderdaad kleiner dan de gekozen $\alpha$, dan wijkt het regressiegewicht significant af van 0. We verwerpen dan de nulhypothese, en concluderen daarmee dat in de gekozen regressievergelijking deze predictor ook in de populatie een zekere extra bijdrage levert aan de voorspelling van $Y$. Is daarentegen de p-waarde van de toets groter dan het gekozen significantieniveau $\alpha$, dan wordt de nulhypothese niet verworpen, en nemen we aan dat de desbetreffende predictor in de populatie geen extra bijdrage levert aan de voorspelling van $Y$ (bovenop de andere predictoren in de regressievergelijking).

De teller van de T-statistiek bepaalt wàt men toetst en de noemer hoe gevoelig men dit doet. In de teller staat $B_j$, en dus toetst men met de T-statistiek een nulhypothese die betrekking heeft op de waarde van $B_j$ in de populatie. En in de noemer staat $\mathrm{SE}(B_j)$, de standaardfout van $B_j$, en dus toetst men gevoeliger naarmate deze standaardfout kleiner is. De formule van de standaardfout is de volgende:

$$\mathrm{SE}(B_j) = \frac{\mathrm{SD}(Y - \widehat{Y})}{\mathrm{SD}(X_j)\sqrt{n(1 - R^2_{j\cdot\mathrm{Rest}})}} \quad .$$

Uit deze formule blijkt dat $\mathrm{SE}(B_j)$ van vier factoren afhangt, waarvan er één in de teller staat en drie in de noemer. Elk van deze vier factoren wordt nu besproken.

1. De standaardfout $\mathrm{SE}(B_j)$ hangt af van de standaardschattingsfout $\mathrm{SD}(Y - \widehat{Y})$. Naarmate de standaardschattingsfout toeneemt, neemt ook de standaardfout $\mathrm{SE}(B_j)$ toe. Dit komt omdat een grote standaardschattingsfout betekent dat er veel ruis zit in de data (schommelingen van de criteriumvariabele $Y$ rond zijn voorspelling $\widehat{Y}$). Naarmate er meer ruis zit in de data, zal ook het regressiegewicht $B_j$ meer variabel zijn over de steekproeven, want $B_j$ wordt berekend op basis van deze data. Naarmate $B_j$ meer varieert over de steekproeven, zal $\mathrm{SE}(B_j)$, de standaarddeviatie van de steekproevenverdeling, groter zijn.

   De standaardschattingsfout $\mathrm{SD}(Y - \widehat{Y})$ die gebruikt wordt in de formule van de standaardfout is de vierkantswortel van de zuivere schatter van $\sigma^2$ (de popu-

[7]In feite geldt dit alleen als aan bepaalde voorwaarden is voldaan. Met name, de verwachte waarden van de scores op de criteriumvariabele moeten een lineaire functie zijn van de predictorvariabelen (lineariteit) en de residuen (hier, de afwijkingen van de criteriumscores ten opzichte van hun verwachte waarde) moeten onafhankelijk normaal verdeeld zijn met gelijke spreiding voor alle scores (onafhankelijkheid, normaliteit, en homoscedasticiteit). Lineariteit en homoscedasticiteit kunnen onderzocht worden door de voorspellingsfouten in een spreidingsdiagram uit te zetten tegen elk van de predictorvariabelen. Afhankelijkheid tussen de residuen doet zich vrijwel uitsluitend voor in een binnen-subjecten design. Voor dit type design werd daarom een apart type regressieanalyse ontwikkeld (zie, Lorch en Myers, 1990) dat hier niet behandeld wordt.

[8]Men kan de T-statistiek ook gebruiken voor een eenzijdige toets, maar hier wordt niet op ingegaan in dit boek. Een eenzijdige toets heeft betrekking op de nulhypothese dat, in de populatie, het regressiegewicht groter of kleiner is dan 0.

latievariantie van de voorspellingsfouten $Y - \widehat{Y}$). Deze zuivere schatter wordt berekend door de som van de gekwadrateerde voorspellingsfouten $\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$ te delen door $n - m - 1$ in plaats van door de steekproefomvang $n$. Op deze manier wordt de variantie van de voorspellingsfouten, $\mathrm{VAR}(Y - \widehat{Y})$, gecorrigeerd voor systematische onderschatting van $\sigma^2$.

2. Het aantal observaties $n$. Naarmate het aantal observaties in de steekproef toeneemt, wordt $\mathrm{SE}(B_j)$ kleiner. We zeggen ook dat $B_j$ nauwkeuriger wordt naarmate het op meer observaties gebaseerd is. Dit is volledig vergelijkbaar met het feit dat een gemiddelde nauwkeuriger wordt naarmate het op meer observaties gebaseerd is. Het fenomeen waar het om gaat is dat, als men een regressieanalyse op een zeer grote steekproef herhaalt op een nieuwe steekproef van dezelfde omvang, de regressiegewichten vrijwel hetzelfde zullen zijn. Als het regressiegewicht weinig varieert over de steekproevenverdeling, dan is $\mathrm{SE}(B_j)$, de standaarddeviatie van deze steekproevenverdeling, klein.

3. De standaardfout $\mathrm{SE}(B_j)$ hangt af van de samenhang tussen $X_j$ en de andere predictoren. Deze samenhang wordt uitgedrukt in termen van een regressieanalyse waarin $X_j$ voorspeld wordt op basis van de andere predictoren. De proportie verklaarde variantie in deze regressieanalyse wordt aangeduid met $R^2_{j\cdot\mathrm{Rest}}$, waarin "Rest" staat voor de andere predictoren. In de formule kan men zien dat $\mathrm{SE}(B_j)$ toeneemt naarmate $R^2_{j\cdot\mathrm{Rest}}$ korter bij 1 ligt ($X_j$ beter voorspelbaar is op basis van de andere predictoren). Dit impliceert dat, in geval van collineariteit, de standaardfouten van de bijna perfect correlerende predictoren zeer groot zullen zijn. Deze predictoren kan men immers bijna perfect voorspellen op basis van een andere predictor.

4. De standaarddeviatie[9] van $X_j$, de predictor waarvan we de bijdrage aan de voorspelling onderzoeken. Als $\mathrm{SD}(X_j)$ met een bepaalde factor toeneemt, dan neemt $\mathrm{SE}(B_j)$ met dezelfde factor af, en omgekeerd, als $\mathrm{SD}(X_j)$ met een bepaalde factor afneemt, dan neemt $\mathrm{SE}(B_j)$ met dezelfde factor toe. Deze afhankelijkheid van $\mathrm{SD}(X_j)$ zorgt ervoor dat de waarde van de T-statistiek niet verandert onder een multiplicatieve schaaltransformatie van de predictor. Bijvoorbeeld, als een predictor niet in minuten maar in seconden uitgedrukt wordt, dan wordt $\mathrm{SD}(X_j)$ 60 keer zo groot en worden $B_j$ en $\mathrm{SE}(B_j)$ 60 keer zo klein en verandert de T-statistiek niet. Men kan de T-toets dus niet gevoeliger maken door de predictor op een fijnere schaal uit te drukken, want hierdoor verandert ook de teller van de T-statistiek met dezelfde multiplicatieve factor.

In de uitvoer van SPSS en andere computerpakketten wordt niet alleen een T-statistiek gegeven voor elk van de regressiegewichten $B_k$, maar ook één voor de regressieconstante $A$. In de uitvoer van SPSS wordt de regressieconstante aangeduid als *(Constant)* en hij staat, enigzins verwarrend, in de kolom die aangeduid wordt met de letter B. Met de T-statistiek voor de regressieconstante kan men de nulhypothese toetsen dat de regressieconstante in de populatie gelijk is aan 0. Er is geen interessante

[9] Deze standaarddeviatie wordt berekend met de formule waarin gedeeld wordt door $n$ (en niet $n-1$).

inhoudelijke onderzoeksvraag die met deze nulhypothese correspondeert, en daarom is men bijna nooit geïnteresseerd in de T-statistiek voor de regressieconstante.

---

*Opgave 17*
In Tabel 4.12 vind je correlaties van vier onderzoeken waarin telkens drie variabelen geobserveerd werden: één criterium en twee predictoren. Een aantal cellen zijn leeg. Vul deze lege cellen in. In de kolommen Onechte Correlatie, Suppressie, en Collineariteit moet je ja of nee invullen, en in de andere kolommen een getal. In elk van de cellen moet je iets schrijven. Voor de kolommen met het ≈-teken is er onvoldoende informatie om het exacte getal te berekenen, en moet je een getal invullen dat ongeveer correct is.

| Onderzoek | $r_{y1}$ | $r_{y2}$ | $r_{12}$ | $r_{y(1\cdot 2)}$ | $r_{y(2\cdot 1)}$ | Onechte Correlatie | Suppressie | Collineariteit |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.00 | 0.60 | 0.60 | | | | | |
| 2 | 0.50 | ≈ | ≈ | ≈ | ≈ | | | Ja |
| 3 | 0.70 | 0.35 | | | 0.00 | | | |
| 4 | 0.62 | | 0.60 | 0.40 | | | | |

Tabel 4.12: Tabel bij opgave 17.

*Opgave 18*
Een klinisch psycholoog wilde een voorspelling maken van de therapietrouw van een groep cliënten met depressieve symptomen. Zijn predictoren waren (1) de mate van overeenstemming tussen de verwachtingen van de cliënt en datgene wat er in feite gebeurde in de therapie (MATCH), en (2) de ernst van de symptomen (ERNST). De therapietrouw werd operationeel gedefinieerd als het aantal dagen dat men aanwezig was op het afgesproken tijdstip (TROUW). De correlaties tussen deze variabelen worden getoond in Tabel 4.13.

| | TROUW | MATCH |
|---|---|---|
| MATCH | 0.710 | |
| ERNST | 0.008 | -0.534 |

Tabel 4.13: Tabel bij opgave 18.

1. Op basis van de lage correlatie tussen ERNST en TROUW (i.c., 0.008) besloot de psycholoog om deze predictor maar niet te gebruiken in de regressievergelijking. Is dit terecht of ten onrechte? Motiveer je antwoord.
2. Geef een inhoudelijke verklaring voor het gevonden patroon van correlaties. Ga er hierbij vanuit dat (1) MATCH op te vatten is als *realiteitszin*, (2) ERNST positief correleert met de niet-geobserveerde variabele *lijdensdruk* (de mate waarin men last heeft van zijn symptomen) en negatief correleert met realiteitszin, en (3) TROUW positief correleert met lijdensdruk én realiteitszin.

*Opgave 19*
In haar boek "Het misverstand opvoeding" stelt J.O. Harris dat de invloed van de ouders op de kinderen verwaarloosbaar zou zijn. Daarentegen zou de invloed van leeftijdsgenoten (peers) aanzienlijk zijn. Een onderzoekster is het hier niet mee eens. Volgens haar voeden ouders hun kind op in overeenstemming met het gedrag van het kind. Dit gedrag wordt enerzijds bepaald

door het aangeboren temperament van het kind en anderzijds door wat het leert van zijn peers. De beïnvloeding die kinderen van hun peers ondervinden wordt op deze manier tegengewerkt (onderdrukt) door de opvoeding. Om haar stelling kracht bij te zetten deed zij een onderzoek naar agressiviteit. Hiervoor observeerde zij 100 tienjarigen gedurende twee weken. De agressie-stimulerende invloed van de peers wordt aangeduid als AGRSTIMPEERS en de mate waarin ouders corrigerend optreden tegen het agressieve gedrag van hun kinderen als AGRCORR. De agressiviteit van de kinderen, aangeduid als AGRK, werd zes jaar later vastgesteld en kan beschouwd worden als het resultaat van de agressie-stimulerende invloed van de peers en de correcties vanwege de ouders. De correlaties tussen deze drie variabelen worden getoond in Tabel 4.14.

| | AGRK | AGRCORR |
|---|---|---|
| AGRCORR | 0.00 | |
| AGRSTIMPEERS | 0.43 | 0.61 |

Tabel 4.14: Tabel bij opgave 19.

1. Zou je AGRCORR opnemen als predictor in een enkelvoudige regressievergelijking met AGRK als afhankelijke variabele? Motiveer je antwoord.
2. Zou je, naast AGRSTIMPEERS, ook AGRCORR opnemen als predictor in een meervoudige regressievergelijking met AGRK als afhankelijke variabele? Motiveer je antwoord.
3. Geef een inhoudelijke verklaring van het al dan niet bijdragen van AGRCORR en AGRSTIMPEERS aan de voorspelling van AGRK. Doe dit aan de hand van een figuur waarin je de richting aangeeft van de verbanden tussen deze drie variabelen.

*Opgave 20*
Een psychogerontologe onderzoekt welke variabelen bepalend zijn voor de mate waarin ouderen in verzorgingstehuizen ertoe neigen zich in hun eigen kamer terug te trekken (gemeten als de variabele TRGTRK). Een andere variabele in haar onderzoek is de mate waarin zo'n oudere contact heeft met medebewoners (INCNTCT, voor "interne contacten"). Tot haar verbazing vindt ze een licht positieve correlatie tussen deze twee variabelen. Uiteraard is ook de leeftijd (LEEFTD) als variabele opgenomen in dit onderzoek. De onderzoekster besluit een regressieanalyse uit te voeren met TRGTRK als criterium- en LEEFTD en INCNTCT als predictoren. De correlaties tussen deze drie variabelen worden gegeven in Tabel 4.15. En een deel van de SPSS-uitvoer wordt gegeven in Tabel 4.16 en Tabel 4.17.

| | TRGTRK | LEEFTD |
|---|---|---|
| LEEFTD | .600 | |
| INCNTCT | .200 | .600 |

Tabel 4.15: Tabel bij opgave 20.

1. Hoe groot is de proportie variantie die verklaard wordt door INCNTCT bovenop LEEFTD?
2. Gegeven de twee andere correlaties in Tabel 4.15 (die tussen INCNTCT en LEEFTD en die tussen LEEFTD en TRGTRK), vind je het dan vreemd dat INCNTCT en TRGTRK positief met elkaar correleren? Motiveer je antwoord.

| | |
|---|---|
| Multiple R | .632 |
| R Square | .400 |
| St. Err. of Est. | 9.451 |

Tabel 4.16: Tabel bij opgave 20.

| Variable | B | SE B | Beta | T | Sig T |
|---|---|---|---|---|---|
| (Constant) | 36.667 | 2.053 | | 17.859 | .000 |
| LEEFTD | .900 | .105 | .750 | 8.571 | .000 |
| INCNTCT | -.350 | .068 | -.250 | -5.147 | .000 |

Tabel 4.17: Tabel bij opgave 20.

3. De predictor INCNTCT blijkt in de regressievergelijking voor TRGTRK een (significant) negatief bèta- en B-gewicht te hebben, terwijl hij volgens de gegeven correlatiematrix toch positief correleert met TRGTRK. Leg uit aan de hand van een formule hoe dit kan.
4. Ga ervan uit dat TRGTRK, behalve door LEEFTD, ook door de niet-geobserveerde variabele *sociale vaardigheid* bepaald wordt; LEEFTD heeft een positief verband met TRGTRK en sociale vaardigheid een negatief verband. En ga ervan uit dat ook INCNTCT bepaald wordt door deze twee variabelen; LEEFTD en sociale vaarigheid hebben beide een positief verband met INCNTCT. (Het positieve verband met LEEFTD is te begrijpen als men ziet dat, met het stijgen van de leeftijd, bejaarden voor hun sociale contacten meer op de nabije omgeving aangewezen zijn.) Hoe kun je dan een inhoudelijk verklaring geven voor het feit dat INCNTCT een negatief regressiegewicht heeft?

*Opgave 21*
Een onderzoeker wil een voorspelling maken van het inkomen dat een pas afgestudeerde economiestudent binnen vijf jaar gaat verdienen. Als predictoren maakt hij gebruik van (a) het gemiddeld eindexamencijfer en (b) de score op een vragenlijst waarin gepeild wordt naar hoe belangrijk men het vindt om veel geld te verdienen. Deze predictoren blijken een nulcorrelatie te hebben met het criterium, zoals blijkt uit een onderzoek waarin de afgestudeerde economiestudenten na vijf jaar weer opgezocht worden en gevraagd wordt naar hun huidig inkomen. Een collega suggereert dat deze nulcorrelaties nog niet hoeven te betekenen dat het onmogelijk is om met deze predictoren een voorspelling te maken van het inkomen binnen vijf jaar. Volgens deze collega is het immers mogelijk dat beide predictoren supressorvariabelen zijn voor elkáár. Hij stelt dat het mogelijk is dat deze predictoren een substantiële bijdrage leveren aan de voorspelling wanneer ze sámen in een regressieanalyse worden opgenomen. Wat vind je van deze stelling?

*Opgave 22*
Een RIAGG wil met behulp van meervoudige regressieanalyse een voorspelling maken van therapiesucces bij alcoholverslaafden. De psychiater en de psycholoog worden het niet eens over welke variabelen men in de regressievergelijking moet opnemen. De psychiater wil alle variabelen in de regressievergelijking opnemen die afzonderlijk hoog correleren met therapiesucces. De psycholoog wil (1) niet álle variabelen opnemen die afzonderlijk hoog correleren en (2) ook variabelen opnemen die vrijwel niet correleren met therapiesucces. Hoe kan de psycholoog gebruikmaken van wat er in dit boek gezegd wordt om zijn standpunt te verdedigen?

*Opgave 23*
De psychologe van Opgave 15 op pagina 340 overweegt wat er zou gebeuren als zij, naast de

schaal voor communicatieve vaardigheden (COMVAARD_A), een tweede schaal voor communicatieve vaardigheden (COMVAARD_B) als predictor zou opnemen. (De extraversievragenlijst en de test voor geschreven taalvaardigheid doen hier niet mee.) COMVAARD_B correleert -0.59 met ZOEKTIJD en 0.99 met COMVAARD_A. Een deel van de SPSS-uitvoer wordt getoond in Tabel 4.18. De psychologe is verbaasd dat, noch COMVAARD_A, noch COMVAARD_B, significant bijdragen aan de voorspelling van ZOEKTIJD. Hoe kun je dit verklaren vertrekkende van de nulhypotheses die hier getoetst worden? (Geef deze nulhypotheses.)

| Variable | B | SE B | Beta | T | Sig T |
|---|---|---|---|---|---|
| (Constant) | 14.641 | 9.314 | | 1.572 | .122 |
| COMVAARD_A | -.418 | .973 | -.799 | -.430 | .669 |
| COMVAARD_B | .127 | .914 | .201 | .139 | .890 |

Tabel 4.18: Tabel bij opgave 23.

*Opgave 24*
Een onderzoeker wil een voorspelling maken van het ziekteverzuim bij de werknemers van een grote verzekeringsmaatschappij. Zijn criteriumvariabele (VERZUIM) is het gemiddeld aantal dagen afwezigheid per jaar wegens kleine lichamelijke problemen (verkoudheid, lage rugpijn, ...). Hij maakt gebruik van de volgende drie predictoren: de rang binnen het bedrijf (directeur, afdelingshoofd, klerk, ...) uitgedrukt in een getal tussen 1 en 10 (RANG), het inkomen per maand (INKOMEN) en de afstand tussen woon- en werkplaats (WOONWERK). RANG en INKOMEN correleren 0.97 met elkaar en WOONWERK correleert 0.00 met zowel RANG als INKOMEN.

1. Uit een analyse met SPSS blijkt dat RANG binnen een enkelvoudige regressieanalyse een T-statistiek heeft die net significant is (p=0.04). Nu voert de onderzoeker een analyse uit met RANG *en* INKOMEN als predictoren. Wat verwacht je van de p-waarden van de T-statistieken van RANG en INKOMEN in deze analyse? Leg dit uit aan de hand van de hypothese die met deze T-statistieken getoetst wordt.
2. RANG verklaart 9 procent van de variantie binnen een enkelvoudige regressieanalyse. De onderzoeker weet dat hij de proportie verklaarde variantie binnen een meervoudige regressieanalyse kan berekenen met behulp van partcorrelaties. Welke partcorrelaties moet hij hiervoor berekenen binnen een meervoudige regressieanalyse met RANG, INKOMEN en WOONWERK? Motiveer je antwoord.

*Opgave 25*
In de Tabel 4.19 en Tabel 4.20 vind je een gedeelte van de SPSS-uitvoer.

| | |
|---|---|
| Multiple R | .89468 |
| R Square | .80046 |
| St. Err. of Est. | 5.48264 |

Tabel 4.19: Tabel bij opgave 25.

1. Welke variabelen moet men met elkaar correleren om het getal achter Multiple R te krijgen?
2. Wat betekent het als er in de kolom Sig T een zeer klein getal (kort bij nul) staat?

| Variable | B | SE B | Beta | T | Sig T |
|---|---|---|---|---|---|
| (Constant) | 28.92980 | 2.06883 | | 13.984 | .0000 |
| PREDICTOR1 | .38977 | .07462 | .14164 | 5.223 | .0000 |
| PREDICTOR2 | .55014 | .05025 | .39523 | 10.949 | .0000 |
| PREDICTOR3 | .86812 | .03735 | .62828 | 23.245 | .0000 |
| PREDICTOR4 | .41142 | .05297 | .28087 | 7.767 | .0000 |

Tabel 4.20: Tabel bij opgave 25.

3. Iemand beweert dat PREDICTOR 3 en PREDICTOR 4 vrijwel perfect met elkaar correleren (boven de 0.99). Is deze tabel hiermee in overeenstemming? Motiveer je antwoord.
4. Kan het zijn dat, ondanks de lage Sig T-waarden, een van de onafhankelijke variabelen een nulcorrelatie heeft met de afhankelijke variabele? Leg uit waarom wel of waarom niet.

*Opgave 26*
Bij een grote uitgeverij wil men onderzoeken hoe de naamsbekendheid binnen het bedrijf totstandkomt. Daartoe werd de naam van elk personeelslid voorgelegd aan alle andere personeelsleden, met de vraag of men deze persoon kende of niet. Daarna werd voor elk personeelslid berekend hoeveel andere personeelsleden hem/haar kenden. Deze aantallen werden gebruikt als criteriumvariabele binnen een meervoudige regressieanalyse. Als predictoren gebruikte men (1) de rangorde binnen de hiërarchie van het bedrijf (10 voor de directeur, 9 voor de vicedirecteur, ...), aangeduid als RANG, (2) het aantal dienstjaren, aangeduid als ANCIEN, en (3) het salaris in duizenden Euro's, aangeduid als SALARIS. Een deel van de SPSS-uitvoer wordt getoond in Tabel 4.21 en Tabel 4.22.

| | |
|---|---|
| Multiple R | .88489 |
| R Square | .78303 |
| St. Err. of Est. | 5.71706 |

Tabel 4.21: Tabel bij opgave 26.

| Variable | B | SE B | Beta | T | Sig T |
|---|---|---|---|---|---|
| (Constant) | 34.96050 | 1.76698 | | 19.785 | .0000 |
| RANG | -.72889 | .61687 | -.50496 | -1.182 | .2383 |
| ANCIEN | .54078 | .05292 | .38851 | 10.219 | .0000 |
| SALARIS | 1.17705 | .63239 | .80357 | 1.861 | .0637 |

Tabel 4.22: Tabel bij opgave 26.

1. Stel dat men het salaris niet in duizenden Euro's, maar in Euro's uitdrukt (bijvoorbeeld, 2 wordt 2000). Wat verandert dan in de kolom B? En wat gebeurt er in de kolom Beta? Motiveer je antwoord.
2. Waar in deze uitvoer vind je informatie over de spreiding van de geobserveerde waarden van de criteriumvariabele rond hun voorspelde waarden? Motiveer je antwoord.
3. Zou je op basis van deze uitvoer (1) RANG *en* SALARIS uit de regressievergelijking verwijderen, (2) slechts één van beide verwijderen, of (3) geen van beide verwijderen? Motiveer je antwoord.

*Opgave 27*
In de Verenigde Staten worden kinderen vaak naar zomerkampen gestuurd. De doelstelling van zo een kamp kan nogal variëren. Er zijn kampen voor sportieve activiteiten zoals surfen of mountainbiken, schilder- of muziekkampen, computerkampen en zelfs kampen waar je naar toe gaat om af te vallen. Een ontwikkelingspsycholoog die voor de AAREC (American Association of Recreative and Educative Camps) werkt, beweert dat het herhaalde verblijf op een kamp erg bevorderlijk is voor de ontwikkeling van sociale vaardigheden, die later in het beroepsleven de kansen op een succesvolle carrière sterk vergroten.

Om zijn hypothese te toetsen, vraagt de psycholoog aan 100 mannen en 100 vrouwen om de volgende informatie:

1. SALARIS: verdiensten per jaar.
2. KAMP: aantal keren dat iemand tijdens zijn of haar jeugd op een zomerkamp geweest is.
3. SALOUD: verdiensten per jaar van de ouders.

De correlaties tussen deze variabelen worden getoond in Tabel 4.23.

| | SALARIS | KAMP |
|---|---|---|
| KAMP | 0.30 | |
| SALOUD | 0.40 | 0.75 |

Tabel 4.23: Tabel bij opgave 27.

1. Zou je KAMP, SALOUD, beide, of geen van beide opnemen in een regressieanalyse om SALARIS te voorspellen? Motiveer je antwoord.
2. Volgens de onderzoeker bevestigen deze data zijn hypothese dat deelname aan zomerkampen tot een hoger salaris leiden. Ben je het met hem eens? Geef de redenering die tot je conclusie (eens of oneens) leidt.

Naast de bovenstaande variabelen, heeft de onderzoeker ook aan iedereen zijn of haar leeftijd gevraagd (verder aangeduid als LEEFTIJD). Het blijkt dat LEEFTIJD een correlatie van 0.35 heeft met SALARIS maar geen correlatie met KAMP of SALOUD.

3. Laat zien hoe je met behulp van partcorrelaties kunt uitrekenen hoe groot de proportie verklaarde variantie is als je SALARIS met KAMP, SALOUD en LEEFTIJD voorspelt.

Nu komt er nog een variabele bij. De onderzoeker heeft namelijk aan iedereen ook het aantal dienstjaren gevraagd (verder aangeduid als DIENSTJAREN). DIENSTJAREN correleert zeer hoog (bijna 1) met LEEFTIJD. Stel dat je SALARIS nu voorspelt met KAMP, SALOUD, LEEFTIJD en DIENSTJAREN.

4. Wat verwacht je van het resultaat van de significantietoetsen voor LEEFTIJD en DIENSTJAREN? Geef aan wat de nulhypotheses zijn die getoetst worden en of ze aanvaard of verworpen worden.

*Opgave 28*
In Tabel 4.24 vind je de gemiddelden en de standaarddeviaties van vier variabelen. En in Tabel 4.25 en Tabel 4.26 vind je een deel van de uitvoer van een regressieanalyse.

1. Zou je alledrie de predictoren in de regressievergelijking laten? Of zou je iets veranderen? Motiveer je antwoord.

| | Mean | Standard Deviation | N |
|---|---|---|---|
| CRITERIUM | 32578.26 | 6505.44 | 46 |
| PRED1 | 15140.22 | 2126.36 | 46 |
| PRED2 | 13.59 | 2.26 | 46 |
| PRED3 | 53.13 | 34.62 | 46 |

Tabel 4.24: Tabel bij opgave 28.

| | |
|---|---|
| Multiple R | .618 |
| R Square | .382 |
| St. Err. of Est. | 5294.59 |

Tabel 4.25: Tabel bij opgave 28.

| | B | SE B | Beta | T | Sig T |
|---|---|---|---|---|---|
| (Constant) | 5829.731 | 6047.032 | | .964 | .341 |
| PRED1 | 1.252 | .441 | .409 | 2.836 | .007 |
| PRED2 | 773.139 | 414.326 | .268 | 1.866 | .069 |
| PRED3 | -50.955 | 23.057 | -.271 | -2.210 | .033 |

Tabel 4.26: Tabel bij opgave 28.

2. Door welke factoren wordt de Standard Error of the Estimate in Tabel 4.25 bepaald?
3. Hoe groot is de verklaarde variantie?
4. Wat gebeurt er met de volgende grootheden als je de scores op PRED1 met 100 vermenigvuldigt: B (PRED1), Beta (PRED1), B (PRED2) en Beta (PRED2)? Leg uit.

*Opgave 29*
Een onderzoeker heeft een meervoudige regressieanalyse uitgevoerd met drie predictoren ($X_1$, $X_2$ en $X_3$). Het criterium wordt aangeduid met $Y$. Het aantal observaties was 75. De standaarddeviaties van deze vier variabelen zijn: $\mathrm{SD}(X_1) = 1.061$, $\mathrm{SD}(X_2) = 0.875$, $\mathrm{SD}(X_3) = 1.170$ en $\mathrm{SD}(Y) = 0.014$. Een gedeelte van de SPSS-uitvoer vind je in Tabel 4.27.

| Variable | B | SE B | Beta | T | Sig. T |
|---|---|---|---|---|---|
| $X_1$ | -1.71E-03 | 0.010 | ? | ? | 0.870 |
| $X_2$ | ? | ? | 0.734 | 0.936 | 0.352 |
| $X_3$ | 3.87E-03 | 1.63E-03 | 0.323 | 3.374 | 0.020 |

Tabel 4.27: Tabel bij opgave 29.

1. Bereken de waarden die in de cellen met de vraagtekens moeten staan. Als je denkt dat je niet genoeg informatie hebt om een waarde te berekenen, schrijf dan NB (Niet Berekenbaar) in de overeenkomstige cel.
2. Schrijf zo volledig mogelijk de nulhypothese op die getoetst wordt in de rij van $X_2$ in Tabel 4.27. Vermeld ook het resultaat van deze toets.
3. Hoe kan het dat, ondanks het kleine regressiegewicht $B_3$ voor $X_3$, deze predictor toch een kleine $p$-waarde heeft?

4. Als je als bijkomende informatie krijgt dat $r_{y1} = 0.609$, $r_{y2} = 0.616$, en dat $X_3$ ongecorreleerd is met de andere predictoren, wat is dan de vermoedelijke verklaring voor de hoge p-waarden die horen bij $X_1$ en $X_2$?
5. Een tweede onderzoeker repliceert dit onderzoek, maar hij gebruikt 7500 observaties. Zal hij tot dezelfde conclusie komen met betrekking tot de significantie van de predictoren?

*Opgave 30*
Een onderzoeker verzamelde scores op twee predictoren ($X_1$ en $X_2$) en een criterium ($Y$) in een steekproef van 100 proefpersonen. In Tabel 4.28 vind je links de correlaties tussen deze drie variabelen en rechts hun gemiddelden en varianties.

| | $Y$ | $X_1$ |
|---|---|---|
| $X_1$ | 0.58 | |
| $X_2$ | 0.09 | 0.06 |

| Variabele | Gemiddelde | Variantie |
|---|---|---|
| $Y$ | 70 | 25 |
| $X_1$ | 5 | 1 |
| $X_2$ | 10 | 100 |

Tabel 4.28: Tabel bij opgave 30.

Levert $X_2$ een significante bijdrage aan de voorspelling van $Y$ bovenop $X_1$? Laat zien welke formules je gebruikt voor je berekeningen. Maak gebruik van het volgende:

1. De standaarddeviatie van de steekproevenverdeling van $B_2$ is gelijk aan 0.0413.
2. De kritieke waarde voor de toetsstatistiek is 1.9847. (De kritieke waarde voor een toetsstatistiek is de waarde van deze toetsstatistiek die correspondeert met een p-waarde van 0.05. Voor een positieve toetsstatistiek, zoals de F-statistiek, geldt dat de p-waarde kleiner is dan 0.05 als de toetsstatistiek groter is dan zijn kritieke waarde.)

---

## 4.9 Het Selecteren van Predictoren

Wanneer we voor de voorspelling van een bepaalde criteriumvariabele $Y$ beschikken over scores op een aanzienlijk aantal (potentiële) predictorvariabelen, kunnen we ons afvragen welke van deze predictoren we het beste kunnen opnemen in een regressievergelijking voor $Y$. Op het eerste gezicht, lijkt deze vraag onzinnig. Het is immers duidelijk dat iedere toegevoegde predictor $R^2$, de proportie verklaarde variantie van $Y$, alleen maar kan verhogen, nooit verlagen. De voorspelling wordt dus alleen maar beter met meer predictoren en daarom zouden we dus altijd alle beschikbare predictoren moeten gebruiken.

De zaak ligt echter niet zo simpel. Stel dat we de regressievergelijking willen gebruiken om voorspellingen te doen voor een groep personen. Voordat we deze voorspellingen kunnen doen, moeten eerst de scores op de predictorvariabelen verzameld worden. Als deze predictoren bestaan uit psychologische tests (bijvoorbeeld, selectietests om succes in een bepaald beroep te voorspellen, of toelatingstests voor een bepaalde opleiding, of tests in allerlei diagnostische settings), dan vereist iedere extra opgenomen

predictor de afname van een extra test voor iedere te voorspellen score. Aan het verzamelen van predictorscores kunnen bepaalde kosten verbonden zijn (tijd, geld, inspanning), en dan is de vraag gewettigd of de door een extra predictor geleverde verbetering van de voorspelling deze kosten wel waard is. We kunnen een gelijkaardige vraag stellen als we met de regressieanalyse een meer theoretische doelstelling hebben (achterhalen welke psychologische variabelen belangrijke determinanten zijn van het psychologische construct dat via onze criteriumvariabele gemeten wordt). Ook in dit geval, willen we een grens kunnen stellen aan welke bijdragen (dus, welke predictoren) nog interessant zijn.

Het kan dus nuttig zijn, gegeven een ruim aanbod van predictoren, hieruit een geschikt deel te selecteren voor opname in de regressievergelijking. Er zijn een aantal procedures ontwikkeld om dit op een automatische wijze te doen. Basiselement van deze procedures is dat er statistische toetsen worden uitgevoerd op in de steekproef verkregen regressiegewichten. Deze toetsen werden in de vorige paragraaf beschreven. Zij toetsen de nulhypothese dat, in de populatie, een predictor geen extra bijdrage levert aan de voorspelling van het criterium bovenop de andere predictoren.

We beschrijven nu kort drie procedures die gebruikmaken van statistische toetsen op regressiegewichten om een geschikte regressievergelijking te construeren. Variaties hierop zijn mogelijk, maar de hier volgende drie zijn in computerpakketten als SPSS beschikbaar.

1. De voorwaartse methode (*forward*).

   Hier beginnen we met nul predictoren en kiezen dan als eerste kandidaat-predictor die variabele die, absoluut gezien, de hoogste correlatie heeft met $Y$ (en daarmee, van alle predictoren, de meeste variantie verklaart). Van deze predictor wordt het regressiegewicht bepaald en op de boven beschreven wijze getoetst. Wijkt dit gewicht significant af van 0, dan wordt de predictor inderdaad in de vergelijking opgenomen. Nu wordt als tweede kandidaat-predictor die variabele gekozen die, uitgezuiverd voor de eerste predictor, de hoogste partcorrelatie met $Y$ heeft. Deze predictor levert samen met de eerste predictor de grootste proportie verklaarde variantie. Dit volgt uit de volgende opsplitsing, die op pagina 334 voorgesteld werd:

   $$R^2_{y\cdot 12} = r^2_{y1} \ + \ r^2_{y(2\cdot 1)} = R^2_{y\cdot 1} \ + \ r^2_{y(2\cdot 1)} \quad ,$$

   Indien via een toets op het regressiegewicht deze predictor ook een bijdrage blijkt te leveren, gaan we verder met het zoeken naar een derde predictor, die nu, uitgezuiverd voor de al aanwezige predictoren, de hoogste partcorrelatie met $Y$ heeft en daarmee de grootste ophoging van $R^2$ veroorzaakt. Dit gaat zo door totdat de als beste kandidaat gekozen variabele bij toetsing een niet-significant regressiegewicht blijkt te hebben en dus (concluderen we) in de populatie geen extra bijdrage meer levert aan de voorspelling. Dan voegen we deze predictor dus niet meer toe en stoppen we met de vergelijking zoals die vóór deze stap was.

2. De achterwaartse methode (*backward*).

   Nu gaan we in omgekeerde richting, maar volgens precies hetzelfde principe. We beginnen met alle predictoren in de regressievergelijking en zoeken dan die pre-

dictor die, uitgezuiverd voor de overige predictoren, de laagste partcorrelatie met $Y$ heeft. De verwijdering van deze predictor zou dus de kleinste verlaging van $R^2$ veroorzaken. Als bij toetsing van het desbetreffende regressiegewicht dit niet significant blijkt te zijn, wordt deze predictor inderdaad verwijderd en zoeken we in de nieuwe vergelijking naar de predictor die nu, uitgezuiverd voor de resterende predictoren, de laagste partcorrelatie met $Y$ heeft. Dit gaat zo door, totdat de predictor met de laagste partcorrelatie (en dus kleinste daling van proportie verklaarde variantie) een significant gewicht blijkt te hebben. Deze predictor blijft dan in de regressievergelijking behouden en het proces stopt.

3. De stapsgewijze methode (*stepwise*).

   Dit is een combinatie van bovenstaande methoden. We beginnen de regressievergelijking te construeren volgens de forward-methode, maar iedere keer als er een predictor is toegevoegd, kijken we via de backward-methode of er nu misschien andere predictoren kunnen worden verwijderd. Het proces stopt dus als er aan de volgende twee voorwaarden is voldaan: in een regressievergelijking is er (a) geen significante predictor toe te voegen en (b) geen niet-significante predictor te verwijderen.

Bij deze beschrijving van de drie selectieprocedures passen enkele relativerende opmerkingen. Ten eerste, deze drie procedures, toegepast op dezelfde data, kunnen verschillende regressievergelijkingen als eindresultaat hebben. In deze zin zijn de procedures dus niet equivalent. We kunnen geen algemene beschrijving geven van de manier waarop de eindresultaten kunnen verschillen omdat dit afhankelijk lijkt te zijn van heel specifieke kenmerken van de data. Uiteraard zullen de regressievergelijkingen die via de verschillende methoden worden bereikt ook vaak wél overeenstemmen.

Ten tweede, men moet beseffen dat deze procedures zich slechts baseren op significantietoetsen. Hiermee kunnen zinnige uitspraken gedaan worden over de vraag of een predictor in de populatie énige bijdrage levert, maar ze toetsen niet of deze bijdrage ook de moeite waard is. Behalve naar het al dan niet significant zijn van een predictor zouden we dan ook moeten kijken naar het verschil in $R^2$ dat gepaard gaat met het al dan niet opnemen van die predictor. Uiteindelijk spelen natuurlijk ook de aan het afnemen van een bepaalde predictor verbonden kosten (zie het begin van deze paragraaf) een rol bij onze beslissing om een bepaalde regressievergelijking te gaan gebruiken.

Ten derde, een beperking van deze selectieprocedures is dat ze gevoelig zijn voor toevalligheden in de data. In alle drie de varianten verloopt de selectieprocedure via stapsgewijze beslissingen. Dit betekent dat beslissingen die in een vroeg stadium toevallig een bepaalde kant uitvallen (bijvoorbeeld, welke van twee kandidaat-predictoren in Stap 2 toevallig net de hoogste stijging in $R^2$ veroorzaakt) in principe doorwerken in het hele verdere verloop (afhankelijk van de beslissing in Stap 2 krijgen we twee verschillende situaties in Stap 3, enzovoort). Dit draagt ertoe bij dat herhaling van dezelfde selectieprocedure in een nieuwe steekproef niet dezelfde regressievergelijking als eindresultaat hoeft te leveren. Het gevaar van een dergelijke afhankelijkheid van toevalligheden in die ene steekproef waarop de regressievergelijking wordt berekend is groter naarmate er meer predictoren in het spel zijn (meer gewichten die optimaal te kiezen zijn voor de gegeven

data) en naarmate de steekproef kleiner is (waardoor enkele toevallige uitschieters in de data een grotere rol kunnen spelen).

## 4.10 Uitvoeren van een Regressieanalyse met de SPSS-procedure Regression

Nu volgt een korte beschrijving van de SPSS-procedure Regression, die men het beste kan lezen vóór een computer waarop SPSS draait. Men vindt deze procedure in het menu *Analyze* bij de categorie *Regression*. Er bestaan verschillende soorten regressieanalyse, maar degene waar wij het nu over hebben vind je onder de optie *Linear* .... Als men deze optie kiest, dan verschijnt het dialoogvenster *Linear Regression*. In dit dialoogvenster kies je (1) de criteriumvariabele door hem naar het vakje *Dependent* over te brengen en (2) de predictoren door ze naar het vakje *Independent(s)* over te brengen.

In het dialoogvenster kan men de predictoren in meerdere *blokken* verdelen. Dit doet men door op de knop *Next* te drukken nadat men het eerste blok predictoren heeft gekozen. Dan verandert de tekst *Block 1 of 1* in *Block 2 of 2* en kan men het tweede blok predictoren kiezen. In dit hoofdstuk hadden we het voorlopig alleen over regressieanalyses waarin alle predictoren in één blok gestopt werden. In de volgende paragraaf, bij de bespreking van de *hiërarchische meervoudige regressieanalyse* zullen we het hebben over een type analyse dat de vraag beantwoordt of een blok predictoren (in tegenstelling tot één enkele predictor) een bijdrage levert aan de voorspelling bovenop een ander blok predictoren. Voorlopig houden we het echter bij regressieanalyses waarin alle predictoren in één blok gestopt worden.

In het dialoogvenster *Linear Regression* kun je bij *Method* een keuze maken uit vier methoden om predictoren te selecteren voor de regressievergelijking: enter, forward, backward en stepwise. Deze laatste drie methoden werden in de vorige paragraaf beschreven. De eerste methode, enter, komt er eenvoudig op neer dat alle geselecteerde predictoren in de regressievergelijking opgenomen worden. Behalve deze vier methoden die bepalen welke predictoren men opneemt, is er ook de methode remove waarmee men predictoren kan weghalen die men in een vorig blok met enter had opgenomen.

In het dialoogvenster onder de knop *Statistics* staan de vakjes *Estimates* en *Model fit* aangekruisd. Voor de opgaven in dit boek is het nodig dat deze kruisjes blijven staan. Het kruisje bij *Model fit* produceert onder andere de $R^2$, de $R^2_{Adjusted}$ en een ANOVA-tabel in de SPSS-uitvoer. De ANOVA-tabel is in de regressieanalyse niet zo belangrijk als in de variantieanalyse, en daarom wordt er in dit hoofdstuk geen aandacht aan besteed.

De uitvoer van een regressieanalyse met de enter-methode is vrij eenvoudig te begrijpen op basis van datgene wat tot nu toe behandeld is. Dit is niet zo voor regressieanalyses met de forward-, backward- en stepwise-procedures. In Tabel 4.29 staat een deel van de uitvoer die geproduceerd wordt door de FORWARD procedure voor het onderzoek dat beschreven wordt in Opgave 32. Deze tabel heeft betrekking op één stap uit de forward-procedure; de stap ná het opnemen van de predictoren OPSYCHOL en OPSYCHIA in de regressievergelijking. In het deel *Included Variables* staan de variabelen die reeds geselecteerd zijn door de FORWARD-procedure. In dit deel van de tabel staat

dezelfde informatie die men ook krijgt met de enter-methode. In het deel *Excluded Variables* staan de variabelen die nog niet geselecteerd zijn. Uit deze laatste variabelen zal de forward-procedure een keuze maken, en zij doet dat op basis van (1) het al dan niet significant zijn van de bijdrage van deze variabelen, en (2) de partcorrelatie van deze variabelen, uitgezuiverd voor de "Included Variables".

——— Included Variables ———

| Variable | B | SE B | Beta | T | Sig T |
|---|---|---|---|---|---|
| (Constant) | 31.50969 | 1.74953 | | 18.010 | .0000 |
| OPSYCHOL | .85372 | .04072 | .60183 | 20.967 | .0000 |
| OPSYCHIA | .83163 | .03995 | .59746 | 20.815 | .0000 |

——— Excluded Variables ———

| Variable | Beta In | Partial | Tolerance | T | Sig T |
|---|---|---|---|---|---|
| SOCSTEU | -.00748 | -.01512 | .99439 | -.260 | .7949 |
| VERTROUW | .25502 | .37232 | .52020 | 6.745 | .0000 |

Tabel 4.29: Tabel bij opgave 32.

In het deel "Excluded Variables" van Tabel 4.29 zijn de volgende kolommen van belang: *Beta In*, *T* en *Sig T*.

1. Onder *Beta In* staat het bètagewicht van de variabele in de regressievergelijking waarin ook de "Included Variables" staan.

2. Onder *T* staat de T-statistiek voor de bijdrage van deze variabele in deze regressievergelijking.

3. Onder *Sig T* staat de p-waarde van deze T-statistiek (onder de nulhypothese dat deze variabele geen bijdrage heeft in de populatie).

In het deel "Excluded Variables" staan ook twee kolommen die verder niet van belang zijn: *Partial* en *Tolerance*.

1. Onder *Partial* staat de *partiële* correlatie. Deze is nauw verwant aan de part correlatie, maar heeft geen echte functie binnen de forward-procedure. Daarom is het vreemd dat hij in de SPSS-uitvoer wordt afgedrukt.

2. Onder *Tolerance* staat een maat voor de graad van collineariteit tussen elk van de predictoren en de rest.

---

*Opgave 31*
Een onderzoeker verzamelt scores op vijf predictoren ($X_1$ t/m $X_5$) en een criterium ($Y$). In Tabel 4.30 vind je de correlaties tussen deze zes variabelen.

1. Indien predictoren geselecteerd zouden worden met de forward-selectieprocedure, welke predictor zou dan als eerste in aanmerking komen om geselecteerd te worden? Beargumenteer.

| | $Y$ | $X_1$ | $X_2$ | $X_3$ | $X_4$ |
|---|---|---|---|---|---|
| $X_1$ | 0.58 | | | | |
| $X_2$ | 0.00 | -0.58 | | | |
| $X_3$ | 0.71 | 0.95 | -0.49 | | |
| $X_4$ | 0.09 | 0.06 | 0.00 | 0.06 | |
| $X_5$ | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |

Tabel 4.30: Tabel bij opgave 31.

2. Indien predictoren geselecteerd zouden worden met de forward-selectieprocedure, welke predictor zou dan als tweede in aanmerking komen om geselecteerd te worden? Beargumenteer.
3. Indien predictoren geselecteerd zouden worden met de backward-selectieprocedure, op basis waarvan wordt dan bepaald welke predictor weggelaten wordt? Beargumenteer.

*Opgave 32*
Er wordt een meervoudige regressieanalyse uitgevoerd met de forward-procedure met als criterium het aantal dagen dat iemand opgenomen is in een psychiatrische inrichting met dagbehandeling, en als predictoren de volgende vier variabelen:

OPSYCHOL Het oordeel van de psycholoog over de verwachte opnameduur (op basis van een interview).

OPSYCHIA Het oordeel van de psychiater over de verwachte opnameduur (op basis van een interview).

SOCSTEU De sociale steun van de omgeving.

VERTROUW Het vertrouwen in de effectiviteit van de behandeling (uitgesproken na afloop van de interviews met de psycholoog en de psychiater).

Een deel van de SPSS-uitvoer staat in Tabel 4.29 en Tabel 4.31.

| | |
|---|---|
| Multiple R | .86945 |
| R Square | .75594 |
| Adjusted R Square | .75430 |
| St. Err. of Est. | 6.04300 |

Tabel 4.31: Tabel bij opgave 32.

1. Achter Multiple R staat de meervoudige correlatie. Welke twee variabelen moet men met elkaar correleren om deze meervoudige correlatie te krijgen?
2. Wordt er nog een derde predictor toegevoegd, en zo ja, welke? Leg uit waarom.
3. Iemand beweert dat het oordeel van de psychiater en het oordeel van de psycholoog (OPSYCHOL en OPSYCHIA) vrijwel perfect met elkaar correleren (rond 0.95). Is Tabel 4.29 hiermee in overeenstemming? Motiveer je antwoord.
4. Kan het zijn dat de variabele VERTROUW, ondanks haar significante bijdrage, toch een nulcorrelatie heeft met het criterium? Leg uit waarom wel of waarom niet.

*Opgave 33*
Een bedrijf wil personeel selecteren voor een helpdeskfunctie. Binnen een zusterbedrijf is net een groot aantal mensen aangenomen voor hetzelfde type functie. Van deze mensen wordt een groot aantal gegevens verzameld en vervolgens wordt een meervoudige regressieanalyse uitgevoerd. De criteriumvariabele PRODUCTIE is een index gebaseerd op productiegegevens (afhandelingssnelheid klanten), een beoordeling door de directe chef en een steekproefsgewijze beoordeling door klanten. De gebruikte predictoren zijn leeftijd (LEEFTIJD), niveau van vooropleiding (VOOROPL), score op een arbeidsmotivatietest (ARBMOT), en een ervaringsindicator (ERVARING). In Tabel 4.32 en Tabel 4.33 staat een deel van de SPSS-uitvoer voor de gegevens van het zusterbedrijf.

| | |
|---|---|
| Multiple R | .63877 |
| R Square | .40803 |
| Adjusted R Square | .35870 |
| St. Err. of Est. | 2.89712 |

Tabel 4.32: Tabel bij opgave 33.

Included Variables

| Variable | B | SE B | Beta | T | Sig T |
|---|---|---|---|---|---|
| (Constant) | 2.72431 | .86425 | | 3.152 | .0016 |
| ARBMOT | .54981 | .05112 | .19763 | 10.755 | .0000 |
| ERVARING | 3.39512 | .96239 | 1.14391 | 3.527 | .0006 |
| VOOROPL | .298765 | .08716 | .76549 | 3.427 | .0008 |

Excluded Variables

| Variable | Beta In | Partial | Tolerance | T | Sig T |
|---|---|---|---|---|---|
| LEEFTIJD | -.08275 | -.115342 | .79413 | -.987 | .8348 |

Tabel 4.33: Tabel bij opgave 33.

1. Het bedrijf waar men de personeelsselectie wil doen (dus niet het zusterbedrijf dat de gegevens leverde voor de regressieanalyse) verzamelde scores op de vier predictoren. In Tabel 4.34 geven we de scores op de predictoren van drie kandidaten. Als dit de enige drie kandidaten zijn, welke kandidaat zou jij dan aannemen en waarom?

| | VOOROPL | ARBMOT | ERVARING | LEEFTIJD |
|---|---|---|---|---|
| Kandidaat 1 | 2 | 20 | 4 | 23 |
| Kandidaat 2 | 3 | 15 | 3 | 20 |
| Kandidaat 3 | 4 | 25 | 1 | 21 |

Tabel 4.34: Tabel bij opgave 33.

2. Wordt in de forward-procedure de variabele LEEFTIJD nog toegevoegd als predictor? Motiveer je antwoord.

*Opgave 34*
In een onderzoek naar de openheid van mensen in een psychotherapeutische situatie wordt, via een meervoudige regressieanalyse met de forward-procedure, onderzocht welke items uit de

zogenaamde TALK-schaal goede predictoren zijn voor de variabele AFFIL, die de bereidheid meet om met andere mensen te praten over de eigen relatie. Een deel van de uitvoer wordt gegeven in Tabel 4.35 en Tabel 4.36.

| | |
|---|---|
| Multiple R | .68399 |
| R Square | .46784 |
| Adjusted R Square | .46034 |
| St. Err. of Est. | .63895 |

Tabel 4.35: Tabel bij opgave 34.

Included Variables

| Variable | B | SE B | Beta | T | Sig T |
|---|---|---|---|---|---|
| (Constant) | .261121 | .165213 | | 1.581 | .1162 |
| TALK10 | .456974 | .059959 | .504992 | 7.621 | .0000 |
| TALK40 | .263154 | .056807 | .306940 | 4.632 | .0000 |

Excluded Variables

| Variable | Beta In | Partial | Tolerance | T | Sig T |
|---|---|---|---|---|---|
| TALK11 | .071052 | .095584 | .825179 | 1.140 | .2561 |
| TALK12 | .043362 | .054561 | .801657 | .649 | .5175 |
| TALK13 | .006800 | .008052 | .746107 | .096 | .9240 |
| TALK41 | .075136 | .101218 | .828577 | 1.208 | .2290 |

Tabel 4.36: Tabel bij opgave 34.

1. Hoeveel procent van de variantie van AFFIL laten de twee predictoren TALK10 en TALK40 nog onverklaard?
2. Gegeven de twee nu opgenomen predictoren, hoeveel bedraagt de correlatie tussen de geobserveerde en de voorspelde AFFIL-scores?
3. Wordt er nog een derde predictor toegevoegd, en zo ja, welke? Verklaar.

*Opgave 35* COMPUTEROPDRACHT

Deze opgave gaat over een gefingeerd onderzoek waarin een klinisch psychologe zich de vraag stelde of het mogelijk is om op basis van een screening op jonge leeftijd te bepalen of kinderen een risico lopen om later depressief te worden. Indien dit mogelijk is, dan kan men bij risico-kinderen extra aandachtig zijn voor de eerste symptomen van depressie en eventueel vroegtijdig ingrijpen. Om te onderzoeken of zo'n voorspelling effectief mogelijk is, zette de psychologe een longitudinaal onderzoek op, waarin zij van een groep van 300 kinderen uit het basisonderwijs een reeks gegevens verzamelde.

Bij de keuze van de predictoren liet zij zich leiden door de opvatting dat het ontstaan van depressie zowel door een emotionele als door een cognitieve component bepaald wordt: de emotionele component is de neiging om zich triest (neerslachtig, futloos) te voelen en de cognitieve component is de neiging om de gebeurtenissen in de omgeving te zien als iets dat niet onder de eigen controle staat (een externe locus of control; zie verder). Hieronder volgt een beschrijving van elke predictor, voorafgegaan door de variabelennaam waaronder deze variabele in het databestand "depr.sav" is opgeslagen.

DEPRVL1 De score op een vragenlijst voor depressieve gevoelens. De items in deze vragenlijst hebben betrekking op gevoelens van neerslachtigheid, triestheid, en dergelijke.

DEPRVL2 De score op een tweede vragenlijst voor depressieve gevoelens.

NDTRIEST Het aantal dagen in de afgelopen twee maanden dat elk van de kinderen beweerde zich triest te voelen. (De kinderen moesten een dagboekje bijhouden, waarin zij op het eind van elke dag o.a. noteerden hoe zij die dag gevoeld hadden.)

IELOCON De score op een vragenlijst voor interne versus externe locus of control; een hoge score betekent een sterke externe locus of control (de gebeurtenissen worden vooral gezien als oncontroleerbaar), en een lage score betekent een sterke interne locus of control (de gebeurtenissen worden vooral gezien als controleerbaar).

PUZZEL De score op een proef waarin de kinderen een aantal puzzels moesten oplossen. Enkele puzzels waren onoplosbaar. De score op deze proef was de totale tijd die de kinderen besteedden aan deze onoplosbare puzzels. De onderzoekster beschouwde de score op deze proef als een maat van doorzettingsvermogen. Zij ging ervan uit dat doorzettingsvermogen uit twee componenten bestaat: (1) een cognitieve component, samenvallend met interne locus of control, en (2) een sociale component, samenvallend met competitiedrang.

PUZZCOMP De score op een proef waarin de kinderen eerst alleen en daarna met zijn tweeën een aantal puzzels moesten oplossen (andere puzzels dan bij de proef die de PUZZEL-score oplevert). In de fase van de proef waarin ze met zijn tweeën aan de puzzels werkten, had ieder kind zijn eigen puzzel maar kon het wel zien waar de ander mee bezig was. Opnieuw waren er enkele onoplosbare puzzels, en weer waren dat degenen waar het om draaide. Concreet, de score op deze proef was de gemiddelde tijd die de kinderen besteedden aan de onoplosbare puzzels die ze in aanwezigheid van de ander oplosten mín de gemiddelde tijd die ze eraan besteedden toen ze alleen werkten. De onderzoekster beschouwde de score op deze proef als een maat voor competitiedrang.

INTELL De score op een intelligentietest.

AANGEPAS Het oordeel van de leerkracht over het sociaal aangepast zijn van het kind.

GLOBDEPR De criteriumvariabele is een globale maat voor depressiviteit. Deze maat is een gewogen som van de score op een vragenlijst waarin gepeild wordt naar het voorkomen van depressieve symptomen (allemaal gedragsaspecten) en de score op een vragenlijst waarin gepeild wordt naar de beleving. Dit criterium werd bepaald toen de kinderen ongeveer de leeftijd van 25 jaar bereikt hadden.

Voer de onderstaande analyses uit met de SPSS-procedure *Regression*. In het venster onder de knop *Statistics* moet je alleen het vakje *Estimates* en *Model fit* aankruisen.

1. Enkelvoudige en meervoudige regressieanalyse.
   (a) Voer eerst een enkelvoudige regressieanalyse uit met GLOBDEPR als criterium en DEPRVL1 als predictor.
   (b) Voer dan een meervoudige regressieanalyse uit met DEPRVL1 en NDTRIEST als predictoren.
   (c) Bekijk hoe de proportie verklaarde variantie ($R^2$) verandert van de eerste naar de tweede analyse. Bekijk ook hoe het regressiegewicht ($B$) van DEPRVL1 verandert van de eerste naar de tweede analyse. Hoe kun je dit verklaren?
2. Collineariteit.
   (a) Voer een meervoudige regressieanalyse uit met GLOBDEPR als criterium en de volgende variabelen als predictoren: DEPRVL1, NDTRIEST, en PUZZEL.

(b) Voer dan dezelfde analyse nogmaals uit met DEPRVL1 vervangen door DEPRVL2.

(c) En voer dan tenslotte de analyse nogmaals uit met zowel DEPRVL1 als DEPRVL2 (naast NDTRIEST en PUZZEL) als predictoren. Vraag bij deze analyse ook de correlaties tussen de variabelen. Deze krijg je als je in het venster onder de knop *Statistics* het vakje bij *Descriptives* aankruist.

(d) Wat concludeer je als je de regressiegewichten van DEPRVL1 en DEPRVL2, en de proporties verklaarde variantie van de drie analyses met elkaar vergelijkt? Breng dit in verband met de correlatie tussen DEPRVL1 en DEPRVL2.

3. Suppressie.

   (a) Met dezelfde criteriumvariabele voeren we nu een meervoudige regressieanalyse uit met de volgende predictoren: DEPRVL1, NDTRIEST, PUZZEL en PUZZCOMP. Vraag weer dezelfde *Descriptives* als in de vorige opdracht.

   (b) Heeft PUZZCOMP een voorspellende waarde voor GLOBDEPR? Beantwoord deze vraag op basis van twee soorten informatie: de correlatie tussen PUZZCOMP en GLOBDEPR en het regressiegewicht van PUZZCOMP. Hoe kun je je bevindingen verklaren?

4. Zelf predictoren selecteren.

   (a) Voer nu een meervoudige regressieanalyse uit met alle beschikbare variabelen als predictoren voor GLOBDEPR.

   (b) Laat dan in opeenvolgende analyses één voor één alle predictoren weg die geen significante bijdrage (op 0.05-niveau) hebben in de voorspelling van GLOBDEPR.

   (c) Bekijk hoe de regressiegewichten en de proportie verklaarde variantie veranderen in de opeenvolgende analyses. Hoe kun je dit verklaren?

   (d) Vind je het, tegen de achtergrond van de opvattingen van de onderzoekster over het ontstaan van depressie, begrijpelijk dat juist deze variabelen geselecteerd worden als predictoren van GLOBDEPR?

   (e) Bewaar alleen de uitvoer van de laatste regressieanalyse (die waarin alle predictoren een significante bijdrage hebben).

5. Predictoren laten selecteren.

   (a) Voer opnieuw een meervoudige regressieanalyse uit met alle mogelijke predictorvariabelen, maar laat de selectie nu over aan SPSS. Maak gebruik van de forward-procedure.

   (b) Bekijk de uitvoer en bepaal of deze selectie overeenkomt met jouw selectie in de vorige opgave. De uitvoer van deze analyse hoef je niet te bewaren.

Verwijder alle delen van de uitvoer die je niet wilt bewaren en druk de rest af.

---

## 4.11 Hiërarchische Meervoudige Regressieanalyse

De hiërarchische meervoudige regressieanalyse (HMRA) is een uitbreiding op de gewone meervoudige regressieanalyse. De HMRA wordt gebruikt om te onderzoeken of een

groép predictoren een bijdrage levert aan de voorspelling bovenop een groep andere predictoren. De HMRA wordt meestal gebruikt voor de beantwoording van één van de volgende twee onderzoeksvragen: (1) de vraag of een bepaalde groep variabelen *storende* variabelen zijn, verantwoordelijk voor het verband tussen het criterium en een groep andere variabelen, en (2) de vraag of een bepaalde groep variabelen *mediërende* variabelen zijn, een noodzakelijke schakel in het verband tussen het criterium en een groep andere variabelen. In wat volgt, wordt de HMRA uitgelegd aan de hand van een onderzoeksvraag die betrekking heeft op storende variabelen. Daarna wordt kort aangegeven hoe dezelfde methode gebruikt kan worden voor een onderzoeksvraag die betrekking heeft op mediërende variabelen.

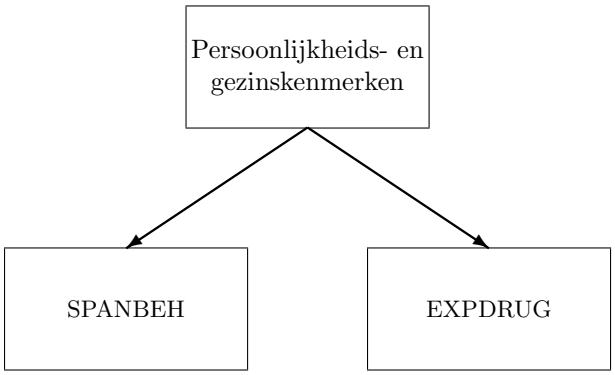
### 4.11.1 HMRA binnen het Onderzoek naar de Rol van Storende Variabelen

In het algemeen, heeft de HMRA betrekking op de vraag of een groep predictoren een bijdrage levert aan de voorspelling bovenop een groep andere predictoren. Voordat we beschrijven hoe de HMRA in het algemeen uitgevoerd wordt, behandelen we eerst een bijzonder geval van de HMRA: we onderzoeken of één predictor een bijdrage levert aan de voorspelling bovenop een groep andere predictoren. Dit kunnen we immers doen met ingrediënten die reeds in het voorgaande beschreven werden: de partcorrelatiecoëfficiënt en de T-toets voor één regressiegewicht.

#### De Bijdrage van één Predictor Bovenop een Groep andere Predictoren

Stel, we zijn geïnteresseerd in de hypothese dat spanningsbehoefte (in het Engels, *sensation seeking tendency*) een bepalende factor is bij het experimenteren met drugs (XTC en andere partydrugs, cocaïne, heroïne, ...). De populatie waarin we geïnteresseerd zijn, zijn niet-verslaafde jongeren van 14 tot 18 jaar. Met één vragenlijst meten we de spanningsbehoefte (SPANBEH) en met een andere vragenlijst stellen we vast hoeveel iemand tot op heden geëxperimenteerd heeft met drugs (EXPDRUG). We verwachten een positieve correlatie tussen SPANBEH en EXPDRUG. Echter, indien we een positieve correlatie vaststellen, dan betekent dit nog niet dat SPANBEH een invloed heeft op EXPDRUG; met andere woorden, dat er een causaal verband is tussen deze twee variabelen. Het is immers goed mogelijk dat een positieve correlatie het resultaat is van de invloed van storende variabelen, zonder dat er sprake is van een causaal verband. Potentiële storende variabelen zijn bepaalde persoonlijkheidskenmerken (bijvoorbeeld, extraversie, gewetensvolheid) en bepaalde kenmerken van het gezinsmilieu (bijvoorbeeld, de mate waarin de ouders duidelijke regels hanteren). Deze variabelen zijn storend als zij zowel met SPANBEH als met EXPDRUG samenhangen. Hierdoor kunnen zij verantwoordelijk zijn voor een eventuele positieve correlatie tussen SPANBEH en EXPDRUG, zonder dat er een causale invloed is van SPANBEH. Deze mogelijke rol van de storende variabelen wordt grafisch voorgesteld in Figuur 4.3. In deze figuur loopt er geen pijl van SPANBEH naar EXPDRUG; het verband tussen deze twee variabelen loopt volledig over de persoonlijkheids- en gezinskenmerken.

De meervoudige regressieanalyse is zeer geschikt om te onderzoeken of een bepaalde predictor een bijdrage levert aan de voorspelling bovenop een groep andere predictoren.




Figuur 4.3: Grafische voorstelling van de invloed van een groep storende variabelen (persoonlijkheids- en gezinskenmerken) op het verband tussen de criteriumvariabele (EXPDRUG) en de variabele waarvan men de invloed wil onderzoeken (SPANBEH).

Dit kan zowel beschrijvend, met de partcorrelatie, als *inferentieel*[10], met de T-toets voor een regressiegewicht. In ons voorbeeld willen we weten of SPANBEH nog iets bijdraagt aan de voorspelling van EXPDRUG bovenop de persoonlijkheids- en gezinskenmerken. De maat voor deze bijdrage is de toename in de proportie verklaarde variantie door het toevoegen van SPANBEH als extra predictor naast de persoonlijkheids- en gezinskenmerken. Deze toename in de proportie verklaarde variantie is het kwadraat van de partcorrelatie van EXPDRUG met SPANBEH uitgezuiverd voor de persoonlijkheids- en gezinskenmerken. Met de T-toets voor het regressiegewicht van SPANBEH kunnen we nagaan of de bijdrage van deze predictor bovenop de persoonlijkheids- en gezinskenmerken significant is. De p-waarde van deze T-toets geeft aan hoe aannemelijk het is dat het populatieregressiegewicht van SPANBEH gelijk is aan 0 (en daarmee ook de populatiepartcorrelatie).

Het fenomeen dat hier onderzocht wordt, werd in het voorgaande aangeduid als onechte correlatie. We gebruiken deze term om aan te geven dat een predictor die substantieel correleert met het criterium toch geen bijdrage levert aan de voorspelling bovenop een groep andere predictoren. In de terminologie van deze paragraaf, duiden we dit fenomeen aan door deze groep andere predictoren te benoemen als storende variabelen voor het verband tussen de eerstgenoemde predictor en het criterium.

### De Bijdrage van een Groep Predictoren Bovenop een Groep andere Predictoren

In het bovenstaande werd de meest eenvoudige variant van HMRA uiteengezet: HMRA waarin men de bijdrage onderzoekt van één predictor bovenop een groep andere predictoren. Om deze eenvoudige variant van HMRA uit te voeren is datgene wat we reeds uit de gewone regressieanalyse kennen genoeg (i.c., de partcorrelatie en de T-toets voor

[10] "Inferentieel" betekent dat we op basis van de observaties iets zullen zeggen over het niet-observeerbare populatieregressiegewicht. We maken deze inferentie op basis van de T-toets.

een regressiegewicht). De situatie wordt anders als men de bijdrage wil onderzoeken van meer dan één predictor bovenop een groep andere predictoren. Bijvoorbeeld, stel dat de onderzoeker één enkele vragenlijstscore als operationalisering van spanningsbehoefte toch wat mager vindt. Als tweede operationalisering voegt hij een situatieproef toe. Deze bestaat erin dat de proefpersoon een taak moet uitvoeren die enige handigheid vereist, en waarbij hij een vervelend geluid te horen krijgt telkens wanneer hij een fout maakt. Na afloop van deze taak, moet de proefpersoon op een schaal van 0 tot 10 aangeven hoe aangenaam hij deze taak vond. Volgens de onderzoeker zal een proefpersoon het vervelende geluid minder onaangenaam vinden naarmate zijn spanningsbehoefte groter is. Het oordeel van de proefpersoon op deze taak wordt aangeduid als SPANAANG.

We zijn nu geïnteresseerd in de gezamenlijke bijdrage van SPANBEH en SPANAANG aan de voorspelling van EXPDRUG bovenop de persoonlijkheids- en gezinskenmerken. Omdat we met twee in plaats van een predictor te maken hebben (naast de persoonlijkheids- en gezinskenmerken), kunnen we de partcorrelatie niet meer gebruiken om de toename in de proportie verklaarde variantie te berekenen. Het kan natuurlijk wel met twee partcorrelaties, maar er bestaat een veel makkelijker manier, en die wordt nu beschreven.

Om uit te leggen hoe we de toename in de proportie verklaarde variantie kunnen berekenen, blijkt het handig te zijn als we de predictoren in twee groepen indelen, die we aanduiden als *Blok 1* en *Blok 2*. In ons voorbeeld, bevat Blok 1 de persoonlijkheids- en gezinskenmerken en Blok 2 de variabelen SPANBEH en SPANAANG. In het algemeen, bevat Blok 1 de (potentiële) storende variabelen en Blok 2 de variabelen in wiens effect men geïnteresseerd is. De bijdrage van Blok 2 aan de proportie verklaarde variantie, bovenop de bijdrage van Blok 1, kan men bepalen op basis van twee regressieanalyses: één analyse met alleen de predictoren uit Blok 1 en één analyse met de predictoren uit Blok 1 én Blok 2. De proportie verklaarde variantie uit de eerste analyse duiden we aan met $R^2_{\text{Blok 1}}$ en die uit de tweede met $R^2_{\text{Blok 1+2}}$. Nu, de toename in de proportie verklaarde variantie die toe te schrijven is aan SPANBEH en SPANAANG, wordt aangeduid met $R^2_{\text{Change}}$, en kan als volgt berekend worden:

$$\boxed{R^2_{\text{Change}} = R^2_{\text{Blok 1+2}} - R^2_{\text{Blok 1}} \quad .}$$

Merk op dat $R^2_{\text{Change}}$ niet negatief kan zijn, want de voorspellingen kunnen niet slechter worden door het toevoegen van de predictoren uit Blok 2.

Ook al is $R^2_{\text{Change}}$ groter dan 0, het blijft de vraag of $R^2_{\text{Change}}$ ook significant groter is dan 0. Met andere woorden, het is de vraag of men op basis van de data de nulhypothese kan verwerpen dat, in de populatie, $R^2_{\text{Change}}$ gelijk is aan 0. Deze nulhypothese is equivalent met de nulhypothese dat, in de populatie, de regressiegewichten van alle predictoren in Blok 2 gelijk zijn aan 0. Deze nulhypothese kan men toetsen met een F-toets, vergelijkbaar met de bekende F-toets van de variantieanalyse. Deze F-toets is gebaseerd op de volgende toetsingsgrootheid:

$$\boxed{F_{\text{Change}} = \frac{R^2_{\text{Change}}/k_{\text{Blok 2}}}{(1 - R^2_{\text{Blok 1+2}})/(n - k_{\text{Blok 1+2}} - 1)} \quad .}$$

In deze formule staat $k_{\text{Blok 2}}$ voor het aantal predictoren in Blok 2 en $k_{\text{Blok 1+2}}$ voor het aantal predictoren in Blok 1 en Blok 2 samen. In ons voorbeeld is $k_{\text{Blok 2}}$ gelijk

aan 2 (SPANBEH en SPANAANG) en hangt $k_{\text{Blok 1+2}}$ af van het aantal persoonlijkheids- en gezinskenmerken dat men gemeten heeft. Onder de nulhypothese ($R^2_{\text{Change}}$ in de populatie is gelijk aan 0) heeft $F_{\text{Change}}$ een $\mathcal{F}$-verdeling met $k_{\text{Blok 2}}$ vrijheidsgraden voor de teller en $(n - k_{\text{Blok 1+2}} - 1)$ vrijheidsgraden voor de noemer. Bij de geobserveerde $F_{\text{Change}}$-waarde wordt op de gebruikelijke manier de p-waarde berekend of in een tabel opgezocht.

De teller van de $F_{\text{Change}}$-statistiek bepaalt wát men toetst en de noemer hoe gevoelig men dit doet. In de teller staat $R^2_{\text{Change}}$, en dus toetst men met de $F_{\text{Change}}$-statistiek een nulhypothese die betrekking heeft op de populatie-$R^2_{\text{Change}}$. Concreet, naarmate $R^2_{\text{Change}}$ groter is, wordt het meer aannemelijk dat ook de populatie-$R^2_{\text{Change}}$ groter is dan 0. De noemer van de $F_{\text{Change}}$-statistiek, die de gevoeligheid van de toets bepaalt, hangt af van $R^2_{\text{Blok 1+2}}$ en de steekproefomvang $n$. (Het aantal predictoren in de twee blokken samen, $k_{\text{Blok 1+2}}$, wordt hier buiten beschouwing gelaten omdat het meestal veel kleiner is dan $n$.) We bespreken nu de twee laatstgenoemde factoren.

1. $F_{\text{Change}}$ stijgt met $R^2_{\text{Blok 1+2}}$. Dit komt omdat een hoge $R^2_{\text{Blok 1+2}}$ betekent dat er weinig ruis zit in de data. Naarmate er minder ruis zit in de data, zal ook $R^2_{\text{Change}}$ minder variabel zijn over de steekproeven ($R^2_{\text{Change}}$ wordt immers berekend op basis van deze data). Daarom is men met een grote $R^2_{\text{Blok 1+2}}$ zekerder van het feit dat een positieve $R^2_{\text{Change}}$ ook in de populatie positief is.

2. $F_{\text{Change}}$ stijgt met $n$, het aantal observaties in de steekproef. Dit komt omdat $R^2_{\text{Change}}$ nauwkeuriger wordt naarmate het op meer observaties gebaseerd is. Daarom is men met een grote $n$ zekerder van het feit dat een positieve $R^2_{\text{Change}}$ ook in de populatie positief is.

Men kan de $F_{\text{Change}}$-statistiek ook berekenen indien men de bijdrage onderzoekt van slechts één predictor bovenop een groep andere predictoren. In dit geval, bevat Blok 2 slechts één predictor. Op basis van deze $F_{\text{Change}}$-statistiek trekt men dezelfde conclusie als op basis van de T-statistiek. De p-waarde van deze $F_{\text{Change}}$-statistiek is exact gelijk aan de (tweezijdige) p-waarde van de T-statistiek. Dit komt omdat de $F_{\text{Change}}$-statistiek gelijk is aan het kwadraat van de T-statistiek. Uit dit laatste volgt dat de p-waarde voor een $F_{\text{Change}}$-statistiek (i.e., de kans onder de nulhypothese dat de $F_{\text{Change}}$-statistiek groter is dan de feitelijk geobserveerde waarde $x$) gelijk is aan de volgende som:

$$\boxed{\begin{array}{c}\text{kans op een T-statistiek}\\ \text{kleiner dan } -\sqrt{x}\end{array}} \quad + \quad \boxed{\begin{array}{c}\text{kans op een T-statistiek}\\ \text{groter dan } +\sqrt{x}\end{array}} \quad .$$

De vierkantswortels $-\sqrt{x}$ en $+\sqrt{x}$ zijn de twee mogelijke T-statistieken die corresponderen met een feitelijk geobserveerde $F_{\text{Change}}$-statistiek gelijk aan $x$.

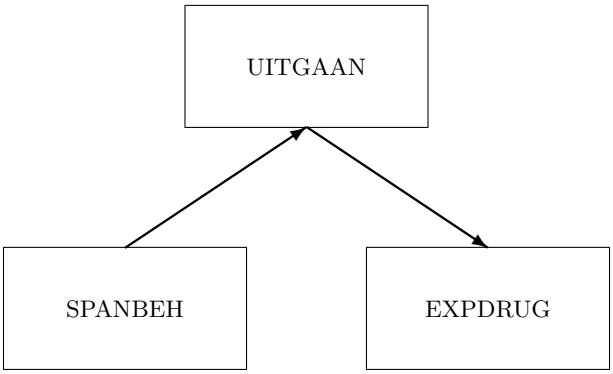
De HMRA is een nuttige techniek om uit te zuiveren voor de invloed van storende variabelen (de variabelen in Blok 1). We noemen $R^2_{\text{Change}}$ de *uitgezuiverde* proportie verklaarde variantie omdat de verklaarde variantie die Blok 2 (de variabelen waarin we geïnteresseerd zijn) gemeenschappelijk heeft met Blok 1 uitgezuiverd is. De niet-uitgezuiverde proportie verklaarde variantie is $R^2_{Blok\,2}$, de proportie van de variantie die we kunnen verklaren met een regressievergelijking waarin alleen de predictoren uit Blok 2 zijn opgenomen.

Men mag het nut van de HMRA echter niet overschatten. Met een HMRA kan men immers alleen uitzuiveren voor die storende variabelen die men ook geregistreerd heeft (en opgenomen in Blok 1). Bij een significante $R^2_{\text{Change}}$ blijft het altijd mogelijk dat er een niet-geregistreerde storende variabele werkzaam is (een variabele die zowel met het criterium als met de variabelen in Blok 2 samenhangt en daardoor verantwoordelijk kan zijn voor de samenhang tussen deze twee). Op basis van een significante $R^2_{\text{Change}}$ kan men daarom alleen besluiten dat de bijdrage van Blok 2 aan de voorspelling niet volledig verklaard kan worden door Blok 1.

### 4.11.2 HMRA binnen het Onderzoek naar de Rol van Mediërende Variabelen

Een mediërende variabele is een noodzakelijke schakel in het verband tussen het criterium en een groep andere variabelen. Dit kunnen we illustreren aan de hand van een nieuw voorbeeldonderzoek naar de determinanten van experimenteren met drugs. Het is leerzaam om weer de variabelen SPANBEH en EXPDRUG te gebruiken, maar nu binnen een ander onderzoek.

Stel, anders dan in het vorige voorbeeldonderzoek, dat we (1) ervan overtuigd zijn dat SPANBEH een invloed heeft op EXPDRUG en (2) de vraag willen beantwoorden of deze invloed volledig loopt via een derde variabele, hoe vaak men zich in het uitgaansleven begeeft (aangeduid als UITGAAN). De achterliggende gedachte bij deze vraag is dat SPANBEH geen rechtstreekse invloed heeft op EXPDRUG, maar er wel voor zorgt dat men meer uitgaat en daardoor in contact komt met drugs. Deze onderzoeksvraag wordt grafisch voorgesteld in Figuur 4.4. Het belangrijke punt in deze figuur is dat er nu een pijl wégloopt van SPANBEH in plaats van er aan te komen. Hiermee wordt aangegeven dat de invloed van SPANBEH volledig via UITGAAN loopt. Dit kan men ook als volgt uitdrukken: indien men UITGAAN constant zou houden, dan is er geen invloed meer van SPANBEH op EXPDRUG. Om van een mediërende variabele te kunnen spreken, moet het dus zó zijn dat de beïnvloeding uitsluitend via UITGAAN loopt.




Figuur 4.4: Grafische voorstelling van de rol van een mediërende variabele (UITGAAN) in het verband tussen de criteriumvariabele (EXPDRUG) en een andere predictor (SPANBEH).

Deze hypothese kan eveneens met een HMRA onderzocht worden. Blok 1 bevat de (potentieel) mediërende variabelen en Blok 2 de variabelen waarvan men vermoedt dat hun invloed gemedieerd wórdt door de variabelen in Blok 1. In ons voorbeeld is UITGAAN de potentiële mediërende variabele en daarom zit deze variabele in Blok 1. De variabele SPANBEH is de variabele waarvan men vermoedt dat zijn invloed gemedieerd wórdt en daarom zit hij in Blok 2. (Men kan spanningsbehoefte ook operationaliseren met twee variabelen, en dan bijvoorbeeld SPANAANG toevoegen aan Blok 2, maar dit verandert de analyse niet wezenlijk.) Nu, indien $R^2_{\text{Change}}$ significant groter is dan 0 dan concluderen we dat SPANBEH een invloed heeft die níet via UITGAAN (de inhoud van Blok 1) loopt. Met andere woorden, een significante $R^2_{\text{Change}}$ betekent dat UITGAAN geen mediërende variabele is.

Merk op dat in het vorige voorbeeldonderzoek (over de rol van de persoonlijkheids- en gezinskenmerken als storende variabele in het verband tussen SPANBEH en EXPDRUG), de variabele SPANBEH ook in Blok 2 gestopt werd. Dus, twee verschillende onderzoeksvragen worden beantwoord met dezelfde analyse:

1. Is het verband tussen SPANBEH en EXPDRUG toe te schrijven aan een storende variabele?
2. Wordt het verband tussen SPANBEH en EXPDRUG gemedieerd door een derde variabele?

Het verschil tussen deze twee onderzoeksvragen zit uitsluitend in de richting van de linkerpijl in Figuur 4.3 en Figuur 4.4: in de eerste figuur loopt deze pijl naar SPANBEH toe en in de tweede figuur loopt hij ervan weg. De richting van de pijlen geeft de richting van het causale verband aan. Deze richting zit echter in het hoofd van de onderzoeker en niet in de data. Omdat de data geen informatie bevatten over de richting van de verbanden, worden de twee genoemde onderzoeksvragen met dezelfde data-analytische techniek beantwoord.

---

*Opgave 36*

Een onderzoeker beweert dat het persoonlijkheidskenmerk "gewetensvolheid" medebepalend is voor het aantal verkeersovertredingen dat iemand maakt. Als criteriumvariabele neemt hij het aantal verkeersovertredingen per jaar gedeeld door het aantal kilometers dat men jaarlijks met de auto rijdt. Hij gebruikt drie vragenlijsten om gewetensvolheid te meten omdat hij vindt dat elk van hen een ander aspect van dit persoonlijkheidskenmerk meet. Deze drie vragenlijstscores vormen samen Blok 2. De onderzoeker is beducht voor de invloed van storende variabelen. Daarom meet hij ook vijf variabelen waarvan hij vermoedt dat ze zowel met gewetensvolheid als met de criteriumvariabele samenhangen. Deze storende variabelen vormen samen Blok 1. Er nemen in totaal 110 proefpersonen deel aan dit onderzoek. De proportie verklaarde variantie door Blok 1 is 0.45 en die door Blok 1 en Blok 2 samen is 0.59.

1. Bereken $R^2_{\text{Change}}$ en $F_{\text{Change}}$. Zoek, in een statistische tabel, wat de bijbehorende kritische F-waarde is op het significantieniveau 0.05. Wat concludeer je?
2. Een collega bekritiseert de onderzoeker omdat hij de variabele LEEFTIJD niet opgenomen heeft bij de potentiële storende variabelen. Achteraf doet hij dat wel, en vindt 0.61 voor $R^2_{\text{Blok 1}}$ en 0.62 voor $R^2_{\text{Blok 1+2}}$. Bereken $R^2_{\text{Change}}$, $F_{\text{Change}}$ en zoek de bijbehorende kritische F-waarde op significantieniveau 0.05. Wat concludeer je nu?

---

## 4.12 Regressieanalyse met Categorische Predictoren

Tot hiertoe hebben we alleen gebruik gemaakt van kwantitatieve predictoren. Kwantitatieve predictoren drukken een bepaalde graad of hoeveelheid uit. Voorbeelden van kwantitatieve predictoren zijn leeftijd, een testscore, en een beoordeling op een 10-puntenschaal. Categorische predictoren, daarentegen, drukken alleen uit dát er een verschil is, maar de getallen waarmee dit gebeurt drukken geen hoeveelheid uit. Voorbeelden van categorische predictoren zijn geslacht, nationaliteit, politieke voorkeur, en studierichting. Voor bijvoorbeeld de variabele GESLACHT kan men mannen aanduiden met het getal 1 en vrouwen met het getal 0. Categorische variabelen worden ook kwalitatieve of nominale variabelen genoemd.

In deze paragraaf wordt beschreven hoe de regressieanalyse uitgebreid kan worden door ook categorische variabelen als predictoren te gebruiken. De analyse die hieruit volgt blijkt formeel identiek te zijn aan de covariantieanalyse. Het verschil tussen de paragraaf over covariantieanalyse en deze paragraaf is de benadering: in de eerstgenoemde paragraaf werd de analyse benaderd vanuit de vraag hoe men kan uitzuiveren voor de invloed van een storende kwantitatieve variabele en hier wordt de analyse benaderd vanuit de vraag hoe men een criteriumvariabele zo goed mogelijk kan voorspellen.

### 4.12.1 Voorbeeld

Stel dat we een voorspelling willen maken van de technische leesvaardigheid van kinderen in Groep 3. Technische leesvaardigheid (TECHNISCH LEZEN) is het correct en vlot kunnen oplezen van losse woordjes. Het moet onderscheiden worden van begrijpende leesvaardigheid, het correct kunnen beantwoorden van vragen over een gelezen tekst. We maken gebruik van twee predictoren waarvan de scores vastgesteld werden op het einde van Groep 2, dus vóór het begin van het formele leesonderwijs. De eerste predictor is ALFABETKENNIS: het aantal letters van het alfabet dat een kind foutloos kan benoemen. Gehele of gedeeltelijke kennis van het alfabet is vaak reeds aanwezig voordat een kind naar Groep 3 gaat. De tweede predictor is FONEEMDISCRIMINATIE: het kunnen manipuleren van fonemen in gesproken woorden en pseudowoorden[11]. Fonemen zijn de elementaire klankeenheden waaruit de gesproken woorden zijn opgebouwd. Eén van de taken waarmee FONEEMDISCRIMINATIE gemeten wordt is foneemsubstitutie. In deze taak vraagt men aan een kind om in een woord (bijvoorbeeld, *roep*) één klank (bijvoorbeeld, de $r$) te vervangen door een andere klank (bijvoorbeeld, de $s$). Uit veel onderzoek is gebleken dat ALFABETKENNIS en FONEEMDISCRIMINATIE de belangrijkste predictoren zijn van TECHNISCH LEZEN.

We willen nu de kwaliteit van de voorspelling verbeteren door ook rekening te houden met het milieu waarin het kind opgroeit. Concreet, we willen rekening houden met het feit of de ouders van het kind autochtoon of allochtoon zijn (aangeduid als AFKOMST). Dit heeft zin omdat een kind wellicht sneller leert lezen als het ook thuis veel oefent. De mate waarin het kind thuis oefent hangt af van de hulp die het van zijn ouders

[11] Een voorbeeld van een pseudowoord is *kloem*. Pseudowoorden zijn foneemreeksen die geen betekenis hebben, maar die wel voldoen aan de regels van het Nederlands voor wat betreft de toegelaten foneemcombinaties. Een voorbeeld van een foneemreeks die niet voldoet aan de regels van het Nederlands is *tvko*.

krijgt, en die hulp is wellicht meer beschikbaar bij autochtone dan bij allochtone ouders. Men kan de variabele AFKOMST op meerdere manieren definiëren. De eenvoudigste manier is door elk kind in te delen in ofwel de autochtone ofwel de allochtone categorie. Dit is een categorische variabele met twee niveaus. Indien men deze tweedeling te ruw vindt, dan kan men de volgende driedeling overwegen: (1) autochtoon, (2) tweede generatie allochtoon (geboren als kind van immigranten), (3) eerste generatie allochtoon (immigrant, geboren in het buitenland). Dit is een categorische variabele met drie niveaus. Voor wat hierna volgt, doet het aantal niveaus van de categorische variabele er niet toe. Daarom zijn beide operationalisaties van de variabele AFKOMST (met twee of met drie niveaus) geschikt als voorbeeld. In de volgende paragraaf wordt beschreven hoe men gebruik kan maken van een categorische variabele om de voorspelling te verbeteren.

### 4.12.2 De Regressievergelijking

We maken gebruik van een categorische variabele bij de voorspelling door deze in de regressievergelijking op te nemen. In ons voorbeeld is AFKOMST de categorische variabele. Het aantal niveaus van een categorische variabele wordt aangeduid met $K$, en in ons voorbeeld is is $K$ gelijk aan 2 of 3. De niveaus van de categorische variabele worden ook groepen genoemd. Nu, het opnemen van een categorische variabele in de regressievergelijking gebeurt door van die ene regressievergelijking $K$ regressievergelijkingen te maken, elk met hun eigen intercept. Stel dat er slechts één kwantitatieve predictor is. Dan wordt de k-de regressievergelijking ($k = 1, \ldots, K$) als volgt geschreven:

$$\boxed{\widehat{Y}_k = A_k + BX_k \quad .}$$

Hierin staat $\widehat{Y}_k$ voor de voorspelde scores van de personen uit de k-de groep en staat $X_k$ voor de scores van deze personen op de kwantitatieve predictor. Merk op dat deze $K$ regressievergelijkingen slechts van elkaar verschillen wat betreft het intercept; het regressiegewicht $B$ is hetzelfde voor alle groepen[12]. Deze regressievergelijking is eenvoudig uit te breiden naar de situatie met twee kwantitatieve predictoren:

$$\widehat{Y}_k = A_k + B_1X_{1k} + B_2X_{2k} \quad .$$

Hierin staan $X_{1k}$ en $X_{2k}$ voor de scores op de twee kwantitatieve predictoren van de personen uit de k-de groep. In ons voorbeeld zijn ALFABETKENNIS en FONEEMDISCRIMINATIE de twee kwantitatieve variabelen.

De beste manier om een idee te krijgen van de rol van het intercept, is je af te vragen wat het verschil is tussen de voorspelde scores van twee personen die dezelfde scores hebben op de kwantitatieve predictoren, maar tot verschillende groepen behoren. Concreet, voor personen die tot groepen $k$ en $l$ behoren, maar dezelfde scores hebben op de kwantitatieve predictoren, is het verschil tussen de voorspelde scores gelijk aan $A_k - A_l$. Hieruit blijkt dat, naarmate de intercepten onderling meer van elkaar verschillen, de voorspelde scores in sterkere mate bepaald worden door de groep waartoe men behoort.

[12] Later zullen we een variant bespreken waarin de regressiegewichten wél verschillen over de groepen.

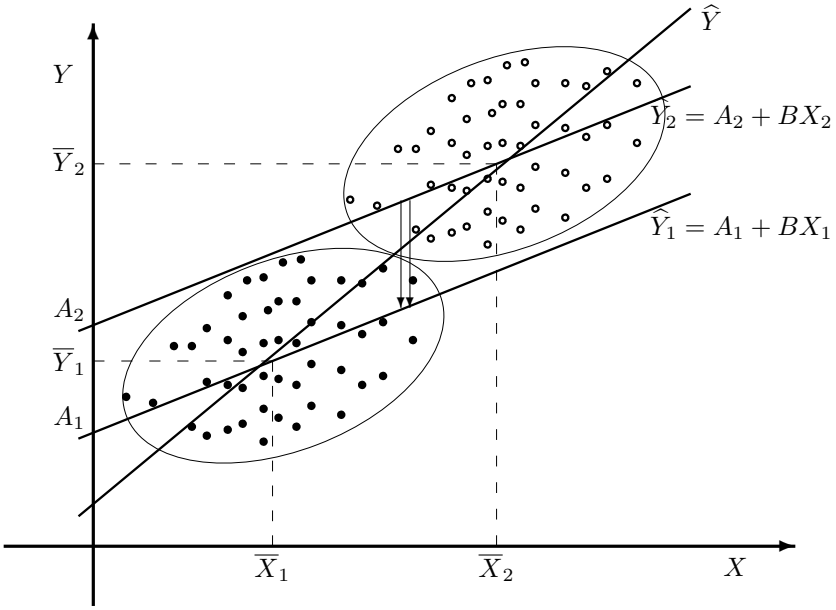
De formules voor de regressiegewichten $B$, $B_1$ en $B_2$ zijn hier ingewikkelder dan wanneer er slechts één gemeenschappelijk intercept is. De formules voor de intercepten $A_k$ ($k = 1, \ldots, K$), daarentegen, zijn analoog aan die voor het gemeenschappelijke intercept $A$. Voor een regressievergelijking met één kwantitatieve predictor wordt $A_k$ als volgt berekend:

$$A_k = \overline{Y}_k - B\overline{X}_k \quad ,$$

waarin $\overline{Y}_k$ de gemiddelde criteriumscore is in de k-de groep, en $\overline{X}_k$ de gemiddelde score op de kwantitatieve predictor in deze groep. En voor een regressievergelijking met twee kwantitatieve predictoren wordt $A_k$ als volgt berekend:

$$A_k = \overline{Y}_k - B_1\overline{X}_{1k} - B_2\overline{X}_{2k} \quad .$$

De formules voor de regressiegewichten hangen niet alleen af van de samenhang tussen de kwantitatieve predictoren onderling, maar ook van de samenhang tussen elk van de kwantitieve predictoren en de categorische predictor. In Figuur 4.5 wordt dit geïllustreerd voor een regressieanalyse met één kwantitatieve en één categorische predictor. De situatie met meerdere kwantitatieve predictoren is volledig analoog, maar is wat moeilijker voor te stellen in een figuur. De categorische predictor in Figuur 4.5 heeft twee niveaus. De scores in het ene niveau worden voorgesteld met dikke punten en die in het andere niveau met kleine cirkeltjes.




Figuur 4.5: De regressielijn voor de voorspelling op basis van alleen de kwantitatieve predictor ($\widehat{Y}$) en de twee regressielijnen voor de voorspelling op basis van de kwantitatieve en de categorische predictor ($\widehat{Y}_1$ en $\widehat{Y}_2$).

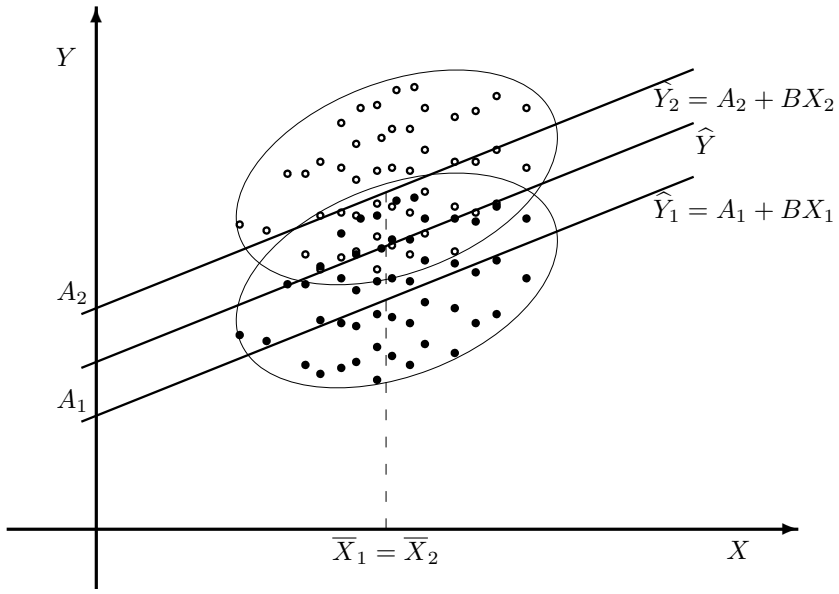
De regressielijn voor de voorspelling met alleen de kwantitatieve predictor ($\widehat{Y}$) wordt zo gekozen dat de samenhang tussen $X$ en $Y$ in de twee puntenwolken samen, zo

goed mogelijk beschreven wordt met een rechte lijn (zoals uitgedrukt in het kleinste-kwadratencriterium). Dit is niet zo voor de voorspelling op basis van de kwantitatieve én de categorische predictor. Daar worden de twee regressielijnen ($\widehat{Y}_1$ en $\widehat{Y}_2$) zo gekozen dat de samenhang bínnen elk van de groepen (i.e., de samenhang tussen de $X_k$- en de $Y_k$-variabelen) zo goed mogelijk beschreven wordt met een rechte lijn. Merk op dat deze twee binnen-groepen regressielijnen ($\widehat{Y}_1$ en $\widehat{Y}_2$) hetzelfde regressiegewicht hebben (ze zijn parallel). Het regressiegewicht van de binnen-groepen regressielijnen is kleiner dan dat van de over-groepen regressielijn ($\widehat{Y}$). Dit komt omdat de samenhang tussen de kwantitatieve predictor $X$ en het criterium $Y$ óver de groepen sterker is dan de samenhang bínnen elk van de groepen. Anders dan bij de binnen-groepen regressielijnen, wordt het regressiegewicht van de over-groepen regressielijn ook bepaald door het patroon in de groepsgemiddelden op de kwantitatieve predictor (i.e., de $\overline{X}_k$'s) en de bijbehorende groepsgemiddelden op het criterium (i.e., de $\overline{Y}_k$'s). Voor de data in Figuur 4.5 is de samenhang óver de groepen positiever dan de samenhang bínnen de groepen omdat $\overline{X}_2$ en $\overline{Y}_2$ beide groot zijn en $\overline{X}_1$ en $\overline{Y}_1$ beide klein.

Het verschillend zijn van de regressiegewichten in de voorspelling mét en de voorspelling zónder categorische predictor, is een gevolg van de samenhang tussen de kwantitatieve en de categorische predictor. Deze samenhang blijkt uit het feit dat het gemiddelde van de kwantitatieve variabele verschillend is binnen elk van de niveaus van de categorische variabele. Dit is goed te zien in Figuur 4.5: het gemiddelde van de kwantitatieve predictor in het tweede niveau van de categorische predictor ($\overline{X}_2$) is groter dan het gemiddelde in het eerste niveau ($\overline{X}_1$). Indien er geen samenhang zou zijn tussen deze twee predictoren, dan zou het regressiegewicht in de voorspelling mét de categorische predictor hetzelfde zijn als in de voorspelling zónder deze predictor. Dit kan men zien als de twee puntenwolken over hun regressielijn naar elkaar toegeschoven worden totdat ze precies boven elkaar liggen. Het resultaat van deze verschuiving wordt getoond in Figuur 4.6. Na de verschuiving zijn de gemiddelden van de kwantitatieve predictor in elk van de twee niveaus van de categorische predictor gelijk aan elkaar. De regressielijn van de voorspelling zónder categorische predictor komt dan precies tussen de twee regressielijnen van de voorspelling mét categorische predictor te liggen. Deze laatste twee regressielijnen veranderen niet door de verschuiving van de puntenwolken.

### 4.12.3 Statistische Significantie

We behandelen nu de vraag hoe men kan bepalen of een predictor een statistisch significante bijdrage levert aan de voorspelling van het criterium bovenop de andere predictoren. We maken hier een onderscheid tussen (1) de bijdrage van een categorische predictor bovenop de andere predictoren en (2) de bijdrage van een kwantitatieve predictor bovenop de andere predictoren, waaronder een categorische. Voor ons voorbeeld, komt de vraag naar de statistische significantie van de categorische predictor erop neer of de predictor AFKOMST een significante bijdrage levert aan de voorspelling van TECHNISCH LEZEN bovenop ALFABETKENNIS en FONEEMDISCRIMINATIE. En de vraag naar de statistische significantie van de kwantitatieve predictor komt erop neer of de predictor ALFABETKENNIS of FONEEMDISCRIMINATIE een significante bijdrage levert aan de voorspelling van TECHNISCH LEZEN bovenop de categorische predictor AFKOMST en




Figuur 4.6: Binnen-groepen regressielijnen ($\widehat{Y}_1$ en $\widehat{Y}_2$) en over-groepen regressielijn ($\widehat{Y}$) voor de situatie met gelijke groepsgemiddelden op de kwantitatieve predictor.

de andere kwantitatieve predictor.

### De Statistische Significantie van een Categorische Predictor

Het bepalen van de statistische significantie van een categorische predictor gebeurt op dezelfde manier als het bepalen van de statistische significantie van een blok predictoren in de HMRA. Ook hier maken we gebruik van de $\mathrm{F}_{\mathrm{Change}}$-statistiek. Het blijkt handig te zijn als we voor deze $\mathrm{F}_{\mathrm{Change}}$-statistiek nu een andere formule gebruiken

$$\boxed{F_{\mathrm{Change}} = \frac{R^2_{\mathrm{Change}}/q_{\mathrm{Blok\,2|Blok\,1}}}{(1 - R^2_{\mathrm{Blok\,1+2}})/(n - q_{\mathrm{Blok\,1+2}})} \quad .}$$

De categorische predictor stoppen we in Blok 2 en alle andere predictoren stoppen we in Blok 1. Dan berekenen we $R^2_{\mathrm{Change}}$ als het verschil tussen $R^2_{\mathrm{Blok\,1+2}}$ en $R^2_{\mathrm{Blok\,1}}$, en gebruiken dit resultaat om $F_{\mathrm{Change}}$ te berekenen. Merk op dat $R^2_{\mathrm{Blok\,1+2}}$ gelijk is aan de proportie verklaarde variantie door de regressievergelijking met álle predictoren (de categorische predictor die men wil toetsen plus alle andere). En $R^2_{\mathrm{Blok\,1}}$ is gelijk aan de proportie verklaarde variantie door de regressievergelijking zónder de categorische predictor die men wil toetsen.

Om de $F_{\mathrm{Change}}$-statistiek te kunnen berekenen, moeten we weten waaraan $q_{\mathrm{Blok\,2|Blok\,1}}$ en $q_{\mathrm{Blok\,1+2}}$ gelijk zijn. Het symbool $q_{\mathrm{Blok\,2|Blok\,1}}$ staat voor het aantal extra parameters

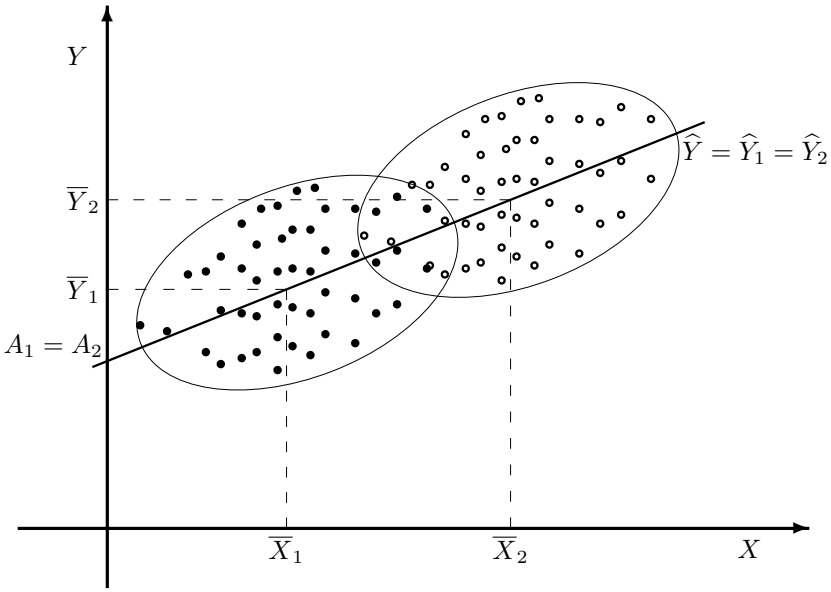
dat nodig is om de categorische predictor in de regressievergelijking op te nemen. De parameters waar het hier over gaat, zijn de intercepten en de regressiegewichten. Het aantal extra parameters dat nodig is om een categorische predictor in de regressievergelijking op te nemen wordt ook het aantal vrijheidsgraden van deze categorische predictor genoemd. Het aantal vrijheidsgraden van een categorische predictor is gelijk aan het aantal groepen min 1 (i.e., $K - 1$). Dit aantal vrijheidsgraden is eenvoudig af te leiden als men de regressievergelijking met één kwantitatieve predictor vergelijkt met de regressievergelijking met één kwantitatieve en één categorische predictor. De eerste regressievergelijking bevat één intercept en de tweede bevat evenveel intercepten als er niveaus zijn in de categorische predictor. Dus, het aantal extra intercepten is gelijk aan het aantal niveaus (groepen) min 1. Merk op dat, indien de predictor in Blok 2 kwantitatief zou zijn in plaats van categorisch, $q_{\text{Blok 2}|\text{Blok 1}}$ dan gelijk zou zijn aan 1. Men heeft immers slechts één extra parameter nodig om een kwantitatieve predictor in de regressievergelijking op te nemen. Daarom zeggen we ook dat een kwantitieve predictor slechts één vrijheidsgraad heeft.

Het symbool $q_{\text{Blok 1+2}}$ in de noemer van de $\text{F}_{\text{Change}}$-statistiek staat voor het totaal aantal parameters in de regressievergelijking met alle predictoren. We noemen $n - q_{\text{Blok 1+2}}$ ook het aantal vrijheidsgraden van de noemer van de $\text{F}_{\text{Change}}$-statistiek. Het totaal aantal parameters in de regressievergelijking met alle predictoren is gelijk aan het aantal kwantitatieve predictoren plus het aantal niveaus van de categorische predictor. Immers, voor elke kwantitatieve predictor is er één regressiegewicht en voor elk niveau van de categorische predictor is er één intercept. Merk op dat, indien de regressievergelijking geen categorische predictor bevat, $q_{\text{Blok 1+2}}$ dan gelijk zou zijn aan het aantal kwantitatieve predictoren plus 1. In dit geval is er immers slechts één gemeenschappelijk intercept.

Met de bovenstaande $\text{F}_{\text{Change}}$-statistiek wordt de nulhypothese getoetst dat, in de populatie, de categorische predictor geen extra bijdrage levert aan de voorspelling van het criterium bovenop de andere predictoren. Indien er slechts één andere predictor is, en deze bovendien kwantitatief is, dan is deze nulhypothese equivalent met de nulhypothese dat, in de populatie, de intercepten van de binnen-groepen regressielijnen gelijk zijn aan elkaar. De gegevens in Figuur 4.5 en 4.6 zijn van die aard dat, indien het aantal observaties voldoende groot is, deze nulhypothese verworpen zal worden.

Het is nuttig om een voorbeeld te geven van een situatie waarin deze nulhypothese níet verworpen zal worden. Dit voorbeeld wordt gegeven in Figuur 4.7. We krijgen deze figuur als we in Figuur 4.5 de twee puntenwolken in verticale richting naar elkaar toeschuiven totdat hun regressielijnen samenvallen. In Figuur 4.7 hebben we te maken met drie regressielijnen die samenvallen: (1) de over-groepen regressielijn $\widehat{Y}$, (2) de binnen-groep regressielijn van de eerste groep $\widehat{Y}_1$ en (3) de binnen-groep regressielijn van de tweede groep $\widehat{Y}_2$. De voorspellingen op basis van deze drie regressievergelijkingen zijn dus gelijk. Omdat de meervoudige correlatie gelijk is aan de gewone correlatie tussen de geobserveerde en de voorspelde criteriumscores, volgt hieruit dat $R^2_{\text{Blok 1+2}}$ (de proportie verklaarde variantie door de regressievergelijking met beide predictoren) gelijk is aan $R^2_{\text{Blok 1}}$ (de proportie verklaarde variantie door de regressievergelijking met alleen de kwantitatieve predictor). Immers, wat ook de regressievergelijking is (met of zonder categorische predictor), de voorspelde criteriumscores blijven hetzelfde, en

daarom blijft ook hun correlatie met de geobserveerde criteriumscores hetzelfde. Dus, in dit voorbeeld is $R^2_{\text{Change}}$ precies gelijk aan 0, en daarom is ook $F_{\text{Change}}$ gelijk aan 0. In de praktijk zal $R^2_{\text{Change}}$ altijd groter zijn dan 0, maar als hij niet véél groter is dan 0 (wat "veel" is, hangt af van $n$, $K$ en $q_{\text{Blok 1+2}}$), zal $F_{\text{Change}}$ niet significant zijn.




Figuur 4.7: Illustratie van de nulhypothese dat de intercepten van de binnen-groepen regressielijnen ($\widehat{Y}_1$ en $\widehat{Y}_2$) gelijk zijn.

Het patroon in Figuur 4.7 is nauw verwant met de onechte correlatie. Het verband tussen de categorische predictor en het criterium $Y$ kan immers volledig verklaard worden door de kwantitatieve predictor $X$. Het verband tussen de categorische predictor en het criterium blijkt uit het verschil tussen $\overline{Y}_1$ en $\overline{Y}_2$. Strikt genomen gaat het hier dus niet om een onechte correlatie, maar om een onecht *verschil*. Dit verschil noemen we onecht omdat de twee puntenwolken door dezelfde regressielijn beschreven worden. Dit impliceert dat het verschil tussen $\overline{Y}_1$ en $\overline{Y}_2$ volledig verklaard kan worden door het verschil tussen $\overline{X}_1$ en $\overline{X}_2$.

De statistische toets voor de bijdrage van een categorische predictor aan de voorspelling is identiek aan de F-toets van de covariantieanalyse. In de covariantieanalyse is men geïnteresseerd in de vraag of er een verschil is tussen de groepen na uitzuivering voor een covariaat. In dit hoofdstuk, noemen we een covariaat een kwantitatieve predictor. Het uitgezuiverde verschil tussen twee groepen is het verschil tussen de voorspelde scores voor deze twee groepen bij één bepaalde waarde van de covariaat. In Figuur 4.5 wordt dit uitgezuiverde verschil aangeduid met een dubbele pijl. We noemen dit verschil uitgezuiverd voor de covariaat omdat het berekend wordt bij één bepaalde waarde van de covariaat. Een eventueel verschil tussen de groepen in gemiddelde covariaatscore, dat medebepalend kan zijn voor het niet-uitgezuiverde verschil $\overline{Y}_2 - \overline{Y}_1$, kan dus geen rol

**Andere Notatie en Andere Berekening**
Het is goed om even stil te staan bij de verschillen in notatie tussen dit hoofdstuk en Hoofdstuk 3 (*Variantie- en Covariantieanalyse*). De voorspelling op basis van het covariantieanalysemodel werd in Hoofdstuk 3 aangeduid met $\widehat{y}_k$, en het werd als volgt berekend:

$$\overline{y}_k + B(x_k - \overline{x}_k) \quad .$$

In dit hoofdstuk worden variabelen aangeduid met hoofdletters in plaats van met kleine letters, zodat men het volgende krijgt:

$$\begin{aligned}\widehat{Y}_k =& \overline{Y} + B(X_k - \overline{X}_k)\\ =& (\overline{Y}_k - B\overline{X}_k) + BX_k \quad .\end{aligned}$$

We weten ook dat het intercept $A_k$ gelijk is aan $\overline{Y}_k - B\overline{X}_k$. Daarom kan $\widehat{Y}_k$ als volgt geschreven worden:

$$\widehat{Y}_k = A_k + BX_k \quad ,$$

hetgeen identiek is met regressievergelijking van de k-de groep, gegeven op pagina 379.
Een ander verschil met Hoofdstuk 3 is de berekening van de toetsstatistiek. In Hoofdstuk 3 berekenden we een F-statistiek door $MS_{A|X}$ te delen door $MS_{Error|X,A}$ en hier berekenen we een F-statistiek op basis van twee proporties verklaarde variantie, $R^2_{\text{Blok 1+2}}$ en $R^2_{\text{Blok 1}}$. Alhoewel men het niet onmiddellijk kan zien aan de formules, geven beide F-statistieken precies hetzelfde resultaat.
In Hoofdstuk 3 werd slechts voor één covariaat uitgezuiverd. Door de covariantieanalyse te formuleren als regressieanalyse met een categorische predictor, is het onmiddellijk duidelijk hoe men kan uitzuiveren voor meer dan één covariaat (kwantitatieve predictor). Dit doet men door meerdere kwantitatieve predictoren op te nemen in Blok 1.

spelen bij het uitgezuiverde verschil. Het uitgezuiverde verschil is gelijk aan het verschil tussen de intercepten voor deze groepen. Dus, de nulhypothese dat de uitgezuiverde populatieverschillen allemaal gelijk zijn aan 0 is equivalent met de nulhypothese dat, in de populatie, de intercepten voor de verschillende groepen gelijk zijn aan elkaar.

We kunnen het bovenstaande illustreren aan de hand van ons voorbeeld. Stel dat autochtone kinderen hoger scoren op TECHNISCH LEZEN dan allochtone kinderen. Stel ook dat we vinden dat de categorische predictor AFKOMST een significante bijdrage levert aan de voorspelling van TECHNISCH LEZEN bovenop ALFABETKENNIS en FONEEMDISCRIMINATIE. In termen van de covariantieanalyse betekent dit dat, na uitzuivering voor ALFABETKENNIS en FONEEMDISCRIMINATIE, autochtone en allochtone kinderen nog steeds van elkaar verschillen wat betreft TECHNISCH LEZEN. Met andere woorden, het verschil in TECHNISCH LEZEN wordt niet volledig verklaard door verschillen in ALFABETKENNIS en FONEEMDISCRIMINATIE.

### De Statistische Significantie van een Kwantitatieve Predictor

Voor het bepalen van de statistische significantie van een kwantitatieve predictor kan men gebruikmaken van twee toetsstatistieken: de T-statistiek en de $F_{\text{Change}}$-statistiek.

Deze twee toetsstatistieken zijn equivalent omdat ze resulteren in dezelfde p-waarde[13]. We beperken ons hier tot de $\mathrm{F}_{\mathrm{Change}}$-statistiek omdat deze ook gebruikt wordt voor het toetsen van een categorische predictor. Concreet, de kwantitatieve predictor stoppen we in Blok 2 en alle andere predictoren stoppen we in Blok 1. Dan berekenen we $R^2_{\mathrm{Change}}$ als het verschil tussen $R^2_{\mathrm{Blok\,1+2}}$ en $R^2_{\mathrm{Blok\,1}}$, en gebruiken dit resultaat om $F_{\mathrm{Change}}$ te berekenen. Zoals steeds, is $R^2_{\mathrm{Blok\,1+2}}$ gelijk aan de proportie verklaarde variantie door de regressievergelijking met álle predictoren en is $R^2_{\mathrm{Blok\,1}}$ gelijk aan de proportie verklaarde variantie door de regressievergelijking zónder de predictor die men wil toetsen (in dit geval, een kwantitatieve predictor).

Het aantal vrijheidsgraden in de teller van de $\mathrm{F}_{\mathrm{Change}}$-statistiek ($q_{\mathrm{Blok\,2|Blok\,1}}$) is gelijk aan het aantal parameters dat nodig is om de kwantitatieve predictor in de regressievergelijking op te nemen. Dit kan met één extra regressiegewicht en daarom is $q_{\mathrm{Blok\,2|Blok\,1}}$ gelijk aan 1. Het aantal vrijheidsgraden in de noemer van de $\mathrm{F}_{\mathrm{Change}}$-statistiek ($n - q_{\mathrm{Blok\,1+2}}$) is hetzelfde als in de toets van een categorische predictor. Daarom blijven we er niet bij stilstaan.

Met deze $\mathrm{F}_{\mathrm{Change}}$-statistiek wordt de nulhypothese getoetst dat, in de populatie, de kwantitatieve predictor geen extra bijdrage levert aan de voorspelling van het criterium bovenop de andere predictoren. Indien er slechts één andere predictor is, en deze bovendien categorisch is, dan is deze nulhypothese equivalent met de nulhypothese dat, in de populatie, de regressiegewichten van de binnen-groepen regressielijnen gelijk zijn aan nul. De gegevens in Figuur 4.5 en 4.6 zijn zo dat, indien het aantal observaties voldoende groot is, deze nulhypothese verworpen zal worden.

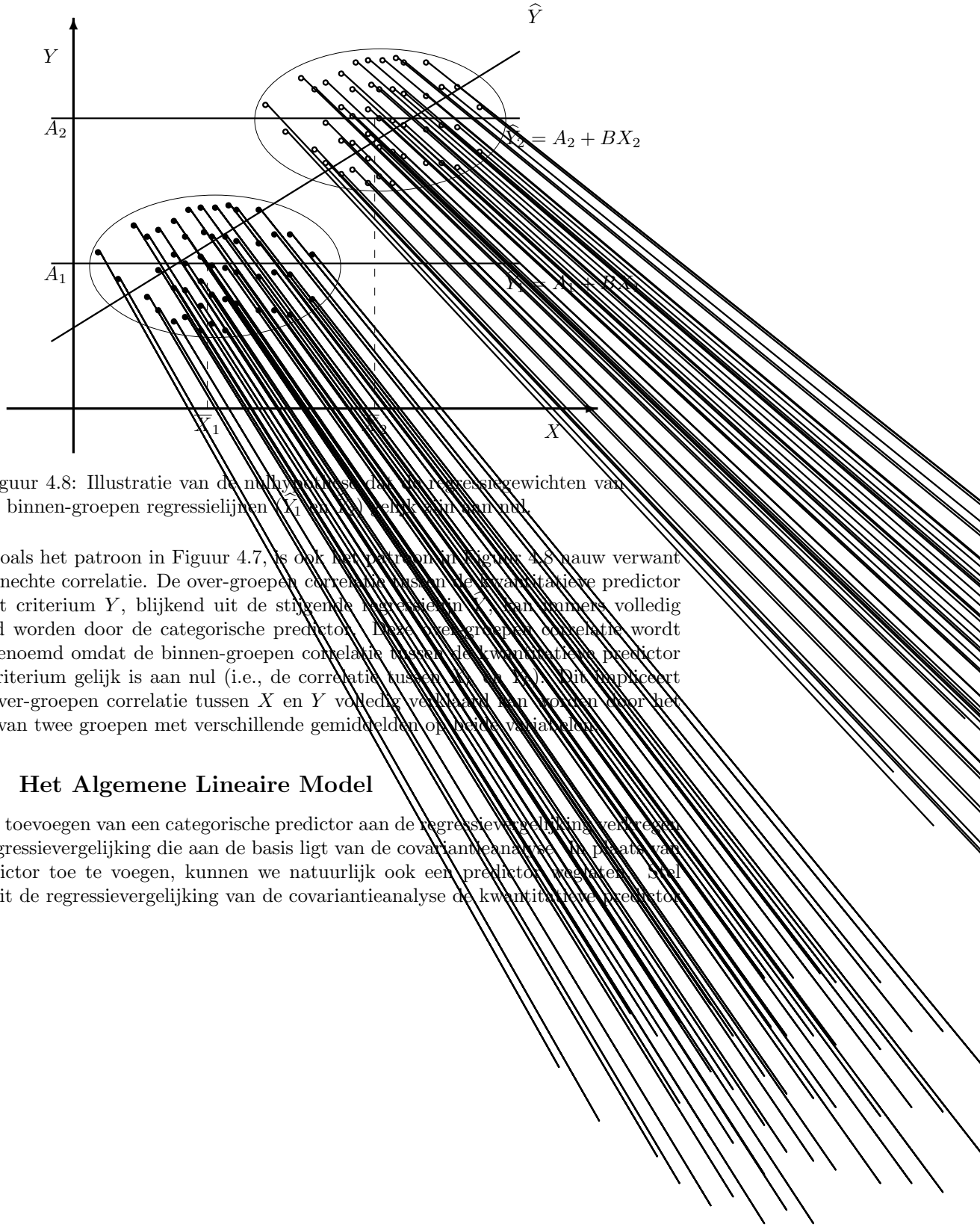
Het is nuttig om een voorbeeld te geven van een situatie waarin deze nulhypothese niet verworpen zal worden. Dit voorbeeld wordt gegeven in Figuur 4.8. We krijgen deze figuur als we in Figuur 4.5 beide puntenwolken ongeveer 25 graden met de richting van de klok meedraaien totdat de binnen-groepen regressielijnen horizontaal lopen. In Figuur 4.8 wordt ook de over-groepen regressielijn $\widehat{Y}$ getoond, maar die is voor de toetsing van de kwantitatieve predictor niet van belang. Wat wél van belang is, zijn de voorspelde criteriumscores op basis van de regressievergelijking met alleen de categorische predictor. Zoals we in de volgende paragraaf zullen zien, is de voorspelling op basis van deze regressievergelijking gelijk aan het groepsgemiddelde $\overline{Y}_k$. Voor de situatie met vlakke binnen-groepen regressielijnen, getoond in Figuur 4.8, is de voorspelling op basis van de regressievergelijking met de categorische én de kwantitatieve predictor óók gelijk aan het groepsgemiddelde. Vlakke binnen-groepen regressielijnen betekenen immers dat $B$ gelijk is aan 0, en hieruit volgt dat

$$\begin{aligned}\widehat{Y}_k &= A_k + BX_k \\ &= A_k \\ &= \overline{Y}_k - B\overline{X}_k \\ &= \overline{Y}_k \quad .\end{aligned}$$

Omdat de meervoudige correlatie gelijk is aan de gewone correlatie tussen de geobserveerde en de voorspelde criteriumscores, volgt uit het voorgaande dat $R^2_{\mathrm{Blok\,1+2}}$ (de

[13]Dit geldt alleen voor een tweezijdige toets. De T-statistiek kan men zowel voor een eenzijdige als voor een tweezijdige toets gebruiken. De $\mathrm{F}_{\mathrm{Change}}$-statistiek, daarentegen, kan alleen gebruikt worden voor een tweezijdige toets.

proportie verklaarde variantie door de regressievergelijking met beide predictoren) gelijk is aan $R^2_{\text{Blok 1}}$ (de proportie verklaarde variantie door de regressievergelijking met alleen de categorische predictor). Dus, in dit voorbeeld is $R^2_{\text{Change}}$ precies gelijk aan 0, en daarom is ook $F_{\text{Change}}$ gelijk aan 0. In de praktijk zal $R^2_{\text{Change}}$ altijd groter zijn dan 0, maar als hij niet véél groter is dan 0 (wat "veel" is, hangt af van $n$ en $q_{\text{Blok 1+2}}$), zal $F_{\text{Change}}$ niet significant zijn. Dit is volledig analoog aan de statistische toets van het effect van de categorische predictor in Figuur 4.7.




Figuur 4.8: Illustratie van de nulhypothese dat de regressiegewichten van de binnen-groepen regressielijnen ($\widehat{Y}_1$ en $\widehat{Y}_2$) gelijk zijn aan nul.

Net zoals het patroon in Figuur 4.7, is ook het patroon in Figuur 4.8 nauw verwant met de onechte correlatie. De over-groepen correlatie tussen de kwantitatieve predictor $X$ en het criterium $Y$, blijkend uit de stijgende regressielijn $\widehat{Y}$, kan immers volledig verklaard worden door de categorische predictor. Deze over-groepen correlatie wordt onecht genoemd omdat de binnen-groepen correlatie tussen de kwantitatieve predictor en het criterium gelijk is aan nul (i.e., de correlatie tussen $X_1$ en $Y_1$). Dit impliceert dat de over-groepen correlatie tussen $X$ en $Y$ volledig verklaard kan worden door het bestaan van twee groepen met verschillende gemiddelden op beide variabelen.

### 4.12.4 Het Algemene Lineaire Model

Door het toevoegen van een categorische predictor aan de regressievergelijking verkregen we de regressievergelijking die aan de basis ligt van de covariantieanalyse. In plaats van een predictor toe te voegen, kunnen we natuurlijk ook een predictor weglaten. Stel dat we uit de regressievergelijking van de covariantieanalyse de kwantitatieve predictor

weglaten. Dan blijft de volgende regressievergelijking over:

$$\widehat{Y}_k = A_k \quad .$$

Men kan laten zien dat de kleinste-kwadratenschatter van $A_k$ gelijk is aan $\overline{Y}_k$, de gemiddelde criteriumscore in de k-de groep. Dit is exact dezelfde voorspelde score als in de variantieanalyse met één factor. In dit hoofdstuk noemen we een factor een categorische predictor. Men kan ook laten zien dat de $\mathrm{F}_{\mathrm{Change}}$-statistiek waarmee de significantie van deze categorische predictor getoetst wordt exact gelijk is aan de gekende F-statistiek van de variantieanalyse.

Als men meerdere categorische predictoren in de regressievergelijking opneemt en geen kwantitatieve predictoren, dan krijgt men exact dezelfde voorspelde scores als in de variantieanalyse met meerdere factoren. Bijvoorbeeld, voor de voorspelling van TECHNISCH LEZEN kan men naast de predictor AFKOMST ook gebruik maken van het opleidingsniveau van de ouders. Stel dat men deze variabele operationaliseert in termen van een aantal categorieën (bijvoorbeeld, alleen middelbaar onderwijs, diploma hoger beroepsonderwijs, academisch diploma), dan ligt het voor de hand om deze variabele (OPLEIDINGSNIVEAU) als een tweede categorische predictor aan de regressievergelijking toe te voegen. Met twee categorische predictoren (factoren) kan men de regressievergelijking als volgt schrijven:

$$\widehat{Y}_{kl} = A_{kl} \quad .$$

In deze formule gebruiken we de index $l$ om de verschillende niveaus van de tweede factor (hier, OPLEIDINGSNIVEAU) aan te duiden. De kleinste-kwadratenschatter van $A_{kl}$ hangt af van het feit of men wel of geen interactie veronderstelt tussen de twee factoren. Als men wél interactie veronderstelt, dan is de kleinste-kwadratenschatter gelijk aan $\overline{Y}_{kl}$, het gemiddelde in groep $(k, l)$ (bijvoorbeeld, autochtoon en ouders met een academisch diploma). Als men géén interactie veronderstelt, dan is de kleinste-kwadratenschatter niet in een eenvoudige formule uit te drukken, althans niet in een ongebalanceerd design.

De regressie-, variantie-, en covariantieanalyse zijn nauw verwant aan elkaar. De modellen (regressievergelijkingen) waarop deze drie analysetechnieken gebaseerd zijn worden vaak onder één noemer gevat als het *algemene lineaire model*. Alle statistische toetsen binnen het algemene lineaire model zijn van hetzelfde type. Concreet, alle statistische toetsen zijn gebaseerd op de $\mathrm{F}_{\mathrm{Change}}$-statistiek. Deze statistiek kan op meerdere manieren berekend worden. Eén manier is met behulp van de proporties verklaarde variantie $R^2_{\mathrm{Blok\,1+2}}$ en $R^2_{\mathrm{Blok\,1}}$, zoals uitgelegd in deze paragraaf. Een andere manier om de $\mathrm{F}_{\mathrm{Change}}$-statistiek te berekenen is met behulp van kwadratensommen. De meeste computerprogramma's voor het algemene lineaire model, zoals GLM-Univariate binnen SPSS, geven in hun uitvoer de berekening van de $\mathrm{F}_{\mathrm{Change}}$-statistiek (die in de uitvoer weliswaar niet zo genoemd wordt) in de vorm van een ANOVA-tabel.

Er bestaan verbanden tussen, enerzijds, de proporties verklaarde variantie $R^2_{\mathrm{Blok\,1+2}}$ en $R^2_{\mathrm{Blok\,1}}$, en anderzijds, de kwadratensommen uit Hoofdstuk 3 (*Variantie- en Covariantieanalyse*). Een eerste verband is dat men $R^2_{\mathrm{Change}}$, het verschil tussen $R^2_{\mathrm{Blok\,1+2}}$ en $R^2_{\mathrm{Blok\,1}}$, als volgt kan berekenen:

$$R^2_{\mathrm{Change}} = R^2_{\mathrm{Blok\,1+2}} - R^2_{\mathrm{Blok\,1}} = \frac{SS_{\mathrm{Blok\,2|Blok\,1}}}{SS_{Totaal}} \quad .$$

Hierin staat $SS_{\text{Blok 2}|\text{Blok 1}}$ voor de effectkwadratensom van de categorische predictor in Blok 2 uitgezuiverd voor de predictoren in Blok 1. De totale kwadratensom $SS_{Totaal}$ is de som van de gekwadrateerde afwijkingen $(Y_i - \overline{Y})^2$, de teller van de formule van VAR($Y$). Men krijgt de verschillende types kwadratensommen door aan Blok 1 verschillende invullingen te geven. In Blok 2 zit telkens het hoofdeffect van een categorische predictor, aangeduid met de letter $A$.

1. Men krijgt de type I kwadratensom $SS_A$ indien Blok 1 géén predictoren bevat.
2. Men krijgt de type II kwadratensom $SS_{A|B}$ indien Blok 1 het hoofdeffect van een tweede categorische predictor bevat.
3. Men krijgt de type III kwadratensom $SS_{A|B,A\times B}$ indien Blok 1 het hoofdeffect van de tweede categorische predictor $B$ en het interactie-effect $A \times B$ bevat.
4. Men krijgt de voor $X$ uitgezuiverde kwadratensom $SS_{A|X}$ indien Blok 1 de kwantitatieve predictor $X$ bevat.

Een tweede verband is dat men $R^2_{\text{Blok 1+2}}$, de proportie verklaarde variantie op basis van alle predictoren samen (Blok 1 én Blok 2), als volgt berekenen uit de errorkwadratensom $SS_{Error|\text{Blok 1+2}}$:

$$R^2_{\text{Blok 1+2}} = \frac{SS_{Totaal} - SS_{Error|\text{Blok 1+2}}}{SS_{Totaal}} \quad .$$

In de notatie voor de errorkwadratensom dient het gedeelte "|Blok 1+2" om aan te geven dat de voorspelling gemaakt wordt op basis van alle predictoren in Blok 1 en Blok 2 samen.

Gebruikmakend van de twee bovenstaande formules, kan men de $\text{F}_{\text{Change}}$-statistiek als volgt schrijven als een functie van kwadratensommen:

$$\text{F}_{\text{Change}} = \frac{SS_{\text{Blok 2}|\text{Blok 1}}/q_{\text{Blok 2}|\text{Blok 1}}}{SS_{Error|\text{Blok 1+2}}/(n - q_{\text{Blok 1+2}})} \quad .$$

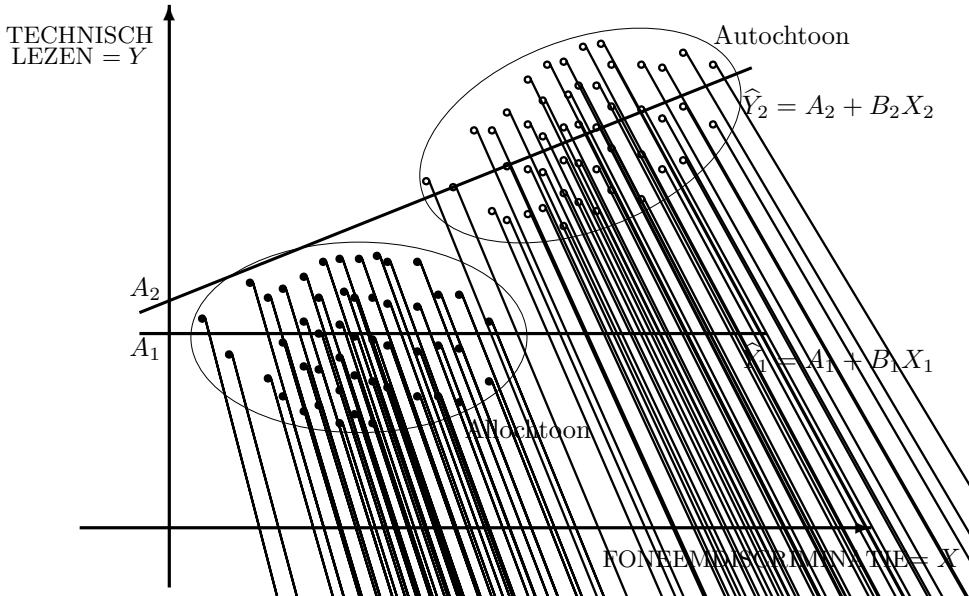
In de teller en de noemer van deze formule staan de gemiddelde kwadratensommen: de gewone kwadratensommen gedeeld door hun aantal vrijheidsgraden. We kennen dit type F-statistiek van de variantie- en de covariantieanalyse.

### 4.12.5 Groepsspecifieke Regressiegewichten

Uit de variantieanalyse kennen we het fenomeen interactie. Interactie is het fenomeen dat het effect van de ene factor (categorische predictor) verschillend is binnen de verschillende niveaus van de andere factor. De vraag is nu of interactie tussen een kwantitatieve en een categorische predictor ook mogelijk is. Dit is inderdaad het geval, en deze vorm van interactie houdt in dat de binnen-groepen regressielijnen verschillende regressiegewichten hebben.

Om dit te illustreren keren we even terug naar het voorbeeld. Stel dat we een voorspelling maken van TECHNISCH LEZEN op basis van de kwantitatieve predictor

FONEEMDISCRIMINATIE en de categorische predictor AFKOMST. (De predictor ALFABETKENNIS wordt dus niet gebruikt.) De gegevens van dit denkbeeldige onderzoek worden getoond in Figuur 4.9, samen met de binnen-groepen regressielijnen. Het belangrijkste in deze figuur is het feit dat de twee binnen-groepen regressielijnen niet parallel zijn. Binnen de groep autochtone kinderen is er een positief verband tussen FONEEMDISCRIMINATIE en TECHNISCH LEZEN, en binnen de groep allochtone kinderen is er geen verband. Een mogelijke verklaring van dit patroon is dat FONEEMDISCRIMINATIE er pas toe doet als voldaan is aan de belangrijkere voorwaarde dat er thuis tijd en aandacht besteed wordt aan het leren lezen. Deze tijd en aandacht is bij de meeste autochtone gezinnen wellicht voldoende aanwezig en daarom wordt daar het verschil gemaakt door FONEEMDISCRIMINATIE. Dus, de gedachte is dat verschillen in aanleg pas goed zichtbaar worden in een omgeving die voor iedereen optimaal of bijna optimaal is. Binnen allochtone gezinnen zijn er wellicht grote verschillen in de tijd en de aandacht die aan het lezen besteed wordt. De variabiliteit in TECHNISCH LEZEN wordt vooral door deze factor bepaald. Gesteld dat er slechts weinig allochtone gezinnen zijn met voldoende aandacht voor het leren lezen, dan zullen de verschillen in FONEEMDISCRIMINATIE slechts weinig tot uiting komen in TECHNISCH LEZEN.




Figuur 4.9: Twee binnen-groepen regressielijnen met verschillende regressiegewichten voor de voorspelling van de TECHNISCH LEZEN op basis van de kwantitatieve predictor FONEEMDISCRIMINATIE en de categorische predictor AFKOMST.

Wanneer er sprake is van interactie tussen een kwantitatieve en een categorische predictor, wordt de categorische predictor vaak een *moderatorvariabele* genoemd. De naam "moderatorvariabele" verwijst naar het feit dat de categorische predictor een moder-

erende (vormbepalende) rol speelt bij het verband tussen de kwantitatieve predictor en het criterium. In ons voorbeeld, bepaalt de categorische predictor AFKOMST of er een positief verband is tussen FONEEMDISCRIMINATIE en TECHNISCH LEZEN (bij de autochtone kinderen) of geen verband (bij de allochtone kinderen).

Alhoewel het ongebruikelijk is, kan de kwantitatieve predictor met evenveel recht een moderatorvariabele genoemd worden. Immers, de kwantitatieve predictor speelt een modererende rol bij het verband tussen de categorische predictor en het criterium. In ons voorbeeld, bepaalt de kwantitatieve predictor FONEEMDISCRIMINATIE of het verschil in TECHNISCH LEZEN tussen autochtone en allochtone kinderen groot of klein is. Voor kinderen met een slechte foneemdiscriminatie is er een vrij klein verschil tussen de twee binnen-groepen regressielijnen (i.e., een vrij zwak verband tussen de categorische predictor en het criterium) en voor kinderen met een goede foneemdiscriminatie is er een groot verschil (i.e., een sterk verband). Merk op dat er geen autochtone kinderen zijn met een zeer slechte foneemdiscriminatie, en ook geen allochtone kinderen met een zeer goede foneemdiscriminatie. Dus, onze vergelijking van autochtone en allochtone leerlingen bij een hoog en een laag niveau van foneemdiscriminatie is gebaseerd op een extrapolatie van de regressielijnen. "Extrapolatie" wil zeggen dat we ons baseren op dat deel van de regressielijn waar geen observaties zijn. Extrapolatie is een risicovolle bezigheid omdat we niet weten of de observaties in dat gebied, gesteld dat we ze toch zouden hebben, ook netjes rond de regressielijn zouden liggen. Of extrapolatie nodig is of niet, hangt af van het patroon in de gegevens. Voor ons voorbeeld is extrapolatie nodig, maar er zijn andere voorbeelden te geven waarvoor dit niet nodig is, omdat de puntenwolken elkaar overlappen.

Keren we nu even terug naar de term "moderatorvariabele". Uit het voorgaande blijkt dat als de categorische predictor een modererende rol speelt bij het verband tussen de kwantitatieve predictor en het criterium, dat dan ook het omgekeerde geldt (de kwantitatieve predictor speelt een modererende rol bij ...). Daarom is het beter om te zeggen dat de twee predictoren elkáárs verband met het criterium modereren. Er is nooit slechts één moderatorvariabele; het zijn er altijd twee.

Voor het statistisch toetsen van de interactie tussen een kwantitatieve en een categorische predictor, hebben we $K$ regressievergelijkingen nodig die toelaten dat de binnengroepen regressielijnen verschillende regressiegewichten hebben. De k-de van deze regressievergelijkingen wordt als volgt geschreven:

$$\boxed{\widehat{Y}_k = A_k + B_k X_k \quad .}$$

In tegenstelling tot de regressievergelijking zónder interactie-effect, verschillen deze $K$ regressievergelijkingen niet alleen wat betreft het intercept maar ook wat betreft het regressiegewicht[14]. Merk op dat het hier gaat over $K$ gewone enkelvoudige regressievergelijkingen, één voor elke groep. Het verschil tussen een enkelvoudige regressieanalyse op de gegevens van alle $K$ groepen samen en de hier voorgestelde regressieanalyse met

[14]Er is hier bewust gekozen voor de wat slordige notatie $B_k$ die, los van de context, op twee manieren opgevat kan worden: (1) het groepsspecifieke regressiegewicht voor de k-de groep in een regressievergelijking met één kwantitatieve predictor of (2) het gemeenschappelijke regressiegewicht van de k-de predictor in een regressievergelijking met meerdere kwantitatieve predictoren. Voor de betekenis van het symbool $B_k$ moet men de context dus goed in de gaten houden.

groepsspecifieke regressiegewichten is dat, in de eerste analyse, alle observaties hetzelfde intercept en regressiegewicht hebben, terwijl in de tweede analyse elke groep zijn eigen intercept en regressiegewicht heeft. De formules voor de groepsspecifieke regressiecoëfficiënten (intercepten en regresssiegewichten) zijn dezelfde als die voor de enkelvoudige regressieanalyse, met dit verschil dat ze nu toegepast worden op de gegevens van de afzonderlijke groepen.

De regressievergelijkingen met groepsspecifieke regressiegewichten zijn eenvoudig uit te breiden naar de situatie met meerdere kwantitatieve predictoren. Het enige wat hiervoor nodig is, zijn extra kwantitatieve predictoren met hun bijbehorende groepsspecifieke regressiegewichten.

De statistische toetsing van de interactie tussen een kwantitatieve en een categorische predictor gebeurt opnieuw met de $F_{\text{Change}}$-statistiek. De interactie tussen de kwantitatieve en de categorische predictor stoppen we in Blok 2 en de kwantitatieve en de categorische predictor zelf (of beter, hun hoofdeffecten) stoppen we in Blok 1. Het is niet onmiddellijk duidelijk hoe die interactie in Blok 2 eruit ziet, en het is op dit moment ook niet nodig om dit precies te weten. Dit wordt later uitgelegd, in de paragraaf over *dummycodering*. Het enige wat hier van belang is, is het resultaat van het wel of niet opnemen van de interactie: als de interactie niet opgenomen wordt als predictor, dan hebben alle binnen-groepen regressielijnen hetzelfde regressiegewicht, en als hij wel opgenomen wordt, dan hebben ze een verschillend regressiegewicht.

Het aantal vrijheidsgraden in de teller van de $F_{\text{Change}}$-statistiek ($q_{\text{Blok 2}|\text{Blok 1}}$) is gelijk aan het aantal parameters dat nodig is om de interactie tussen de kwantitatieve en de categorische predictor in de regressievergelijking op te nemen. Het aantal parameters dat hiervoor nodig is, is gelijk aan het aantal groepen min 1 (i.e., $K - 1$). Dit aantal is eenvoudig af te leiden als men de regressievergelijking zonder interactie (i.e., met één gemeenschappelijk regressiegewicht voor de binnen-groepen regressielijnen) vergelijkt met de regressievergelijking mét interactie (i.e., met $K$ groepsspecifieke regressiegewichten voor de binnen-groepen regressielijnen). Dus, het aantal extra parameters is gelijk aan het aantal groepen min 1.

Het aantal vrijheidsgraden in de noemer van de $F_{\text{Change}}$-statistiek is gelijk aan het aantal observaties ($n$) min het aantal parameters in de regressievergelijking met alle predictoren (dus, mét de interactie). Dit aantal parameters is gelijk aan $K + K = 2 \times K$. Immers, deze regressievergelijking bevat $K$ groepsspecifieke intercepten en $K$ groepsspecifieke regressiegewichten.

Met deze $F_{\text{Change}}$-statistiek wordt de nulhypothese getoetst dat, in de populatie, alle binnen-groepen regressielijnen hetzelfde regressiegewicht hebben. Deze nulhypothese wordt ook de hypothese van homogene of parallelle binnen-groepen regressielijnen genoemd.

De $F_{\text{Change}}$-statistiek geeft ons informatie over het al dan niet gelijk zijn van de binnen-groepen regressiegewichten. In het geval van een significant resultaat, geeft zij ons echter geen informatie over de manier waarop deze regressiegewichten van elkaar verschillen. Deze informatie haalt men uit een vergelijking van de groepsspecifieke regressiegewichten. Voor ons voorbeeld in Figuur 4.9 zullen we dan vaststellen dat het regressiegewicht voor de autochtone kinderen ($B_2$) duidelijk positief is (ongeveer 0.45) en het regressiegewicht voor de allochtone kinderen ($B_1$) ongeveer gelijk is aan 0.

De groepsspecifieke regressiegewichten zijn een interessante bron van informatie voor een onderzoeker die het verband tussen twee kwantitatieve variabelen (i.c., een kwantitatieve predictor en een criterium) inhoudelijk wil verklaren. Men mag hun belang echter niet overschatten. Voor een inhoudelijke verklaring van het verband tussen een kwantitatieve predictor $X$ en een criterium $Y$ moet men immers ook rekening houden met de verschillen tussen de groepsgemiddelden op de kwantitatieve predictor (i.e., de $\overline{X}_k$'s) en het criterium (i.e., de $\overline{Y}_k$'s). Voor ons voorbeeld in Figuur 4.9 blijkt dat autochtone kinderen een veel betere foneemdiscriminatie hebben dan de allochtone kinderen (i.e., $\overline{X}_2 > \overline{X}_1$) en ook dat hun technische leesvaardigheid veel beter is (i.e., $\overline{Y}_2 > \overline{Y}_1$). Dit doet vermoeden dat foneemdiscriminatie een belangrijke rol speelt bij het leren lezen. Dit vermoeden wordt bevestigd door de stijgende regressielijn binnen de groep autochtone kinderen, maar het wordt tegengesproken door de vlakke regressielijn binnen de groep allochtone kinderen. Het is de taak van de onderzoeker om een verklaring te bieden voor dit patroon in de gegevens. Deze verklaring moet aannemelijk maken dat de tegenspraak slechts schijnbaar is. Een mogelijke verklaring is gebaseerd op de gedachte dat verschillen in aanleg (hier, foneemdiscriminatie) pas goed zichtbaar worden in een omgeving die voor iedereen optimaal of bijna optimaal is.

### 4.12.6 Een Overzicht van de Verschillende Types F-toetsen binnen het Algemene Lineaire Model

Binnen het algemene lineaire model kan men nulhypotheses toetsen die betrekking hebben op kwantitatieve predictoren, categorische predictoren, en op beide. We vertrekken hier van de situatie met één kwantitatieve en één categorische predictor, en behandelen de hypotheses die in termen van deze predictoren geformuleerd kunnen worden. Voor de situatie met meerdere kwantitatieve en categorische predictoren kan men meer complexe nulhypotheses formuleren. De wijze waarop men ze toetst is echter niet wezenlijk anders dan in het eenvoudige geval met één kwantitatieve en één categorische predictor.

#### Eén Kwantitatieve en Eén Categorische Predictor

We willen de volgende nulhypotheses toetsen:

a. In de populatie is het regressiegewicht van de kwantitatieve predictor gelijk aan nul.

b. In de populatie zijn er geen verschillen tussen de niveaus van de categorische predictor.

c. In de populatie zijn de intercepten van de binnen-groepen regressielijnen gelijk.

d. In de populatie zijn de regressiegewichten van de binnen-groepen regressielijnen gelijk aan nul.

e. In de populatie zijn de regressiegewichten van de binnen-groepen regressielijnen gelijk.

Van deze vijf nulhypotheses hebben er vier betrekking op regressielijnen: a, c, d en e. In nulhypothese b, daarentegen, is er alleen sprake van een categorische predictor, en daarom hebben we hier niet te maken met regressielijnen; er is pas een regressielijn als er een kwantitatieve predictor is. Nulhypothese b wordt getoetst in de variantieanalyse

met één factor (categorische predictor). De zinssnede "geen verschillen tussen de niveaus van de categorische predictor" betekent "geen verschillen tussen de verwachte waarden binnen elk van de niveaus van de factor", de nulhypothese van de variantieanalyse.

Om elk van deze vijf nulhypotheses te toetsen, berekenen we telkens een $R^2_{\text{Change}}$ die correspondeert met de nulhypothese die we willen toetsen. Deze $R^2_{\text{Change}}$ is het verschil tussen de proporties verklaarde variantie op basis van twee regressievergelijkingen. In totaal, hebben we voor het toetsen van deze vijf nulhypotheses ook vijf regressievergelijkingen nodig:

1. De regressievergelijking zonder predictoren:

$$\widehat{Y} = A \quad .$$

2. De regressievergelijking met één kwantitatieve predictor:

$$\widehat{Y} = A + BX \quad .$$

3. De regressievergelijking met één categorische predictor:

$$\widehat{Y}_k = A_k \quad .$$

4. De regressievergelijking met één kwantitatieve en één categorische predictor:

$$\widehat{Y}_k = A_k + BX_k \quad .$$

5. De regressievergelijking met één kwantitatieve en één categorische predictor en groepsspecifieke regressiegewichten:

$$\widehat{Y}_k = A_k + B_k X_k \quad .$$

Hierbij moet opgemerkt worden dat de regressievergelijking met één categorische predictor, strikt genomen, geen regressievergelijking is omdat er hier geen sprake is van een kwantitatieve predictor. Het symbool $A_k$ in de derde regressievergelijking is daarom ook geen intercept van een regressielijn; het is het gemiddelde op het criterium binnen het k-de niveau van de categorische predictor (de k-de groep).

Alle toetsen gebeuren met de $F_{\text{Change}}$-statistiek, die we hier voor het gemak herhalen:

$$\boxed{F_{\text{Change}} = \frac{R^2_{\text{Change}}/q_{\text{Blok 2}|\text{Blok 1}}}{(1 - R^2_{\text{Blok 1+2}})/(n - q_{\text{Blok 1+2}})} \quad .}$$

In Tabel 4.37 beschrijven we vijf $F_{\text{Change}}$-statistieken in termen van de volgende kenmerken:

- De nulhypothese die ermee getoetst wordt.
- De regressievergelijkingen die betrokken zijn in de berekening van $R^2_{\text{Change}}$ en $R^2_{\text{Blok 1+2}}$.

| | Nulhypothese | Regressie-vergelijkingen | Blok 1 | Blok 2 | $q_{\text{Blok 2}\|\text{Blok 1}}$ | $q_{\text{Blok 1+2}}$ |
|---|---|---|---|---|---|---|
| a | Het regressie-gewicht van de kwantitatieve predictor is gelijk aan nul. | 1 vs. 2 | niets | kwanti-tatief | 1 | 2 |
| b | Er zijn geen ver-schillen tussen de niveaus van de categorische predictor. | 1 vs. 3 | niets | catego-risch | $K-1$ | $K$ |
| c | De intercepten van de binnen-groepen regres-sielijnen zijn ge-lijk. | 2 vs. 4 | kwanti-tatief | catego-risch | $K-1$ | $K+1$ |
| d | De regressiege-wichten van de binnen-groepen regressielijnen zijn gelijk aan nul. | 3 vs. 4 | catego-risch | kwanti-tatief | 1 | $K+1$ |
| e | De regressiege-wichten van de binnen-groepen regressielijnen zijn gelijk. | 4 vs. 5 | kwanti-tatief + catego-risch | interac-tie | $K-1$ | $2K$ |

Tabel 4.37: De vijf nulhypotheses die men kan toetsen in een onderzoek met één kwantitatieve en één categorische predictor met aanduiding van (1) de regressievergelijkingen die betrokken zijn in de berekening van $R^2_{\text{Change}}$ en $R^2_{\text{Blok 1+2}}$, (2) de predictoren in Blok 1 en Blok 2, en (3) de aantallen parameters die nodig zijn om de $F_{\text{Change}}$-statistiek te berekenen ($q_{\text{Blok 2}|\text{Blok 1}}$ en $q_{\text{Blok 1+2}}$).

- De predictoren in Blok 1 en Blok 2.
- Het aantal extra parameters dat men nodig heeft om de predictor in Blok 2 in de regressievergelijking op te nemen ($q_{\text{Blok 2}|\text{Blok 1}}$).
- Het aantal parameters in de regressievergelijking met de predictoren in Blok 1 én Blok 2 ($q_{\text{Blok 1+2}}$).

De toetsing van nulhypotheses in c, d, en e kwam reeds eerder aan bod in deze paragraaf. De nulhypothese in a is dezelfde als degene die getoetst wordt met de T-statistiek die in paragraaf 4.8 beschreven werd. De nulhypothese in b werd niet behandeld in deze paragraaf. Het is de nulhypothese die getoetst wordt in de variantieanalyse met één factor.

### Twee Kwantitatieve en Eén Categorische Predictor

Tabel 4.37 is voor een onderzoek met één kwantitatieve en één categorische predictor. Het is echter niet moeilijk om deze tabel uit te breiden naar onderzoeken met meerdere kwantitatieve en categorische predictoren. Bijvoorbeeld, in een onderzoek met twee kwantitatieve ($X_1$ en $X_2$) en één categorische predictor kan men de volgende nulhypotheses toetsen:

A. In de populatie is het regressiegewicht van de kwantitatieve predictor $X_1$ (of $X_2$) gelijk aan nul.

   Deze nulhypothese wordt getoetst met kwantitatieve predictor $X_2$ in Blok 1 en kwantitatieve predictor $X_1$ in Blok 2 (of omgekeerd, $X_1$ in Blok 1 en $X_2$ in Blok 2).

B. In de populatie zijn er geen verschillen tussen de niveaus van de categorische predictor.

   Deze nulhypothese wordt getoetst zoals als de tweede rij van Tabel 4.37 wordt aangegeven.

C. In de populatie zijn de intercepten van de binnen-groepen regressie*vlakken* gelijk.

   In deze nulhypothese wordt gesproken van regressie*vlakken* in plaats van regressie*lijnen* omdat de regressievergelijking $\widehat{Y}_k = A_k + B_1X_{1k} + B_2X_{2k}$ een vlak beschrijft in een driedimensionele ruimte. De drie assen van deze ruimte zijn de criteriumvariabele $Y$ en de twee kwantitatieve predictoren $X_1$ en $X_2$. De index k geeft aan dat elk groep zijn eigen regressievlak heeft. Het regressievlak van de k-de groep snijdt de $Y$-as in het punt $A_k$. Dit punt wordt het intercept van het k-de regressievlak genoemd.

   Deze nulhypothese (dat de intercepten van de binnen-groepen regressievlakken gelijk zijn) wordt getoetst met de categorische predictor in Blok 2 en de kwantitatieve predictoren $X_1$ en $X_2$ in Blok 1.

D. In de populatie zijn de regressiegewichten van de binnen-groepen regressievlakken gelijk aan nul.

Hier moeten we een onderscheid maken tussen de $X_1$- en de $X_2$-richting van de regressievlakken. De nulhypothese gaat over één van deze twee richtingen, $X_1$ of $X_2$, die men zelf mag kiezen. De nulhypothese is dat, in de gekozen richting, het regressievlak horizontaal loopt.

Deze nulhypothese wordt getoetst met kwantitatieve predictor van de gekozen richting in Blok 2 en de andere kwantitatieve predictor in Blok 1, samen met de categorische predictor.

E. In de populatie zijn de regressiegewichten van de binnen-groepen regressievlakken gelijk.

Weer moeten we hier een onderscheid maken tussen de $X_1$- en de $X_2$-richting van de regressievlakken. De nulhypothese gaat over één van deze twee richtingen. De nulhypothese is dat, in de gekozen richting, de hellingsgraad van het regressievlak dezelfde is voor alle groepen.

Deze nulhypothese wordt getoetst met, in Blok 2, de interactie tussen de categorische predictor en de kwantitatieve predictor van de gekozen richting, en in Blok 1, de categorische predictor en de twee kwantitatieve predictoren.

Het aantal extra parameters dat men nodig heeft om een predictor in Blok 2 in de regressievergelijking op te nemen ($q_{\text{Blok 2}|\text{Blok 1}}$) hangt niet af van het aantal extra kwantitatieve predictoren in Blok 1. Voor het bepalen van $q_{\text{Blok 2}|\text{Blok 1}}$ kan men dus nog steeds gebruik maken van Tabel 4.37. Daarentegen, voor het bepalen van $q_{\text{Blok 1+2}}$, het totaal aantal parameters in de regressievergelijking met de predictoren in Blok 1 én Blok 2, moet men wel rekening houden met het aantal kwantitatieve predictoren in Blok 1. Dit aantal wordt vermeerderd met 1 voor elke extra kwantitatieve predictor die in Blok 1 wordt opgenomen.

### Eén Kwantitatieve en Twee Categorische Predictoren

De uitbreiding voor een onderzoek met twee categorische en één kwantitatieve predictor is analoog aan de beschreven uitbreiding voor een onderzoek met één categorische en twee kwantitatieve predictoren. Voor deze uitbreiding moeten we een onderscheid maken tussen gewogen en ongewogen verwachte waarden. Zoals beschreven werd in Hoofdstuk 3 (*Variantie- en Covariantieanalyse*, corresponderen gewogen verwachte waarden met een analyse waarin níet uitgezuiverd wordt voor een andere categorische predictor (geen categorische predictor in Blok 1), en ongewogen verwachte waarden met een analyse waarin hier wel voor uitgezuiverd wordt (wel een categorische predictor in Blok 1). Het verschil tussen de analyses in Hoofdstuk 3 en de analyses waar we hier over hebben, is dat we hier ook met een kwantitatieve predictor te maken hebben. Voor deze kwantitatieve predictor kan uitgezuiverd worden (door hem in Blok 1 te stoppen samen met een categorische predictor) of hij kan zelf uitgezuiverd worden (door hem in Blok 2 te stoppen). In de analyses met twee categorische en één kwantitatieve predictor moet men rekening houden met het feit dat, voor elke extra categorische predictor die in Blok 1 wordt opgenomen, $q_{\text{Blok 1+2}}$ vermeerderd wordt met $L - 1$ (het aantal niveaus van deze extra categorische predictor min 1).

### 4.12.7 Dummycodering

Een regressieanalyse met een categorische predictor werd tot nu toe voorgesteld met $K$ regressievergelijkingen, één voor elke groep. Deze $K$ regressievergelijkingen werden soms, een beetje slordig, ook aangeduid als dé regressievergelijking voor de situatie met een categorische en een kwantitatieve predictor. Dummycodering is een manier om een regressieanalyse met categorische predictoren uit te drukken in één in plaats van $K$ regressievergelijkingen.

Voordat we uitleggen wat dummycodering is, is het goed om eerst de enkelvoudige regressievergelijking op een nieuwe manier voor te stellen. In een tweede stap, wordt deze regressievergelijking uitgebreid met één categorische predictor. Pas in deze tweede stap hebben we dummycodering nodig. Als men dummycodering combineert met een zogenaamd globaal intercept, dan doet zich een onbepaaldheidsprobleem voor. In wat volgt wordt precies uitgelegd wat dit onbepaaldheidsprobleem inhoudt en ook hoe het opgelost kan worden. Tenslotte, in een derde stap, wordt de regressievergelijking uitgebreid met groepsspecifieke regressiegewichten. Ook deze uitbreiding is voor stellen met behulp van dummycodering. Als men deze dummycodering combineert met een zogenaamd globaal regressiegewicht, dan doet zich hier opnieuw een onbepaaldheidsprobleem voor. Ook dit onbepaaldheidsprobleem is op te lossen.

#### Enkelvoudige Regressieanalyse

We stellen nu de enkelvoudige regressievergelijking op een nieuwe manier voor. We doen dit aan de hand van het getallenvoorbeeld in Tabel 4.38. Het nieuwe element in deze tabel is de variabele die voor elke proefpersoon de waarde 1 heeft, en daarom aangeduid wordt met **1** (een vetgedrukte 1). Deze variabele correspondeert met het intercept $A$ en wordt daarom de *interceptvariabele* genoemd. De voorspelde criteriumscore $\widehat{Y}$ wordt berekend door (1) de regressiecoëfficiënten ($A$ en $B$) te vermenigvuldigen met hun corresponderende variabelen ($A$ met de interceptvariabele **1** en $B$ met de kwantitatieve predictor $X$) en (2) de resulterende variabelen bij elkaar op te tellen.

| pp | **1** | $X$ | $\widehat{Y}$ | $Y$ |
|---|---|---|---|---|
| 1 | 1 | 1 | 2.01 | 3 |
| 2 | 1 | 5 | 3.65 | 4 |
| 3 | 1 | 9 | 5.29 | 5 |
| 4 | 1 | 8 | 4.88 | 2 |
| 5 | 1 | 5 | 3.65 | 2 |
| 6 | 1 | 2 | 2.42 | 2 |
| 7 | 1 | 9 | 5.29 | 8 |
| 8 | 1 | 5 | 3.65 | 5 |
| 9 | 1 | 2 | 2.42 | 2 |
| | 1.60 | 0.41 | | |
| | $A$ | $B$ | | |

Tabel 4.38: Getallenvoorbeeld van een enkelvoudige regressieanalyse waarin gebruikgemaakt wordt van een interceptvariabele.

### Regressieanalyse met een Categorische Predictor

Als men een regressieanalyse met een categorische predictor op een gelijkaardige manier wil voorstellen als de enkelvoudige regressieanalyse met alleen een kwantitatieve predictor (zie Tabel 4.38), dan moet men de tabel op zo'n manier invullen dat elke groep zijn eigen intercept krijgt. Dit kan door gebruik te maken van groepsspecifieke interceptvariabelen. Dit gebeurt in Tabel 4.39, gebruikmakend van hetzelfde getallenvoorbeeld als in Tabel 4.38 maar nu met als extra informatie de categorische variabele, aangeduid als GROEP. De groepsspecifieke interceptvariabelen worden aangeduid met $D_1$, $D_2$ en $D_3$. De voorspelde criteriumscore $\widehat{Y}$ wordt op dezelfde manier berekend als in Tabel 4.38, namelijk door (1) de regressiecoëfficiënten ($A_1$, $A_2$, $A_3$ en $B$) te vermenigvuldigen met hun corresponderende variabelen en (2) de resulterende variabelen bij elkaar op te tellen.

| pp | GROEP | $D_1$ | $D_2$ | $D_3$ | $X$ | $\widehat{Y}$ | $Y$ |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 0 | 1 | 2.45 | 3 |
| 2 | 1 | 1 | 0 | 0 | 5 | 4.01 | 4 |
| 3 | 1 | 1 | 0 | 0 | 9 | 5.57 | 5 |
| 4 | 2 | 0 | 1 | 0 | 8 | 3.18 | 2 |
| 5 | 2 | 0 | 1 | 0 | 5 | 2.01 | 2 |
| 6 | 2 | 0 | 1 | 0 | 2 | 0.84 | 2 |
| 7 | 3 | 0 | 0 | 1 | 9 | 6.44 | 8 |
| 8 | 3 | 0 | 0 | 1 | 5 | 4.88 | 5 |
| 9 | 3 | 0 | 0 | 1 | 2 | 3.71 | 2 |
| | | 2.06 | 0.06 | 2.93 | 0.39 | | |
| | | $A_1$ | $A_2$ | $A_3$ | $B$ | | |

Tabel 4.39: Getallenvoorbeeld van een regressieanalyse met een categorische predictor waarin gebruikgemaakt wordt van groepsspecifieke interceptvariabelen.

Dummycodering is het voorstellen van een categorische variabele als $K$ groepsspecifieke interceptvariabelen, zoals gebeurt in Tabel 4.39. Door gebruik te maken van deze groepsspecifieke interceptvariabelen, kan men de $K$ groepsspecifieke regressievergelijkingen vervangen door één enkele regressievergelijking. Voor een categorische predictor met drie niveaus, overeenstemmend met het getallenvoorbeeld in Tabel 4.39, kan deze ene regressievergelijking als volgt geschreven worden:

$$\widehat{Y} = A_1D_1 + A_2D_2 + A_3D_3 + BX \quad .$$

Het resultaat van de som $A_1D_1+A_2D_2+A_3D_3$ is dat voor elke proefpersoon het intercept van zijn of haar groep geselecteerd wordt. Immers, elke proefpersoon heeft slechts één interceptvariabele waarop hij of zij de waarde 1 heeft, namelijk die interceptvariabele die codeert voor het lidmaatschap van de groep waartoe de proefpersoon behoort.

Sommige computerprogramma's, zoals SPSS, voegen standaard het globale intercept $A$ toe aan de regressievergelijking. Dit creëert een probleem omdat, indien men daarnaast ook de groepsspecifieke intercepten in de regressievergelijking opneemt, de

intercepten dan onbepaald zijn. Het gaat hier over de volgende regressievergelijking:

$$\widehat{Y} = A + A_1 D_1 + A_2 D_2 + A_3 D_3 + BX \quad .$$

Nu, $A$, $A_1$, $A_2$ en $A_3$ zijn onbepaald omdat $\widehat{Y}$ (de variabele die we zó kiezen dat aan het kleinste-kwadratencriterium voldaan is) niet verandert als we bij $A$, bijvoorbeeld, 10 punten optellen en tegelijkertijd van $A_1$, $A_2$ en $A_3$ 10 punten aftrekken. In het Engels, noemt men dit een *trade-off* tussen het globale intercept $A$ en de groepsspecifieke intercepten $A_1$, $A_2$ en $A_3$. Daarom zijn de waarden voor $A$, $A_1$, $A_2$ en $A_3$ die aan het kleinste-kwadratencriterium voldoen niet uniek of, met andere woorden, onbepaald.

Als oplossing voor dit onbepaaldheidsprobleem, kiezen de meeste computerprogramma's ervoor om niet langer te werken met de set van coëfficiënten $A$, $A_1$, $A_2$ en $A_3$. Deze coëfficiënten worden vervangen door de coëfficiënten $\mathcal{A}$ (spreek uit "ronde A"), $\mathcal{A}_1$, $\mathcal{A}_2$ en $\mathcal{A}_3$, waarop de restrictie ligt dat $\mathcal{A}_3$ gelijk is aan 0. Deze restrictie op één van de coëfficiënten legt geen restrictie op $\widehat{Y}$, zoals geïllustreerd wordt in het getallenvoorbeeld in Tabel 4.40. De coëfficiënten die horen bij de variabelen $\mathbf{1}$, $D_1$, $D_2$ en $D_3$ zijn de coëfficiënten $\mathcal{A}$, $\mathcal{A}_1$, $\mathcal{A}_2$ en $\mathcal{A}_3$. Men kan deze coëfficiënten berekenen uit de groepsspecifieke intercepten $A_1$, $A_2$ en $A_3$ in Tabel 4.39. Uit een vergelijking van Tabel 4.39 en Tabel 4.40 blijkt dat het globale intercept $\mathcal{A}$ in 4.40 gelijk is aan het groepsspecifieke intercept $A_3$ in Tabel 4.39. Uit deze vergelijking van tabellen blijkt ook dat het groepsspecifieke intercept $A_1$ in Tabel 4.39 gelijk is aan $\mathcal{A} + \mathcal{A}_1$ in Tabel 4.40. Uit deze gelijkheid kan $\mathcal{A}_1$ als volgt berekend worden:

$$\begin{aligned} A_1 =& \mathcal{A} + \mathcal{A}_1 \\ \mathcal{A}_1 =& A_1 - \mathcal{A} \\ \mathcal{A}_1 =& A_1 - A_3 \quad . \end{aligned}$$

We noemen $\mathcal{A}_1$ een *contrast* omdat het gelijk is aan het verschil tussen twee andere coëfficiënten, de twee groepsspecifieke intercepten $A_1$ en $A_3$. Op dezelfde manier als voor $\mathcal{A}_1$, kan men laten zien dat $\mathcal{A}_2$ berekend kan worden als het verschil tussen $A_2$ en $A_3$:

$$\mathcal{A}_2 = A_2 - A_3$$

Dus, $\mathcal{A}_2$ is ook een contrast.

In de praktijk, heeft men meestal de omgekeerde formules nodig, de formules waarmee men de groepsspecifieke intercepten $A_1$, $A_2$ en $A_3$ kan berekenen uit $\mathcal{A}$, $\mathcal{A}_1$ en $\mathcal{A}_2$. Dit komt omdat deze laatste coëfficiënten gegeven worden in de uitvoer van de meeste computerprogramma's, en men voor de inhoudelijke interpretatie meestal de groepsspecifieke intercepten nodig heeft. Men kan deze groepsspecifieke intercepten als volgt berekenen:

$$\begin{aligned} A_1 =& \mathcal{A} + \mathcal{A}_1 \\ A_2 =& \mathcal{A} + \mathcal{A}_2 \\ A_3 =& \mathcal{A} \quad . \end{aligned}$$

In de uitvoer van sommige computerprogramma's worden T-statistieken gegeven voor alle coëfficiënten waarop geen restrictie ligt. Dit zijn $\mathcal{A}$, $\mathcal{A}_1$, $\mathcal{A}_2$ en $B$. Met deze

| pp | GROEP | **1** | $D_1$ | $D_2$ | $D_3$ | $X$ | $\widehat{Y}$ | $Y$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 0 | 0 | 1 | 2.45 | 3 |
| 2 | 1 | 1 | 1 | 0 | 0 | 5 | 4.01 | 4 |
| 3 | 1 | 1 | 1 | 0 | 0 | 9 | 5.57 | 5 |
| 4 | 2 | 1 | 0 | 1 | 0 | 8 | 3.18 | 2 |
| 5 | 2 | 1 | 0 | 1 | 0 | 5 | 2.01 | 2 |
| 6 | 2 | 1 | 0 | 1 | 0 | 2 | 0.84 | 2 |
| 7 | 3 | 1 | 0 | 0 | 1 | 9 | 6.44 | 8 |
| 8 | 3 | 1 | 0 | 0 | 1 | 5 | 4.88 | 5 |
| 9 | 3 | 1 | 0 | 0 | 1 | 2 | 3.71 | 2 |
| | | 2.93 | -0.87 | -2.87 | 0 | 0.39 | | |
| | | $\mathcal{A}$ | $\mathcal{A}_1$ | $\mathcal{A}_2$ | $\mathcal{A}_3$ | $B$ | | |

Tabel 4.40: Getallenvoorbeeld van een regressieanalyse met een categorische predictor waarin gebruik gemaakt wordt van de globale interceptvariabele, de groepsspecifieke interceptvariabelen, en waar een restrictie is opgelegd aan de coëfficiënt $\mathcal{A}_3$.

T-statistieken toetst men de nulhypothese dat, in de populatie, deze coëfficiënten gelijk zijn aan 0. Merk op dat, met de T-statistiek voor $\mathcal{A}_1$, de nulhypothese getoetst wordt dat, in de populatie, de groepsspecifieke intercepten van Groep 1 en Groep 3 gelijk zijn. En met de T-statistiek voor $\mathcal{A}_2$ wordt de nulhypothese getoetst dat, in de populatie, de groepsspecifieke intercepten van Groep 2 en Groep 3 gelijk zijn.

De restrictie dat de $\mathcal{A}_k$-coëfficiënt van de laatste groep gelijk is aan 0, is een arbitraire restrictie. Men kan bijvoorbeeld ook de restrictie opleggen dat de $\mathcal{A}_k$-coëfficiënt van de éérste groep gelijk is aan 0, of de restrictie dat de sóm van de $\mathcal{A}_k$-coëfficiënten gelijk is aan 0, een restrictie die vaak gebruikt wordt in de variantieanalyse. Met een andere restrictie krijgt men weliswaar andere coëfficiënten, maar de voorspelde criteriumscore $\widehat{Y}$ verandert niet. En omdat $R^2$ (o.a., $R^2_{\text{Blok 1}}$ en $R^2_{\text{Blok 1+2}}$) gelijk is aan het kwadraat van de correlatie tussen de voorspelde en de geobserveerde criteriumscore, verandert ook deze $R^2$ niet. En dus veranderen ook de $F_{\text{Change}}$-statistieken niet, want zij zijn een functie van $R^2_{\text{Blok 1}}$ en $R^2_{\text{Blok 1+2}}$.

### Regressieanalyse met Groepsspecifieke Regressiegewichten

We beschrijven nu hoe dummycodering gebruikt wordt binnen de regressieanalyse met groepsspecifieke regressiegewichten. De regressievergelijking met groepsspecifieke regressiegewichten wordt eerst op een analoge manier voorgesteld als de de regressievergelijking met groepsspecifieke intercepten in Tabel 4.39. Dit gebeurt in Tabel 4.41, en we maken hierbij gebruik van groepsspecifieke predictoren. De groepsspecifieke predictoren worden aangeduid met $(XD)_1$, $(XD)_2$ en $(XD)_3$. De voorspelde criteriumscore wordt berekend door (1) de regressiecoëfficiënten ($A_1$, $A_2$, $A_3$, $B_1$, $B_2$ en $B_3$) te vermenigvuldigen met hun corresponderende variabelen en (2) de resulterende variabelen bij elkaar op te tellen. Terzijde, merk op dat in Groep 1 en Groep 2, de geobserveerde en de voorspelde criteriumscores identiek zijn. Dit komt omdat, binnen deze twee groepen,

de geobserveerde criteriumscores een lineaire functie zijn van hun bijbehorende predictorscores.

| pp | GROEP | $D_1$ | $D_2$ | $D_3$ | $(XD)_1$ | $(XD)_2$ | $(XD)_3$ | $\widehat{Y}$ | $Y$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 3 | 3 |
| 2 | 1 | 1 | 0 | 0 | 5 | 0 | 0 | 4 | 4 |
| 3 | 1 | 1 | 0 | 0 | 9 | 0 | 0 | 5 | 5 |
| 4 | 2 | 0 | 1 | 0 | 0 | 8 | 0 | 2 | 2 |
| 5 | 2 | 0 | 1 | 0 | 0 | 5 | 0 | 2 | 2 |
| 6 | 2 | 0 | 1 | 0 | 0 | 2 | 0 | 2 | 2 |
| 7 | 3 | 0 | 0 | 1 | 0 | 0 | 9 | 8.11 | 8 |
| 8 | 3 | 0 | 0 | 1 | 0 | 0 | 5 | 4.71 | 5 |
| 9 | 3 | 0 | 0 | 1 | 0 | 0 | 2 | 2.16 | 2 |
| | | 2.75 | 2 | 0.46 | 0.25 | 0 | 0.85 | | |
| | | $A_1$ | $A_2$ | $A_3$ | $B_1$ | $B_2$ | $B_3$ | | |

Tabel 4.41: Getallenvoorbeeld van een regressieanalyse met groepsspecifieke regressiegewichten waarin gebruikgemaakt wordt van groepsspecifieke interceptvariabelen en groepsspecifieke predictoren.

Door gebruik te maken van de groepsspecifieke interceptvariabelen en predictoren, kan men de $K$ groepsspecifieke regressievergelijkingen vervangen door één enkele regressievergelijking. Voor een categorische predictor met drie niveaus kan deze ene regressievergelijking als volgt geschreven worden:

$$\widehat{Y} = A_1D_1 + A_2D_2 + A_3D_3 + B_1(XD)_1 + B_2(XD)_2 + B_3(XD)_3 \quad .$$

Het resultaat van de som in het rechterlid van de vergelijking is dat voor elke proefpersoon het intercept én het regressiegewicht van zijn of haar groep geselecteerd wordt.

Sommige computerprogramma's voegen standaard, behalve het globale intercept $A$, ook de term $BX$ (het globale regressiegewicht maal de predictor $X$) toe aan de regressievergelijking. Behalve het reeds behandelde probleem van de onbepaalde intercepten, wordt nu ook het probleem van onbepaalde regressiegewichten gecreëerd indien men, naast $A$ en $BX$, ook de groepsspecifieke termen $A_1$, $B_1(XD)_1$, $A_2$, $B_2(XD)_2$, enzovoort, in de regressievergelijking opneemt. Immers, er bestaat een trade-off tussen het globale regressiegewicht $B$ en de groepsspecifieke regressiegewichten $B_1$, $B_2$ en $B_3$; als men bij de ene, bijvoorbeeld, 10 optelt en tegelijkertijd van de anderen 10 aftrekt, dan verandert $\widehat{Y}$ niet. Daarom zijn de waarden voor $B$, $B_1$, $B_2$ en $B_3$ die aan het kleinste-kwadratencriterium voldoen niet uniek.

Als oplossing voor dit onbepaaldheidsprobleem, kiezen de meeste computerprogramma's ervoor om de set van regressiegewichten $B$, $B_1$, $B_2$ en $B_3$ te vervangen door de gewichten $\mathcal{B}$, $\mathcal{B}_1$, $\mathcal{B}_2$ en $\mathcal{B}_3$, waarop de restrictie ligt dat $\mathcal{B}_3$ gelijk is aan 0. Deze restrictie op één van de gewichten legt geen restrictie op $\widehat{Y}$, zoals geïllustreerd wordt in het getallenvoorbeeld in Tabel 4.42. De regressiegewichten die horen bij de variabelen $X$, $(XD)_1$, $(XD)_2$ en $(XD)_3$ zijn de gewichten $\mathcal{B}$, $\mathcal{B}_1$, $\mathcal{B}_2$ en $\mathcal{B}_3$. Men kan deze gewichten berekenen uit de groepsspecifieke regressiegewichten $B_1$, $B_2$ en $B_3$ in Tabel 4.41. Uit

een vergelijking van Tabel 4.41 en Tabel 4.42 blijkt dat het globale regressiegewicht $\mathcal{B}$ in Tabel 4.42 gelijk is aan het groepsspecifieke regressiegewicht $B_3$ in Tabel 4.41. Uit deze vergelijking van tabellen blijkt ook dat het groepsspecifieke regressiegewicht $B_1$ in Tabel 4.41 gelijk is aan $\mathcal{B} + \mathcal{B}_1$ in Tabel 4.42. Uit deze gelijkheid kan $\mathcal{B}_1$ als volgt berekend worden:

$$\begin{aligned} B_1 =& \mathcal{B} + \mathcal{B}_1 \\ \mathcal{B}_1 =& B_1 - \mathcal{B} \\ \mathcal{B}_1 =& B_1 - B_3 \quad . \end{aligned}$$

Dus, $\mathcal{B}_1$ is een contrast. Op dezelfde manier als voor $\mathcal{B}_1$, kan men laten zien dat $\mathcal{B}_2$ berekend kan worden als het verschil tussen $B_2$ en $B_3$:

$$\mathcal{B}_2 = B_2 - B_3$$

Dus, $\mathcal{B}_2$ is ook een contrast.

In de praktijk, heeft men meestal de omgekeerde formules nodig, de formules waarmee men de groepsspecifieke regressiegewichten $B_1$, $B_2$ en $B_3$ kan berekenen uit $\mathcal{B}$, $\mathcal{B}_1$ en $\mathcal{B}_2$. Immers, deze laatste coëfficiënten worden gegeven in de uitvoer van de meeste computerprogramma's. Voor de inhoudelijke interpretatie heeft men meestal de groepsspecifieke regressiegewichten nodig. Men kan deze groepsspecifieke regressiegewichten als volgt berekenen:

$$\begin{aligned} B_1 =& \mathcal{B} + \mathcal{B}_1 \\ B_2 =& \mathcal{B} + \mathcal{B}_2 \\ B_3 =& \mathcal{B} \quad . \end{aligned}$$

In de uitvoer van sommige computerprogramma's worden T-statistieken gegeven voor alle coëfficiënten waarop geen restrictie ligt (hier, $\mathcal{A}$, $\mathcal{A}_1$, $\mathcal{A}_2$, $\mathcal{B}$, $\mathcal{B}_1$ en $\mathcal{B}_2$), waarmee men de nulhypothese toetst dat, in de populatie, deze coëfficiënten gelijk zijn aan 0. Merk op dat, met de T-statistieken voor $\mathcal{B}_1$ en $\mathcal{B}_2$, nulhypotheses getoetst worden die betrekking hebben op *verschillen* tussen groepsspecifieke regressiegewichten.

### 4.12.8 Uitvoeren van een Regressieanalyse met de SPSS-procedure GLM-Univariate

In Hoofdstuk 3, waarin de variantie- en de covariantieanalyse behandeld werd, gaven we reeds een beschrijving van GLM-Univariate. Deze beschrijving zullen we hier niet herhalen. GLM-Univariate is een procedure om de data te analyseren met het algemene lineaire model. Dit betekent dat men deze procedure zowel kan gebruiken voor regressie-, variantie-, en covariantieanalyse. Alle kwantitatieve predictoren moet men invullen onder *Covariate(s)* in het dialoogvenster van GLM-Univariate en alle categorische predictoren onder *Fixed Factor(s)*.

In de uitvoer van GLM-Univariate wordt altijd een ANOVA-tabel afgedrukt met daarin de kwadratensommen en de F-statistieken. Dit werd reeds beschreven in Hoofdstuk 3. Hier richten we ons op de regressiecoëfficiënten die in de uitvoer afgedrukt

| pp | Groep | **1** | $D_1$ | $D_2$ | $D_3$ | $X$ | $(XD)_1$ | $(XD)_2$ | $(XD)_3$ | $\widehat{Y}$ | $Y$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 3 | 3 |
| 2 | 1 | 1 | 1 | 0 | 0 | 5 | 5 | 0 | 0 | 4 | 4 |
| 3 | 1 | 1 | 1 | 0 | 0 | 9 | 9 | 0 | 0 | 5 | 5 |
| 4 | 2 | 1 | 0 | 1 | 0 | 8 | 0 | 8 | 0 | 2 | 2 |
| 5 | 2 | 1 | 0 | 1 | 0 | 5 | 0 | 5 | 0 | 2 | 2 |
| 6 | 2 | 1 | 0 | 1 | 0 | 2 | 0 | 2 | 0 | 2 | 2 |
| 7 | 3 | 1 | 0 | 0 | 1 | 9 | 0 | 0 | 9 | 8.11 | 8 |
| 8 | 3 | 1 | 0 | 0 | 1 | 5 | 0 | 0 | 5 | 4.71 | 5 |
| 9 | 3 | 1 | 0 | 0 | 1 | 2 | 0 | 0 | 2 | 2.16 | 2 |
| | | 0.46 | 2.29 | 1.54 | 0 | 0.85 | -0.60 | -0.85 | 0 | | |
| | | $\mathcal{A}$ | $\mathcal{A}_1$ | $\mathcal{A}_2$ | $\mathcal{A}_3$ | $\mathcal{B}$ | $\mathcal{B}_1$ | $\mathcal{B}_2$ | $\mathcal{B}_3$ | | |

Tabel 4.42: Getallenvoorbeeld van een regressieanalyse met groepsspecifieke regressiegewichten waarin gebruikgemaakt wordt van de globale interceptvariabele, de globale predictor $X$, en de groepsspecifieke interceptvariabelen en predictoren. Er is een restrictie opgelegd aan de coëfficiënten $\mathcal{A}_3$ en $\mathcal{B}_3$.

worden indien men in het dialoogvenster onder de knop *Options* het vakje *Parameter Estimates* afvinkt.

In Tabel 4.43 staat een deel van de uitvoer van GLM-Univariate voor het onderzoek dat beschreven wordt in Opgave 37. De data van dit onderzoek staan in het bestand "tevreden.sav". We bespreken hier een aantal analyses van deze data. Het is leerzaam deze analyses ook zelf te doen. In Tabel 4.43 staat de uitvoer van een enkelvoudige regressieanalyse met de kwantitatieve predictor TEVRVROE en criterium TEVRNU. Onderaan Tabel 4.43 staat de $R^2$ voor deze analyse. In de uitvoer van GLM-Univariate staat de $R^2$ onder de ANOVA-tabel en niet onder de tabel met regressiecoëfficiënten, zoals hier.

Dependent Variable: TEVRNU

| Parameter | B | Std. Error | t | Sig |
|---|---|---|---|---|
| Intercept | -.131 | .563 | -.234 | .817 |
| TEVRVROE | .798 | .073 | 10.875 | .000 |

$R^2 = 0.757$

Tabel 4.43: Tabel bij een enkelvoudige regressieanalyse voor het onderzoek in opgave 37.

In Tabel 4.44 staat een deel van de uitvoer van een regressieanalyse waarin, behalve van de kwantitatieve predictor TEVRVROE, ook de categorische predictor NRELAT is opgenomen. In het dialoogvenster van GLM-Univariate wordt deze categorische predictor ingevuld onder *Fixed Factor(s)*. In Tabel 4.44 kan men de groepsspecifieke intercepten niet zomaar aflezen. Dit komt omdat GLM-Univariate elk groepsspecifiek intercept $A_k$ vervangt door de som $\mathcal{A} + \mathcal{A}_k$, zoals uitgelegd in de paragraaf over dummycodering, en in de uitvoer alleen deze laatste coëfficiënten geeft. Concreet, $\mathcal{A}$ is de

coëfficiënt achter *Intercept* in Tabel 4.44, $\mathcal{A}_1$ is de coëfficiënt achter [NRELAT=1] en $\mathcal{A}_2$ is de coëfficiënt achter [NRELAT=2], waarop de restrictie ligt dat hij gelijk is aan 0. GLM-Univariate hanteert immers de standaardrestrictie dat de $\mathcal{A}_k$-coëfficiënt van de laatste groep gelijkgesteld wordt aan 0. Met deze standaardrestrictie is het groepsspecifieke intercept $A_k$ gelijk aan $\mathcal{A} + \mathcal{A}_k$. Hieruit volgt dat het groepsspecifieke intercept van de laatste groep gelijk is aan $\mathcal{A}$ (in ons voorbeeld, 0.033). De groepsspecifieke intercepten van de andere groepen vindt men door de $\mathcal{A}_k$-coëfficiënt op te tellen bij $\mathcal{A}$ (in ons voorbeeld, $A_1 = 0.033 + 0.247 = 0.280$).

Dependent Variable: TEVRNU

| Parameter | B | Std. Error | t | Sig |
|---|---|---|---|---|
| Intercept | .033 | .585 | .056 | .955 |
| TEVRVROE | .761 | .082 | 9.291 | .000 |
| [NRELAT=1] | .247 | .240 | 1.029 | .310 |
| [NRELAT=2] | 0 | | | |

$R^2 = 0.764$

Tabel 4.44: Tabel bij een regressieanalyse met één kwantitatieve en één categorische predictor voor het onderzoek in opgave 37.

In Tabel 4.45 staat een deel van de uitvoer van GLM-Univariate voor een analyse waarin alleen de categorische predictor NRELAT gebruikt wordt. Dus, onder *Covariate(s)* in het dialoogvenster van GLM-Univariate staat nu níets en onder het vakje *Fixed Factor(s)* staat NRELAT. Een dergelijke analyse wordt ook een één-factoriële variantieanalyse genoemd.

Dependent Variable: TEVRNU

| Parameter | B | Std. Error | t | Sig |
|---|---|---|---|---|
| Intercept | 5.286 | .267 | 19.773 | .000 |
| [NRELAT=1] | 1.241 | .388 | 3.199 | .000 |
| [NRELAT=2] | 0 | | | |

$R^2 = 0.212$

Tabel 4.45: Tabel bij een regressieanalyse met één categorische predictor voor het onderzoek in opgave 37.

In Tabel 4.46 staat een deel van de uitvoer van een regressieanalyse met groepsspecifieke regressiegewichten. Voor deze analyse moet men aangeven dat er een interactie is tussen de kwantitatieve en de categorische predictor. Dit doet men door onder de knop *Model* te kiezen voor *Custom* in plaats van *Full Factorial* en dan aan te geven welke predictoren en welke interacties in de regressievergelijking (i.e., het model) opgenomen moeten worden. Men kiest een predictor door zijn naam in het linkerlijstje te selecteren en dan op de knop onder *Build Term(s)* te drukken. Dan verschijnt zijn naam ook in het rechterlijstje onder *Custom*. Doe dit voor alle predictoren die men in de regressievergelijking op wil nemen. Om nu ook de interactie tussen een kwantitatieve en een categorische predictor in de regressievergelijking op te nemen moet men (1) ervoor zorgen dat het venstertje onder de knop *Build Term(s)* het woord *Interaction* bevat, (2)

de kwantitatieve én de categorische predictor selecteren en (3) op de knop onder *Build Term(s)* drukken. Terzijde, het is ook mogelijk om de interactie tussen twee kwantitatieve predictoren in de regressievergelijking op te nemen. In de volgende paragraaf beantwoorden we de vraag waarom dit interessant kan zijn en bekijken we vervolgens hoe men de resultaten van zo'n analyse moet interpreteren. Nu, de manier waarop men, binnen GLM-Univariate, de interactie tussen twee kwantitatieve predictoren aan de regressievergelijking toevoegt, is hetzelfde als de manier waarop men dit doet voor de interactie tussen een kwantitatieve en een categorische predictor: (1) zorg ervoor dat het venstertje onder de knop *Build Term(s)* het woord *Interaction* bevat, (2) selecteer de twee kwantitatieve predictoren en (3) druk op de knop onder *Build Term(s)*.

In Tabel 4.46 kan men de groepsspecifieke regressiegewichten niet zomaar aflezen. Dit komt omdat GLM-Univariate elk groepsspecifiek regressiegewicht $B_k$ vervangt door de som $\mathcal{B} + \mathcal{B}_k$, zoals uitgelegd in de paragraaf over dummycodering, en in de uitvoer alleen deze laatste coëfficiënten geeft. Concreet, $\mathcal{B}$ is de coëfficiënt achter TEVRVROE in Tabel 4.46, $\mathcal{B}_1$ is de coëfficiënt achter [NRELAT=1] $*$ TEVRVROE en $\mathcal{B}_2$ is de coëfficiënt achter [NRELAT=2] $*$ TEVRVROE, waarop de restrictie ligt dat hij gelijk is aan 0. Opnieuw is dit alleen het geval indien men niets verandert achter de knop *Contrasts* en dus de standaardrestrictie krijgt. Met deze standaardrestrictie is het groepsspecifieke regressiegewicht $B_k$ gelijk aan aan $\mathcal{B} + \mathcal{B}_k$. Hieruit volgt dat het groepsspecifieke regressiegewicht van de laatste groep gelijk is aan $\mathcal{B}$ (in ons voorbeeld, 0.786). De groepsspecifieke regressiegewichten van de andere groepen vind je door de $\mathcal{B}_k$-coëfficiënt op te tellen bij $\mathcal{B}$ (in ons voorbeeld, $B_1 = 0.782 + (-0.049) = 0.733$).

Dependent Variable: TEVRNU

| Parameter | B | Std. Error | t | Sig |
|---|---|---|---|---|
| Intercept | -.141 | .831 | -.169 | .866 |
| [NRELAT=1] | .620 | 1.275 | .486 | .630 |
| [NRELAT=2] | 0 | | | |
| TEVRVROE | .786 | .118 | 6.637 | .000 |
| [NRELAT=1] $*$ TEVRVROE | -.049 | .166 | -.298 | .768 |
| [NRELAT=2] $*$ TEVRVROE | 0 | | | |

$R^2 = 0.764$

Tabel 4.46: Tabel bij een regressieanalyse met groepsspecifieke regressiegewichten voor het onderzoek in opgave 37.

---

*Opgave 37* COMPUTEROPDRACHT

Een klinisch psychologe deed een onderzoek naar de tevredenheid van vrouwen met hun relatie. Zij deed dit onderzoek bij een groep van 40 vrouwen die allemaal ongeveer vijf jaar samenleefden met hun partner. De criteriumvariabele in haar onderzoek is de score op een korte vragenlijst. In deze vragenlijst wordt gepeild naar de tevredenheid met verschillende aspecten binnen de relatie (taakverdeling, intimiteit, zorgzaamheid, ...). De score op deze vragenlijst is een getal tussen 0 en 10. Deze vragenlijst werd twee keer afgenomen: één keer net voor de vrouwen gingen samenwonen met hun partner en één keer vijf jaar na deze datum. De meest recente vragenlijstscore is het criterium. Dit criterium wordt aangeduid met TEVRNU (tevredenheid nu). De vragenlijstscore van vijf jaar eerder wordt aangeduid met TEVRVROE (tevredenheid

vroeger). Deze tweede variabele wordt gebruikt als predictor in de voorspelling van het criterium TEVRNU. Indien men TEVRNU goed kan voorspellen op basis van TEVRVROE dan betekent dit dat de tevredenheid met de relatie een stabiel kenmerk is.

Naast TEVRVROE beschikte de onderzoekster ook nog over een tweede predictor. Zij wist van elke vrouw of zij in het verleden (in de periode voordat zij met haar huidige partner ging samenleven) reeds met een andere partner had samengeleefd. Deze variabele wordt aangeduid met NRELAT (aantal relaties). De variabele NRELAT is een categorische predictor die slechts twee waarden kan aannemen: de waarde 1 als de huidige relatie de eerste relatie is, en de waarde 2 als de huidige relatie reeds de tweede relatie is. (Er zaten geen vrouwen in het onderzoek voor wie de huidige relatie reeds de derde was.)

De data van dit onderzoek vind je in het bestand "tevreden.sav". Op deze data werden een aantal analyses uitgevoerd en ze werden besproken in de paragraaf over het uitvoeren van een regressieanalyse met GLM-Univariate. Nu worden een aantal vragen gesteld die je kunt beantwoorden op basis van deze analyses.

1. Levert NRELAT een significante bijdrage aan de voorspelling van TEVRNU bovenop TEVRVROE? Beantwoord deze vraag met behulp van een $F_{Change}$-statistiek. Deze statistiek moet je met de hand berekenen op basis van twee $R^2$-statistieken die men kan vinden in de uitvoer van GLM-Univariate.
2. Levert TEVRVROE een significante bijdrage aan de voorspelling van TEVRNU bovenop NRELAT? Beantwoord deze vraag met behulp van een $F_{Change}$-statistiek.
3. Maakt het, voor twee vrouwen met dezelfde mate van tevredenheid vijf jaar geleden, wat uit voor hun huidige tevredenheid of hun huidige relatie de eerste of de tweede is?
4. Maakt het, voor twee vrouwen die een tweede relatie begonnen zijn, wat uit voor hun huidige tevredenheid hoe tevreden zij vijf jaar geleden waren?
5. Geef in een figuur weer wat het patroon is in de gegevens van dit onderzoek. Doe dit op dezelfde manier als in de andere figuren in deze paragraaf (i.e., met ellipsen en regressielijnen). Zorg ervoor dat deze figuur in overeenstemming is met de gemiddelden in Figuur 4.47.

| | | Gemiddelde | |
|---|---|---|---|
| | N | TEVRNU | TEVRVROE |
| NRELAT=1 | 19 | 6.526 | 8.211 |
| NRELAT=2 | 21 | 5.268 | 6.905 |

Tabel 4.47: Tabel bij opgave 37.

*Opgave 38*

Een ontwikkelingspsychologe wil weten of er een verband is tussen de opvoedingsstijl van de ouders en de mate van agressiviteit van de kinderen. Op basis van de literatuur maakt de onderzoekster een onderscheid tussen de volgende drie opvoedingsstijlen (OPVSTIJL): (1) een autoritaire opvoedingsstijl waarbij kinderen geen inspraak hebben bij het bepalen van regels, (2) een democratische opvoedingsstijl waarbij er overleg is tussen ouders en kinderen bij het bepalen van regels, en (3) een permissieve opvoedingsstijl waarbij kinderen volledig zelf bepalen wat kan en wat niet. De mate van agressiviteit (AGRESSIE) wordt beoordeeld door de leerkracht van het kind. Een collega wijst de onderzoekster op het feit dat de socio-economische status van de ouders (SES) in de literatuur vaak genoemd wordt als bepalende factor bij het ontstaan van agressief gedrag bij kinderen. De SES is een kwantitatieve variabele. Er werden 210 kinderen onderzocht. Een deel van de SPSS-uitvoer wordt gegeven in Tabel 4.48 en Tabel 4.49.

Dependent Variable: AGRESSIE

| Parameter | B | Std. Error | t | Sig |
|---|---|---|---|---|
| Intercept | 54.144 | 2.193 | 24.688 | .000 |
| SES | -.219 | .017 | -13.052 | .000 |

$R^2 = 0.450$

Tabel 4.48: Tabel bij opgave 38.

Dependent Variable: AGRESSIE

| Parameter | B | Std. Error | t | Sig |
|---|---|---|---|---|
| Intercept | 53.517 | 2.270 | 23.579 | .000 |
| SES | -.215 | .016 | -13.320 | .000 |
| [OPVSTIJL=AUTOR] | 4.119 | 1.722 | 2.392 | .000 |
| [OPVSTIJL=DEMOC] | -3.730 | 1.728 | -2.158 | .032 |
| [OPVSTIJL=PERMIS] | 0 | | | |

$R^2 = 0.501$

Tabel 4.49: Tabel bij opgave 38.

1. De onderzoekster beweert dat er een unieke bijdrage is van OPVSTIJL bovenop SES in de voorspelling van AGRESSIE. Zij baseert zich op het feit dat de $\mathrm{F}_{\mathrm{Change}}$-statistiek significant was op het 0.01-niveau. Bereken deze $\mathrm{F}_{\mathrm{Change}}$-statistiek op basis van de gegevens in Tabel 4.48 en Tabel 4.49.
2. Geef de formules voor de binnen-groepen regressielijnen.
3. Teken de binnen-groepen regressielijnen in een grafiek en geef daarin aan welk verschil er getoetst wordt met de $\mathrm{F}_{\mathrm{Change}}$-statistiek in deelopgave 1.
4. Waar of niet waar en leg uit: SES is een storende variabele die verantwoordelijk is voor het verband tussen OPVSTIJL en AGRESSIE.

Stel dat de onderzoekster ook wil onderzoeken of het verband tussen SES en AGRESSIE verschillend is voor de verschillende niveaus van OPVSTIJL.

5. Welke analyse moet de onderzoekster uitvoeren om dit te onderzoeken? Concreet, wat zit er in Blok1 en Blok2?
6. Beschrijf zo concreet mogelijk de toetsstatistiek die nodig is om de nulhypothese te toetsen dat het verband tussen SES en AGRESSIE hetzelfde is voor de verschillende niveaus van OPVSTIJL. Concreet, hoe worden $R^2_{\mathrm{Change}}$ en $R^2_{Blok\,1+2}$ berekend, en wat zijn de vrijheidsgraden van de teller en de noemer van de $F_{\mathrm{Change}}$-statistiek?

*Opgave 39* COMPUTEROPDRACHT

In deze opgave heranalyseren we de data van een onderzoek naar de mentale ontwikkeling bij kinderen met een laag geboortegewicht (minder dan 2000 gr.). Je vindt deze data in het bestand "mentontw.sav". De afhankelijke variabele in dit onderzoek is het IQ zoals gemeten met de WISC (de Wechsler Intelligence Scale for Children). Er zijn twee predictoren in dit onderzoek: het aantal dagen hospitalisatie, aangeduid met HOSPDUUR, en het opleidingsniveau van de ouders, aangeduid met OPLNIV. Het opleidingsniveau van de ouders werd ingedeeld in drie categorieën: (1) geen van beide ouders heeft hoger onderwijs gevolgd (OPLNIV=1), (2) één van beide ouders heeft een diploma van het niet-universitair hoger onderwijs (OPLNIV=2), of (3) één van beide ouders heeft een universitair diploma (OPLNIV=3).

1. Heeft OPLNIV een significante bijdrage aan de voorspelling van IQ bovenop HOSPDUUR? Beantwoord deze vraag met behulp van een $F_{Change}$-statistiek. Deze statistiek kan je met de hand berekenen op basis van twee $R^2$-statistieken die men kan vinden in de uitvoer van GLM-Univariate.
2. Voor de onderstaande vragen, moet je gebruikmaken van de uitvoer van de analyse met zowel HOSPDUUR als OPLNIV als predictoren.
   (a) Welke IQ-score voorspel je voor een kind dat 25 dagen gehospitaliseerd is geweest en waarvan de ouders
      i. tot het laagste opleidingsniveau behoren?
      ii. tot het hoogste opleidingsniveau behoren?
   (b) Wat is het verschil in de voorspelde IQ-score van twee kinderen waarvan de ouders hetzelfde opleidingsniveau hebben maar waarvan één kind twee weken langer gehospitaliseerd is geweest?
3. Hebben de regressielijnen binnen de drie niveaus van OPLNIV hetzelfde regressiegewicht? Met andere woorden, zijn de regressielijnen parallel? Beantwoord deze vraag met behulp van een $F_{Change}$-statistiek.
4. Wat is, voor de analyse met de niet-parallelle regressielijnen, het verschil in voorspelde IQ-score voor twee kinderen die beiden 25 dagen gehospitaliseerd waren maar waarvan één kind ouders heeft met het hoogste opleidingsniveau en het andere kind ouders met het laagste opleidingsniveau?

*Opgave 40*
Een onderwijspsycholoog was geïnteresseerd in de factoren die bepalend zijn voor de moeilijkheidsgraad van de items van een test voor begrijpend lezen. Hij onderzocht een test bestaande uit 87 items. Voor elk item werd de moeilijkheidsgraad uitgedrukt in een getal tussen -5 (zeer gemakkelijk) en +5 (zeer moeilijk). De moeilijkheidsgraden van deze 87 items, aangeduid als MOEILIJK, fungeerden als criteriumvariabele in twee regressieanalyses. De predictoren in deze regressieanalyses waren kenmerken van de items. Alle items waren tekstfragmenten (uit boeken, tijdschriften, brieven, ...) waarover vragen gesteld werden. De itemkenmerken/predictoren waren de volgende:

1. NIVEAU: het niveau van informatieverwerking dat nodig was om de vraag te beantwoorden. De waarden van deze variabele varieerden tussen 1 (informatie die letterlijk in de tekst staat terugvinden) en 4 (informatie in de tekst beoordelen).
2. NZIN: het aantal zinnen waaruit het tekstfragment bestond.
3. NSTAP: het aantal stappen in de redenering dat nodig was om de vraag te beantwoorden. De variabele NSTAP werd bepaald door twee beoordelaars die alle items nauwkeurig bestudeerden.
4. TEKST: wetenschappelijke tekst (TEKST=1), ambtelijke tekst (TEKST=2) of literaire tekst (TEKST=3).

De predictoren NIVEAU, NZIN en NSTAP werden als kwantitatieve variabelen beschouwd en de predictor TEKST als een categorische variabele.

Er werden twee regressieanalyses uitgevoerd: één waarbij MOEILIJK voorspeld werd met NIVEAU, NZIN en NSTAP, en één waarbij, naast NIVEAU, NZIN en NSTAP, ook TEKST als predictor werd opgenomen. De resultaten van deze analyses worden gegeven in Tabel 4.50 en Tabel 4.51.

Dependent Variable: MOEILIJK

| Parameter | B | Std. Error | t | Sig |
|---|---|---|---|---|
| Intercept | -.572 | .157 | -3.644 | .000 |
| NIVEAU | .169 | .052 | 3.281 | .002 |
| NSTAP | 7.817E-02 | .029 | 2.651 | .010 |
| NZIN | -.241 | .056 | -4.304 | .000 |

$R^2 = .336$

Tabel 4.50: Tabel bij opgave 40.

Dependent Variable: MOEILIJK

| Parameter | B | Std. Error | t | Sig |
|---|---|---|---|---|
| Intercept | -.369 | .182 | -2.034 | .045 |
| NSTAP | 7.517E-02 | .029 | 2.579 | .012 |
| NIVEAU | .131 | .054 | 2.407 | .018 |
| NZIN | -.219 | .063 | -3.476 | .001 |
| [TEKST=1.00] | -.260 | .124 | -2.099 | .039 |
| [TEKST=2.00] | -.144 | .128 | -1.131 | .261 |
| [TEKST=3.00] | 0 | | | |

$R^2 = .371$

Tabel 4.51: Tabel bij opgave 40.

1. Wat is, op basis van de resultaten in Tabel 4.50, de voorspelde moeilijkheidsgraad van een item met de volgende kenmerken: een vraag van verwerkingsniveau 2 over een tekstfragment met 4 zinnen met drie redeneringsstappen? Laat je berekeningen zien.
2. Toets of TEKST een bijdrage levert aan de voorspelling van MOEILIJK bovenop NIVEAU, NZIN en NSTAP. Laat je berekeningen zien.
3. Uit Tabel 4.51 blijkt dat het regressiegewicht bij [TEKST=2.00] niet significant afwijkt van 0. Formuleer zo precies mogelijk wat je hieruit kunt concluderen over het verschil tussen de groepsspecifieke regressielijnen van wetenschappelijke, ambtelijke en literaire teksten.
4. Een collega van de onderzoekers stelde dat de relatie tussen de moeilijkheidsgraad (MOEILIJK) en het aantal redeneerstappen (NSTAP) verschillend is voor wetenschappelijke, ambtelijke en literaire teksten.
   (a) Welke analyse moet men uitvoeren om deze hypothese te onderzoeken? Beschrijf deze analyse in termen van de regressievergelijking(en) waarmee men MOEILIJK voorspelt.
   (b) Geef de formule van de toetsstatistiek die men moet uitrekenen om de nulhypothese te toetsen dat de relatie tussen MOEILIJK en NSTAP dezelfde is voor wetenschappelijke, ambtelijke en literaire teksten. Zeg ook hoe je de verschillende ingrediënten van deze formule moet berekenen.
5. De collega maakte zijn hypothese nog explicieter door te stellen dat de correlatie tussen MOEILIJK en NSTAP (a) positief is voor wetenschappelijke teksten, (b) negatief voor ambtelijke teksten, en (c) nul voor literaire teksten. Maak een grafische voorstelling van deze hypothese.

## 4.13 Regressieanalyse met Interactie Tussen Twee Kwantitatieve Predictoren

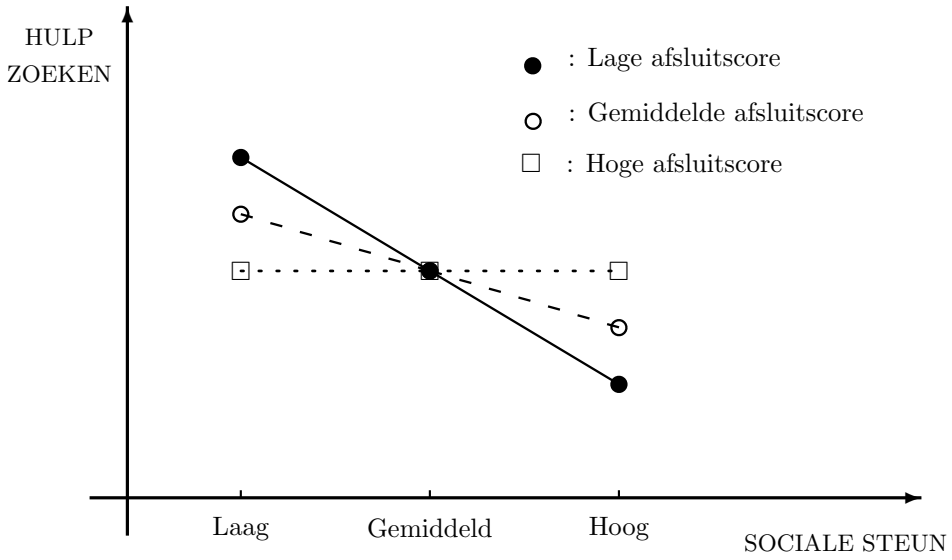
### 4.13.1 Voorbeeld

In de vorige paragraaf hebben we stilgestaan bij de interactie tussen een kwantitatieve en een categorische predictor. Nu staan we stil bij de interactie tussen twee kwantitatieve predictoren. We illustreren dit aan de hand van een onderzoek van Cepeda-Benito en Short (1998) (verder afgekort als C&S). C&S waren geïnteresseerd in de factoren die maken dat iemand psychotherapeutische hulp zoekt. Zij waren in het bijzonder geïnteresseerd in de variabelen SOCIALE STEUN en de neiging zichzelf af te sluiten, verder verkort aangeduid als AFSLUITEN. C&S verwachtten een interactie tussen deze twee variabelen. Concreet, zij verwachtten een negatief verband tussen SOCIALE STEUN en de neiging om psychotherapeutische hulp te zoeken (aangeduid als HULP ZOEKEN), maar alleen voor personen die zichzelf weinig of niet afsluiten. De gedachte hierachter is dat iemand die zichzelf niet afsluit voor de personen in zijn of haar omgeving, bij hen terecht kan in moeilijke periodes. Dus, voor zo iemand hangt het al dan niet zoeken van psychotherapeutische hulp af van de hoeveelheid sociale steun in zijn of haar omgeving. Dus, binnen de groep die zichzelf niet afsluit voor de omgeving, verwachtten C&S een negatief verband tussen HULP ZOEKEN en SOCIALE STEUN. Binnen de groep die zichzelf wél afsluit voor de omgeving, daarentegen, verwachtten C&S geen verband tussen HULP ZOEKEN en SOCIALE STEUN.

In principe kan men de hypothese van C&S onderzoeken met een regressieanalyse met als criterium HULP ZOEKEN en als predictoren SOCIALE STEUN en AFSLUITEN. Stel dat AFSLUITEN een categorische predictor is (bijvoorbeeld, met de twee niveaus wel en niet afsluiten), dan kunnen we gebruik maken van de regressieanalyse met groepsspecifieke regressiegewichten. In het onderzoek van C&S was AFSLUITEN echter een kwantitatieve predictor. Concreet, de predictor AFSLUITEN was de score op de *Self Concealment Scale*, een vragenlijst bestaande uit 10 items die men moest beantwoorden met een cijfer tussen 1 ("volledig oneens") en 5 ("volledig eens"). Ook de predictor SOCIALE STEUN en het criterium HULP ZOEKEN waren vragenlijstscores. De predictor SOCIALE STEUN was de score op een vragenlijst die peilde naar de uitgebreidheid van het sociale netwerk (18 vragen, te beantwoorden met "Ja" of "Nee"). En het criterium was de score op een vragenlijst bestaande uit 17 verschillende psychische problemen waarvoor men diende aan te geven (op schaal van 1 tot 6) hoe waarschijnlijk het was dat men voor dit type probleem psychotherapeutische hulp zou zoeken. Indien men geïnteresseerd is in wat mensen echt doen (i.c., zoeken ze hulp of niet), dan kan men zich natuurlijk afvragen of het zinvol is een vragenlijst te gebruiken waarin men moet aangeven wat men zóu doen. We komen hier later op terug, en werken voorlopig verder met de variabelen uit het onderzoek van C&S.

Om te beschrijven hoe we de interactie tussen SOCIALE STEUN en AFSLUITEN kunnen onderzoeken, gaan we er eenvoudigheidshalve vanuit dat de twee predictoren slechts drie mogelijke waarden kunnen aannemen: een lage, een gemiddelde en een hoge.

Bovendien gaan we er vanuit dat het verschil tussen de gemiddelde en de lage score gelijk is aan het verschil tussen de hoge en de gemiddelde score. Deze vereenvoudiging brengt geen wezenlijke verandering van het probleem met zich mee. In Figuur 4.10 wordt een geïdealiseerd patroon van gegevens getoond dat in overeenstemming is met de hypothese van C&S. Voor personen met een lage score op AFSLUITEN is er een negatief verband tussen SOCIALE STEUN en HULP ZOEKEN, voor personen met een hoge score is er geen verband, en voor personen met een gemiddelde score is er een licht negatief verband. Indien AFSLUITEN een categorische predictor was, dan zouden we deze drie regressielijnen binnen-groepen regressielijnen noemen. AFSLUITEN is echter een kwantitatieve predictor, en daarom noemen we ze *binnen-score* regressielijnen. Een binnen-score regressielijn beschrijft het verband tussen het criterium en één predictor voor een bepaalde vaste score op de andere predictor (in ons voorbeeld, een lage, een gemiddelde of een hoge score op AFSLUITEN).



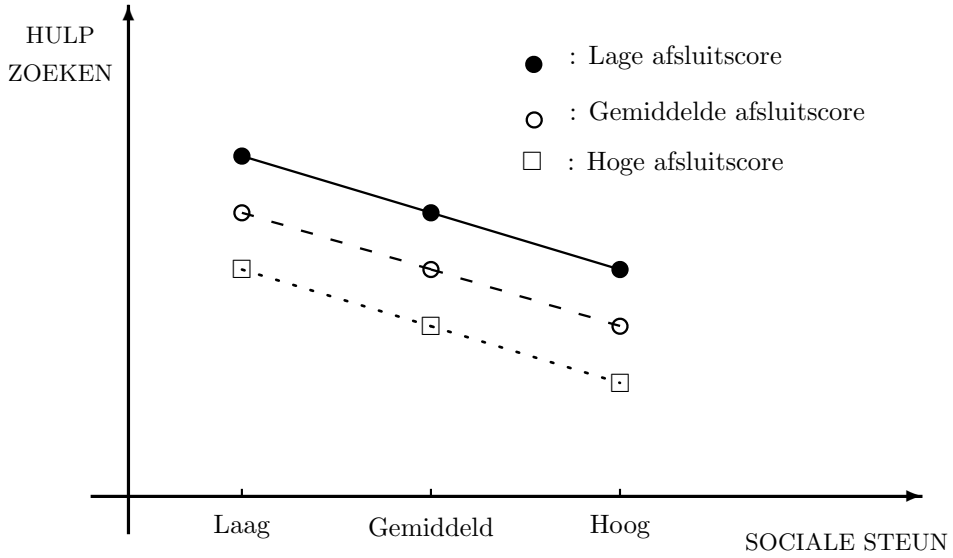

Figuur 4.10: Vereenvoudigde voorstelling van de hypothese van Cepeda-Benito en Short over het verband tussen de predictoren SOCIALE STEUN en AFSLUITEN en het criterium HULP ZOEKEN. Zie tekst.

Het patroon in Figuur 4.10 kan níet beschreven worden door een regressievergelijking met als enige predictoren SOCIALE STEUN en AFSLUITEN:

$$\widehat{Y} = A + B_1X_1 + B_2X_2 \quad .$$

In deze vergelijking staat $X_1$ voor SOCIALE STEUN en $X_2$ voor AFSLUITEN. Deze vergelijking wordt verder aangeduid als de regressievergelijking zónder interactie. In Figuur 4.11 wordt een patroon getoond dat wél beschreven kan worden door deze regressievergelijking zonder interactie. We krijgen dit patroon door negatieve regressiegewichten

$B_1$ en $B_2$ te kiezen, maar deze keuze is niet essentieel voor het punt dat we hier willen maken[15]. Het belangrijke punt in Figuur 4.11 is dat de binnen-score regressielijnen *parallel* zijn.




Figuur 4.11: Patroon in de gegevens dat beschreven kan worden door een regressievergelijking zonder interactie. Zie tekst.

We zullen nu laten zien dat de parallelliteit van de binnen-score regressielijnen volgt uit de regressievergelijking. We onderzoeken de binnen-score regressielijn voor het verband tussen HULP ZOEKEN en de predictor SOCIALE STEUN ($X_1$) bij een vaste waarde van de andere predictor AFSLUITEN ($X_2$). Stel dat we deze vaste waarde van $X_2$ aanduiden met $a$. Dan kan de regressievergelijking als volgt geschreven worden:

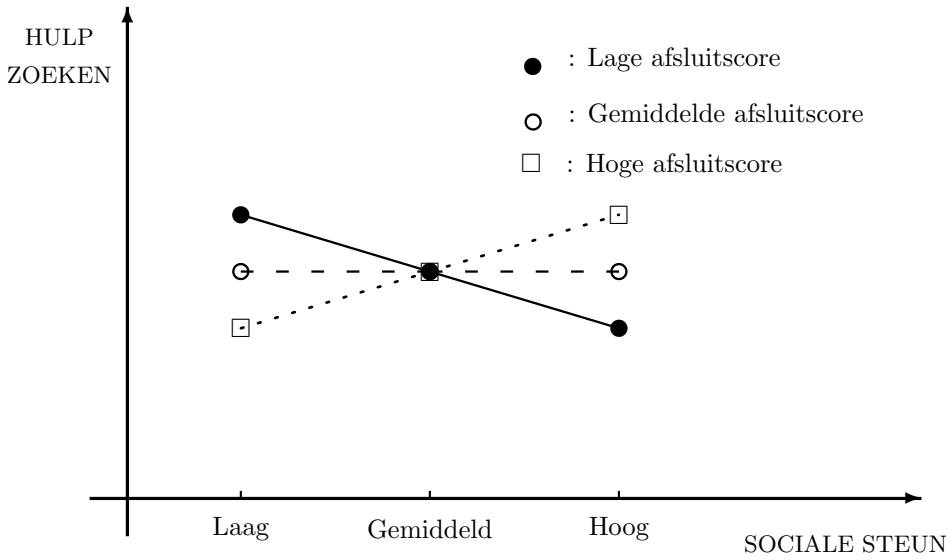
$$\widehat{Y} = (A + B_2 a) + B_1 X_1 \quad .$$

Dit is een enkelvoudige regressievergelijking met intercept $A + B_2 a$ en regressiegewicht $B_1$. Dus, de binnen-score regressielijnen voor de predictor $X_1$ bij telkens andere vaste waarden van $X_2$ (i.e., andere waarden voor $a$) hebben een verschillend intercept (nl., $A + B_2 a$) maar hetzelfde regressiegewicht (nl., $B_1$). Dit stemt overeen met Figuur 4.11: de drie binnen-score regressielijnen voor SOCIALE STEUN, elk bij een andere waarde van AFSLUITEN (laag, gemiddeld, hoog), hebben elk een ander intercept maar hetzelfde regressiegewicht (want ze dalen alledrie even sterk). Omdat ze alledrie hetzelfde regressiegewicht hebben, zijn ze parallel.

[15]Indien $B_1$ positief zou zijn, dan zouden de drie binnen-score regressielijnen stijgend zijn, en indien $B_2$ positief zou zijn dan zou de volgorde van de binnen-score regressielijnen (van boven naar onder) omgekeerd zijn (de hoge afsluitscore komt nu boven en de lage onder).

Vertrekkende van de regressievergelijking zonder interactie, kunnen we ook kijken naar de binnen-score regressielijnen voor het verband tussen HULP ZOEKEN en AFSLUITEN voor een bepaalde vaste score op SOCIALE STEUN. Ook deze binnen-score regressielijnen zijn parallel.

Om uit te leggen hoe men de interactie tussen twee kwantitatieve predictoren onderzoekt is het handig om te vertrekken van een patroon in de gegevens dat nóg meer vereenvoudigd is dan het patroon in Figuur 4.10. Het nadeel van het patroon in Figuur 4.10 is dat, óver de drie niveaus van AFSLUITEN, SOCIALE STEUN een negatief verband heeft met HULP ZOEKEN. Dus, het patroon in Figuur 4.10 wordt ten dele verklaard door de samenhang tussen SOCIALE STEUN en HULP ZOEKEN. We zoeken naar een patroon dat *in het geheel niet* verklaard kan worden door de regressievergelijking zonder interactie. Dit patroon wordt getoond in Figuur 4.12. Over de drie niveaus van AFSLUITEN, is er nu geen verband meer tussen SOCIALE STEUN en HULP ZOEKEN. En net als in Figuur 4.10, is er ook geen verband tussen AFSLUITEN en HULP ZOEKEN. De variantie in Figuur 4.12 kan dus in het geheel niet verklaard worden door de regressievergelijking zonder interactie. Merk op dat Figuur 4.12 niet in overeenstemming is met de hypothese van C&S, want voor de personen met een hoge score op AFSLUITEN is er nu een positief verband tussen SOCIALE STEUN en het HULP ZOEKEN. Dit werd niet voorspeld door C&S.




Figuur 4.12: Vereenvoudigde voorstelling van de hypothese dat alleen de interactie tussen de predictoren SOCIALE STEUN en AFSLUITEN een bijdrage levert aan de voorspelling van het criterium (HULP ZOEKEN). Zie tekst.

### 4.13.2 De Regressievergelijking met een Interactievariabele

We richten ons nu op de vraag hoe men de interactie tussen twee kwantitatieve predictoren kan onderzoeken. We doen dit aan de hand van verzonnen data van 9 proefpersonen. Deze data worden gegeven in Tabel 4.52. Indien men deze data in een figuur zou uitzetten, dan krijgt men precies het patroon van Figuur 4.12. De lage, gemiddelde en hoge predictorscores corresponderen met de waarden $-1$, 0 en $+1$. In Tabel 4.52 staan niet de originele maar de *gecentreerde* scores op de predictoren SOCIALE STEUN en AFSLUITEN. Centreren van een variabele betekent dat men van elke score het gemiddelde aftrekt. De originele scores voor SOCIALE STEUN hadden gelijk kunnen zijn aan 3 (voor de eerste drie proefpersonen), 4 (voor de volgende drie) en 5 (voor de laatste drie). Het gemiddelde van SOCIALE STEUN zou dan gelijk zijn aan 4 en de gecentreerde scores zouden gelijk zijn aan, respectievelijk, $-1$, 0 en $+1$. Om de interactie tussen kwantitatieve predictoren te onderzoeken maakt men bijna altijd gebruik van gecentreerde predictoren. Dit wordt later uitgelegd. Gecentreerde variabelen worden in dit boek aangeduid met vet gedrukte hoofdletters ($\boldsymbol{X}_1$ en $\boldsymbol{X}_2$).

| pp | HULP ZOEKEN $Y$ | SOCIALE STEUN $\boldsymbol{X}_1$ | AFSLUITEN $\boldsymbol{X}_2$ | Interactie $\boldsymbol{X}_1\boldsymbol{X}_2$ |
|---|---|---|---|---|
| 1 | 6 | $-1$ | $-1$ | $+1$ |
| 2 | 4 | $-1$ | 0 | 0 |
| 3 | 2 | $-1$ | $+1$ | $-1$ |
| 4 | 4 | 0 | $-1$ | 0 |
| 5 | 4 | 0 | 0 | 0 |
| 6 | 4 | 0 | $+1$ | 0 |
| 7 | 2 | $+1$ | $-1$ | $-1$ |
| 8 | 4 | $+1$ | 0 | 0 |
| 9 | 6 | $+1$ | $+1$ | $+1$ |
| Gem. | 4 | 0 | 0 | 0 |
| SD | 1.333 | 0.816 | 0.816 | 0.666 |

Correlaties

| | $Y$ | $\boldsymbol{X}_1$ | $\boldsymbol{X}_2$ |
|---|---|---|---|
| $\boldsymbol{X}_1$ | 0.000 | | |
| $\boldsymbol{X}_2$ | 0.000 | 0.000 | |
| $\boldsymbol{X}_1\boldsymbol{X}_2$ | 1.000 | 0.000 | 0.000 |

Tabel 4.52: Verzonnen data van een onderzoek naar de voorspelling van HULP ZOEKEN op basis van de predictoren SOCIALE STEUN en AFSLUITEN en hun interactie. Zie tekst.

De belangrijkste informatie in Tabel 4.52 is de laatste kolom die de zogenaamde *interactievariabele* bevat, aangeduid met $\boldsymbol{X}_1\boldsymbol{X}_2$. Men krijgt deze interactievariabele door de gecentreerde predictorscores elementsgewijs met elkaar te vermenigvuldigen. Deze manier van berekenen resulteert in een positieve score op de interactievariabele als beide predictorscores in dezelfde richting afwijken van het gemiddelde (i.e., als de gecentreerde

predictorscores hetzelfde teken hebben), een negatieve score als ze in tegengestelde richting afwijken van het gemiddelde (i.e., als de gecentreerde predictorscores een verschillend teken hebben), en een score gelijk aan 0 als één van de predictorscores gelijk is aan het gemiddelde. Dus, de interactievariabele geeft weer óf en in welke mate de twee predictoren in dezelfde richting wijzen. Indien een criteriumvariabele sterk correleert met deze interactievariabele, dan betekent dit dat de criteriumvariabele sterk samenhangt met de mate waarin de twee predictoren in dezelfde richting wijzen. Dit correspondeert met wat men intuïtief onder interactie verstaat.

Om de interactie tussen twee kwantitatieve predictoren te onderzoeken maakt men gebruik van de volgende regressievergelijking:

$$\widehat{Y} = A + B_1\boldsymbol{X}_1 + B_2\boldsymbol{X}_2 + B_{1\times 2}\boldsymbol{X}_1\boldsymbol{X}_2 \quad ,$$

waarin het regressiegewicht van de interactievariabele aangeduid wordt met $B_{1\times 2}$. Voor de gegevens in Tabel 4.52 is het eenvoudig om de regressiegewichten $B_1$, $B_2$ en $B_{1\times 2}$ te bepalen. De drie predictoren, $\boldsymbol{X}_1$, $\boldsymbol{X}_2$ en $\boldsymbol{X}_1\boldsymbol{X}_2$, zijn immers ongecorreleerd. Daarom zijn hun regressiegewichten gelijk aan die van de enkelvoudige regressievergelijking. Van de drie predictoren heeft alleen $\boldsymbol{X}_1\boldsymbol{X}_2$ een correlatie met het criterium $Y$, namelijk een correlatie van 1. Hieruit volgt dat (1) $B_1$ en $B_2$ gelijk zijn aan 0, (2) $B_{1\times 2}$ gelijk is aan $\mathrm{SD}(Y)/\mathrm{SD}(\boldsymbol{X}_1\boldsymbol{X}_2) = 2$, (3) $A$ gelijk is aan $(\overline{Y} - B_{1\times 2}\overline{\boldsymbol{X}}_{1\times 2}) = 4$, en (4) de proportie verklaarde variantie gelijk is aan 1. Dit laatste kunnen we controleren door de regressiegewichten en het intercept in de regressievergelijking in te vullen:

$$\widehat{Y} = 4 + 2\boldsymbol{X}_1\boldsymbol{X}_2 \quad .$$

Het is eenvoudig te controleren in Tabel 4.52 dat de voorspelde criteriumscores $\widehat{Y}$ precies gelijk zijn aan de geobserveerde criteriumscores $Y$. Dus, de geobserveerde variantie in het criterium wordt volledig verklaard door de interactievariabele $\boldsymbol{X}_1\boldsymbol{X}_2$.

Het is verleidelijk om uit het voorgaande te concluderen dat men de interactie tussen twee kwantitatieve predictoren kan onderzoeken met een enkelvoudige regressievergelijking met de interactievariabele $\boldsymbol{X}_1\boldsymbol{X}_2$ als enige predictor. Alhoewel dit waar is voor de gegevens in Tabel 4.52, is het in zijn algemeenheid (én in vrijwel alle praktijksituaties) onwaar. We illustreren dit met een voorbeeld. Stel dat we zes extra proefpersonen observeren. Hiervan hebben twee proefpersonen dezelfde scores (op het criterium en beide predictoren) als proefpersoon 1, twee hebben dezelfde scores als proefpersoon 4, en twee hebben dezelfde scores als proefpersoon 7. In termen van de symbolen in Figuur 4.12 betekent dit dat er twee extra proefpersonen zijn op elk zwart cirkeltje. Merk op dat alleen de verdeling van de proefpersonen over het patroon verandert; het patroon zelf verandert niet. De data van de 15 proefpersonen staan in Tabel 4.53. Merk op dat de gecentreerde afsluitscores verschillen van de scores in Tabel 4.52. Dit komt omdat de zes extra proefpersonen de laagst mogelijk afsluitscore hebben. Hierdoor wordt de gemiddelde afsluitscore kleiner en worden de individuele gecentreerde afsluitscores groter (want er wordt minder afgetrokken van de oorspronkelijke afsluitscores).

In vergelijking met Tabel 4.52 zijn de belangrijkste punten in Tabel 4.53 (1) dat de correlatie tussen de interactievariabele en het criterium niet langer gelijk is aan 1 en (2) dat de correlatie tussen SOCIALE STEUN en het criterium nu negatief is. Een enkelvoudige regressievergelijking met alleen de interactievariabele kan het criterium

| pp | HULP ZOEKEN $Y$ | SOCIALE STEUN $\boldsymbol{X}_1$ | AFSLUITEN $\boldsymbol{X}_2$ | Interactie $\boldsymbol{X}_1\boldsymbol{X}_2$ |
|---|---|---|---|---|
| 1 | 6 | −1 | −0.6 | +0.6 |
| 1bis | 6 | −1 | −0.6 | +0.6 |
| 1tris | 6 | −1 | −0.6 | +0.6 |
| 2 | 4 | −1 | +0.4 | −0.4 |
| 3 | 2 | −1 | +1.4 | −1.4 |
| 4 | 4 | 0 | −0.6 | 0 |
| 4bis | 4 | 0 | −0.6 | 0 |
| 4tris | 4 | 0 | −0.6 | 0 |
| 5 | 4 | 0 | +0.4 | 0 |
| 6 | 4 | 0 | +1.4 | 0 |
| 7 | 2 | +1 | −0.6 | −0.6 |
| 7bis | 2 | +1 | −0.6 | −0.6 |
| 7tris | 2 | +1 | −0.6 | −0.6 |
| 8 | 4 | +1 | +0.4 | +0.4 |
| 9 | 6 | +1 | +1.4 | +1.4 |
| Gem. | 4 | 0 | 0 | 0 |
| SD | 1.461 | 0.816 | 0.800 | 0.653 |

Correlaties

| | $Y$ | $\boldsymbol{X}_1$ | $\boldsymbol{X}_2$ |
|---|---|---|---|
| $\boldsymbol{X}_1$ | −0.447 | | |
| $\boldsymbol{X}_2$ | 0.000 | 0.000 | |
| $\boldsymbol{X}_1\boldsymbol{X}_2$ | 0.894 | 0.000 | 0.000 |

Tabel 4.53: Data van Tabel 4.52 plus data van zes extra proefpersonen. Zie tekst.

dus niet perfect voorspellen. Dus, ook al hebben de gegevens het patroon dat wij interactie noemen (afgebeeld in Figuur 4.12), de interactievariabele kan dit patroon niet perfect beschrijven. Dit heeft te maken met het feit dat de verdeling van de proefpersonen over dit patroon ongelijk is (i.c., verhoudingsgewijs meer proefpersonen met een lage afsluitscore). De oplossing voor dit probleem is eenvoudig: men neemt gewoon de predictoren $\boldsymbol{X}_1$ en $\boldsymbol{X}_2$ mee op in de regressievergelijking. Voor ons voorbeeld zijn de regressiegewichten $B_1$, $B_2$, en $B_{1\times 2}$ eenvoudig te berekenen. Dit komt omdat de drie predictoren, $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, en $\boldsymbol{X}_1\boldsymbol{X}_2$, onderling ongecorreleerd zijn. Daarom zijn $B_1$, $B_2$, en $B_{1\times 2}$ gelijk aan de regressiegewichten van de enkelvoudige regressievergelijkingen met, respectievelijk, $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, en $\boldsymbol{X}_1\boldsymbol{X}_2$:

$$\begin{aligned}
B_1 =& r_{Y\boldsymbol{X}_1}\frac{\mathrm{SD}(Y)}{\mathrm{SD}(\boldsymbol{X}_1)} = -0.447 \times \frac{1.461}{0.816} = 0.8 \\
B_2 =& r_{Y\boldsymbol{X}_2}\frac{\mathrm{SD}(Y)}{\mathrm{SD}(\boldsymbol{X}_2)} = 0.000 \times \frac{1.461}{0.800} = 0.0 \\
B_{1\times 2} =& r_{Y(\boldsymbol{X}_1\boldsymbol{X}_2)}\frac{\mathrm{SD}(Y)}{\mathrm{SD}(\boldsymbol{X}_1\boldsymbol{X}_2)} = 0.894 \times \frac{1.461}{0.653} = 2 \quad .
\end{aligned}$$

En omdat het gemiddelde van elk van de predictoren gelijk is aan 0, is ook het intercept $A$ eenvoudig te berekenen:

$$\begin{aligned}
A &= \overline{Y} - B_1\overline{\boldsymbol{X}_1} - B_2\overline{\boldsymbol{X}_2} - B_{1\times 2}\overline{\boldsymbol{X}_1\boldsymbol{X}_2} \\
&= \overline{Y} - B_1 \times 0 - B_2 \times 0 - B_{1\times 2} \times 0 \\
&= 4 \quad .
\end{aligned}$$

We krijgen dan de volgende regressievergelijking:

$$\widehat{Y} = 4 - 0.8\boldsymbol{X}_1 + 2\boldsymbol{X}_1\boldsymbol{X}_2 \quad .$$

Het is eenvoudig te controleren in Tabel 4.53 dat de voorspelde criteriumscores $\widehat{Y}$ precies gelijk zijn aan de geobserveerde criteriumscores $Y$. Dus, de variantie in het criterium wordt opnieuw volledig verklaard, maar nu door twéé predictoren, $\boldsymbol{X}_1$ en $\boldsymbol{X}_1\boldsymbol{X}_2$. Omdat de predictoren onderling ongecorreleerd zijn, kan men de proportie verklaarde variantie ook berekenen als $(-0.446)^2 + 0.894^2$, hetgeen gelijk is aan 1 (op afrondingsfouten na).

Het patroon in Figuur 4.12 kan altijd perfect verklaard worden door een regressievergelijking met de interactievariabele én de predictoren waarvan zij is afgeleid. (Hierbij gaan we ervan uit dat het verschil tussen een hoge en een gemiddelde score gelijk is aan het verschil tussen een gemiddelde en een lage score, en dat dit geldt voor beide predictoren.) Een ongelijke verdeling over de verschillende niveaus van deze predictoren verandert hier niets aan. Er zijn zelfs extreem ongelijke verdelingen denkbaar (die ook voorkomen in de praktijk) zoals die waarbij geen van de proefpersonen een hoge score heeft op SOCIALE STEUN. In dit geval, zijn de binnen-score regressielijnen nog steeds dezelfde, maar voor het gedeelte van deze binnen-score regressielijnen dat betrekking heeft op personen met een hoge score op SOCIALE STEUN zijn er geen observaties. We drukken dit ook uit door te zeggen dat dit gedeelte van de binnen-score regressielijnen gebaseerd is op extrapolatie.

### 4.13.3 Inhoudelijke Interpretatie van de Binnen-score Regressielijnen

Voor de inhoudelijke interpretatie van de resultaten van een regressieanalyse met interactie tussen de kwantitatieve predictoren is het van belang om te weten welk deel van de binnen-score regressielijnen effectief de observaties beschrijft en niet alleen gebaseerd is op extrapolatie. Deze informatie is eenvoudig te geven. Stel, we zijn geïnteresseerd in de binnen-score regressielijnen voor het verband tussen HULP ZOEKEN en AFSLUITEN gegeven een bepaalde vaste waarde voor SOCIALE STEUN. Veronderstel dat er geen personen zijn met een hoge score op SOCIALE STEUN en ook geen personen met een hoge score op AFSLUITEN. De interpretatie van de interactie gebeurt dan op basis van slechts een gedeelte van Figuur 4.12, namelijk dat gedeelte waarvoor men ook observaties heeft. Dit is het gedeelte linksboven in deze figuur. Op basis van dit gedeelte van de figuur besluit men (1) dat er geen verschil is tussen personen met een lage en een gemiddelde score op AFSLUITEN indien de score op SOCIALE STEUN gemiddeld is, en (2) dat personen met een lage score op AFSLUITEN een hogere criteriumscore hebben dan personen met een hoge score op AFSLUITEN indien de score op SOCIALE STEUN laag is.

In de praktijk heeft men meestal niet te maken met predictoren die slechts drie verschillende waarden kunnen aannemen, zoals in het getallenvoorbeeld dat hierboven besproken werd. Om te beschrijven hoe men dan te werk gaat, richten we ons op de binnen-score regressielijnen voor het verband tussen het criterium en predictor $\boldsymbol{X}_1$ gegeven een bepaalde vaste waarde voor predictor $\boldsymbol{X}_2$. Als vaste waarden van $\boldsymbol{X}_2$ neemt men meestal de volgende: (1) het gemiddelde (hetgeen gelijk is aan 0 omdat $\boldsymbol{X}_2$ gecentreerd is), (2) één standaarddeviatie boven het gemiddelde, en (3) één standaarddeviatie onder het gemiddelde. Op of in de buurt van deze waarden van $\boldsymbol{X}_2$ zijn er altijd voldoende observaties. In veel onderzoeken zijn er zelfs voldoende observaties op twéé standaarddeviaties boven en onder het gemiddelde van $\boldsymbol{X}_2$ en kan men ook deze waarden gebruiken.

Stel dat de vaste waarde van $\boldsymbol{X}_2$ aangeduid wordt met $a$. De binnen-score regressielijn wordt dan gegeven door de volgende vergelijking:

$$\begin{aligned}\widehat{Y} &= A + B_1\boldsymbol{X}_1 + B_2 a + B_{1\times 2}\boldsymbol{X}_1 a \\ &= A + B_2 a + (B_1 + B_{1\times 2}a)\boldsymbol{X}_1 \quad .\end{aligned}$$

Dit is een enkelvoudige regressievergelijking met regressiegewicht $(B_1 + B_{1\times 2}a)$ en intercept $(A + B_2 a)$. Deze regressievergelijking berekenen we voor de bovengenoemde waarden van $a$ en we tekenen de bijbehorende regressielijnen in een figuur. Om ervoor te zorgen dat deze regressielijnen effectief de observaties beschrijven, tekenen we ze alleen voor dát gebied van $\boldsymbol{X}_1$-waarden waarin we observaties hebben. Dit is altijd het geval voor het gebied tussen één standaarddeviatie onder en één standaarddeviatie boven het gemiddelde van $\boldsymbol{X}_1$, en vaak kan men ook twee standaarddeviaties nemen.

Aan de hand van de regressievergelijking van de binnen-score regressielijn kunnen we uitleggen hoe het regressiegewicht $B_1$ geïnterpreteerd moet worden. Als we deze regressievergelijking evalueren op het gemiddelde van predictor $\boldsymbol{X}_2$ (i.e., met $a = 0$), dan krijgen we

$$\widehat{Y} = A + B_1\boldsymbol{X}_1 \quad .$$

Dus, $B_1$ is het regressiegewicht van de binnen-score regressielijn voor een proefpersoon met een gemiddelde score op de andere predictor (hier, $\boldsymbol{X}_2$). Men is vaak in deze binnen-score regressielijn geïnteresseerd omdat zij ons zegt of het verband tussen $\boldsymbol{X}_1$ en het criterium gemiddeld genomen (i.e., gemiddeld over de waarden van $\boldsymbol{X}_2$) stijgend of dalend is. Vaak wil men ook weten of dit gemiddelde verband significant afwijkt van nul. Hier ligt het grote voordeel van een analyse met gecentreerde predictoren: in deze analyse toetst men met de vertrouwde T-statistiek van het regressiegewicht $B_1$ de nulhypothese dat, in de populatie, er gemiddeld genomen geen verband is tussen $\boldsymbol{X}_1$ en het criterium.

Soms is men geïnteresseerd in de statistische significantie van $B_1 + B_{1\times 2}a$, het regressiegewicht van de binnen-score regressielijn voor een bepaalde waarde van $a$. Voor $a$ gelijk aan 0 (i.e., op het gemiddelde van $\boldsymbol{X}_2$), kan men deze vraag beantwoorden op basis van een analyse met gecentreerde predictoren. Voor andere waarden van $a$, kan men ook een T-statistiek berekenen, maar deze wordt niet automatisch gegeven door SPSS-procedures zoals Regression of GLM-Univariate. Het berekenen van deze T-statistiek met de computer kan alleen op een vrij omslachtige manier. We zullen die manier hier niet beschrijven. Hoe je het moet doen wordt beschreven door Aiken en West (1991) die een heel boek gewijd hebben aan regressieanalyse met interacties tussen de predictoren.

### 4.13.4 De Regressievergelijking met Niet-gecentreerde Predictoren

Uit het bovenstaande kan men de indruk krijgen dat gecentreerde predictoren essentieel zijn voor het onderzoeken van de interactie tussen twee kwantitatieve predictoren. Deze indruk is onterecht. Dit kunnen we laten zien door de regressievergelijking met gecentreerde predictoren te herschrijven in termen van de originele niet-gecentreerde predictoren. De originele predictoren duiden we hier aan met $X_1$ en $X_2$. In termen van de originele predictoren, zijn de gecentreerde predictoren $\boldsymbol{X}_1$ en $\boldsymbol{X}_2$ gelijk aan, respectievelijk, $X_1 - \overline{X}_1$ en $X_2 - \overline{X}_2$. De regressievergelijking kan nu als volgt herschreven worden:

$$\begin{aligned}\widehat{Y} =& A + B_1\boldsymbol{X}_1 + B_2\boldsymbol{X}_2 + B_{1\times 2}\boldsymbol{X}_1\boldsymbol{X}_2\\ =& A + B_1(X_1 - \overline{X}_1) + B_2(X_2 - \overline{X}_2) + B_{1\times 2}(X_1 - \overline{X}_1)(X_2 - \overline{X}_2)\\ =& A - B_1\overline{X}_1 - B_2\overline{X}_2 + B_{1\times 2}\overline{X}_1\overline{X}_2 + (B_1 - B_{1\times 2}\overline{X}_2)X_1 +\\ & (B_2 - B_{1\times 2}\overline{X}_1)X_2 + B_{1\times 2}X_1X_2 \quad .\end{aligned}$$

Hieruit kunnen we besluiten dat, indien we een regressieanalyse zouden uitvoeren met de originele predictoren $X_1$, $X_2$ en $X_1X_2$, de regressiecoëfficiënten als volgt gerelateerd zijn aan de regressiecoëfficiënten van de vergelijking met gecentreerde predictoren ($A$, $B_1$, $B_2$ en $B_{1\times 2}$):

1. Het intercept is gelijk aan $(A - B_1\overline{X}_1 - B_2\overline{X}_2 + B_{1\times 2}\overline{X}_1\overline{X}_2)$.

2. Het regressiegewicht van $X_1$ is gelijk aan $(B_1 - B_{1\times 2}\overline{X}_2)$.

3. Het regressiegewicht van $X_2$ is gelijk aan $(B_2 - B_{1\times 2}\overline{X}_1)$.

4. Het regressiegewicht van $X_1X_2$ is gelijk aan $B_{1\times 2}$.

We kunnen hieruit het volgende besluiten:

1. Het regressiegewicht van de interactievariabele hangt niet af van het al dan niet gecentreerd zijn van de predictoren. Dus, onze conclusie over het al dan niet aanwezig zijn van interactie in de gegevens (op basis van een T-toets op $B_{1\times 2}$) hangt niet af van het gemiddelde van de predictoren.

2. Het regressiegewicht van de originele predictor $X_1$ kan níet geïnterpreteerd worden als de regressiegewicht van de binnen-score regressielijn voor een proefpersoon met een gemiddelde score op de andere predictor (hier, $X_2$). Deze interpretatie is voorbehouden aan de regressievergelijking met gecentreerde predictoren.

Ook al verschillen de regressiecoëfficiënten van de analyses met gecentreerde en originele predictoren, de binnen-score regressielijnen zijn hetzelfde. Dit kan men laten zien door de vaste waarde $a + \overline{X}_2$ (die correspondeert met de vaste waarde $a$ in de situatie met gecentreerde predictoren) in te vullen voor $X_2$ en dan de regressievergelijking te vereenvoudigen (schrappen van termen die tegen elkaar wegvallen en groeperen van de resterende termen). Dus, de informatie waarop de inhoudelijke interpretatie van de interactie gebaseerd is (i.e., de binnen-score regressielijnen) hangt niet af van het al dan niet gecentreerd zijn van de predictoren.

### 4.13.5 Het Belang van de Regressieanalyse met Interactie Tussen Twee Kwantitatieve Predictoren

Het onderzoeken van de interactie tussen twee kwantitatieve predictoren is interessant omdat het betrekking heeft op een patroon in de gegevens dat meer om het lijf heeft dan het simpele "hoe hoger de score op predictor $X$, hoe hoger/lager de score op het criterium", karakteristiek voor de regressievergelijking zonder interactievariabele. Als de binnen-score regressielijnen verschillende regressiegewichten hebben (niet parallel lopen) dan dringt zich de vraag op waarom dit zo is. Waarom is het verband tussen een predictor en het criterium anders voor verschillende waarden van een andere predictor?

Men mag het belang van de interactie tussen kwantitatieve predictoren ook niet overschatten. Het is immers vaak zo dat de interactievariabele slechts weinig variantie verklaart bovenop de afzonderlijke predictoren. Bijvoorbeeld, in het onderzoek van C&S verklaarde de interactie tussen SOCIALE STEUN en AFSLUITEN slechts 1.2 procent van de variantie in het criterium HULP ZOEKEN bovenop zeven andere predictoren (waaronder SOCIALE STEUN en AFSLUITEN). Deze toename in de proportie verklaarde variantie was wel significant, maar omwille van het grote aantal proefpersonen (i.c., 716) mag dit geen verbazing wekken.

Er zijn situaties waarin men een precieze hypothese heeft over de aard van de interactie tussen de predictoren. Bijvoorbeeld, C&S verwachtten een negatief verband tussen SOCIALE STEUN en het HULP ZOEKEN bij personen met een lage score op AFSLUITEN en géén verband bij personen met een hoge score op AFSLUITEN. In de lijn van

deze hypothese, kan men ervoor kiezen om de kwantitatieve variabele AFSLUITEN te dichotomiseren, waardoor het een categorische variabele wordt (met de niveaus hoog en laag). In plaats van een regressieanalyse met een interactievariabele als extra predictor doet men dan een regressieanalyse met groepsspecifieke regressiegewichten (waarbij de groepen corresponderen met de niveaus hoog en laag). Deze laatste analyse heeft als voordeel dat zij wat eenvoudiger is, maar als nadeel dat zij door de dichotomisering wat informatie weggooit.

---

*Opgave 41*
Waar of niet waar en leg uit.

1. Als men wil toetsen of de interactie tussen twee kwantitatieve predictoren statistisch significant is, dan maakt het niet uit of de predictoren gecentreerd zijn of niet.
2. Als men de statistische significantie wil onderzoeken van het regressiegewicht van de binnen-score regressielijn voor een persoon met een gemiddelde score op de andere predictor, dan maakt het niet uit of de predictoren gecentreerd zijn of niet.
3. Als de binnen-score regressielijnen parallel zijn, dan is het regressiegewicht van de interactiepredictor gelijk aan 0.

*Opgave 42* COMPUTEROPDRACHT
Een arbeids- en organisatiepsychologe is geïnteresseerd in de determinanten van arbeidstevredenheid. Enige jaren geleden onderzocht zij de bijdrage van de predictoren autonomie en werkdruk aan de voorspelling van de arbeidstevredenheid. Autonomie is de mate waarin een werknemer zelf kan beslissen over de invulling van zijn of haar taak. Een lage score op deze variabele betekent dat de werknemer weinig of geen inbreng heeft in de wijze waarop hij zijn dagtaak invult en een hoge score betekent dat hij dit vooral zelf bepaalt. Werkdruk is de mate waarin men onder druk staat om een bepaalde prestatie te leveren. Een lage score op deze variabele betekent dat de werknemer aan weinig of geen prestatienormen moet voldoen en een hoge score betekent dat hij aan hoge prestatienormen moet voldoen. Beide predictoren en het criterium (arbeidstevredenheid) werden vastgesteld met behulp van een vragenlijst. De twee predictoren worden aangeduid met AUTONOM en WERKDRUK, en het criterium wordt aangeduid met ARBTEVR. Je vindt de data van dit onderzoek in het bestand “arbtevr.sav”.

1. Is er een significante bijdrage van de interactievariabele bovenop AUTONOM en WERKDRUK? Je kan deze vraag beantwoorden op basis van een analyse met GLM-Univariate. In deze analyse moet je onder de knop *Model* kiezen voor *Custom*. Daarna selecteer je AUTONOM en WERKDRUK in het linkerlijstje en breng je deze variabelen door middel van de knop onder *Build Term(s)* één voor één over naar het rechterlijstje. De interactievariabele wordt aan het model toegevoegd door AUTONOM en WERKDRUK śamen te selecteren en daarna over te brengen naar het rechterlijstje.
2. Voer dezelfde analyse uit als hierboven maar nu met gecentreerde predictoren. Wat is het verschil met de uitvoer van de analyse met de niet-gecentreerde predictoren?

   Deze gecentreerde predictoren moet je zelf berekenen en dit doe je met een *Compute*-opdracht (die je vindt onder *Transform*). Bij de berekening van de gecentreerde predictoren heb je de gemiddelden van de predictoren nodig. Deze gemiddelden kun je berekenen met de procedure *Descriptives* onder *Descriptive Statistics*.
3. Teken de binnen-score regressielijn voor een persoon die een gemiddelde score behaalt op WERKDRUK.
4. Geef een interpretatie van de data. Maak hierbij gebruik van de binnen-score regressielijnen voor (1) een persoon met een gemiddelde score op WERKDRUK, (2) een persoon met

een score op WERKDRUK die één standaarddeviatie boven het gemiddelde ligt, en (3) een persoon met een score op WERKDRUK die één standaarddeviatie onder het gemiddelde ligt. Teken deze binnen-score regressielijnen voor de AUTONOM-scores in het interval tussen één standaarddeviatie onder en één standaarddeviatie boven het gemiddelde. De standaarddeviaties van AUTONOM en WERKDRUK vind je in de uitvoer van de procedure *Descriptives*.

---

## 4.14 Logistische Regressieanalyse

Logistische regressieanalyse is een vorm van regressieanalyse voor een dichotome criteriumvariabele. Een dichotome variabele kan slechts twee waarden aannemen, meestal aangeduid met 0 en 1. Zeer veel interessante criteriumvariabelen zijn dichotoom en daarom is het vreemd dat de logistische regressieanalyse nog steeds vrij weinig gebruikt wordt. Bijvoorbeeld, Cepeda-Benito en Short (1998) waren geïnteresseerd in de redenen waarom mensen psychotherapeutische hulp zoeken. De meest voor de hand liggende operationalisatie van deze variabele is gewoon vragen of iemand in psychotherapie (geweest) is of niet. Dit is een dichotome variabele. Dit is echter niet de operationalisatie die Cepeda-Benito en Short gebruikten. Zij maakten gebruik van een vragenlijst bestaande uit 17 verschillende psychische problemen waarvoor men diende aan te geven (op schaal van 1 tot 6) hoe waarschijnlijk het is dat men voor dit type probleem psychotherapeutische hulp zou zoeken. De criteriumvariabele in het onderzoek van Cepeda-Benito en Short is dus de zelf-ingeschatte kans op het zoeken van psychotherapeutische hulp en níet of men dat ook feitelijk doet.

Vragenlijstonderzoek zoals dat van Cepeda-Benito en Short is zeer populair binnen de psychologie. Vaak is dit jammer, want we zijn meestal in het feitelijke gedrag geïnteresseerd en niet in een mening. De logistische regressieanalyse kan ons hier een beetje bij helpen, want zij is juist bedoeld voor een dichotome criteriumvariabele, en of men een bepaald gedrag stelt of niet is nu juist een dichotome variabele. Men kan hierbij denken aan trouwen, een echtscheiding aangaan, verslaafd raken aan drugs, in de gevangenis terechtkomen, werkloos worden, in de ziektewet terechtkomen, promotie maken, ontslagen worden, een universitair diploma behalen, enzovoort.

De gewone regressieanalyse, die in de vorige paragrafen behandeld werd, is niet geschikt voor een dichotoom criterium. Dit is omdat de gewone regressieanalyse veronderstelt dat er een lineair (rechtlijnig) verband is tussen elk van de predictoren en het criterium. Bijvoorbeeld, als de score op een bepaalde predictor met 10 punten toeneemt, dan neemt de voorspelde criteriumscore $\widehat{Y}$ met $B \times 10$ punten toe of af ($B$ is het regressiegewicht van deze predictor). Dus, de verandering in $\widehat{Y}$ is proportioneel aan de verandering in de predictor. Dit is wat men bedoelt met een *lineair* verband.

Voor een dichotoom criterium, dat slechts de waarden 0 en 1 kan aannemen, is het vreemd om te denken in termen van een lineair verband met de predictor. Om dit te illustreren, stel je voor dat men toch de gewone regressieanalyse gebruikt voor een dichotoom criterium. De voorspelde criteriumscores $\widehat{Y}_i$ (voor alle personen $i = 1, \ldots, n$) kunnen in de volgende drie categorieën ingedeeld worden: (1) voorspelde scores van 0 tot

1 (0 en 1 inclusief), (2) voorspelde scores kleiner dan 0 en (3) voorspelde scores groter dan 1. De eerste categorie stelt de minste problemen. De geobserveerde criteriumscores kunnen weliswaar niet tussen 0 en 1 liggen, maar een voorspelde criteriumscore tussen 0 en 1 kan wel opgevat worden als de káns op een criteriumscore gelijk aan 1. Dit is een interpretatie waarmee goed te werken is in de praktijk. De twee andere categorieën voorspelde criteriumscores stellen ons echter voor onoverkomelijke interpretatieproblemen. Voorspelde criteriumscores kleiner dan 0 of groter dan 1 kunnen immers nooit opgevat worden als kansen.

In wat volgt, beschrijven we hoe de logistische regressie een oplossing biedt voor het boven geschetste probleem. De kern van deze oplossing is dat men niet langer het criterium zelf probeert te voorspellen, maar dat men uitrekent wat de kansen zijn van de twee mogelijke waarden van het criterium (i.c., 1 en 0) gegeven bepaalde scores op de predictoren. Na een korte beschrijving van deze oplossing bespreken we een aantal meer technische aspecten van de logistische regressie, zoals het berekenen en statistisch toetsen van de regressiecoëfficiënten, enkele eigenschappen van deze regressiecoëfficiënten, en het evalueren van de effectgrootte.

### 4.14.1 Kansen in plaats van Scores

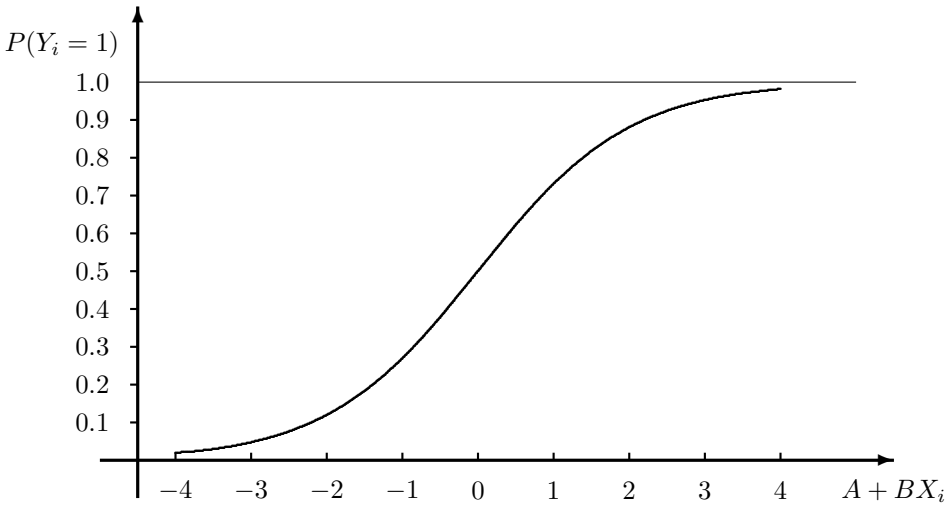
De kern van de logistische regressieanalyse is dat men vertrekt van de kans dat de criteriumvariabele gelijk is aan 1 gegeven bepaalde scores op de predictoren. (We zouden net zo goed kunnen vertrekken van de kans dat de criteriumvariabele gelijk is 0. Deze kans is immers complementair aan de kans dat de criteriumvariabele gelijk is aan 1.) In de logistische regressieanalyse probeert men kansen te vinden, één voor elke persoon, die zo kort mogelijk liggen bij de criteriumscores. Iets concreter, men probeert kansen te vinden die groot zijn (kort bij 1 liggen) als de criteriumscore $Y_i$ gelijk is aan 1 en die klein zijn (kort bij 0 liggen) als de criteriumscore $Y_i$ gelijk is aan 0. De kans op $Y_i = 1$ wordt verder aangeduid met $P(Y_i = 1)$.

De meest eenvoudige logistische regressieanalyse is de enkelvoudige, waarin men slechts één predictor gebruikt. De volgende formule ligt aan de basis van de enkelvoudige logistische regressieanalyse:

$$P(Y_i = 1) = \frac{e^{A+BX_i}}{1 + e^{A+BX_i}} \quad .$$

In wat volgt, wordt deze formule aangeduid als de kansformule van de enkelvoudige logistische regressieanalyse. In teller en noemer van het rechterlid van deze formule staat een macht met als grondtal $e$ ($e = 2.718$). De functie in deze formule noemt men de logistische functie en haar verloop wordt getoond in Figuur 4.13. De functiewaarde is dicht bij 1 als $A + BX_i$ groot positief is en is dicht bij 0 als $A + BX_i$ groot negatief is. De functiewaarde is gelijk aan 0.5 als $A + BX_i$ gelijk is aan 0. Met het gebruik van deze formule is gegarandeerd dat $P(Y_i = 1)$ altijd tussen 0 en 1 ligt, zoals dat hoort bij een kans.

We moeten hier opmerken dat $A + BX_i$ in de kansformule van de enkelvoudige regressieanalyse níet gelijk is aan de $A+BX_i$ in de gewone regressieanalyse. In de gewone regressieanalyse worden $A$ en $B$ bepaald door de som (over $i$) van de gekwadrateerde




Figuur 4.13: Het verloop van de logistische functie met argument $A+BX_i$.

afwijkingen $(Y_i - \widehat{Y}_i)^2$ te minimaliseren. De logistische regressiecoëfficiënten $A$ en $B$ (zelfde symbolen maar andere getallen) worden echter op een andere manier bepaald. Deze berekening is het onderwerp van de volgende paragraaf.

Het blijkt handig te zijn als we de kansformule van de enkelvoudige logistische regressieanalyse op een verkorte manier kunnen schrijven. Omdat het resultaat van deze formule alleen afhangt van $A + BX_i$, kunnen we haar ook als volgt schrijven:

$$P(Y_i = 1) = \text{Logist}(A + BX_i) \quad .$$

In deze formule staat $\text{Logist}(A + BX_i)$ voor de logistische functie met als argument $A + BX_i$.

Met de formule voor $P(Y_i = 1)$ hebben we in één klap ook een formule voor $P(Y_i = 0)$, de kans dat het criterium gelijk is aan 0. Omdat de som van $P(Y_i = 1)$ en $P(Y_i = 0)$ gelijk is aan 1, is $P(Y_i = 0)$ gelijk aan $[1 - P(Y_i = 1)]$. Vertrekkende hiervan is het eenvoudig om te laten zien dat

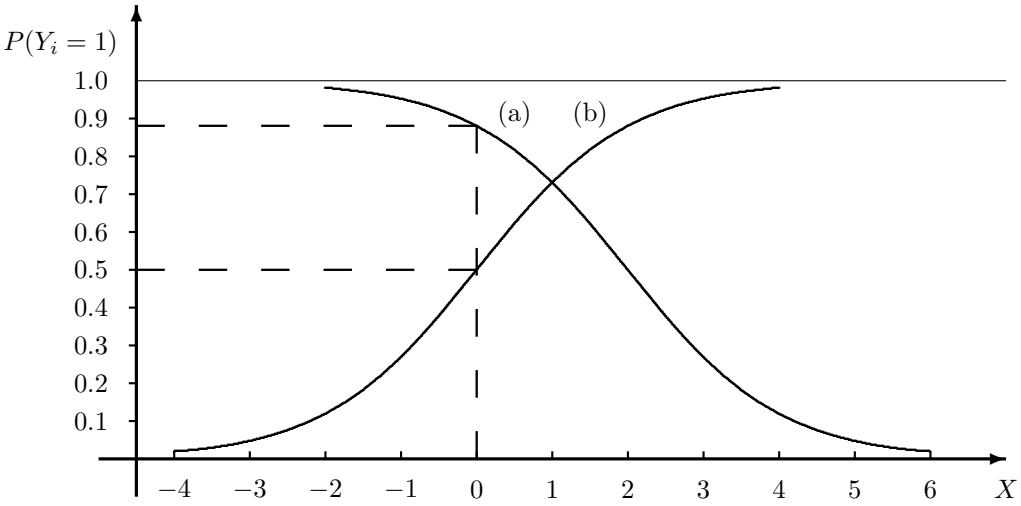
$$P(Y_i = 0) = \frac{1}{1 + e^{A+BX_i}} \quad .$$

Omdat $P(Y_i = 0)$ een functie is van $P(Y_i = 1)$, is het niet nodig om beide formules evenveel aandacht te geven. Immers, als we het hebben over een bepaalde eigenschap van $P(Y_i = 1)$, dan volgt hieruit onmiddellijk dat een corresponderende eigenschap opgaat voor $P(Y_i = 0)$. Bijvoorbeeld, uit Figuur (4.13) blijkt dat $P(Y_i = 1)$ een stijgende functie is van $A + BX_i$. Hieruit volgt onmiddellijk dat $P(Y_i = 0)$ een dalende functie is van $A + BX_i$. Immers, $P(Y_i = 0)$ is gelijk aan $1 - P(Y_i = 1)$.

Het feit dat $P(Y_i = 1)$ een stijgende functie is van $A + BX_i$ heeft een belangrijke implicatie voor de interpretatie van het logistische regressiegewicht $B$. Als dit re-

gressiegewicht positief is, dan is $P(Y_i = 1)$ een stijgende functie van $X_i$ (en $P(Y_i = 0)$ is dan een dalende functie). Immers, als $B$ positief is, dan is $A+BX_i$ een stijgende functie van $X_i$, en bijgevolg is ook $P(Y_i = 1)$ een stijgende functie van $X_i$. Indien $B$ echter negatief is, dan is $A + BX_i$ een dalende functie van $X_i$, en bijgevolg is ook $P(Y_i = 1)$ een dalende functie van $X_i$. Dus, het teken van $B$ bepaalt de richting van het verband tussen de predictor $X$ en de kansen $P(Y_i = 1)$ (voor $i = 1, \ldots, n$). De interpretatie van het logistische regressiegewicht $B$ is volledig analoog aan die van het gewone regressiegewicht: het gewone regressiegewicht bepaalt de bijdrage van de predictor aan de voorspelling van het criterium en het logistische regressiegewicht bepaalt de bijdrage van de predictor $X$ aan de kansen $P(Y_i = 1)$.

We kunnen de vorige alinea illustreren aan de hand van de zogenaamde *logistische regressielijnen* die het verband beschrijven tussen de predictor $X$ en de kansen $P(Y_i = 1)$. In Figuur 4.14 staat onder (a) een dalende en onder (b) een stijgende logistische regressielijn. Voor de dalende logistische regressielijn is $B$ negatief; hoe groter de predictor $X$, hoe kleiner de kans $P(Y_i = 1)$. En voor de stijgende logistische regressielijn is $B$ positief; hoe groter de predictor $X$, hoe groter de kans $P(Y_i = 1)$. De grootte van het regressiegewicht bepaalt ook de steilheid van de logistische regressielijn: hoe groter de absolute waarde van $B$ (i.e., $|B|$), hoe steiler de logistische regressielijn (stijgend of dalend, afhankelijk van het teken van $B$). In Figuur 4.14 zijn de regressiegewichten van de twee logistische regressielijnen even groot in absolute waarde, en daarom zijn de logistische regressielijnen even steil.




Figuur 4.14: Een dalende (a) en een stijgende (b) logistische regressielijn.

De logistische regressieconstante $A$ bepaalt de plaats van de logistische regressielijn ten opzichte van de X-as: hoe groter $A$, hoe meer de logistische regressielijn naar rechts verschoven is. Dit komt omdat de logistische regressieconstante $A$ de hoogte van de logistische regressielijn boven $X$=0 bepaalt. De kans $P(Y_i = 1)$ boven $X$=0 is immers

gelijk aan

$$\frac{e^A}{1+e^A} \quad ,$$

hetgeen alleen afhangt van $A$. De regressieconstanten van de dalende en de stijgende logistische regressielijn in Figuur 4.14 zijn gelijk aan, respectievelijk, 2 en 0. En de hoogtes boven het punt $X$=0 van deze twee logistische regressielijnen zijn gelijk aan, respectievelijk, 0.88 en 0.5.

De enkelvoudige logistische regressieanalyse kan op een eenvoudige manier veralgemeend worden tot de meervoudige logistische regressieanalyse. Het enige dat hiervoor nodig is, is het vervangen van $A + BX_i$ in de kansformule van de enkelvoudige logistische regressieanalyse door $A + B_1X_{1i} + B_2X_{2i} + \ldots + B_mX_{mi}$, waarin $m$ het aantal predictoren is en $X_{ji}$ de score van de i-de persoon op de j-de predictor. De interpretatie van de logistische regressiegewichten is weer volledig analoog aan die van de gewone regressiegewichten: een regressiegewicht $B_j$ in de meervoudige logistische regressie bepaalt de bijdrage van de predictor $X_j$ aan de kansen $P(Y_i = 1)$ gegeven de welbepaalde collectie overige predictoren. Dus, het logistische regressiegewicht $B_j$ hangt mede af van welke andere predictoren gebruikt worden om het criterium te voorspellen.

### 4.14.2 Het Berekenen van de Logistische Regressiecoëfficiënten

De logistische regressiecoëfficiënten worden op een andere manier berekend dan de gewone regressiecoëfficiënten. In plaats van het kleinste-kwadratencriterium maken we nu gebruik van het *maximum likelihood*[16] (ML) *criterium*. Men kan het ML-criterium gebruiken als men een kansformule heeft voor de data. De data zijn hier de dichotome criteriumscores $Y_i$ en de kansformule werd gegeven op pagina 424. Het begrip *likelihood* is één van de belangrijkste begrippen uit de wiskundige statistiek. Hier beperken we ons tot de concrete manier waarop we het ML-criterium kunnen gebruiken om de logistische regressiecoëfficiënten te berekenen. Om dit uit te leggen, blijkt het handig te zijn om $P(Y_i = 1)$ te vervangen door de verkorte notatie $\widehat{\pi}_i$. (De Griekse letter $\pi$ spreekt men uit als "pie". Deze letter wordt hier níet gebruikt als symbool voor het getal 3.14159 ....) Met het dakje boven de $\pi$ wordt aangegeven dat de status van $\widehat{\pi}_i$ dezelfde is als die van $\widehat{Y}_i$ in de gewone regressie[17].

Bij het toepassen van het ML-criterium vertrekken we van de zogenaamde *loglikelihood*, die we aanduiden met $L$.

$$L = \sum_{i=1}^{n} [Y_i \ln \widehat{\pi}_i + (1 - Y_i) \ln(1 - \widehat{\pi}_i)] \quad .$$

In deze formule staat *ln* voor het natuurlijke logaritme (i.e., het logaritme met grondtal $e$). Voor een goed begrip van deze formule is het voldoende om te weten dat het natuurlijke logaritme een stijgende functie is. Dit wil zeggen, als het argument van deze functie ($\widehat{\pi}_i$ of $(1 - \widehat{\pi}_i)$) groter wordt, dan wordt ook de functiewaarde groter. En

[16] De Nederlandse vertaling van *maximum likelihood* is *maximale aannemelijkheid*. De term *maximum likelihood* is echter zo ingeburgerd dat we hier de Engelse benaming gebruiken.

[17] De symbolen $\widehat{\pi}_i$ en $\widehat{Y}_i$ staan voor een schatting van de verwachte criteriumscore van een persoon die precies dezelfde predictorscores heeft als de i-de persoon in deze steekproef.

omgekeerd, geldt ook dat als het argument kleiner wordt, dat dan ook de functiewaarde kleiner wordt.

Het is goed om de termen tussen vierkante haken in de bovenstaande formule apart te bekijken voor het geval $Y_i = 1$ en $Y_i = 0$. Als $Y_i = 1$, dan vereenvoudigen deze termen tot $\ln \widehat{\pi}_i$, en als $Y_i = 0$ dan vereenvoudigen zij tot $\ln(1 - \widehat{\pi}_i)$. Dus, de loglikelihood $L$ is gelijk aan de som van de $\ln \widehat{\pi}_i$'s over alle observaties met $Y_i = 1$ plus de som van de $\ln(1 - \widehat{\pi}_i)$'s over alle observaties met $Y_i = 0$:

$$L = \sum_{i:Y_i=1} \ln \widehat{\pi}_i + \sum_{i:Y_i=0} \ln(1 - \widehat{\pi}_i) \quad .$$

In deze formule staat "$i : Y_i = 1$" voor "alle observaties met $Y_i = 1$" en "$i : Y_i = 0$" voor "alle observaties met $Y_i = 0$". Merk op dat $\widehat{\pi}_i$ gelijk is aan $P(Y_i = 1)$ en dat $(1 - \widehat{\pi}_i)$ gelijk is aan $P(Y_i = 0)$. Het moet dan duidelijk zijn dat de loglikelihood $L$ groter is naarmate (1) de observaties met $Y_i = 1$ een grotere kans $\widehat{\pi}_i$ hebben, en (2) de observaties met $Y_i = 0$ een kleinere kans $\widehat{\pi}_i$ (en dus een grotere kans $1 - \widehat{\pi}_i$).

Voor een goed begrip van de loglikelihood moet men $\widehat{\pi}_i$ niet als een getal op zich beschouwen maar als een functie van de regressiecoëfficiënten $A$ en $B_j$ $(j = 1, \ldots, m)$. De kans $\widehat{\pi}_i$ is de logistische functie van $A + B_1X_{1i} + B_2X_{2i} + \ldots + B_mX_{mi}$. Behalve van de regressiecoëfficiënten $A$ en $B_j$ $(j = 1, \ldots, m)$, is $\widehat{\pi}_i$ ook een functie van de predictorscores $X_{ji}$ $(j = 1, \ldots, m)$, maar in tegenstelling tot de regressiecoëfficiënten zijn deze predictorscores al bekend. De regressiecoëfficiënten moeten nog berekend worden uit de data. Dit brengt ons bij het ML criterium:

> **Maximum Likelihood Criterium:**
> Kies de regressiecoëfficiënten $A$ en $B_j$ $(j = 1, \ldots, m)$ zó dat de loglikelihood
> $$L = \sum_{i=1}^{n} [Y_i \ln \widehat{\pi}_i + (1 - Y_i) \ln(1 - \widehat{\pi}_i)]$$
> maximaal is.

We weten dat de loglikelihood stijgt naarmate de observaties met $Y_i = 1$ een grotere kans $\widehat{\pi}_i$ hebben, en de observaties met $Y_i = 0$ een kleinere kans $\widehat{\pi}_i$. Dus, het ML-criterium komt erop neer dat we de regressiecoëfficiënten $A$ en $B_j$ zó kiezen dat de hiermee berekende kansen $\widehat{\pi}_i$ deze samenhang met de $Y_i$-waarden zo sterk mogelijk maakt.

We hebben nu een criterium waaraan de regressiecoëfficiënten moeten voldoen, maar we hebben nog geen formules waarmee we ze kunnen berekenen. Enkele bijzondere gevallen daargelaten, kunnen we deze formules ook niet geven. Maar toch is het mogelijk om de regressiecoëfficiënten te vinden die aan het ML-criterium voldoen. Uitleggen hoe men dat precies doet zou ons echter te ver voeren[18].

[18] Voor de geïnteresseerde, de berekeningsmethode is gebaseerd op de idee van opeenvolgende verbeteringen. Concreet, men begint met bepaalde startwaarden voor de regressiecoëfficiënten en evalueert dan hoe sterk de hiermee berekende kansen $\widehat{\pi}_i$ samenhangen met de $Y_i$-waarden. Tegelijkertijd met deze evaluatie krijgt men ook informatie over hoe de regressiecoëfficiënten nog verbeterd kunnen worden (welke moeten groter en welke moeten kleiner worden) zodat ze beter voldoen aan het ML-criterium. De regressiecoëfficiënten worden dan aangepast en men evalueert ze opnieuw. Wanneer blijkt dat de regressiecoëfficiënten niet meer verbeterd kunnen worden dan stopt men. Omdat deze berekeningsmethode veel tijd kost, maakt men hiervoor altijd gebruik van een computerprogramma.

### 4.14.3 Eigenschappen van de Logistische Regressiecoëfficiënten

We vergelijken nu de eigenschappen van de logistische regressiecoëfficiënten met die van de gewone regressiecoëfficiënten. Zoals reeds eerder werd aangegeven, is de interpretatie van de logistische regressiecoëfficiënten volledig analoog aan die van de gewone regressiecoëfficiënten. Concreet, een logistisch regressiegewicht $B_j$ bepaalt de bijdrage van de predictor $X_j$ aan de kansen $P(Y_i = 1)$ gegeven de welbepaalde collectie overige predictoren. Er zijn ook enkele belangrijke verschillen tussen de logistische en de gewone regressiecoëfficiënten en daar richten we hier de aandacht op.

Een eerste verschil met de gewone regressiecoëfficiënten is dat logistische regressiecoëfficiënten soms oneindig zijn (plus of min oneindig). Dit is vervelend omdat de meeste computerprogramma's (zoals SPSS) niet met het getal oneindig kunnen rekenen. Het computerprogramma stopt dan voordat aan het ML-criterium voldaan is en levert ons geen bruikbare uitvoer (regressiecoëfficiënten, p-waarden, e.d.). Het programma zal dan ook een waarschuwing geven. De SPSS-procedure *Logistic Regression* doet dit met de volgende cryptische boodschap: "The parameter covariance matrix cannot be computed. Remaining statistics will be omitted."

Een tweede verschil tussen de gewone en de logistische regressiecoëfficiënten is dat logistische regressiegewichten veranderen als men een predictor toevoegt die ongecorreleerd is met de andere predictoren. Dit gebeurt niet bij gewone regressiegewichten. Gelukkig gaat het hier om een kleine verandering, die bovendien alsmaar kleiner wordt naarmate het aantal observaties toeneemt.

Het feit dat ongecorreleerde predictoren toch een invloed hebben op elkaars logistische regressiegewichten betekent niet dat men zich geen zorgen meer hoeft te maken over de correlaties tussen de predictoren. Immers, de máte waarin predictoren een invloed hebben op elkaars logistische regressiegewichten wordt wél bepaald door hun onderlinge correlatie. Het toevoegen van een sterk gecorreleerde predictor heeft een veel sterkere invloed op de andere regressiegewichten dan het toevoegen van een zwak gecorreleerde predictor. Bovendien is het zo dat, net zoals in de gewone regressieanalyse, de logistische regressiegewichten van sterk gecorreleerde predictoren slechts zelden statistisch significant zijn.

### 4.14.4 De Kwaliteit van de Voorspelling op basis van een Logistische Regressieanalyse

#### De Proportionele Reductie in de Kwadratische Fout

In de gewone regressieanalyse wordt de kwaliteit van de voorspelling uitgedrukt in de proportie verklaarde variantie. Het is nuttig om een onderscheid te maken tussen (1) de proportie verklaarde variantie door alle predictoren samen en (2) de extra proportie verklaarde variantie door één predictor bovenop alle andere predictoren. We beginnen met de eerste. De proportie verklaarde variantie door alle predictoren samen kan men als volgt schrijven:

$$\frac{\mathrm{VAR}(Y) - \mathrm{VAR}(Y - \widehat{Y})}{\mathrm{VAR}(Y)} \quad .$$

De teller van deze breuk is de reductie van de gemiddelde kwadratische fout door gebruik te maken van de predictoren. Dit kan men als volgt zien: (1) $\text{VAR}(Y)$ is de gemiddelde kwadratische fout na de voorspelling zonder predictoren (i.e., het gemiddelde van de gekwadrateerde afwijkingen $(Y_i - \overline{Y})^2$) en (2) $\text{VAR}(Y - \widehat{Y})$ is de gemiddelde kwadratische fout na de voorspelling op basis van alle predictoren (i.e., het gemiddelde van de gekwadrateerde afwijkingen $(Y_i - \widehat{Y}_i)^2$). Door het verschil $\text{VAR}(Y) - \text{VAR}(Y - \widehat{Y})$ te delen door $\text{VAR}(Y)$, krijgt men de *proportionele* reductie in kwadratische fout. Dit wordt vaak aangeduid als PRE (in het Engels, *proportional reduction in error*). In de gewone regressieanalyse zijn er twee handige manieren om de PRE te berekenen:

1. Omdat $\text{VAR}(Y - \widehat{Y})$ gelijk is aan $\text{VAR}(Y) - \text{VAR}(\widehat{Y})$, het verschil tussen de geobserveerde en de verklaarde variantie, is de teller van de formule van de PRE gelijk aan $\text{VAR}(\widehat{Y})$. Daarom is het niet nodig om de voorspellingsfout $Y - \widehat{Y}$ effectief te berekenen en daarvan de variantie te nemen.

2. De PRE kan berekend worden door het kwadraat te nemen van de meervoudige correlatie R, die gelijk is aan $r_{Y\widehat{Y}}$, de correlatie tussen de geobserveerde en de voorspelde criteriumscores.

Het ligt voor de hand om in de logistische regressieanalyse eveneens de PRE te berekenen. Omdat $\widehat{\pi}_i$ het analogon is van $\widehat{Y}_i$, gebeurt dit op de volgende manier:

$$\boxed{\text{PRE} = \frac{\text{VAR}(Y) - \text{VAR}(Y - \widehat{\pi})}{\text{VAR}(Y)} \quad .}$$

Anders dan in de gewone regressieanalyse, kan men hier de PRE niet berekenen via $\text{VAR}(\widehat{\pi})$ en $r_{Y\widehat{\pi}}$[19]. Dus, om de PRE te berekenen, moet men $Y - \widehat{\pi}$ effectief berekenen en daarvan de variantie nemen. Het moet opgemerkt worden dat SPSS en andere computerpakketten de PRE niet automatisch berekenen. Deze pakketten leveren meestal wel de benodigde ingrediënten om de PRE via een aantal extra operaties (zoals het berekenen van de variabele $Y - \widehat{\pi}$) zelf uit te rekenen. Het computerpakket blijft dus nuttig, alhoewel men de PRE niet kan verkrijgen met één enkele klik op een knop. Computerpakketten berekenen soms wél automatisch statistieken met verwarrende namen zoals *Cox & Snell* $R^2$ en *Nagelkerke* $R^2$. Deze statistieken hebben niets te maken met een meervoudige correlatie en zijn ook niet te interpreteren als een proportionele reductie in de kwadratische fout[20].

### De Proportionele Reductie in de Kwadratische Fout door één Predictor Bovenop de Andere Predictoren

We keren nu terug naar de gewone regressieanalyse en daarbinnen naar de extra proportie verklaarde variantie door predictor $X_j$ bovenop alle andere predictoren. Deze proportie

[19] Dit komt omdat $\widehat{\pi}$ en $Y - \widehat{\pi}$ gecorreleerd zijn. Het zou te ver voeren om hier dieper op in te gaan.

[20] De Cox & Snell $R^2$ en de Nagelkerke $R^2$ zijn beide een functie van twee loglikelihoods: (1) de loglikelihood als er een voorspelling gemaakt wordt zonder predictoren en (2) de loglikelihood als er een voorspelling gemaakt wordt op basis van álle predictoren. Deze twee statistieken zijn veel moeilijker te interpreteren dan de PRE.

verklaarde variantie kan men als volgt schrijven:

$$\frac{\text{VAR}(Y - \widehat{Y}^{Zonder\ X_j}) - \text{VAR}(Y - \widehat{Y})}{\text{VAR}(Y)} \quad .$$

In deze formule staat $\widehat{Y}^{Zonder\ X_j}$ voor de voorspelde criteriumscore op basis van alle predictoren behálve $X_j$. De teller van deze breuk is de reductie van de gemiddelde kwadratische fout door gebruik te maken van predictor $X_j$ bovenop alle andere predictoren. Delen door VAR($Y$) levert ons de proportionele reductie in kwadratische fout, die we aanduiden als PRE$_j$. In de gewone regressieanalyse zijn er twee handige manieren om de PRE$_j$ te berekenen:

1. Omdat VAR($Y - \widehat{Y}$) gelijk is aan VAR($Y$) − VAR($\widehat{Y}$) en VAR($Y - \widehat{Y}^{Zonder\ X_j}$) gelijk is aan VAR($Y$) − VAR($\widehat{Y}^{Zonder\ X_j}$), is de teller van de formule van de PRE$_j$ gelijk aan VAR($\widehat{Y}$) − VAR($\widehat{Y}^{Zonder\ X_j}$). Daarom is het niet nodig om de voorspellingsfouten $Y - \widehat{Y}$ en $Y - \widehat{Y}^{Zonder\ X_j}$ effectief te berekenen en daarvan de variantie te nemen.

2. De PRE$_j$ kan berekend worden door het kwadraat te nemen van de partcorrelatie $r_{Y(j\cdot\text{Rest})}$, de correlatie tussen $Y$ en het voor de andere predictoren uitgezuiverde deel van $X_j$ (aangeduid met $j \cdot$ *Rest*).

Het ligt voor de hand om in de logistische regressieanalyse eveneens de PRE$_j$ te berekenen. Omdat $\widehat{\pi}_i$ correspondeert met $\widehat{Y}_i$, gebeurt dit op de volgende manier:

$$\boxed{\text{PRE}_j = \frac{\text{VAR}(Y - \widehat{\pi}^{Zonder\ X_j}) - \text{VAR}(Y - \widehat{\pi})}{\text{VAR}(Y)} \quad .}$$

In deze formule staat $\widehat{\pi}^{Zonder\ X_j}$ voor de kans $P(Y_i = 1)$ bepaald op basis van alle predictoren behálve $X_j$. Anders dan in de gewone regressie, kan men hier de PRE$_j$ niet berekenen via VAR($\widehat{\pi}$), VAR($\widehat{\pi}^{Zonder\ X_j}$) en $r_{Y(j\cdot\text{Rest})}$. Dus, om de PRE$_j$ te berekenen, moet men $Y - \widehat{\pi}$ en $Y - \widehat{\pi}^{Zonder\ X_j}$ effectief berekenen en daarvan de variantie nemen. De PRE$_j$'s worden niet automatisch berekend door computerpakketten zoals SPSS. Deze pakketten leveren echter wel de benodigde ingrediënten om deze PRE$_j$'s via een aantal extra operaties zelf uit te rekenen. SPSS berekent wel automatisch een statistiek die met de letter $R$ aangeduid wordt, maar die niets te maken heeft met de meervoudige correlatie en ook niet te interpreteren is als een proportionele reductie in de kwadratische fout.

### Het Percentage Correcte Voorspellingen

We beschrijven nu een manier om de kwaliteit van de voorspelling te evalueren die gebaseerd is op het vergelijken van geobserveerde en voorspelde criteriumscores. De voorspelde criteriumscores worden afgeleid van de kansen $\widehat{\pi}$. Concreet, de voorspelde criteriumscore voor de i-de persoon is gelijk aan 1 als $\widehat{\pi}_i$ groter of gelijk is aan 0.5 en gelijk aan 0 als $\widehat{\pi}_i$ kleiner is dan 0.5. Dus, de voorspelde criteriumscore is de meest

waarschijnlijke criteriumscore gegeven de kans $\widehat{\pi}_i$. Op deze manier krijgen we voor elke observatie één geobserveerde criteriumscore, aangeduid met $Y_i$, en één voorspelde, bepaald op basis van $\widehat{\pi}_i$. Nu, een goede voorspelling betekent dat de geobserveerde en voorspelde criteriumscores voor de meeste observaties gelijk zijn (beide 0 of beide 1). Deze informatie kan men in een tabel weergeven, en in Tabel 4.54 wordt hiervan een voorbeeld gegeven.

| | | Voorspelde Criteriumscore 1 | Voorspelde Criteriumscore 0 | |
|---|---|---|---|---|
| Geobserveerde Criteriumscore | 1 | 15 | 15 | 50 % correct |
| | 0 | 14 | 56 | 80 % correct |
| | | | | 71 % correct |

Tabel 4.54: Classificatie van geobserveerde en voorspelde criteriumscores met aanduiding van het percentage correcte voorspellingen voor elk van de twee mogelijke criteriumscores (0 en 1) en voor alle observaties samen.

In Tabel 4.54 wordt het percentage correcte voorspellingen gegeven, apart voor de criteriumscore 1 en 0, en voor alle observaties samen. Bij het evalueren van het percentage correcte voorspellingen moet men altijd vertrekken van het percentage correcte voorspellingen dat men zou maken zónder predictoren. Men kan laten zien dat, in een logistische regressieanalyse zonder predictoren, $\widehat{\pi}_i$ gelijk is aan de proportie observaties met criteriumscore gelijk aan 1. Dus, in dit geval, is $\widehat{\pi}_i$ hetzelfde voor elke observatie $i$. Voor de gegevens in Tabel 4.54 betekent dit dat, bij een voorspelling zonder predictoren, elke $\widehat{\pi}_i$ gelijk is aan 0.30, want van de 100 observaties zijn er 30 met een criteriumscore gelijk aan 1. Dus, de voorspelde criteriumscore is gelijk aan 0 voor alle 100 observaties. Dit levert 70 correcte voorspellingen op. Dus, het percentage correcte voorspellingen zónder predictoren (70 procent) is in dit voorbeeld slechts één minder dan het percentage correcte voorspellingen mét predictoren (71 procent).

Men mag het belang van het percentage correcte voorspellingen niet overschatten. Het is immers vaak zo dat het percentage correcte voorspellingen zónder predictoren al zeer hoog is, zodat er weinig ruimte voor verbetering overblijft. Bijvoorbeeld, indien 90 procent van de observaties een criteriumscore gelijk aan 1 heeft, dan is ook het percentage correcte voorspellingen zonder predictoren 90 procent. Dit geldt ook indien slechts 10 procent van de observaties een criteriumscore gelijk aan 1 heeft. Met een perfecte voorspelling hebben we in beide gevallen een verbetering van slechts 10 procent.

### 4.14.5 Het Evalueren van de Effectgrootte

Er zijn twee manieren om de bijdrage van een predictor aan de voorspelling te evalueren. De eerste werd in de vorige paragraaf besproken, en hij geeft aan hoeveel de voorspelling verbetert door het opnemen van een extra predictor. Deze bijdrage van de predictor wordt gekwantificeerd door de $\text{PRE}_j$. De tweede manier, die nu wordt besproken, geeft aan hoeveel de $\widehat{\pi}$-waarde verandert door een verandering in de predictor waarvan men de bijdrage wil evalueren. Deze verandering in $\widehat{\pi}$-waarde wordt de *effectgrootte* genoemd. Om dit wat meer in detail uit te leggen is het goed om een onderscheid te maken tussen categorische en kwantitatieve predictoren.

#### Categorische Predictoren

We vertrekken hier van het onderzoek van Herrnstein en Murray (1994), beschreven in hun boek "The Bell Curve". Deze auteurs waren onder andere geïnteresseerd in de verschillen tussen blanken, zwarten en Latijns-Amerikanen (de drie belangrijkste rassen in de Noord-Amerikaanse samenleving) op een groot aantal criteriumvariabelen, waaronder zeer veel dichotome. De dichotome criteriumvariabele die we hier als voorbeeld gebruiken is het antwoord op de vraag "Bent U ooit meer dan een maand werkloos geweest?" Een ja-antwoord wordt met een 1 gecodeerd en een nee-antwoord met een 0.

Herrnstein en Murray (verder afgekort als H&M) waren geïnteresseerd in de verschillen tussen de rassen op deze en andere criteriumvariabelen na uitzuivering voor een aantal andere variabelen. H&M waren vooral geïnteresseerd in de verschillen tussen de rassen die nog overblijven als men voor IQ heeft uitgezuiverd. De stelling van H&M is immers dat verschillen tussen de rassen op indicatoren van maatschappelijk succes (zoals het behalen van een diploma, werkloos worden, het krijgen van een goede baan, etc.) voor een groot gedeelte toe te schrijven zijn aan IQ-verschillen tussen de rassen. Met andere woorden, H&M stellen dat IQ een belangrijke mediërende variabele is voor het verband tussen de categorische predictor RAS en allerlei indicatoren van maatschappelijk succes (de criteriumvariabelen).

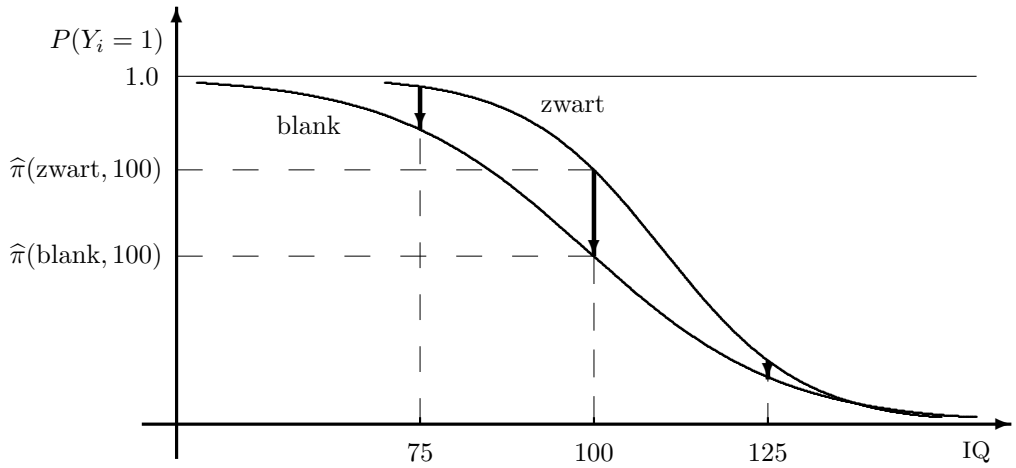
Omwille van de eenvoud, beperken we ons hier tot de categorische variabele RAS met slechts twee niveaus, blank en zwart. Bovendien zuiveren we alleen uit voor de kwantitatieve predictor IQ, alhoewel H&M ook uitzuiverden voor de kwantitatieve predictoren leeftijd en socio-economische status van de ouders. Het blijkt handig te zijn als we de kansen $\widehat{\pi}_i$ niet langer indexeren met $i$, maar in plaats daarvan $\widehat{\pi}$ opvatten als een *functie* met als argumenten de variabelen RAS en IQ. Om dit expliciet te maken, schrijven we in het vervolg $\widehat{\pi}(\text{RAS}, \text{IQ})$. De kans op een criteriumscore gelijk aan 1 voor, bijvoorbeeld, een zwarte persoon met een IQ van 109 schrijven we als $\widehat{\pi}(\text{zwart}, 109)$. Deze kans wordt als volgt berekend:

$$\widehat{\pi}(\text{zwart}, 109) = \text{Logist}(A_{\text{zwart}} + B_{\text{zwart}} \times 109) \quad .$$

Het is belangrijk op te merken dat het logistisch regressiegewicht $B_{\text{zwart}}$ geïndexeerd is met "zwart", waarmee wordt aangegeven dat dit regressiegewicht specifiek is voor de zwarten. In de gewone regressieanalyse hadden we het reeds over groepsspecifieke regressiegewichten, waarmee aangegeven werd dat het verband tussen de kwantitatieve predictor (hier, IQ) en het criterium verschillend kan zijn voor de verschillende niveaus

van de categorische predictor (hier, RAS). Dit werd eerder ook aangeduid door te zeggen dat er een interactie is tussen de kwantitatieve en de categorische predictor. De categorische predictor werd dan aangeduid als moderatorvariabele.

Om uit te leggen hoe men de effectgrootte van RAS kan evalueren met behulp van de kansen $\widehat{\pi}(\text{RAS}, \text{IQ})$, vertrekken we van Figuur 4.15. In deze figuur staan de logistische regressielijnen die het verband beschrijven tussen de kwantitatieve predictor IQ en de kans dat de score op het criterium gelijk is aan 1 (hier, de kans dat een persoon ooit langer dan een maand werkloos is geweest). Deze regressielijnen zijn specifiek voor zwarten en blanken, de twee niveaus van de categorische predictor RAS. De regressielijnen zijn dalend omdat de regressiegewichten ($B_{\text{blank}}$ en $B_{\text{zwart}}$) negatief zijn, hetgeen begrijpelijk is voor ons voorbeeld (met IQ als predictor en het langer dan een maand werkloos zijn als criterium).




Figuur 4.15: Logistische regressielijnen voor de voorspelling van een dichotoom criterium op basis van de kwantitatieve predictor IQ en de categorische predictor Ras (met de niveaus blank en zwart) met aanduiding van de effectgrootte van Ras bij IQ-score 75, 100 en 125 (neerwaarts gerichte pijlen).

Het feit dat de binnen-groepen regressiegewichten verschillend zijn, zorgt ervoor dat de regressielijnen naar elkaar toelopen en, bij een voldoende grote IQ-score, elkaar kruisen. Dit komt omdat, per eenheid toename in IQ, het argument van de logistische functie meer afneemt voor de ene groep (i.c., de zwarten, met als argument $A_{\text{zwart}} + B_{\text{zwart}}\text{IQ}$) dan voor de andere (i.c., de blanken, met als argument $A_{\text{blank}} + B_{\text{blank}}\text{IQ}$). Indien de binnen-groepen regressiegewichten wel gelijk zouden zijn, dan zouden de regressielijnen elkaar níet kruisen, maar in horizontale richting ten opzichte van elkaar verschoven zijn.

Het verschil tussen blanken en zwarten na uitzuivering voor IQ wordt in Figuur 4.15 aangegeven met een neerwaarts gerichte pijl. Dit uitgezuiverde verschil is het verschil tussen $\widehat{\pi}(\text{zwart}, \text{IQ})$ en $\widehat{\pi}(\text{blank}, \text{IQ})$ voor een vaste IQ-score. Dit verschil kan men op-

vatten als het verschil tussen een zwarte en een blanke persoon met hetzelfde IQ in de kans om ooit langer dan een maand werkloos geweest te zijn. Het is de vraag bij welke IQ-score men het effect van RAS moet evalueren. De meeste onderzoekers evalueren het effect van een categorische predictor op het gemiddelde van de kwantitatieve predictor. In ons voorbeeld is de kwantitatieve predictor het IQ, en daarvan weten we dat het gemiddelde in de populatie gelijk is aan 100. Voor de grafiek in Figuur 4.15 resulteert dit in een kans van 0.74 voor een zwarte en een kans van 0.48 voor een blanke (beiden met een IQ van 100). Dus, het verschil tussen blanken en zwarten na uitzuivering voor IQ is 0.26.

Het probleem bij deze manier om het effect van een categorische variabele te evalueren is dat zowel de grootte als de richting van het uitgezuiverde verschil afhangen van de score op de kwantitatieve variabele waarop men dit effect evalueert. Dit wordt geïllustreerd in Figuur 4.15 door met pijltjes het uitgezuiverde verschil aan te duiden voor drie verschillende IQ-scores (75, 100 en 125). Het uitgezuiverde verschil blijkt het grootst te zijn voor de gemiddelde IQ-score 100. Indien men het uitgezuiverde verschil zou berekenen voor een IQ-score van 150, dan zou men zelfs een ómgekeerd uitgezuiverd verschil vinden. Dit uitgezuiverde verschil is echter zeer klein: bij een IQ-score van 150 hebben blanken een iets grotere kans dan zwarten om ooit langer dan een maand werkloos geweest te zijn. Omdat een IQ-score van 150 zeer zelden voorkomt (in de steekproef misschien zelfs helemaal niet) moet men hier echter niet veel belang aan hechten. Voor de praktijk is het belangrijk om te kijken naar die uitgezuiverde verschillen (i.e., bij die IQ-scores) waar men de meeste observaties heeft.

Bij gelijke binnen-groepen regressiegewichten kruisen de regressielijnen elkaar niet; ze zijn immers in horizontale richting ten opzichte van elkaar verschoven. Dus, de richting van het uitgezuiverde verschil is altijd hetzelfde. De grootte van dit verschil verandert echter als men het berekent voor andere waarden van de kwantitatieve variabele (hier, IQ). Het is altijd zo dat, voor extreem grote en extreem kleine waarden van de kwantitatieve variabele het uitgezuiverde verschil vrijwel gelijk is aan 0. Dit komt omdat, bij deze extreme waarden, $P(Y_i = 1)$ zijn maximum (i.e., 1), respectievelijk, minimum (i.e., 0) bereikt. Dit geldt voor alle niveaus van de categorische variabele. Dus, een eventueel verschil tussen de regressielijnen bij een bepaalde waarde van de kwantitatieve variabele, verdwijnt altijd naarmate deze waarde extremer wordt.

### Kwantitatieve Predictoren

We beschrijven nu hoe de effectgrootte geëvalueerd wordt bij een kwantitatieve predictor. Om dit te illustreren vertrekken we opnieuw van het onderzoek van H&M. Deze auteurs waren niet alleen geïnteresseerd in de bijdrage aan de voorspelling van de categorische predictor RAS bovenop de kwantitatieve predictor IQ, maar ook in de bijdrage van IQ zelf. Met andere woorden, IQ werd niet alleen gebruikt als een variabele waarvoor men uitzuivert maar ook als een variabele waarvan men de eigen bijdrage wil kennen. H&M waren vooral geïnteresseerd in de bijdrage van IQ bovenop een andere kwantitatieve variabele, de socio-economische status van de ouders (aangeduid als SES). De variabele SES is gebaseerd op het opleidingsniveau van de ouders, de status van het beroep van de ouders, en het inkomen van de ouders. Elk van deze drie ingrediënten wordt in een getal uitgedrukt, en daarna worden zij via een gewogen som gecombineerd tot de SES.

Elk van deze ingrediënten draagt positief bij tot de SES: hoe hoger het opleidingsniveau, de beroepsstatus en het gezinsinkomen, hoe hoger de SES. In het onderzoek van H&M is SES een gestandaardiseerde variabele, waarvan de meeste waarden tussen $-3$ en $+3$ liggen en het gemiddelde gelijk is aan 0.

De interesse van H&M in de bijdrage van IQ bovenop SES is te begrijpen vanuit hun stelling dat maatschappelijk succes voor een groot gedeelte bepaald wordt door het IQ. Van een eventueel verband tussen IQ en maatschappelijk succes moet men natuurlijk aantonen dat het niet toe te schrijven is aan een storende variabele, zoals de SES van de ouders (een maat voor het maatschappelijk succes van de ouders), die zowel met IQ als met het eigen maatschappelijk succes kan samenhangen. Daarom ligt het voor de hand om IQ én SES te gebruiken als predictoren in een logistische regressieanalyse. Deze analyse stelt ons in staat om de bijdrage van IQ bovenop SES te evalueren.

Het blijkt handig te zijn als we $\widehat{\pi}$ opvatten als een functie met als argumenten de variabelen SES en IQ. Om dit expliciet te maken, schrijven we in het vervolg $\widehat{\pi}(\text{SES}, \text{IQ})$. De kans op een criteriumscore gelijk aan 1 voor, bijvoorbeeld, een persoon met een SES van 2 en een IQ van 109 schrijven we als $\widehat{\pi}(2, 109)$. Deze kans wordt als volgt berekend:

$$\widehat{\pi}(2, 109) = \text{Logist}(A + B_1 \times 2 + B_2 \times 109) \quad .$$

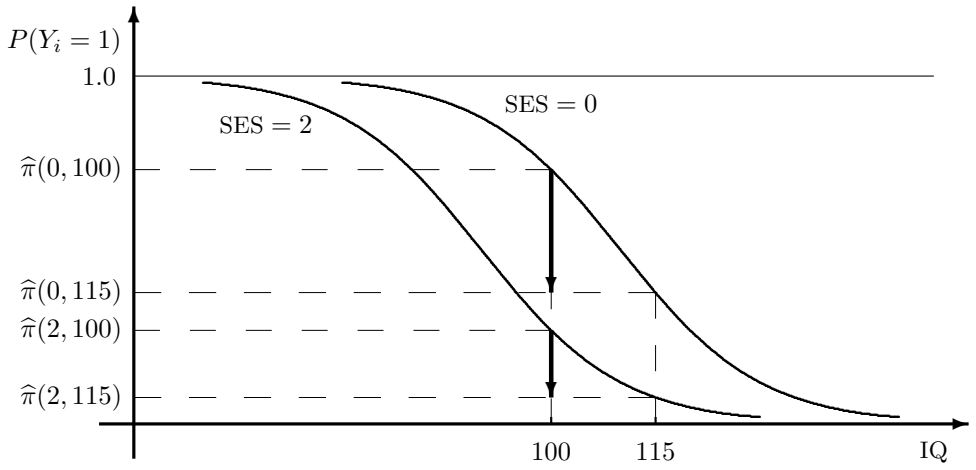
In deze formule is $B_1$ het regressiegewicht van SES en $B_2$ het regressiegewicht van IQ. In wat volgt, bekijken we $\widehat{\pi}(\text{SES}, \text{IQ})$ als een functie van IQ voor een vaste waarde van SES. Met SES $= 2$, is deze functie de volgende:

$$\widehat{\pi}(2, \text{IQ}) = \text{Logist}(A + B_1 \times 2 + B_2 \times \text{IQ}) \quad .$$

Deze formule heeft dezelfde structuur als de kansformule van de enkelvoudige logistische regressieanalyse met alleen de predictor IQ. Het enige verschil met de enkelvoudige logistische regressieanalyse is de regressieconstante, die hier gelijk is aan $A + B_1 \times 2$. We krijgen een andere regressieconstante als we een andere waarde kiezen voor SES. Bijvoorbeeld, met SES $= 0$ krijgen we de regressieconstante $A$.

Om uit te leggen hoe men de effectgrootte van IQ kan evalueren met behulp van de kansen $\widehat{\pi}(\text{SES}, \text{IQ})$, vertrekken we van Figuur 4.16. In deze figuur staan de logistische binnen-score regressielijnen die het verband beschrijven tussen IQ en de kans dat de score op het criterium gelijk is aan 1 (in ons voorbeeld, de kans dat een persoon ooit langer dan een maand werkloos is geweest). De linkerregressielijn is voor SES $= 2$ en de rechter is voor SES $= 0$. Deze twee regressielijnen lopen parallel omdat ze hetzelfde regressiegewicht hebben (namelijk, $B_2$) en slechts van elkaar verschillen wat betreft hun regressieconstante (namelijk, $A + B_1 \times 2$ versus $A$). De regressielijnen zijn dalend omdat het regressiegewicht van IQ (hier, $B_2$) negatief is, hetgeen begrijpelijk is voor ons voorbeeld (het criterium is langer dan een maand werkloos geweest zijn). En de regressielijn voor SES $= 2$ ligt links van de regressielijn voor SES $= 0$ omdat het regressiegewicht voor SES (hier, $B_1$) negatief is (hoe hoger de SES, hoe lager de kans dat men ooit langer dan een maand werkloos is geweest).

De effectgrootte van IQ wordt in Figuur 4.16 aangegeven met een neerwaarts gerichte pijl. Deze effectgrootte is op te vatten als een voor SES uitgezuiverd verschil tussen de kansen $\widehat{\pi}$ voor twee verschillende IQ-scores. In Figuur 4.16 wordt het verschil berekend tussen de kansen voor de IQ-scores 100 en 115. Deze IQ-scores corresponderen met,




Figuur 4.16: Logistische binnen-score regressielijnen voor de voorspelling van een dichotoom criterium op basis van de kwantitatieve predictoren SES en IQ voor twee vaste waarden van SES (i.c., SES = 2 en SES = 0) met aanduiding van de effectgrootte van IQ bij elk van deze twee waarden.

respectievelijk, het gemiddelde IQ en een standaarddeviatie boven het gemiddelde IQ. Dus, het uitgezuiverde verschil is het verschil tussen $\widehat{\pi}(\text{SES}, 100)$ en $\widehat{\pi}(\text{SES}, 115)$ voor een vaste SES-score. Dit verschil kan men opvatten als het verschil tussen iemand met een IQ van 100 en iemand met een IQ van 115 met dezelfde SES in de kans om ooit langer dan een maand werkloos geweest te zijn.

Het is de vraag bij welke SES men het effect van IQ moet evalueren. De meeste onderzoekers evalueren het effect van een kwantitatieve predictor op de gemiddelden van de kwantitatieve predictor waarvoor men uitzuivert. In ons voorbeeld wordt uitgezuiverd voor SES en omdat dit een gestandaardiseerde variabele is, is zijn gemiddelde gelijk aan 0. Deze keuze voor het gemiddelde is te verdedigen op basis van het argument dat er wellicht veel personen zijn met een SES op en rond het gemiddelde. Men mag echter niet vergeten dat, bij een andere SES-score, ook het uitgezuiverde verschil tussen de twee IQ-scores anders is. Dit wordt geïllustreerd in Figuur 4.16. Het verschil tussen $\widehat{\pi}(\text{SES}, 100)$ en $\widehat{\pi}(\text{SES}, 115)$ is groter voor SES = 0 dan voor SES = 2.

Men kan zich eveneens afvragen of men het effect van IQ wel moet evalueren op de scores 100 (het gemiddelde IQ) en 115 (een standaarddeviatie boven het gemiddelde IQ). Men krijgt immers meer informatie als men zich niet beperkt tot slechts twee IQ-scores, maar de kansen vergelijkt voor een veel groter aantal IQ-scores, zoals het hele gebied van twee standaarddeviaties onder tot twee standaarddeviaties boven het gemiddelde. Deze werkwijze komt erop neer dat men als het ware een stuk uit de regressielijnen knipt, namelijk het stuk dat correspondeert met het gebied van twee standaarddeviaties onder tot twee standaarddeviaties boven het gemiddelde. Dit is ook de werkwijze van H&M.

### 4.14.6 Statistische Toetsen

De statistische toetsen in de logistische regressieanalyse zijn analoog aan die van de gewone regressieanalyse. In de gewone regressieanalyse maakten we gebruik van de T- en de $F_{Change}$-statistiek. Met de T-statistiek kunnen we toetsen of, in de populatie, een bepaald regressiegewicht gelijk is aan 0. De $F_{Change}$-statistiek is algemener, omdat hij gebruikt kan worden om een nulhypothese te toetsen die betrekking heeft op meer dan een parameter. Bijvoorbeeld, de $F_{Change}$-statistiek kan gebruikt worden om de nulhypothese te toetsen dat, in de populatie, de regressiegewichten van een groep kwantitatieve predictoren allemaal gelijk zijn aan 0. Zo wordt de $F_{Change}$-statistiek gebruikt in de HMRA. De $F_{Change}$-statistiek kan ook gebruikt worden om de nulhypothese te toetsen dat, in de populatie, de binnen-groepen regressielijnen hetzelfde intercept of hetzelfde regressiegewicht hebben. Deze hypotheses worden getoetst in de regressieanalyse met categorische predictoren.

#### De Waldstatistiek

Het analogon van de T-statistiek in de logistische regressieanalyse is de *Waldstatistiek*, die hier aangeduid wordt met $W$.

$$\boxed{W = \left[\frac{B_j}{\mathrm{SE}(B_j)}\right]^2 \quad .}$$

Dus, de Waldstatistiek is gewoon het kwadraat van de T-statistiek. Onder de nulhypothese dat, in de populatie, het logistisch regressiegewicht $B_j$ gelijk is aan 0, volgt de Waldstatistiek *asymptotisch* een Chi-kwadraatverdeling met 1 vrijheidsgraad. De Chi-kwadraatverdeling behoort tot dezelfde familie als die waartoe ook de $\mathcal{T}$-, en de $\mathcal{F}$-verdeling behoren. Het enige wat men ervan moet weten is dat zij gebruikt wordt om een p-waarde te berekenen (i.e., de kans, onder deze verdeling, om een toetsstatistiek te observeren die nog extremer is dan de geobserveerde toetsstatistiek).

De Waldstatistiek is niet exact Chi-kwadraat verdeeld (in tegenstelling tot de T-statistiek in de gewone regressieanalyse die wel exact $\mathcal{T}$ verdeeld is). De Waldstatistiek is wel asymptotisch Chi-kwadraat verdeeld. Met "asymptotisch" bedoelen we dat, naarmate de steekproefomvang toeneemt (i.e., $n$ groter wordt), de verdeling van de Waldstatistiek onder de nulhypothese steeds beter op een Chi-kwadraatverdeling gaat lijken. De implicatie hiervan voor de praktijk, is dat men bij een kleine steekproef geen vertrouwen mag hebben in de asymptotische p-waarde (i.e., de p-waarde onder de asymptotische verdeling). De vraag is dan wat een kleine steekproef is (10, 50, 200, of nog meer observaties?). Op deze vraag kunnen we, jammer genoeg, geen goed antwoord geven[21]. We volstaan hier met een vuistregel: men mag vertrouwen hebben in de asymptotische p-waarde van de Waldstatistiek als het aantal observaties groter is dan 25 keer het aantal parameters. Dus, voor een logistische regressieanalyse met drie kwantitatieve predictoren heeft men minimaal 100 observaties nodig (i.e., één intercept plus drie regressiegewichten maal 25).

[21] Het beantwoorden van deze vraag is zo moeilijk omdat de minimale steekproefomvang afhangt van de grootte van de logistische regressiecoëfficiënten in de populatie, die onbekend zijn.

De Waldstatistiek die hier besproken werd, wordt ook de univariate Waldstatistiek genoemd. Daarnaast bestaat er ook een multivariate Waldstatistiek, die we hier niet in detail bespreken, maar die wel berekend wordt door sommige computerpakketten. De multivariate Waldstatistiek wordt gebruikt om een nulhypothese te toetsen die betrekking heeft op meer dan één parameter. In de computeruitvoer herken je de multivariate Waldstatistiek aan het feit dat hij meer dan één vrijheidsgraad heeft; hij heeft net zoveel vrijheidsgraden als het aantal parameters waarop hij betrekking heeft. In de uitvoer van de SPSS-procedure Logistic Regression vind je een multivariate Waldstatistiek voor elke categorische predictor met meer dan twee niveaus. Met deze Waldstatistiek toets je de nulhypothese dat, in de populatie, alle groepsspecifieke intercepten gelijk zijn. Het aantal vrijheidsgraden van deze statistiek is gelijk aan het aantal niveaus van de categorische predictor min 1. In wat volgt, bespreken we een toetsstatistiek waarmee men dezelfde nulhypotheses kan toetsen als met de multivariate Waldstatistiek.

### De Likelihood-Ratio Statistiek

Het analogon van de $\mathrm{F}_{\mathrm{Change}}$-statistiek in de logistische regressieanalyse is de *likelihood-ratio* (LR) statistiek. De $\mathrm{F}_{\mathrm{Change}}$-statistiek is een functie van het verschil tussen twee $R^2$-en: $R^2_{\mathrm{Blok\,1+2}}$ en $R^2_{\mathrm{Blok\,1}}$. Dit verschil wordt aangeduid met $R^2_{\mathrm{Change}}$. Analoog hiermee, is de LR-statistiek een functie van het verschil tussen twee loglikelihoods: $L_{\mathrm{Blok\,1+2}}$ en $L_{\mathrm{Blok\,1}}$, die verder uitgelegd worden. De formule van de loglikelihood werd gegeven op pagina 427, en voor het gemak wordt hij hier herhaald:

$$L = \sum_{i=1}^{n} [Y_i \ln \widehat{\pi}_i + (1 - Y_i) \ln(1 - \widehat{\pi}_i)] \quad .$$

Door deze loglikelihood te maximaliseren verkrijgen we de logistische regressiecoëfficiënten. In wat volgt, hebben we het alleen over de loglikelihood op zijn maximale waarde, dit wil zeggen, de loglikelihood die is berekend met de regressiecoëfficiënten die voldoen aan het ML-criterium.

De $\widehat{\pi}_i$'s in de formule van de loglikelihood worden berekend met de predictorscores. Dit impliceert dat, met andere predictoren, men ook een andere loglikelihood krijgt. Dit gegeven ligt aan de basis van het onderscheid tussen de twee loglikelihoods $L_{\mathrm{Blok\,1}}$ en $L_{\mathrm{Blok\,1+2}}$. Concreet, $L_{\mathrm{Blok\,1}}$ is de maximale loglikelihood van de logistische regressieanalyse met de predictoren uit Blok 1, en $L_{\mathrm{Blok\,1+2}}$ is de maximale loglikelihood van de logistische regressieanalyse met de predictoren uit Blok 1 en Blok 2. Nu, analoog aan de $\mathrm{F}_{\mathrm{Change}}$-statistiek, is de LR-statistiek een functie van $L_{\mathrm{Change}}$, het verschil tussen $L_{\mathrm{Blok\,1+2}}$ én $L_{\mathrm{Blok\,1}}$:

$$\boxed{L_{\mathrm{Change}} = L_{\mathrm{Blok\,1+2}} - L_{\mathrm{Blok\,1}} \quad .}$$

Merk op dat $L_{\mathrm{Change}}$ niet negatief kan zijn, want de samenhang tussen de kansen $\widehat{\pi}_i$ en de observaties $Y_i$ kan niet zwakker worden door het toevoegen van de predictoren uit Blok 2. De vraag is echter of $L_{\mathrm{Change}}$ ook significant groter is dan 0. Met andere woorden, het is de vraag of men op basis van de data de nulhypothese kan verwerpen dat, in de populatie, $L_{\mathrm{Change}}$ gelijk is aan 0. Deze nulhypothese is equivalent met de nulhypothese dat, in de populatie, de regressiegewichten van alle predictoren in Blok 2

gelijk zijn aan 0. Deze nulhypothese kan men toetsen met de LR-toets, die gebaseerd is op de volgende toetsstatistiek:

$$\boxed{LR = 2L_{\text{Change}} \quad .}$$

Dus, de LR-statistiek is gelijk aan twee keer $L_{\text{Change}}$. Onder de nulhypothese dat, in de populatie, de regressiegewichten van alle predictoren in Blok 2 gelijk zijn aan 0, heeft $L_{\text{Change}}$ asymptotisch een Chi-kwadraatverdeling met een aantal vrijheidsgraden gelijk aan het aantal extra parameters dat men nodig heeft om de predictoren in Blok 2 in de logistische regressieanalyse op te nemen. Deze vrijheidsgraden worden op dezelfde manier bepaald als in de gewone regressieanalyse. Bijvoorbeeld, indien Blok 2 alleen uit kwantitatieve predictoren bestaat, is het aantal vrijheidsgraden gelijk aan het aantal predictoren in Blok 2. Indien Blok 2 echter uit één categorische predictor bestaat, is het aantal vrijheidsgraden gelijk aan het aantal niveaus van deze predictor min 1.

Merk op dat men de LR-statistiek ook kan gebruiken om één enkel regressiegewicht te toetsen. Dit doet men door in Blok 2 slechts één kwantitatieve predictor op te nemen. In dit geval is, onder de nulhypothese, de LR-statistiek asymptotisch Chi-kwadraat verdeeld met één vrijheidsgraad.

Bij de geobserveerde waarde van de LR-statistiek wordt op de gebruikelijke manier de p-waarde berekend of in een tabel opgezocht. Om dezelfde reden als bij de Wald-statistiek, is deze p-waarde slechts een asymptotische p-waarde.

Met de Wald- en de LR-statistiek beschikken we over twee toetsstatistieken waarmee we dezelfde nulhypotheses kunnen toetsen. Als beide toetsstatistieken tot dezelfde conclusie leiden (de p-waarden alletwee kleiner dan het significantieniveau of alletwee groter) dan is er niets aan de hand. Als zij echter tot een verschillende conclusie leiden (met één toetsstatistiek wordt de nulhypothese verworpen en met de andere niet), dan is het niet duidelijk wat nu gerapporteerd moet worden. Voor dit probleem bestaat geen goede oplossing. Men kan alleen hopen dat het zich niet voordoet.

Sommige computerpakketten berekenen naast de Wald- en de LR-statistiek, ook de *scorestatistiek*. Met deze statistiek kan men in principe dezelfde nulhypotheses toetsen als met de LR-statistiek. In de praktijk gebeurt dit echter niet altijd. Zo toetsen de scorestatistieken die berekend worden door de SPSS-procedure Logistic Regression een nulhypothese die betrekking heeft op logistische regressiecoëfficiënten in een analyse met slechts één predictor. Anders uitgedrukt, de nulhypothese die getoetst wordt met deze scorestatistiek heeft betrekking op de bijdrage van een predictor aan de voorspelling zonder uitzuivering voor eventuele andere predictoren. Omdat men in de praktijk bijna altijd geïnteresseerd is in de bijdrage van een predictor na uitzuivering voor de andere predictoren, zijn de scorestatistieken die berekend worden door Logistic Regression zelden van belang.

### 4.14.7 Uitvoeren van een Logistische Regressieanalyse met de SPSS-procedure Logistic Regression

Nu volgt een korte beschrijving van de SPSS-procedure Logistic Regression die men best kan lezen voor een computer waarop SPSS draait. Men vindt Logistic Regression onder de naam *Binary Logistic* als onderdeel van de categorie *Regression* in het menu onder

de knop *Analyze*. Werken met deze procedure lijkt veel op het uitvoeren van een HMRA met de procedure Regression. Concreet, de predictoren worden in blokken ingedeeld, en van elk nieuw blok wordt getoetst of het een significante bijdrage heeft bovenop de andere predictoren.

In het dialoogvenster van Logistic Regression kies je (1) de criteriumvariabele door hem naar het vakje *Dependent* over te brengen en (2) de predictoren door ze naar het vakje *Covariates* over te brengen. De naam "Covariates" is verwarrend want binnen de procedure Logistic Regression kan dit zowel een kwantitatieve als een categorische variabele zijn. Dit wijkt af van de naamgeving binnen de procedure GLM-Univariate; hier wordt met "Covariate" een kwantitative predictor bedoeld. Als eenmaal alle predictoren opgenomen zijn in het venster onder *Covariates*, moet je na een druk op de knop *Categorical* in een dialoogvenster aangeven welke predictoren categorisch zijn. In dit dialoogvenster breng je alle categorische predictoren over van het linker- naar het rechterlijstje (van *Covariates* naar *Categorical Covariates*).

Logistic Regression produceert veel uitvoer. Hiervan is zeker niet alles even belangrijk. De uitvoer begint altijd met de resultaten van Blok 0. Dit blok wordt automatisch toegevoegd, en het correspondeert met het logistisch regressiemodel zonder predictoren (dus, met alleen het intercept). Op zich, is dit logistisch regressiemodel niet belangrijk, maar SPSS heeft het wel nodig om een LR-statistiek te berekenen waarmee men kan toetsen of de predictoren in Blok 1 een statistisch significante bijdrage leveren aan de voorspelling. Concreet, voor het toetsen van de bijdrage van de predictoren in Blok 1 wordt een LR-statistiek berekend op basis van het verschil $L_{\text{Blok 1}} - L_{\text{Blok 0}}$. Hierin staat $L_{\text{Blok 0}}$ voor de maximale loglikelihood van de logistische regressieanalyse met alleen het intercept. Tussen haakjes, SPSS gebruikt $L_{\text{Blok 0}}$ voor de berekening van de LR-statistiek, maar drukt deze loglikelihood niet automatisch af in de uitvoer.

Van de uitvoer onder Blok 0 is de *Classification Table* het belangrijkst. Hierin kun je aflezen wat het percentage correcte voorspellingen is dat men zou maken zónder predictoren. Het is nuttig om dit percentage te vergelijken met het percentage correcte voorspellingen op basis van een analyse mét predictoren.

De uitvoer onder Blok 0 bevat ook twee tabellen met toetsstatistieken: één tabel onder *Variables in the Equation* en één onder *Variables not in the Equation*. In de eerste tabel vind je de Waldtoets van de nulhypothese dat het intercept gelijk is aan 0. In deze toets is men bijna nooit geïnteresseerd en daarom gaan we er niet verder op in. In de tweede tabel staan de scorestatistieken van alle predictoren in Blok 1. Met deze scorestatistieken wordt de nulhypothese getoetst dat een bepaalde predictor geen bijdrage levert aan de voorspelling bovenop het intercept (i.e., de inhoud van Blok 0). Met "bovenop het intercept" wordt bedoeld "bovenop niets". Ook als er meerdere scorestatistieken in deze tabel staan (hetgeen het geval is als er meerdere predictoren zitten in Blok 1) dan wordt nog steeds dezelfde nulhypothese getoetst, en dus níet de nulhypothese dat een predictor geen bijdrage levert bovenop de andere predictoren in de tabel.

De uitvoer onder Blok 1 hangt af van de wijze waarop men predictoren opneemt in het logistische regressiemodel. Met de methode enter stopt men alle predictoren uit Blok 1 tegelijk in het regressiemodel. In plaats van de methode enter, kan men ook kiezen voor een automatische selectie van predictoren uit het blok. Dit gebeurt met varianten van

de forward- en de backward-selectieprocedures die we kennen van de SPSS-procedure voor gewone regressieanalyse. Wij beperken ons hier tot de methode enter.

In de tabel onder *Omnibus Test of Model Coefficients* vind je de resultaten van enkele LR-toetsen. De LR-statistieken zelf staan in de kolom *Chi-square*.

1. Met de LR-statistiek achter *Block* wordt de nulhypothese getoetst dat alle predictoren in dit blok sámen niets bijdragen aan de voorspelling van het criterium bovenop eventuele andere predictoren in de vorige blokken.

2. Met de LR-statistiek achter *Model* wordt de nulhypothese getoetst dat alle predictoren in dit en alle vorige blokken samen niets bijdragen aan de voorspelling van het criterium.

3. Indien men kiest voor de methode enter, dan is de LR-statistiek achter *Step* altijd gelijk aan de LR-statistiek achter *Block*. Omdat we ons beperken tot de methode enter, gaan we hier niet verder in op het verschil tussen *Step* en *Block*.

Anders dan bij de uitvoer onder Blok 0, wordt bij Blok 1 en alle volgende blokken wel de waarde van de maximale loglikelihood gegeven. Men vindt deze waarde (eigenlijk -2 keer deze waarde) in de tabel onder *Model Summary*. De *Cox & Snell R Square* en de *Nagelkerke R Square* laten we buiten beschouwing.

Anders dan bij de uitvoer onder Blok 0, staat er bij Blok 1 en alle volgende blokken wel interessante informatie in de tabel onder *Variables in the Equation*. In deze tabel staan nu de resultaten van de Waldtoetsen voor elk van de predictoren in dit blok. Met de Waldstatistieken in deze tabel wordt de nulhypothese getoetst dat een bepaalde predictor geen bijdrage levert aan de voorspelling bovenop alle andere predictoren in dit en alle vorige blokken. Indien een blok slechts één predictor bevat, dan toetst men met deze Waldstatistiek dezelfde nulhypothese als met de LR-statistiek onder *Omnibus Test of Model Coefficients*. Indien deze twee toetsen tot een verschillende conclusie leiden, dan baseert men zich het best op het resultaat van de LR-toets. Indien een blok meerdere predictoren bevat, dan toetst men met de Waldstatistiek iets anders dan met de LR-statistiek onder *Omnibus Test of Model Coefficients*. Deze LR-statistiek heeft dan immers betrekking op de bijdrage van álle predictoren in dit blok.

Als er een categorische predictor zit in Blok 1, dan is de informatie onder *Variables in the Equation* wat ingewikkelder dan als er alleen kwantitatieve predictoren in dit blok zitten. Dit bekijken we aan de hand van een voorbeeld. In Tabel 4.55 staat een deel van de uitvoer van een logistische regressieanalyse met een categorische en een kwantitatieve predictor. De predictoren zijn JOBCAT (jobcategorie, een categorische predictor met drie niveaus) en SALARIS (de kwantitatieve predictor). Net zoals in de uitvoer van GLM-Univariate, kan men ook hier de groepsspecifieke intercepten $A_k$ niet zomaar aflezen. Dit komt omdat SPSS elk groepsspecifiek intercept $A_k$ vervangt door de som $\mathcal{A} + \mathcal{A}_k$, zoals uitgelegd in de paragraaf over dummycodering, en in de uitvoer alleen deze laatste coëfficiënten (de sommen $\mathcal{A} + \mathcal{A}_k$ voor $k = 1, \ldots, K$) geeft. (Dit geldt alleen als men niets verandert onder *Change Contrast* in het dialoogvenster dat verschijnt na een druk op de knop *Categorical*.) Concreet, $\mathcal{A}$ is de coëfficiënt achter *Constant* in Tabel 4.55, $\mathcal{A}_1$ is de coëfficiënt achter JOBCAT(1) en $\mathcal{A}_2$ is de coëfficiënt achter JOBCAT(2). Als men niets verandert onder *Change Contrast*, dan wordt de $\mathcal{A}_k$-coëfficiënt van de laatste groep

(hier, de derde) gelijkgesteld aan 0. Met deze standaardrestrictie is het groepsspecifieke intercept $A_k$ gelijk is aan $\mathcal{A}+\mathcal{A}_k$. Hieruit volgt dat het groepsspecifieke intercept van de laatste groep gelijk is aan $\mathcal{A}$ (in ons voorbeeld, $A_3 = \mathcal{A} = -2.433$). De groepsspecifieke intercepten van de andere groepen vindt men door de $\mathcal{A}_k$-coëfficiënt op te tellen bij $\mathcal{A}$ (in ons voorbeeld, $A_1 = -2.433 + 1.546 = -0.887$ en $A_2 = -2.433 + 2.646 = 0.312$).

Variables in the Equation

| | B | S.E. | Wald | df | Sig. |
|---|---|---|---|---|---|
| JOBCAT | | | 13.229 | 2 | .001 |
| JOBCAT(1) | 1.546 | .735 | 4.419 | 1 | .036 |
| JOBCAT(2) | 2.646 | .798 | 10.981 | 1 | .001 |
| SALARIS | -0.007 | 0.013 | 0.299 | 1 | 0.548 |
| Constant | -2.433 | 1.123 | 4.692 | 1 | .030 |

Tabel 4.55: Uitvoer van een logistische regressieanalyse met een categorische (JOBCAT) en een kwantitatieve (SALARIS) predictor.

Met de Waldstatistieken achter JOBCAT(1) en JOBCAT(2) kan men de nulhypothese toetsen dat er geen verschil is tussen het intercept van de corresponderende groep en het intercept van de laatste groep. In deze nulhypotheses is men meestal niet geïnteresseerd. Men is meestal wél geïnteresseerd in de nulhypothese dat alle groepsspecifieke intercepten gelijk zijn. Deze nulhypothese is equivalent met de nulhypothese dat de categorische predictor geen extra bijdrage levert aan de voorspelling van het criterium bovenop de andere predictoren. Behalve met de LR-statistiek, kan men deze nulhypothese dus ook toetsen met de multivariate Waldstatistiek. Deze multivariate Waldstatistiek vind je in de eerste rij van Tabel 4.55. Deze statistiek heeft asymptotisch een Chi-kwadraatverdeling met aantal vrijheidsgraden gelijk aan het aantal groepen min 1.

Met Logistic Regression kan men ook een analyse uitvoeren waarin de interactie tussen twee predictoren gebruikt wordt om het dichotome criterium te voorspellen. De twee soorten interacties waar we het in dit hoofdstuk over hadden zijn (1) een interactie tussen een kwantitatieve en een categorische predictor, resulterend in een logistische regressieanalyse met groepsspecifieke regressiegewichten, en (2) een interactie tussen twee kwantitatieve predictoren. Een interactie wordt toegevoegd aan de lijst van predictoren door in het dialoogvenster van Logistic Regression twee variabelen tegelijkertijd te selecteren. Hierdoor wordt een knop met $> a * b >$ beschikbaar, en door erop te klikken wordt de interactie tussen de twee geselecteerde variabelen toegevoegd aan de lijst van predictoren.

In de uitvoer van een logistische regressieanalyse met groepsspecifieke regressiegewichten wordt men met hetzelfde probleem geconfronteerd als in de uitvoer van de corresponderende gewone regressieanalyse (geproduceerd door GLM-Univariate): men kan hier de groepsspecifieke regressiegewichten $B_k$ niet zomaar aflezen. Dit komt omdat SPSS elk groepsspecifiek regressiegewicht $B_k$ vervangt door de som $\mathcal{B} + \mathcal{B}_k$, zoals uitgelegd in de paragraaf over dummycodering, en omdat SPSS in de uitvoer alleen deze laatste coëfficiënten geeft. Met deze laatste coëfficiënten kan men de groepsspecifieke regressiegewichten natuurlijk wel uitrekenen.

Voor het evalueren van de kwaliteit van de voorspelling is het nuttig om de proportionele reductie in kwadratische fout te berekenen. Deze statistiek bestaat in twee

varianten: één variant waarmee de kwaliteit van de voorspelling op basis van alle predictoren samen geëvalueerd wordt (aangeduid met PRE) en één variant waarmee de extra bijdrage van één predictor aan de voorspelling geëvalueerd wordt (aangeduid met $\text{PRE}_j$). Om deze statistieken te berekenen heeft men de verschilvariabele $Y - \widehat{\pi}$ nodig. Deze variabele is het verschil tussen de geobserveerde criteriumscores en de kansen $\widehat{\pi}$. SPSS noemt deze variabele een *residual* en wij noemen het hier een residu. Men kan dit residu laten wegschrijven in het databestand door onder de knop *Save* de optie *Residuals-Unstandardized* aan te kruisen. Elk residu dat men laat wegschrijven wordt in de laatste kolom van het databestand toegevoegd. Daarna is het eenvoudig om $\text{VAR}(Y)$ en $\text{VAR}(Y - \widehat{\pi})$ (de ingrediënten van PRE en $\text{PRE}_j$) te berekenen. Dit kan, bijvoorbeeld, met de procedure *Descriptives* die je vindt onder *Descriptive Statistics*. Om *Descriptives* de variantie van een variabele te laten berekenen moet je onder *Options* een kruisje zetten bij *Variance*.

---

*Opgave 43*
Waar of niet waar en leg uit.

1. Als de binnen-groepen logistische regressiegewichten gelijk zijn, dan is de *grootte* van het uitgezuiverde verschil tussen twee groepen onafhankelijk van de waarde van de kwantitatieve variabele.
2. Als de binnen-groepen logistische regressiegewichten gelijk zijn, dan is de *richting* van het uitgezuiverde verschil tussen de groepen onafhankelijk van de waarde van de kwantitatieve variabele.
3. Indien de Waldstatistiek van een logistische regressiecoëfficiënt een zeer kleine p-waarde heeft (bijvoorbeeld, 0.000000000001), dan is ook de effectgrootte van de bijbehorende predictor substantieel.
4. Indien de effectgrootte van een predictor substantieel is, dan is de bijdrage van deze predictor aan de voorspelling statistisch significant.
5. Met de Waldstatistiek kan men dezelfde nulhypotheses toetsen als met de LR-statistiek.
6. Indien het percentage correcte voorspellingen op basis van een logistische regressieanalyse (zoals af te lezen uit de classificatietabel in de SPSS-uitvoer) groot is (bijvoorbeeld, 95%) dan leveren de predictoren een substantiële bijdrage aan de voorspelling.

*Opgave 44* COMPUTEROPDRACHT
Een onderzoeker wil voorspellen of een jongere wel of geen universitair diploma zal behalen. Hij maakt gebruik van twee predictoren: het IQ van de jongere en de socio-economische status (SES) van zijn of haar ouders. Anders dan in het begin van deze paragraaf, beschouwen we SES hier als een *categorische* variabele (met vijf niveaus). Dit is niet vreemd, want ook in de literatuur is er geen eensgezindheid over het schaalniveau van deze variabele; sommige auteurs beschouwen SES als een kwantitatieve variabele (zoals H&M, de auteurs van "The Bell Curve") en anderen beschouwen het als een categorische. Net zoals elke discussie over het schaalniveau van een operationele definitie, geldt ook hier dat óf men een variabele als een kwantitatieve variabele kan beschouwen of niet, dit een zaak is van aannemelijkheid en niet van harde evidentie. Als men SES als een categorische variabele beschouwt, dan heeft men het voordeel dat er geen restrictie ligt op de vorm van het verband tussen deze variabele en de kansen $\widehat{\pi}_i$; indien men SES als een kwantitatieve variabele beschouwt, dan moet dit verband monotoon (stijgend of dalend) zijn.

De onderzoeker is geïnteresseerd in de vraag of de socio-economische status van de ouders een bijdrage levert aan de voorspelling van het criterium bovenop het IQ. Dit criterium is het

wel of niet behalen van een universitair diploma, en wordt hier aangeduid met UNIVDIPL. De data van dit onderzoek vind je in het bestand "diploma.sav".

1. Toets de nulhypothese dat SES geen bijdrage levert aan de voorspelling van UNIVDIPL bovenop IQ. Men kan deze nulhypothese op twee manieren toetsen: met een LR-statistiek en met een multivariate Waldstatistiek. Doe het op beide manieren. Wat is je conclusie?
2. Wat is de relatie tussen SES en UNIVDIPL? Baseer je antwoord op de groepsspecifieke intercepten.
3. Bereken de PRE van de volgende modellen:
   (a) Het model met alleen IQ.
   (b) Het model met alleen SES.
   (c) Het model met IQ én SES.
4. Bereken $\text{PRE}_{\text{SES}}$ bovenop IQ.
5. Wat is, voor een jongere met een IQ van 110, de kans om een universitair diploma te behalen als
   (a) de SES van zijn ouders gelijk is aan 1
   (b) de SES van zijn ouders gelijk is aan 3
   (c) de SES van zijn ouders gelijk is aan 5.

   Gebruik voor het berekenen hiervan een zakrekenmachine.
6. Bereken dezelfde kansen als in de vorige opgave maar nu voor een jongere met een IQ van 125. Vergelijk deze kansen met de kansen uit de vorige opgave?
7. Gebruik de kansen die je in de vorige opgaven berekend hebt om het effect van IQ te evalueren. Doe dit door een jongere met een IQ van 110 te vergelijken met een jongere met een IQ van 125. Bereken het verschil tussen de kansen van de twee jongeren op het behalen van een universitair diploma. Doe dit voor een SES gelijk aan 1, een SES gelijk aan 3, en een SES gelijk aan 5.

*Opgave 45*
Een klinisch psycholoog is geïnteresseerd in het effect van verschillende behandelingen van alcoholisme. De onafhankelijke variabele in dit onderzoek is de variabele THERAPIE die de volgende drie niveaus heeft: (a) deelnemen aan bijeenkomsten van de Anonieme Alcoholisten, (b) cognitieve gedragstherapie, en (c) een behandeling met medicijnen. De afhankelijke variabele is het feit dat men twee jaar na de start van de behandeling nog steeds geen alcohol drinkt of dat men hervallen is (aangeduid met, respectievelijk, 1 en 0). Deze afhankelijke variabele wordt ABSTINEN (van *abstinentie*, het zich onthouden van) genoemd. De psycholoog doet een logistische regressieanalyse met als predictor THERAPIE en als criterium ABSTINEN. In Tabel 4.56 en Tabel 4.57 staat een deel van SPSS-uitvoer.

Omnibus Tests of Model Coefficients

| | | Chi-square | df | Sig. |
|---|---|---|---|---|
| Step 1 | Step | 14.651 | 2 | .001 |
| | Block | 14.651 | 2 | .001 |
| | Model | 19.496 | 3 | .000 |

Tabel 4.56: Tabel bij opgave 45.

Variables in the Equation

| | B | S.E. | Wald | df | Sig. |
|---|---|---|---|---|---|
| THERAPIE | | | 14.735 | 2 | .001 |
| THERAPIE(1) | 1.253 | .362 | 11.985 | 1 | .001 |
| THERAPIE(2) | .245 | .405 | .367 | 1 | .545 |
| Constant | -.405 | .289 | 1.973 | 1 | .160 |

Tabel 4.57: Tabel bij opgave 45.

1. Welke statistiek staat achter *Block* in Tabel 4.56 en welke nulhypothese wordt ermee getoetst?
2. Welke conclusie kan men trekken uit dit onderzoek over het effect van THERAPIE op basis van de gegevens in Tabel 4.57? Waarop baseer je je conclusie?

Volgens een collega van de onderzoeker is het niet de inhoud van de gevolgde therapie die ertoe doet, maar wel het aantal bijeenkomsten waarop men aanwezig is geweest (aantal bijeenkomsten van de Anonieme Alcoholisten, aantal gesprekken met de gedragstherapeut, aantal gesprekken met de arts/psychiater). Dit aantal bijeenkomsten (aangeduid met BIJEENKO) wordt als extra predictor in de logistische regressieanalyse opgenomen. De uitvoer van deze analyse wordt getoond in Tabel 4.58.

Variables in the Equation

| | B | S.E. | Wald | df | Sig. |
|---|---|---|---|---|---|
| BIJEENKO | -.056 | .028 | 4.038 | 1 | .044 |
| THERAPIE | | | 14.050 | 2 | .001 |
| THERAPIE(1) | 1.251 | .366 | 11.684 | 1 | .001 |
| THERAPIE(2) | .280 | .410 | .467 | 1 | .494 |
| Constant | .044 | .365 | .015 | 1 | .903 |

Tabel 4.58: Tabel bij opgave 45.

3. Op basis van de gegevens in Tabel 4.58, denk je dat de collega van de onderzoeker gelijk heeft? Motiveer je antwoord.
4. Wat is de kans voor iemand die behandeld is met medicijnen en waarvan het aantal gevolgde bijeenkomsten één standaarddeviatie boven het gemiddelde ligt, dat hij of zij na twee jaar nog steeds geen alcohol drinkt? Het gemiddelde aantal bijeenkomsten is 8.155 en de standaarddeviatie van het aantal bijeenkomsten is 5.43. Hou rekening met het feit dat er voor het laatste niveau van een categorische predictor (hier, de medicamenteuze behandeling) geen parameter in de tabel met coëfficiënten staat.
5. Stel dat je geïnteresseerd bent in de proportionele reductie in de kwadratische fout door THERAPIE op te nemen bovenop BIJEENKO. Welke getallen moet je SPSS laten uitrekenen als je deze proportionele reductie in kwadratische fout wilt berekenen?

*Opgave 46* COMPUTEROPDRACHT

In deze opgave heranalyseren we weer de data van het onderzoek naar de mentale ontwikkeling bij kinderen met een laag geboortegewicht. De onderzoekers wilden weten of men kan voorspellen of een kind in het Bijzonder Onderwijs terecht komt of niet. De data van dit deel van het onderzoek vindt men in het bestand “bijzond.sav”. De criteriumvariabele, aangeduid als

BIJZOND, heeft de waarde 1 als het kind op het einde van het tweede leerjaar in het Bijzonder Onderwijs zit, en de waarde 0 als dat niet zo is. Er zijn twee kwantitatieve predictoren, gewicht en hospitalisatieduur, aangeduid als GEWICHT en HOSPDUUR. Er is één categorische predictor, het opleidingsniveau van de ouders, aangeduid als OPLNIV. Het opleidingsniveau van de ouders werd ingedeeld in drie categorieën: geen van beide ouders heeft hoger onderwijs gevolgd (OPLNIV=1), één van beide ouders heeft een diploma van het niet-universitair hoger onderwijs (OPLNIV=2), of één van beide ouders heeft een universitair diploma (OPLNIV=3).

1. Levert GEWICHT een statistisch significante bijdrage aan de voorspelling van BIJZOND?
2. Levert GEWICHT een statistisch significante bijdrage aan de voorspelling van BIJZOND bovenop HOSPDUUR?
3. Wat is de kans voor een baby met een gemiddeld geboortegewicht (binnen deze groep baby's met een laag geboortegewicht) om in het Bijzonder Onderwijs terecht te komen als de hospitalisatieduur van de baby
   (a) één standaarddeviatie onder de gemiddelde hospitalisatieduur ligt?
   (b) één standaarddeviatie boven de gemiddelde hospitalisatieduur ligt?
4. Levert OPLNIV een statistisch significante bijdrage aan de voorspelling van BIJZOND bovenop HOSPDUUR en GEWICHT?
5. Bereken de PRE voor de voorspelling op basis van het model met HOSPDUUR, GEWICHT, en OPLNIV.
6. Bereken $\text{PRE}_{\text{OPLNIV}}$ bovenop HOSPDUUR en GEWICHT.
7. Hoeveel stijgt het percentage correcte voorspellingen door het gebruik van OPLNIV als predictor naast HOSPDUUR en GEWICHT?

*Opgave 47* ZELFSTUDIEOPDRACHT
Zelden heeft een boek zoveel opschudding veroorzaakt als "The Bell Curve" (TBC) van R. J. Herrnstein en C. Murray (1994). Het centrale begrip in TBC is intelligentie. Ten behoeve van hun onderzoek stelden H&M intelligentie gelijk aan de score op een IQ-test, en daarom wordt deze variabele hier ook aangeduid met IQ. De centrale redenering in TBC werd door Taylor (1995) samengevat in het volgende syllogisme:

1. Eerste premisse: IQ is voor het grootste gedeelte erfelijk bepaald.
2. Tweede premisse: IQ is positief gecorreleerd met tal van indicatoren van maatschappelijk succes, zoals een prestigieuze baan, een hoog inkomen en een hoog opleidingsniveau; en het is negatief gecorreleerd met tal van indicatoren van maatschappelijk falen, zoals criminele activiteiten.
3. Conclusie: Bijgevolg is maatschappelijk succes (en maatschappelijk falen) voor het grootste gedeelte erfelijk bepaald.

In het kielzog van deze centrale redenering, verdedigden H&M ook de stelling dat, omwille van de sterke erfelijke bepaling van IQ, deze weinig of niet beïnvloed kan worden door onderwijskundige en andere interventies. H&M beweerden ook dat de Amerikaanse samenleving evolueert naar een samenleving van cognitieve kasten, waarin de cognitief zwakken in de lage kasten terecht komen en, omwille van de sterke erfelijke bepaling van IQ, er ook niet uitkomen. Deze lage kasten bevatten een groot gedeelte van de Afro-Amerikaanse populatie. H&M beweerden dat de grote hoeveelheid geld en energie die gespendeerd wordt aan het wegwerken van deze ongelijkheid, slecht besteed is, omdat IQ toch voor het grootste gedeelte erfelijk bepaald is.

Tenslotte, beweerden H&M ook dat het genetisch kapitaal dat bepalend is voor het IQ, geleidelijk aan uitgeput raakt, omdat mensen met een laag IQ gemiddeld meer kinderen hebben dan mensen met een hoog IQ.

In deze zelfstudieopdracht beperken we ons tot de volgende drie indicatoren van maatschappelijk falen: onder de armoedegrens terechtkomen, het vroegtijdig verlaten van de middelbare school, en werkloos worden. Dit is slechts een deel van de variabelen die in TBC aan bod komen, maar ze zijn representatief voor zowel het type analyses waarover gerapporteerd wordt in TBC als de resultaten van deze analyses.

Deze zelfstudieopdracht omvat de volgende activiteiten: (1) bestuderen van relevante gedeelten uit TBC en daaraan gerelateerde literatuur, (2) heranalyseren van de data waarop H&M hun conclusies baseerden, en (3) conclusies trekken op basis van deze heranalyses. Voor wat betreft de eerste activiteit, de literatuurstudie, moet een onderscheid gemaakt worden tussen de volgende teksten:

1. Inleidingen. TBC begint met een algemene inleiding, *Preface* genoemd, en ook elk van de vier delen (*Part I* tot en met *Part IV*) heeft een eigen inleiding (zonder kopje). Deze zelfstudieopdracht gaat over Part II en Part III, en daarom moet men, naast de algemene inleiding, de inleidingen tot deze twee delen lezen.

2. Samenvattingen. Elk deel en elk hoofdstuk van TBC begint met een samenvatting. Deze samenvattingen zijn gemakkelijk te herkennen aan het schuine lettertype. Dit is een snelle manier om te weten te komen waar een deel of een hoofdstuk over gaat.

3. Hoofdstukken. Deze zelfstudieopdracht bestaat uit twee delen. Het eerste deel gaat over de samenhang tussen, enerzijds, de drie indicatoren van maatschappelijk falen, en anderzijds, de variabelen IQ, SES, opleidingsniveau en leeftijd. De heranalyses op deze variabelen hebben betrekking op het eerste gedeelte van Hoofdstuk 5 (tot bovenaan p. 137), het eerste gedeelte van Hoofdstuk 6 (tot aan p. 150) en Hoofdstuk 7 (vooral de paragraaf *Unemployment*). Dit moet men lezen.

   Het tweede deel van de zelfstudieopdracht gaat over raciale verschillen in IQ en de mogelijke gevolgen hiervan voor allerlei indicatoren van maatschappelijk succes/falen. De heranalyses hebben betrekking op het eerste gedeelte van Hoofdstuk 14 (tot bovenaan p. 329). Dit moet men lezen. Voordat men begint met Hoofdstuk 14, is het goed om de samenvatting van Hoofdstuk 13 te lezen.

4. Appendices. Voordat men met de heranalyses begint, is het nodig om gedetailleerde informatie te hebben over de variabelen die in deze analyses gebruikt worden. Deze informatie vindt men in de appendices. In Appendix 2 wordt bijvoorbeeld uitgelegd hoe de variabelen IQ, SES en opleidingsniveau (in het Engels, *educational attainment*) bepaald werden.

   In een deel van Appendix 3 (namelijk, pp. 588-591) presenteerden H&M evidentie voor hun stelling dat IQ niet of nauwelijks bepaald wordt door het aantal jaren opleiding (de operationele definitie van opleidingsniveau, en verder aangeduid als JARENOPL), maar dat het omgekeerde wel geldt, namelijk dat JARENOPL in sterke mate bepaald wordt door IQ. Deze stelling vormt de achtergrond van de beslissing van H&M om IQ, SES en JARENOPL nooit samen als predictor te gebruiken. Als je beslist om dit wel te doen –en daar kun je goede redenen voor hebben– moet je dit stuk van de appendix zeker lezen.

   In Appendix 4, tenslotte, vindt men de resultaten van een aantal analyses die men ook zelf moet doen. Aan de hand van deze resultaten kan men controleren of men alles goed gedaan heeft. Let wel, de resultaten in deze appendix zijn slechts een deel van de resultaten die men zelf moet produceren; alleen de resultaten van de basisanalyse worden in de appendix gegeven.

5. Kritische beoordeling. Devlin, Fienberg, Resnick en Roeder (1995) schreven een kritische beoordeling (in het Engels, *review*, een term die ook hier gebruikt zal worden) van TBC. Dat was ook wel nodig, want H&M trokken verregaande conclusies op basis van hun onderzoek. Deze review is slechts één van de vele (er zijn zelfs twee boeken geschreven als reactie op TBC), maar hij blinkt uit in bondigheid. Deze review moet je lezen. Vooral paragraaf 5 (*The NLSY Study*) is van belang voor de heranalyses.

De data die men nodig heeft voor de heranalyses zijn opgeslagen in "Bellcurve.sav". Dit is slechts een deel van de data die H&M gebruikten. Het volledige databestand, opgeslagen in "nation.sav", is afkomstig van de *National Longitudinal Study of Youth*, een enquête van Amerikaanse jongeren tussen 25 en 33 jaar. Dit volledige databestand is echter minder representatief voor de Amerikaanse samenleving dan het databestand in "Bellcurve.sav". Dit komt omdat bepaalde sociale klassen oververtegenwoordigd waren in het volledige databestand. H&M losten dit probleem op door te werken met een *gewogen* logistische regressieanalyse, die corrigeert voor deze oververtegenwoordiging. Omdat gewogen regressieanalyse niet behandeld wordt in dit boek, beperken we ons tot de representatieve steekproef in "Bellcurve.sav". De tabellen in de eigenlijke tekst van TBC (níet de tabellen in Appendix 4) bevatten het resultaat van de gewogen logistische regressieanalyses op het databestand in "nation.sav", en wijken dus een beetje af van de resultaten van de gewone logistische regressieanalyses op het databestand in "Bellcurve.sav". In Appendix 4, rapporteren H&M echter ook de resultaten van gewone logistische regressieanalyses op de representatieve steekproef in "Bellcurve.sav". Deze resultaten kan men gebruiken om te controleren of men alles goed gedaan heeft.

De namen van de variabelen in "Bellcurve.sav" worden gegeven in Tabel 4.59, samen met een korte beschrijving. Dezelfde informatie, maar dan voor het volledige databestand in "nation.sav", kan men vinden in het bestand "nationlogbook.doc". Dit is handig voor diegenen die heranalyses willen uitvoeren op variabelen die niet aan bod komen in deze zelfstudieopdracht.

**Heranalyses Deel 1**

Het eerste deel van de zelfstudie gaat over de samenhang tussen, enerzijds, de drie indicatoren van maatschappelijk falen (onder de armoedegrens terechtkomen, het vroegtijdig verlaten van de middelbare school, en werkloos worden), en anderzijds, de variabelen IQ, SES, JARENOPL en LEEFTIJD. Voor elke indicator van maatschappelijk falen moeten de volgende opdrachten uitgevoerd worden:

1. Onderzoek of IQ een statistisch significante predictor is bovenop de variabelen SES en LEEFTIJD.
2. Onderzoek of IQ een belangrijke predictor is bovenop de variabelen SES en LEEFTIJD. Maak hierbij gebruik van de effectgroottes van deze variabelen en de proportionele reductie in kwadratische fout waarvoor deze predictor verantwoordelijk is.
3. H&M hebben ervoor gekozen om IQ en SES niet samen met JARENOPL (de operationele definitie van opleidingsniveau) in de regressieanalyse op te nemen. In de inleiding tot *Part II* geven H&M hun redenen om dit niet te doen. Tegen de achtergrond van deze redenen, staat de stelling van H&M dat IQ een stabiele eigenschap is die slechts in beperkte mate beïnvloed wordt door JARENOPL. In Appendix 3 geven H&M enkele argumenten voor deze stelling.
   (a) Ben je eens met de beslissing van H&M om het aantal jaren opleiding niet als predictor in de regressieanalyse op te nemen? Motiveer je standpunt.
   (b) Is het zinvol om JARENOPL als extra predictor te gebruiken bij elk van de drie indicatoren? Motiveer je standpunt.

| Variabele | Beschrijving |
|---|---|
| z_age | Leeftijd (gestandaardiseerd). Oorspronkelijke naam: zAge. |
| z_ses | Socio-economische status van de ouders (gestandaardiseerd). Oorspronkelijke naam: zSES. |
| z_afqt | Score op de AFQT, een intelligentietest (gestandaardiseerd). Oorspronkelijke naam: zAFQT89. |
| z_educ | Aantal jaren opleiding (gestandaardiseerd). Oorspronkelijke naam: zEd90. |
| race | Ras, met de niveaus blank (*White*), zwart (*Black*) en Latijns-Amerikaans (*Latino*). Oorspronkelijke naam: Race4. |
| unemp | Dichotome variabele die aangeeft of iemand langer dan vier weken werkloos was in 1989 (unemp=1) of niet (unemp=0). Oorspronkelijke naam: 1MoUnemp89. |
| dropout | Dichotome variabele die aangeeft of iemand de middelbare school voortijdig verlaten heeft (dropout=1) of niet (dropout=0). Berekend op basis van de oorspronkelijke variabele *FinDegree*. |
| poverty | Dichotome variabele die aangeeft of iemand onder de armoedegrens leefde in 1989 (poverty=1) of niet (poverty=0). Oorspronkelijke naam: Pov89. |
| sex | Geslacht. Oorspronkelijke naam: Sex. |
| unempsam | Dichotome variabele die aangeeft of iemand gedurende het volledige jaar werk had of werk zocht (unempsam=1) of niet (unempsam=0), waarmee onder andere studenten en dienstplichtigen uitgesloten kunnen worden van de analyse. Oorspronkelijke naam: UnempSamp. |
| empschl | Dichotome variabele die aangeeft of iemand de school reeds verlaten heeft (empschl=1) of nog niet (empschl=0). Oorspronkelijke naam: Emp<>Schl. |
| gedvhsgr | Variabele die aangeeft of iemand een middelbare schooldiploma op de gewone manier behaald heeft (gedvhsgr=HSGrad) of via een andere weg, zoals het Staatsexamen in Nederland en de Middenjury in Vlaanderen (gedvhsgr=GED). Oorspronkelijke naam: GEDvHSGrad. |

Tabel 4.59: Naam en korte beschrijving van de variabelen in “Bellcurve.sav”.

(c) Laat zien, met behulp van logistische regressieanalyses, wat het gevolg is van het opnemen van JARENOPL als predictor. Op basis van deze analyses moet je de volgende vragen kunnen beantwoorden:

i. Wat vind je van de stelling dat, bovenop JARENOPL, IQ er niet meer toe doet?

ii. Wat vind je van de stelling dat, bovenop IQ, JARENOPL er niet meer toe doet?

Bij elke analyse, neem ook SES en LEEFTIJD op als predictoren.

Bij het uitvoeren van de analyses moet je observaties selecteren. Binnen SPSS, gebeurt het selecteren van observaties met de optie *Select Cases* onder het menu *Data*. Als men deze optie kiest, krijgt men een keuzevenster waarin men de optie *If condition is satisfied* moet kiezen, zodat men via de knop *If ...* de selectierestrictie kan invullen. Alle analyses gebeuren op de gegevens van de blanke proefpersonen. Daarnaast zijn er ook nog selectiecriteria die specifiek zijn voor de drie indicatoren van maatschappelijk falen:

1. Onder de armoedegrens leven (*poverty*). Voer de analyse uit op blanken die niet meer studeren. Dus, de restrictie is de volgende: *race='White' & empschl=1*.
2. Vroegtijdig de middelbare school verlaten (*dropout*). Voer de analyse uit op blanken die geen middelbare schooldiploma hebben behaald via het Staatsexamen of de Middenjury. Dus, de restrictie is de volgende: *race='White' & gedvhsgr<>'GED'*.
3. Werkloos zijn (*unemp*). Voer de analyse uit op de blanke mannen die werk zochten. Dus, de restrictie is de volgende: *race='White' & sex='Male' & unempsam=1*.

**Heranalyses Deel 2**

Het tweede deel van de zelfstudie gaat over verschillen tussen de rassen (blank, zwart en Latijns-Amerikaans) wat betreft de drie indicatoren van maatschappelijk falen die ook in het eerste deel centraal staan. Het is een onomstreden feit dat er raciale verschillen zijn wat betreft de verschillende indicatoren van maatschappelijk succes/falen: blanken doen het duidelijk beter dan zwarten, en de Latijns-Amerikanen zitten daar tussenin. H&M gaan uit van een verband tussen IQ en de verschillende indicatoren van maatschappelijk succes/falen, en vragen zich dan af in hoeverre deze raciale verschillen te verklaren zijn (in de betekenis van "gemedieerd worden") door verschillen in IQ. We zijn dus geïnteresseerd in uitgezuiverde raciale verschillen. Voor elke indicator van maatschappelijk falen moeten de volgende opdrachten uitgevoerd worden:

1. Onderzoek of er raciale verschillen zijn wat betreft de gegeven indicator van maatschappelijk falen?
2. Onderzoek of deze raciale verschillen gemedieerd worden door IQ en LEEFTIJD.
3. Onderzoek of deze raciale verschillen gemedieerd worden door IQ, SES en LEEFTIJD. Is er een belangrijk verschil met het resultaat van de vorige analyse?
4. Onderzoek of RAS een belangrijke predictor is bovenop IQ SES en LEEFTIJD. Maak hierbij gebruik van de effectgrootte van RAS en de proportionele reductie in kwadratische fout waarvoor deze predictor verantwoordelijk is.
5. (Alleen voor werkloosheid en armoede.) Herhaal de vorige drie analyses, maar nu met JARENOPL in plaats van IQ. Wat is de beste mediërende variabele voor het verband tussen RAS en de gegeven indicator van maatschappelijk falen, IQ of JARENOPL?
6. Onderzoek of het zinvol is om analyses uit te voeren met rasspecifieke logistische regressiegewichten. Doe dit voor IQ en SES. Wat betekent een eventuele significante interactie?

Net zoals bij de analyses voor het eerste deel van deze zelfstudieopdracht, moet men ook nu observaties selecteren. Anders dan bij het eerste deel, worden nu níet de blanke proefpersonen geselecteerd. De selectiecriteria zijn specifiek voor de drie indicatoren van maatschappelijk falen:

1. Onder de armoedegrens leven (*poverty*). Voer de analyse uit op de proefpersonen die niet meer studeren. Dus, de restrictie is de volgende: *empschl=1*.
2. Vroegtijdig de middelbare school verlaten (*dropout*). Voer de analyse uit op de proefpersonen die geen middelbare schooldiploma hebben behaald via het Staatsexamen of de Middenjury. Dus, de restrictie is de volgende: *gedvhsgr<>'GED'*.
3. Werkloos zijn (*unemp*). Voer de analyse uit op de mannelijke proefpersonen die werk zochten. Dus, de restrictie is de volgende: *sex='Male' & unempsam=1*.

---

## 4.15 Appendix bij Hoofdstuk 4

### Het Gemiddelde als Beste Voorspeller in de Situatie Zonder Predictoren.

We willen de volgende expressie minimaliseren:

$$\frac{1}{n}\sum_{i=1}^{n}[Y_i - A]^2 \quad .$$

Omdat de factor $\frac{1}{n}$ geen invloed heeft op het resultaat van deze minimalisering, laten we hem verder gewoon weg. Nu passen we het trucje toe om tussen de haakjes $-\overline{Y}+\overline{Y}$ toe te voegen. Dit betekent dat we netto dus niets toevoegen. Dit kunnen we dan als volgt uitwerken:

$$\begin{aligned}\sum_{i=1}^{n}[Y_i - A]^2 &= \sum_{i=1}^{n}[Y_i - \overline{Y} + \overline{Y} - A]^2 \\ &= \sum_{i=1}^{n}[(Y_i - \overline{Y}) + (\overline{Y} - A)]^2 \\ &= \sum_{i=1}^{n}(Y_i - \overline{Y})^2 + \sum_{i=1}^{n}(\overline{Y} - A)^2 + 2\sum_{i=1}^{n}(Y_i - \overline{Y})(\overline{Y} - A) \quad .\end{aligned}$$

De laatste term in deze formule is gelijk aan 0. Dit zien we als we eerst $(\overline{Y} - A)$ voor het somteken brengen en daarna het gedeelte vanaf het somteken als volgt uitwerken:

$$\begin{aligned}\sum_{i=1}^{n}(Y_i - \overline{Y}) &= \sum_{i=1}^{n}Y_i - \sum_{i=1}^{n}\overline{Y} \\ &= n\overline{Y} - n\overline{Y} \\ &= 0 \quad ,\end{aligned}$$

waarin we gebruikmaken van het feit dat $\overline{Y}$ gelijk is aan $\frac{1}{n}\sum_{i=1}^{n} Y_i$.

De te minimaliseren expressie kan nu als volgt geschreven worden:

$$\sum_{i=1}^{n}[Y_i - A]^2 = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 + n(\overline{Y} - A)^2 \quad .$$

Alleen de laatste term hangt af van $A$, en hij wordt geminimaliseerd als we $A$ gelijkstellen aan $\overline{Y}$, het gemiddelde.

## De Variantie van een Variabele Vermenigvuldigd met een Constante

We vertrekken van een variabele $Z$ met variantie $\text{VAR}(Z)$ en een willekeurige constante $c$. We leiden nu een formule af van de variantie van de nieuwe variabele $cZ$ ($c$ keer $Z$). We vertrekken van de definitie van $\text{VAR}(cZ)$, de variantie van de variabele $cZ$:

$$\text{VAR}(cZ) = \frac{1}{n}\sum_{i=1}^{n}(cZ_i - \overline{cZ})^2 \quad .$$

We maken nu gebruik van het feit dat $\overline{cZ}$, het gemiddelde van de variabele $cZ$, gelijk is aan $c\overline{Z}$ ($c$ keer $\overline{Z}$, het gemiddelde van $Z$). Dan kan $\text{VAR}(cZ)$ als volgt uitgewerkt worden:

$$\begin{aligned}\text{VAR}(cZ) &= \frac{1}{n}\sum_{i=1}^{n}(cZ_i - c\overline{Z})^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}c^2(Z_i - \overline{Z})^2 \\ &= c^2\left[\frac{1}{n}\sum_{i=1}^{n}(Z_i - \overline{Z})^2\right] \quad ,\end{aligned}$$

omdat $c^2$ niet afhangt van $i$. Het gedeelte van de formule tussen de vierkante haken is $\text{VAR}(Z)$, de variantie van $Z$. Substitutie van $\text{VAR}(Z)$ levert ons het resultaat van deze afleiding:

$$\text{VAR}(cZ) = c^2\text{VAR}(Z) \quad .$$

Dus, de variantie van $cZ$ is gelijk aan $c^2$ keer de variantie van $Z$.

## Afleiding van de Gestandaardiseerde Enkelvoudige Regressievergelijking en de Gemiddelde Kwadratische Fout

We recapituleren eerst de eigenschappen van standaardscores die in deze afleiding gebruikt gaan worden: standaardscores hebben een gemiddelde gelijk aan 0 en een variantie gelijk aan 1. Dus, bijvoorbeeld, voor $\mathbf{x}$:

$$\overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i = 0$$

en

$$\mathrm{VAR}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})^2 = 1 \quad .$$

Deze laatste uitdrukking kan, vanwege $\overline{\mathbf{x}} = 0$, nog vereenvoudigd worden tot

$$\mathrm{VAR}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^2 = 1 \quad .$$

Geheel analoog gelden dezelfde uitdrukkingen voor de variabele $\mathbf{y}$:

$$\overline{\mathbf{y}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i = 0 \qquad \text{en} \qquad \mathrm{VAR}(\mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i^2 = 1 \quad .$$

Tenslotte kan de correlatie tussen $X$ en $Y$ mooi compact worden uitgedrukt in termen van standaardscores. Per definitie is deze correlatie immers gelijk aan het gemiddelde kruisproduct van de standaardscores:

$$r_{xy} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{y}_i \quad .$$

Hiermee gewapend gaan we nu onderzoeken voor welke keuze van $a$ en $b$ de gemiddelde kwadratische fout $\frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \widehat{\mathbf{y}}_i)^2$ zo klein mogelijk wordt. Daartoe gaan we in deze uitdrukking eerst de vergelijking voor $\widehat{\mathbf{y}}$ substitueren en dan het kwadraat en de sommatie uitwerken:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \widehat{\mathbf{y}}_i)^2 &= \frac{1}{n}\sum_{i=1}^{n}[\mathbf{y}_i - (a + b\,\mathbf{x}_i)]^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}[\mathbf{y}_i^2 - 2(a + b\,\mathbf{x}_i)\mathbf{y}_i + (a + b\,\mathbf{x}_i)^2] \\
&= \frac{1}{n}\sum_{i=1}^{n}[\mathbf{y}_i^2 - 2a\mathbf{y}_i - 2b\mathbf{x}_i\mathbf{y}_i + a^2 + 2ab\mathbf{x}_i + b^2\mathbf{x}_i^2] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i^2 - 2a\frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i - 2b\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{y}_i + \frac{1}{n}na^2 \\
&\quad + 2ab\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i + b^2\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^2 \quad .
\end{aligned}$$

In deze laatste formule staan een aantal bekende grootheden:

1. $\frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i$ en $\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$ zijn het gemiddelde van, respectievelijk, $\mathbf{y}$ en $\mathbf{x}$. Ze worden verder aangeduid met $\overline{\mathbf{y}}$ en $\overline{\mathbf{x}}$. Omdat $\mathbf{y}$ en $\mathbf{x}$ gestandaardiseerde variabelen zijn, zijn deze gemiddelden gelijk aan 0.

2. $\frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i^2$ en $\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^2$ zijn de variantie van, respectievelijk, $\mathbf{y}$ en $\mathbf{x}$. Ze worden verder aangeduid met $\mathrm{VAR}(\mathbf{y})$ en $\mathrm{VAR}(\mathbf{x})$. Omdat $\mathbf{y}$ en $\mathbf{x}$ gestandaardiseerde variabelen zijn, zijn deze varianties gelijk aan 1.

3. $\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{y}_i$ is de covariantie tussen $\mathbf{x}$ en $\mathbf{y}$. Omdat $\mathbf{y}$ en $\mathbf{x}$ gestandaardiseerde variabelen zijn, is deze covariantie gelijk aan de correlatie tussen $X$ en $Y$, verder aangeduid met $r_{xy}$. Op pagina 122 in de appendix van hoofdstuk 2 wordt wat uitleg gegeven over de covariantie en over zijn relatie met de correlatie.

We kunnen nu de gemiddelde kwadratische fout als volgt herschrijven:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \widehat{\mathbf{y}}_i)^2 &= \mathrm{VAR}(\mathbf{y}) - 2a\overline{\mathbf{y}} - 2br_{xy} + a^2 + 2ab\overline{\mathbf{x}} + b^2\mathrm{VAR}(\mathbf{x}) \\ &= 1 - 2a\cdot 0 - 2br_{xy} + a^2 + 2ab\cdot 0 + b^2\cdot 1 \\ &= 1 - 2br_{xy} + a^2 + b^2 \quad .\end{aligned}$$

Nu volgt het trucje om in de laatste uitdrukking $-r_{xy}^2 + r_{xy}^2$ toe te voegen met als enig doel om daar een volledig kwadraat te vormen. Netto voegen we dus niets toe. Dus:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \widehat{\mathbf{y}}_i)^2 &= 1 - r_{xy}^2 + r_{xy}^2 - 2br_{xy} + b^2 + a^2 \\ &= 1 - r_{xy}^2 + (r_{xy} - b)^2 + a^2 \quad .\end{aligned}$$

Nu is duidelijk voor welke $a$ en $b$ deze uitdrukking minimaal wordt: $a$ komt alleen maar voor in de term $a^2$, terwijl $b$ alleen maar voorkomt in de term $(r_{xy} - b)^2$. Allebei deze termen zijn kwadraten, dus groter dan nul tenzij het getal waarvan het kwadraat genomen wordt gelijk is aan nul, want dan is ook het kwadraat gelijk aan nul. De uitdrukking wordt dus minimaal als (1) $a^2 = 0$, dit wil zeggen $a = 0$, en (2) $(r_{xy} - b)^2 = 0$, dit wil zeggen $r_{xy} - b = 0$, oftewel $b = r_{xy}$. Zo volgt uit het kleinste-kwadratencriterium de gestandaardiseerde regressievergelijking:

$$\widehat{\mathbf{y}} = r_{xy}\mathbf{x} \quad .$$

De formule voor de gemiddelde kwadratische fout is een onmiddellijke toegift van de bovenstaande afleiding van de gestandaardiseerde regressievergelijking. Daar hadden we gevonden

$$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \widehat{\mathbf{y}}_i)^2 = 1 - r_{xy}^2 + (r_{xy} - b)^2 + a^2 \quad .$$

Hierin moeten we nu slechts de optimale keuzes $a = 0$ en $b = r_{xy}$ invullen. Dus,

$$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \widehat{\mathbf{y}}_i)^2 = 1 - r_{xy}^2 \quad .$$

## Afleiding van de Ongestandaardiseerde Enkelvoudige Regressievergelijking en de Gemiddelde Kwadratische Fout

Deze afleiding bestaat erin dat we de gestandaardiseerde regressievergelijking $\widehat{\mathbf{y}} = r_{xy}\mathbf{x}$ en de grootte van de bijbehorende gemiddelde kwadratische fout terugvertalen naar de

originele scores $X$ en $Y$ door overal $\mathbf{x}$ en $\widehat{\mathbf{y}}$ herschrijven als een functie van de corresponderende originele scores. Het vertrekpunt bij de herschrijfoperatie is dat de voorspelde standaardscores $\widehat{\mathbf{y}}$ als volgt gerelateerd zijn aan de voorspelde originele scores $\widehat{Y}$:

$$\widehat{\mathbf{y}} = \frac{\widehat{Y} - \overline{Y}}{\mathrm{SD}(Y)} \quad .$$

Van deze formule, die $\widehat{\mathbf{y}}$ uitdrukt in termen van $\widehat{Y}$,

kunnen we een andere formule afleiden die, omgekeerd, $\widehat{Y}$ uitdrukt in termen van $\widehat{\mathbf{y}}$:

$$\widehat{Y} = \widehat{\mathbf{y}}\mathrm{SD}(Y) + \overline{Y} \quad .$$

Nu substitueren we eerst $\widehat{\mathbf{y}}$ door $r_{xy}\mathbf{x}$ en daarna $\mathbf{x}$ door haar definitie in termen van de ongestandaardiseerde variabele $X$:

$$\begin{aligned}\widehat{Y} =& r_{xy}\mathbf{x}\mathrm{SD}(Y) + \overline{Y} \\ =& r_{xy}\left[\frac{X - \overline{X}}{\mathrm{SD}(X)}\right]\mathrm{SD}(Y) + \overline{Y} \quad .\end{aligned}$$

Door $\mathrm{SD}(Y)$ binnen de haakjes te brengen, kan deze formule als volgt geschreven worden:

$$\widehat{Y} = \left[\overline{Y} - r_{xy}\frac{\mathrm{SD}(Y)}{\mathrm{SD}(X)}\overline{X}\right] + \left[r_{xy}\frac{\mathrm{SD}(Y)}{\mathrm{SD}(X)}\right]X \quad .$$

Dus voor de ongestandaardiseerde regressievergelijking

$$\widehat{Y} = A + B\,X \quad ,$$

vinden we als oplossing:

$$\begin{aligned}B =& r_{xy}\frac{\mathrm{SD}(Y)}{\mathrm{SD}(X)} \quad , \\ A =& \overline{Y} - B\,\overline{X} \quad ,\end{aligned}$$

het regressiegewicht en de regressieconstante van de ongestandaardiseerde regressievergelijking.

Ook de gemiddelde kwadratische fout voor de originele scores $Y$ is te halen uit het resultaat voor de standaardscores. We hadden

$$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \widehat{\mathbf{y}}_i)^2 = 1 - r_{xy}^2 \quad ,$$

en we maken gebruik van het volgende:

$$\mathbf{y}_i - \widehat{\mathbf{y}}_i = \frac{Y_i - \overline{Y}}{\mathrm{SD}(Y)} - \frac{\widehat{Y}_i - \overline{Y}}{\mathrm{SD}(Y)} = \frac{Y_i - \widehat{Y}_i}{\mathrm{SD}(Y)} \quad .$$

Dan kan de gemiddelde kwadratische fout als volgt geschreven worden:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{(Y_i-\widehat{Y}_i)^2}{\mathrm{SD}(Y)^2}=1-r_{xy}^2 \quad .$$

Door nu links en rechts te vermenigvuldigen met $\mathrm{SD}(Y)^2=\mathrm{VAR}(Y)$ krijgen we:

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i-\widehat{Y}_i)^2=(1-r_{xy}^2)\mathrm{VAR}(Y) \quad ,$$

de gemiddelde kwadratische fout voor de originele scores $Y$.

## Afleiding van de Formule van de Variantie van een Som van Variabelen

We willen de variantie kennen van de variabele $Y$ die geschreven kan worden als de som van de variabelen $U$ en $V$:

$$Y=U+V \quad .$$

We vertrekken van de formule voor $\mathrm{VAR}(Y)$:

$$\mathrm{VAR}(Y)=\frac{1}{n}\sum_{i=1}^{n}[Y_i-\overline{Y}]^2 \quad .$$

In deze formule vervangen we $Y_i$ door $(U_i+V_i)$ en $\overline{Y}$ door $(\overline{U}+\overline{V})$:

$$\mathrm{VAR}(Y)=\frac{1}{n}\sum_{i=1}^{n}[(U_i+V_i)-(\overline{U}+\overline{V})]^2 \quad .$$

Dit kunnen we als volgt uitwerken:

$$\begin{aligned}\mathrm{VAR}(Y)&=\frac{1}{n}\sum_{i=1}^{n}[(U_i-\overline{U})+(V_i-\overline{V})]^2\\&=\frac{1}{n}\sum_{i=1}^{n}[(U_i-\overline{U})^2+(V_i-\overline{V})^2+2(U_i-\overline{U})(V_i-\overline{V})]\\&=\frac{1}{n}\sum_{i=1}^{n}(U_i-\overline{U})^2+\frac{1}{n}\sum_{i=1}^{n}(V_i-\overline{V})^2+2\frac{1}{n}\sum_{i=1}^{n}(U_i-\overline{U})(V_i-\overline{V}) \quad .\end{aligned}$$

Rechts van het laatste gelijkheidsteken herkennen we achtereenvolgens $\mathrm{VAR}(U)$, $\mathrm{VAR}(V)$ en $2\mathrm{COV}(U,V)$. Dus,

$$\mathrm{VAR}(Y)=\mathrm{VAR}(U)+\mathrm{VAR}(V)+2\mathrm{COV}(U,V) \quad .$$

Als de correlatie tussen $U$ en $V$ (aangeduid met $r_{UV}$) gelijk is aan 0, dan is ook de covariantie tussen $U$ en $V$ gelijk aan 0 ($r_{UV}$ is immers gelijk aan $\mathrm{COV}(U,V)$ gedeeld door $\mathrm{SD}(U)\mathrm{SD}(V)$, het product van de standaarddeviaties van $U$ en $V$). In dit geval, is $\mathrm{VAR}(Y)$ gelijk aan de som van $\mathrm{VAR}(U)$ en $\mathrm{VAR}(V)$:

$$\mathrm{VAR}(Y)=\mathrm{VAR}(U)+\mathrm{VAR}(V) \quad .$$

## De Gelijkheid van VAR($Y - \widehat{Y}$) en de Gemiddelde Kwadratische Fout

De variantie van een variabele $Z$ wordt als volgt berekend:

$$\text{VAR}(Z) = \frac{1}{n}\sum_{i=1}^{n}(Z_i - \overline{Z})^2 \quad .$$

Nu nemen we $Z$ gelijk aan $Y - \widehat{Y}$. Om $\text{VAR}(Y - \widehat{Y})$ te bepalen, moeten we het gemiddelde van $Y - \widehat{Y}$ kennen. Dit gemiddelde duiden we aan met $\overline{Y - \widehat{Y}}$. We kunnen laten zien dat dit gemiddelde gelijk is aan 0:

$$\begin{aligned}
\overline{Y - \widehat{Y}} &= \overline{Y} - \overline{\widehat{Y}} \\
&= \overline{Y} - (\overline{A + BX}) \\
&= \overline{Y} - A - B\overline{X} \\
&= \overline{Y} - (\overline{Y} - B\overline{X}) - B\overline{X} \\
&= \overline{Y} - \overline{Y} + B\overline{X} - B\overline{X} \\
&= 0 \quad .
\end{aligned}$$

We kunnen $\text{VAR}(Y - \widehat{Y})$ nu als volgt schrijven:

$$\text{VAR}(Y - \widehat{Y}) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 \quad .$$

Het rechterlid van deze formule is precies de gemiddelde kwadratische fout.

## Instabiele Bètagewichten en Stabiele Voorspelling

Het wekt misschien verbazing dat, als de bètagewichten instabiel zijn, de predictoren samen toch behoorlijk wat variantie kunnen verklaren. De proportie verklaarde variantie is immers gelijk aan $r^2_{\mathbf{y}\widehat{\mathbf{y}}}$, en indien instabiele bètagewichten gepaard zouden gaan met instabiele voorspelde criteriumscores $\widehat{\mathbf{y}}$, dan zou $r^2_{\mathbf{y}\widehat{\mathbf{y}}}$ klein moeten zijn. Dit klopt echter niet want, ondanks de instabiele bètagewichten, hoeven de voorspelde criteriumscores $\widehat{\mathbf{y}}$ helemaal niet instabiel te zijn. Dit kunnen we illustreren aan de hand van het getallenvoorbeeld in Tabel 4.11 op pagina 349. Voor het gemak, beperken we ons tot gestandaardiseerde variabelen.

1. Stel dat we $\mathbf{y}$ voorspellen op basis van alleen $\mathbf{x}_1$. De voorspelde criteriumscores duiden we aan met $\mathbf{y}^{(1)}$. Hierin geeft het superscript $^{(1)}$ aan dat $\mathbf{y}$ voorspeld wordt op basis van de eerste predictor. De voorspelde criteriumscores $\mathbf{y}^{(1)}$ worden als volgt berekend:

$$\begin{aligned}
\mathbf{y}^{(1)} &= r_{\mathbf{y}1}\mathbf{x}_1 \\
&= 0.58\mathbf{x}_1 \quad .
\end{aligned}$$

2. Stel nu dat we $\mathbf{y}$ voorspellen op basis van alleen $\mathbf{x}_2$, de andere predictor. De voorspelde criteriumscores $\mathbf{y}^{(2)}$ worden als volgt berekend:

$$\begin{aligned}\mathbf{y}^{(2)} &= r_{\mathbf{y}2}\mathbf{x}_2 \\ &= 0.57\mathbf{x}_2 \quad .\end{aligned}$$

   We weten dat de correlatie tussen $\mathbf{x}_1$ en $\mathbf{x}_2$ bijna gelijk is aan 1. Nu, twee gestandaardiseerde variabelen die onderling vrijwel perfect correleren zijn ook vrijwel gelijk aan elkaar. Dus, $\mathbf{x}_1$ en $\mathbf{x}_2$ zijn vrijwel gelijk aan elkaar. En omdat $\mathbf{y}^{(1)}$ gelijk is aan $0.58\mathbf{x}_1$ en $\mathbf{y}^{(2)}$ gelijk is aan $0.57\mathbf{x}_2$, volgt hieruit dat ook $\mathbf{y}^{(1)}$ en $\mathbf{y}^{(2)}$ vrijwel gelijk zijn aan elkaar. Dus, de twee hoog gecorreleerde predictoren leiden tot vrijwel dezelfde voorspelde criteriumscores.

3. Stel nu dat we $\mathbf{y}$ voorspellen op basis van $\mathbf{x}_1$ en $\mathbf{x}_2$ sámen. De voorspelde criteriumscores $\mathbf{y}^{(1,2)}$ worden als volgt berekend:

$$\begin{aligned}\mathbf{y}^{(1,2)} &= b_1\mathbf{x}_1 + b_2\mathbf{x}_2 \\ &= 0.79\mathbf{x}_1 + (-0.21)\mathbf{x}_2 \quad .\end{aligned}$$

   Omdat $\mathbf{x}_1$ en $\mathbf{x}_2$ vrijwel gelijk zijn aan elkaar, geldt nu ook het volgende:

$$\begin{aligned}\mathbf{y}^{(1,2)} &\approx 0.79\mathbf{x}_1 + (-0.21)\mathbf{x}_1 \\ &= 0.58\mathbf{x}_1 \\ &= \mathbf{y}^{(1)} \approx \mathbf{y}^{(2)} \quad .\end{aligned}$$

   Hierin staat $\approx$ voor "is ongeveer gelijk aan".

   Het belangrijke punt in het bovenstaande is dat de som van de bètagewichten $b_1 + b_2$ $(= 0.79 - 0.21 = 0.58)$ ongeveer gelijk is aan $r_{\mathbf{y}1}$ (0.58) en $r_{\mathbf{y}2}$ (0.57). Bij zeer sterk gecorreleerde predictoren is dat altijd het geval. De individuele bètagewichten mogen dan wel instabiel zijn, hun som is dat niet, en daarom zijn de voorspelde criteriumscores $\mathbf{y}^{(1,2)}$ dat ook niet. Ter illustratie, als $r_{\mathbf{y}1}$ en $r_{\mathbf{y}2}$ gelijk zijn aan, respectievelijk, 0.57 en 0.58 (dus, het omgekeerde van het eerste paar correlaties), dan zijn de bètagewichten gelijk aan, respectievelijk, -0.21 en 0.79 (dus, het omgekeerde van het eerste paar bètagewichten), maar blijft de som $b_1 + b_2$ $(= -0.21 + 0.79 = 0.58)$ hetzelfde.

# Hoofdstuk 5

# Factoranalyse

## Inhoudsopgave


**5.1 Het Model met Eén Gemeenschappelijke Factor . . . . . 463**
5.1.1 Het Model en zijn Implicaties voor de Correlaties . . . . . 463
5.1.2 Schattingen van de Factorladingen . . . . . 467
**5.2 Het Model met Twee Gemeenschappelijke Factoren . . . 470**
5.2.1 Het Model en zijn Implicaties voor de Correlaties . . . . . 470
5.2.2 Schattingen van de Factorladingen . . . . . 472
**5.3 Ladingen, Communaliteit en Eigenwaarde . . . . . 474**
5.3.1 Ladingen als Correlaties . . . . . 474
5.3.2 De Communaliteit van een Variabele . . . . . 475
5.3.3 De Eigenwaarde van een Factor . . . . . 476
5.3.4 Grafische Weergave van de Factoroplossing . . . . . 477
**5.4 Hoofdcomponentenanalyse . . . . . 484**
5.4.1 Het Model en zijn Implicaties voor de Correlaties . . . . . 485
5.4.2 Schattingen van de Factorladingen . . . . . 486
5.4.3 Een Vergelijking met het Factormodel met Ongecorreleerde Residuen . . . . . 488
**5.5 De Onbepaaldheid van een Factoroplossing . . . . . 491**
5.5.1 Illustratie en Omschrijving . . . . . 491
5.5.2 Het Belang van de Onbepaaldheid . . . . . 494
**5.6 Roteren naar Interpreteerbare Factoren . . . . . 496**
5.6.1 Enkele Richtlijnen bij de Interpretatie . . . . . 496
5.6.2 Oblieke Rotatie . . . . . 498
**5.7 Het Model met Meer dan Twee Gemeenschappelijke Factoren . . . . . 506**
5.7.1 Veralgemening van de Formules . . . . . 506
5.7.2 Grafische Weergave van de Factoroplossing . . . . . 507
5.7.3 Roteren met Meer dan Twee Factoren . . . . . 507

**5.8 De Keuze van het Aantal Gemeenschappelijke Factoren . 508**
5.8.1 Inspectie van de Residuele Correlaties. . . . . 509
5.8.2 Inspectie van de Eigenwaarden . . . . . 510
5.8.3 De Interpreteerbaarheid van de Gevonden factoroplossing. . . 510
5.8.4 De Uitslag van een Statistische Toets . . . . . 511
**5.9 Het Bepalen van de Factorscores . . . . . 512**
**5.10 Uitvoeren van een Factoranalyse met de SPSS-procedure Factor Analysis . . . . . 513**
**5.11 Appendix bij Hoofdstuk 5 . . . . . 522**
**Literatuur . . . . . 524**

De tweede multivariate analysetechniek die we gaan bekijken is de factoranalyse. In zekere zin is deze techniek nauw verwant met de regressieanalyse, maar het doel is hier heel anders en dit zorgt voor wezenlijke verschillen. Bij regressieanalyse draait het om voorspellingen:

1. We willen een geconstateerde samenhang tussen een aantal geobserveerde variabelen gebruiken om een van deze variabelen (de criteriumvariabele) optimaal te voorspellen uit de overige geobserveerde variabelen (de predictoren).

2. We willen de vraag beantwoorden of een bepaalde variabele een bijdrage levert aan de voorspelling van de criteriumvariabele bovenop een aantal andere variabelen.

In de factoranalyse is de doelstelling meer theoretisch. Hier willen we de geobserveerde variabelen verklaren door ze terug te voeren op een aantal gemeenschappelijke latente (niet-geobserveerde) variabelen. Dit wil zeggen, men veronderstelt dat alle geobserveerde variabelen ten minste voor een deel zijn opgebouwd uit een klein aantal onderliggende variabelen, factoren genaamd.

Van de geobserveerde variabelen kunnen we meerdere aspecten verklaren, waarvan we er hier twee zullen beschouwen: (1) hun onderlinge samenhang, zoals uitgedrukt door hun correlaties, en (2) hun variantie. Als we in de onderlinge samenhang geïnteresseerd zijn, dan gebruiken we de factoren om de correlaties tussen de geobserveerde variabelen te verklaren. Wat concreter: de correlatie tussen elk tweetal geobserveerde variabelen wordt volledig verklaard door het feit dat ze bepaalde factoren als gemeenschappelijk bestanddeel hebben. Dit heeft tot gevolg dat, als het effect van de gemeenschappelijke factoren wordt uitgezuiverd (dit wil zeggen, de invloed van deze factoren wordt uitgeschakeld) er geen correlatie meer mag overblijven.

Als we in het verklaren van variantie geïnteresseerd zijn, dan gaan we, net zoals in de regressieanalyse, de proportie verklaarde variantie maximaliseren. Het verschil met de regressieanalyse is echter dat onze predictoren nu niet op voorhand gegeven zijn. Het probleem is dus niet meer om de optimale gewichten (de regressiegewichten) te kiezen bij gegeven predictoren, maar wel om de predictoren (de factoren) zó te kiezen dat ze zoveel mogelijk variantie verklaren. De techniek waarmee deze maximale-variantie-verklarende factoren bepaald worden, is de *hoofdcomponentenanalyse* (in het Engels, *principal components analysis* of PCA).

In de praktijk wordt de factoranalyse meestal gebruikt bij de constructie van tests en vragenlijsten. Bij test- en vragenlijstconstructie vertrekt men vaak van een groot aantal items die men aan een groep personen voorlegt. Op deze data wordt dan een factoranalyse uitgevoerd. Op basis van deze factoranalyse verdeelt men de items in groepjes, ook *schalen* genoemd. Items die tot dezelfde schaal behoren meten allemaal dezelfde gemeenschappelijke factor. Op basis van de inhoud van de items die tot dezelfde schaal behoren, kan men deze schaal ook inhoudelijk benoemen (ook vaak *interpreteren* genoemd). Samengevat, met behulp van de factoranalyse kan men een inhoudelijk heterogene itemset opsplitsen in inhoudelijk homogene schalen.

In wat volgt, maken we gebruik van een getallenvoorbeeld dat is ontleend aan Jöreskog en Sörbom (1979). De getallen zijn afkomstig uit een hypothetisch onderzoek waarin van een groot aantal proefpersonen zes tests werden afgenomen. De scores op deze tests worden aangeduid met $X_1$, $X_2$, $X_3$, $X_4$, $X_5$, en $X_6$. Het blijkt voor het vervolg geen beperking maar juist handig te zijn ervan uit te gaan dat de variabelen gestandaardiseerd zijn. Dus, we veronderstellen

$$\overline{X}_i = 0 \quad \text{en} \quad \text{VAR}(X_i) = 1 \quad \text{voor } i = 1, \ldots, 6.$$

De berekening van de correlatie voor ieder paar variabelen resulteert in de correlatiematrix die getoond wordt in Tabel 5.1.

| | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ |
|---|---|---|---|---|---|---|
| $X_1$ | 1.00 | | | | | |
| $X_2$ | 0.72 | 1.00 | | | | |
| $X_3$ | 0.38 | 0.34 | 1.00 | | | |
| $X_4$ | 0.32 | 0.29 | 0.42 | 1.00 | | |
| $X_5$ | 0.27 | 0.24 | 0.35 | 0.30 | 1.00 | |
| $X_6$ | 0.27 | 0.24 | 0.13 | 0.11 | 0.09 | 1.00 |

Tabel 5.1: Een correlatiematrix met zes variabelen.

In de paragrafen 5.1 tot en met 5.3 behandelen we de variant van de factoranalyse die de correlaties zo goed mogelijk probeert te verklaren. Daarna, in paragraaf 5.4, behandelen we de variant van de factoranalyse die de variantie zo goed mogelijk probeert te verklaren.

## 5.1 Het Model met Eén Gemeenschappelijke Factor

### 5.1.1 Het Model en zijn Implicaties voor de Correlaties

We willen nu de correlaties in de bovenstaande matrix verklaren door de werking van een klein aantal latente factoren die $X_1$ tot en met $X_6$ gemeenschappelijk hebben. Aangezien we het model zo eenvoudig mogelijk willen houden, kunnen we om te beginnen vragen of deze correlaties verklaard worden door één enkele onderliggende gemeenschappelijke factor. Deze hypothetische factor[1] noemen we $F$. Onze vraag komt erop neer dat we voor

[1] Iedere persoon heeft een bepaalde, niet-geobserveerde score op deze factor. Als we $p$ gebruiken als index voor de personen, dan kunnen we de factorscore van persoon $p$ aanduiden met $F_p$.

het zestal geobserveerde variabelen een opsplitsing zoeken volgens onderstaand schema:

$$\boxed{\begin{array}{lcl} X_1 & = & a_1F + U_1 \\ X_2 & = & a_2F + U_2 \\ X_3 & = & a_3F + U_3 \\ X_4 & = & a_4F + U_4 \\ X_5 & = & a_5F + U_5 \\ X_6 & = & a_6F + U_6 \quad . \end{array}}$$

In deze vergelijkingen wordt, voor $i = 1, \ldots, 6$, de variabele $X_i$ opgesplitst in een door de factor $F$ verklaard deel $a_iF$ en een residu-deel $U_i$. De variabelen $U_i$ zijn dus ook latente variabelen. Zij representeren het gedeelte van $X_i$ dat niet door de gemeenschappelijke factor $F$ wordt verklaard. De variabele $U_i$ wordt daarom ook de *unieke factor* behorende bij $X_i$ genoemd. In zekere zin staan hierboven zes regressievergelijkingen voor zes criteriumvariabelen $X_1$ tot en met $X_6$, met in elke vergelijking slechts één (en wel steeds dezelfde) predictor $F$. Conform deze interpretatie zijn de getalletjes $a_1$ tot en met $a_6$ de regressiegewichten van de latente predictor $F$ voor de zes criteriumvariabelen $X_i$. In de factoranalyse is echter een andere terminologie gebruikelijk. In deze context wordt $a_i$ de *lading* van variabele $X_i$ op de factor $F$ genoemd. Een andere overeenkomst met de regressieanalyse is dat de gemeenschappelijke factor $F$ ongecorreleerd is met alle residuen:

$$\boxed{r(F, U_i) = 0 \quad \text{voor} \quad i = 1, \ldots, 6.}$$

Dit volgt, zoals in de regressieanalyse, uit het feit dat $U_i$ per definitie het deel van $X_i$ is dat niet met $F$ samenhangt.

Er zijn ook belangrijke verschillen met de regressieanalyse. De factor $F$ is geen echte predictor omdat de scores van de subjecten op deze variabele niet gegeven zijn. Bijgevolg zijn ook de regressiegewichten $a_i$ vooralsnog onbekende getallen; ze kunnen niet worden bepaald door een daadwerkelijke regressie van $X_i$ op $F$ uit te voeren. Omdat $F$ niet gegeven is, hebben we de vrijheid zelf het nulpunt en de meeteenheid van deze variabele te kiezen. Het blijkt handig te zijn $F$ gestandaardiseerd te nemen. Dus, we kiezen

$$\overline{F} = 0 \quad \text{en} \quad \text{VAR}(F) = 1.$$

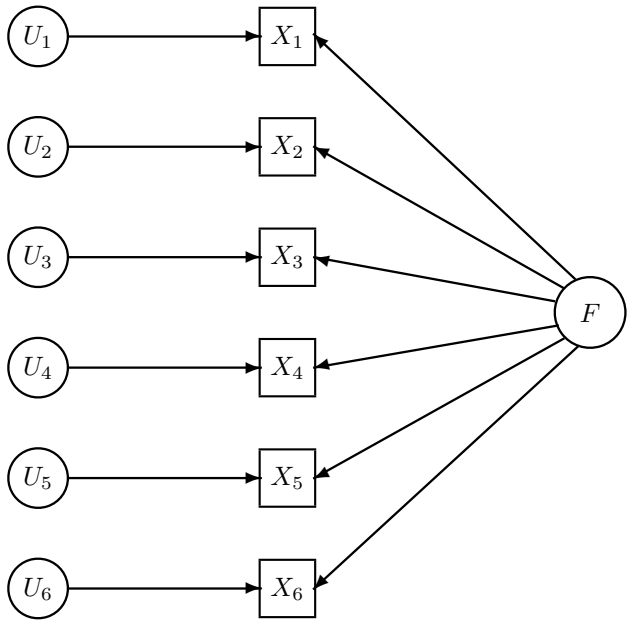
Omdat de geobserveerde variabelen $X_i$ eveneens gestandaardiseerd zijn, volgt hieruit dat $\overline{U_i} = 0$ en $\text{VAR}(U_i) \leq 1$ (op deze ongelijkheid komen we terug in de paragraaf over de communaliteit).

Een volgend verschil met de regressieanalyse is dat hier de zes regressievergelijkingen met elkaar zijn verbonden: dezelfde variabele dient als predictor in alle vergelijkingen. We veronderstellen nu dat deze gemeenschappelijke factor $F$ volledig verantwoordelijk is voor de geobserveerde correlatie tussen de $X$-variabelen. Bijvoorbeeld, de gedeelten van $X_1$ en $X_2$ die niets met $F$ gemeen hebben mogen niet bijdragen aan de correlatie tussen $X_1$ en $X_2$. Dit betekent dat $U_1$ en $U_2$ onderling ongecorreleerd moeten zijn. Op dezelfde wijze volgt dat alle residuvariabelen $U_i$ onderling ongecorreleerd zijn. Dus,

$$\boxed{r(U_i, U_j) = 0 \quad \text{als} \quad i \neq j \quad .}$$

Dit wordt ook de veronderstelling van ongecorreleerde residuen genoemd. De twee varianten van de factoranalyse (correlaties verklaren en variantie verklaren) verschillen van elkaar wat betreft deze veronderstelling: in de variant van de factoranalyse die zoveel mogelijk variantie probeert te verklaren zijn de residuen wel gecorreleerd.

Men kan het factormodel ook grafisch voorstellen. Dit gebeurt in Figuur 5.1. In deze figuur worden de geobserveerde variabelen $(X_1, \ldots, X_6)$ voorgesteld met een vierkant en de latente variabelen $(F, U_1, \ldots, U_6)$ met een cirkel. Met een pijl wordt aangegeven dat de variabele waar de pijl aankomt voor een deel is opgebouwd uit de variabele vanwaar de pijl vertrekt. Om helemaal volledig te zijn, zou er bij sommige pijlen ook nog een getal moeten staan dat aangeeft in welke mate een variabele bijdraagt aan een andere variabele. Dit geldt voor alle pijlen die van $F$ vertrekken; deze zouden allemaal gewogen moeten worden met hun bijbehorende lading.




Figuur 5.1: Grafische voorstelling van het factormodel met één gemeenschappelijke factor.

De regressiegewichten of ladingen zijn voorlopig nog onbekend. De extra eis van ongecorreleerde residuen roept echter de vraag op of er wel getalletjes $a_1$ tot en met $a_6$ bestaan zodat aan deze voorwaarde is voldaan. Dit gaan we onderzoeken door te bekijken wat er in zo'n geval voor de $a_i$ moet gelden. We bekijken bij wijze van voorbeeld de eerste twee vergelijkingen betreffende $X_1$ en $X_2$. Volgens onze aannames zijn de verschillende latente variabelen die aan de rechterkant van deze twee vergelijkingen voorkomen onderling ongecorreleerd: $r(F, U_1) = 0$, $r(F, U_2) = 0$, en $r(U_1, U_2) = 0$. Zoals verderop uitgelegd wordt, volgt uit deze aannames dat de correlatie tussen $X_1$ en $X_2$ gelijk is aan het product van de ladingen van deze twee variabelen op de ene gemeenschappelijke

factor, dus

$$\boxed{r(X_1, X_2) = a_1 \cdot a_2 \quad .}$$

In de correlatiematrix in Tabel 5.1 zien we dat $r(X_1, X_2) = 0.72$. Dus moet gelden: $a_1 \cdot a_2 = 0.72$. Zo volgt ook dat bijvoorbeeld $r(X_1, X_3) = a_1 \cdot a_3 = 0.38$ en $r(X_4, X_6) = a_4 \cdot a_6 = 0.11$.

Het is niet moeilijk om te laten zien dat de bovenstaande formule klopt. Hiervoor hebben we enkele eenvoudige formules voor covarianties nodig. Voor we deze formules geven, moeten we opmerken dat de covariantie, net zoals de correlatie, een maat is voor de samenhang tussen twee variabelen. In de appendix op pagina 122 wordt uitgelegd wat de relatie is tussen de covariantie en de correlatie. Voor wat volgt, is het belangrijk om te weten dat, voor gestandaardiseerde variabelen zoals $X_1$, $X_2$, ..., $X_6$, de covariantie gelijk is aan de correlatie.

Nu, een eerste eigenschap van covarianties is dat de covariantie tussen twee sommen van variabelen gelijk is aan de som van de covarianties tussen de samenstellende variabelen (i.e., de variabelen in de sommen). Deze eigenschap wordt afgeleid in de appendix op pagina 522. Dit passen we nu toe op de correlatie tussen $X_1$ en $X_2$ die, zoals we weten, gelijk is aan hun covariantie. Het model zegt dat $X_1$ gelijk is aan $a_1F + U_1$ en dat $X_2$ gelijk is aan $a_2F + U_2$. Hieruit volgt dat

$$\begin{aligned} r(X_1, X_2) =& \mathrm{COV}(X_1, X_2) \\ =& \mathrm{COV}(a_1F, a_2F) + \mathrm{COV}(a_1F, U_2) + \mathrm{COV}(U_1, a_2F) + \mathrm{COV}(U_1, U_2) \quad . \end{aligned}$$

Deze formule kunnen we vereenvoudigen. De laatste drie termen zijn namelijk gelijk aan 0:

1. Omdat de correlatie $r(U_1, U_2)$ gelijk is aan 0, is ook de corresponderende covariantie $\mathrm{COV}(U_1, U_2)$ gelijk aan 0. (Een nulcorrelatie impliceert altijd een nulcovariantie, en omgekeerd. Dit komt omdat de correlatie gelijk is aan de covariantie gedeeld door het product van de standaarddeviaties.)

2. Omdat de correlatie $r(F, U_2)$ gelijk is aan 0, is ook de correlatie $r(a_1F, U_2)$ gelijk aan 0 (een correlatie verandert niet als een variabele met een constante vermenigvuldigd wordt), en daarom is ook de covariantie $\mathrm{COV}(a_1F, U_2)$ gelijk aan 0.

3. Uit een analoge redenering als in 2., volgt dat ook $\mathrm{COV}(U_1, a_2F)$ gelijk is aan 0.

Nu kan $r(X_1, X_2)$ als volgt geschreven worden:

$$r(X_1, X_2) = \mathrm{COV}(a_1F, a_2F) \quad .$$

Deze formule kunnen we nog verder vereenvoudigen. Zoals bekend, heeft de covariantie de eigenschap dat men bij variabelen die met een constante vermenigvuldigd worden de constanten buiten de haakjes mag brengen. Deze eigenschap wordt afgeleid in de appendix op pagina 523. Dus, $\mathrm{COV}(a_1F, a_2F)$ is gelijk aan $a_1a_2\mathrm{COV}(F, F)$. De covariantie van een variabele met zichzelf is gelijk aan de variantie van die variabele. Dus,

$\text{COV}(F, F)$ is gelijk aan $\text{VAR}(F)$. En omdat $F$ gestandaardiseerd is, is $\text{VAR}(F)$ gelijk aan 1. Dus, het eindresultaat van onze vereenvoudiging is:

$$r(X_1, X_2) = a_1 \cdot a_2 \quad ,$$

hetgeen aangetoond moest worden.

### 5.1.2 Schattingen van de Factorladingen

Om de zes getallen $a_1$ tot en met $a_6$ te bepalen hebben we dus 15 vergelijkingen waarin een correlatiecoëfficiënt gelijkgesteld wordt aan een product van factorladingen, want er staan 15 correlatiecoëfficiënten in Tabel 5.1. Dit betekent dat de $a_i$ overgedetermineerd zijn: in zijn algemeenheid is het niet mogelijk om zes waarden voor de $a_i$ te vinden zodat tegelijkertijd aan alle 15 vergelijkingen voldaan is. Dit lijkt slecht nieuws, omdat het betekent dat het model met één gemeenschappelijke factor in het algemeen niet perfect zal opgaan. In feite is het goed nieuws, want het betekent dat dit één-factormodel niet triviaal is, niet willekeurig welke correlatiematrix kan verklaren. Het model legt blijkbaar beperkingen op aan de data (de waargenomen correlaties) en we moeten nu onderzoeken in hoeverre onze data aan deze beperkingen voldoen. Kortweg, we moeten (1) gezien onze data de beste schattingen vinden voor de getallen $a_1$ tot en met $a_6$ en (2) bekijken of deze beste schattingen goed genoeg zijn, dit wil zeggen of ze de waargenomen correlaties goed genoeg reproduceren.

De beste schattingen voor de $a_i$ zijn die welke de geobserveerde correlaties over het geheel genomen het beste reproduceren. Dit nogal vage criterium kan preciezer gemaakt worden met behulp van de zogenaamde *gereproduceerde correlaties*. De gereproduceerde correlatie tussen $X_1$ en $X_2$ wordt berekend op basis van geschatte of nog te schatten factorladingen en met behulp van de correlatieformule die uit het model volgt (i.e., $a_1 \cdot a_2$). Deze gereproduceerde correlatie wordt aangeduid met het symbool $\widehat{r}$, dus met een dakje toegevoegd. We willen de factorladingen zó kiezen dat de gereproduceerde correlaties zo kort mogelijk liggen bij de geobserveerde correlaties. Het verschil tussen een geobserveerde en een gereproduceerde correlatie wordt een *residuele correlatie* genoemd. Bijvoorbeeld, de residuele correlatie voor de variabelen $X_1$ en $X_2$ wordt berekend als $r(X_1, X_2) - \widehat{r}(X_1, X_2)$. Merk op dat een residuele correlatie eigenlijk geen correlatie is: hij wordt niet berekend met de formule voor een correlatiecoëfficiënt, maar is gelijk aan het verschil tussen een geobserveerde en een gereproduceerde correlatie.

Het criterium van zo klein mogelijke residuele correlaties kan op meerdere manieren ingevuld worden, en we bespreken er hier twee. De eerste invulling is in termen van het kleinste-kwadratencriterium: kies de factorladingen zodat de som van de gekwadrateerde residuele correlaties geminimaliseerd wordt. Ofwel,

Kies de factorladingen $a_1, \ldots, a_6$ zodat

$$\sum_{i<j} [r(X_i, X_j) - \widehat{r}(X_i, X_j)]^2 = \sum_{i<j} [r(X_i, X_j) - a_i \cdot a_j]^2 \qquad \text{minimaal is} \quad ,$$

waarin de som loopt over alle $(i, j)$-paren waarin $i$ kleiner is dan $j$. In deze som wordt elke correlatie slechts één keer geteld en niet twee keer. Dit laatste zou het geval zijn indien

de som over alle $(i, j)$-paren zou lopen. Merk op dat het kleinste-kwadratencriterium hier toegepast wordt op correlaties en niet op de oorspronkelijke observaties, zoals in de regressieanalyse. In het Engels, wordt het bovenstaande criterium *unweighted least-squares* (ULS) of *ordinary least-squares* genoemd.

Het is goed om even stil te staan bij het feit dat het symbool $r(X_i, X_j)$ in meerdere betekenissen gebruikt wordt.

1. In de gekwadrateerde afwijking $[r(X_i, X_j) - a_i \cdot a_j]^2$ staat $r(X_i, X_j)$ voor de *geobserveerde* correlatie tussen de variabelen $X_i$ en $X_j$. Bijvoorbeeld, de geobserveerde correlatie tussen de variabelen $X_1$ en $X_2$ is 0.72.

2. In de formule op pagina 466 staat $r(X_i, X_j)$ voor de *populatie*correlatie tussen de variabelen $X_i$ en $X_j$. De populatiecorrelatie is de correlatie die men zoú observeren indien men een oneindig aantal observaties (personen) heeft. In de praktijk heeft men natuurlijk nooit een oneindig aantal observaties en daarom wordt de populatiecorrelatie ook de *theoretische* correlatie genoemd. Indien het één-factormodel opgaat, dan kan elke populatiecorrelatie $r(X_i, X_j)$ geschreven worden als het product van de factorladingen $a_i$ en $a_j$. Dit geldt niet voor de geobserveerde correlaties. Omdat het aantal observaties altijd eindig is, wijken de geobserveerde correlaties vrijwel altijd af van de populatiecorrelaties.

In welke betekenis $r(X_i, X_j)$ gebruikt wordt, als geobserveerde of als populatiecorrelatie, is altijd duidelijk uit de context.

De tweede invulling van het criterium van zo klein mogelijke residuele correlaties, naast het ULS-criterium, is in termen van het *Maximum Likelihood* (ML) criterium. Het zou ons te ver voeren om hier dieper op in te gaan, en zeker om een vergelijking te maken met het ULS-criterium. Het ML-criterium wordt tegenwoordig het meeste gebruikt. Dit impliceert echter niet dat we het ULS-criterium maar moeten vergeten, want er zijn situaties waarin het ULS-criterium te verkiezen is. We komen hier later nog op terug.

Voor onze data levert het ML-criterium de volgende oplossing:

$$\begin{aligned} X_1 &= 0.87F + U_1 \\ X_2 &= 0.81F + U_2 \\ X_3 &= 0.47F + U_3 \\ X_4 &= 0.40F + U_4 \\ X_5 &= 0.34F + U_5 \\ X_6 &= 0.30F + U_6 \quad . \end{aligned}$$

Dit betekent dus: $a_1 = 0.87$, $a_2 = 0.81$, enzovoort.

Nu moeten we ons de vraag stellen hoe goed deze oplossing is. Dat kunnen we zien door de gereproduceerde correlaties te vergelijken met de geobserveerde correlaties. We zien bijvoorbeeld dat

$$\widehat{r}(X_1, X_2) = a_1 \cdot a_2 = 0.87 \cdot 0.81 = 0.70 \quad ,$$

terwijl $r(X_1, X_2) = 0.72$. Hier is er dus sprake van een goede overeenkomst tussen de gereproduceerde en de geobserveerde waarde. Maar we hebben bijvoorbeeld ook

$\widehat{r}(X_3, X_5) = a_3 \cdot a_5 = 0.47 \cdot 0.34 = 0.16$, terwijl $r(X_3, X_5) = 0.35$. Voor deze twee variabelen is er dus een slechte overeenkomst. In Tabel 5.2 staat links de matrix met de gereproduceerde correlaties; in de rechtermatrix staat, voor elk paar variabelen, de residuele correlatie, het verschil tussen de geobserveerde en de gereproduceerde correlatie. (Vergelijk de linkermatrix met de correlatiematrix in Tabel 5.1.) Uit de rechtermatrix blijkt dat vele correlaties goed gereproduceerd worden (kleine residuen), maar niet alle. Met name de correlaties tussen de variabelen $X_3$, $X_4$ en $X_5$ laten nog grote afwijkingen zien.

Gereproduceerd:

| | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ |
|---|---|---|---|---|---|
| $X_1$ | | | | | |
| $X_2$ | 0.70 | | | | |
| $X_3$ | 0.41 | 0.38 | | | |
| $X_4$ | 0.35 | 0.33 | 0.19 | | |
| $X_5$ | 0.30 | 0.28 | 0.16 | 0.14 | |
| $X_6$ | 0.26 | 0.24 | 0.14 | 0.12 | 0.10 |

Residueel:

| | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ |
|---|---|---|---|---|---|
| $X_1$ | | | | | |
| $X_2$ | 0.02 | | | | |
| $X_3$ | -0.03 | -0.04 | | | |
| $X_4$ | -0.03 | -0.04 | 0.23 | | |
| $X_5$ | -0.03 | -0.04 | 0.19 | 0.16 | |
| $X_6$ | 0.01 | -0.00 | -0.01 | -0.01 | -0.01 |

Tabel 5.2: Matrix met gereproduceerde en residuele correlaties op basis van het één-factormodel.

Er zijn twee redenen waarom de geobserveerde en de gereproduceerde correlaties kunnen verschillen van elkaar en de residuele correlaties dus niet gelijk zijn aan 0:

1. De eerste reden is dat het één-factormodel niet opgaat. Misschien is het wel zo dat er twee factoren nodig zijn om de geobserveerde correlaties te verklaren. Bijvoorbeeld, de hoge residuele correlaties tussen de variabelen $X_3$, $X_4$ en $X_5$ uit ons getallenvoorbeeld kunnen toe te schrijven zijn aan het feit dat deze drie variabelen nog een tweede factor gemeenschappelijk hebben. In de volgende paragraaf laten we zien dat een model met twee gemeenschappelijke factoren de geobserveerde correlaties tussen deze drie variabelen wel kan verklaren.

2. De tweede reden is dat de geobserveerde correlaties altijd berekend worden op een eindig aantal observaties. Misschien is het wel zo dat de populatiecorrelaties (i.e., de correlaties berekend op een oneindig aantal observaties) wél voldoen aan het één-factormodel. Omdat de geobserveerde correlaties vrijwel altijd afwijken van de populatiecorrelaties zullen zij niet perfect voldoen aan het één-factormodel. Het gaat hier om toevallige afwijkingen van het één-factormodel, die afnemen naarmate het aantal observaties toeneemt. Hier ligt het belangrijke verschil met de eerste reden die hierboven genoemd werd. Volgens deze eerste reden is het verschil tussen de geobserveerde en de gereproduceerde correlaties het gevolg van een systematische afwijking van het één-factormodel. Deze afwijking neemt niet af naarmate het aantal observaties toeneemt.

## 5.2 Het Model met Twee Gemeenschappelijke Factoren

### 5.2.1 Het Model en zijn Implicaties voor de Correlaties

Vanwege de grote residuele correlaties beschouwen we de poging om alle correlaties met één factor te verklaren als mislukt. Maar als het met één factor niet lukt, dan misschien wel met twee. Dus, nu breiden we ons model uit en veronderstellen dat er twee onderliggende gemeenschappelijke factoren in het spel zijn. Deze nieuwe factoren noemen we $F_1$ en $F_2$. De aanname is nu dat de correlaties tussen de geobserveerde variabelen volledig te verklaren zijn door deze twee factoren. Dit leidt tot de volgende zes regressievergelijkingen met twee gemeenschappelijke factoren:

$$\boxed{\begin{array}{lcl} X_1 & = & a_{11}F_1 + a_{12}F_2 + U_1 \\ X_2 & = & a_{21}F_1 + a_{22}F_2 + U_2 \\ X_3 & = & a_{31}F_1 + a_{32}F_2 + U_3 \\ X_4 & = & a_{41}F_1 + a_{42}F_2 + U_4 \\ X_5 & = & a_{51}F_1 + a_{52}F_2 + U_5 \\ X_6 & = & a_{61}F_1 + a_{62}F_2 + U_6 \quad . \end{array}}$$

In deze vergelijkingen staan dus rechts van het =–teken latente variabelen $F_1$, $F_2$ en $U_1$ tot en met $U_6$. De gemeenschappelijke factoren $F_1$ en $F_2$ nemen we weer gestandaardiseerd. Dus,

$$\overline{F_1} = 0, \quad \text{VAR}(F_1) = 1; \quad \overline{F_2} = 0, \quad \text{VAR}(F_2) = 1 \quad .$$

Bovendien hebben we bij deze latente variabelen nog de vrijheid om ze onderling ongecorreleerd te nemen. Dus,

$$\boxed{r(F_1, F_2) = 0 \quad .}$$

In de factoranalytische terminologie heten zulke ongecorreleerde factoren *orthogonaal* (letterlijk, *rechthoekig*; we zullen later zien waar deze benaming vandaan komt). De veronderstelling van een nulcorrelatie tussen de latente variabelen betekent geen restrictie op de data. We zullen later zien dat we ze ook gecorreleerd kunnen kiezen en dat dit voor de mate van overeenstemming met de data (de *fit*) niets uitmaakt. Net zoals in het model met één gemeenschappelijke factor, zijn de gemeenschappelijke factoren geassocieerd met regressiegewichten of ladingen: het getal $a_{ij}$ ($i = 1, \ldots, 6$; $j = 1, 2$) is de lading van variabele $X_i$ op factor $F_j$. Verder is er voor elke variabele $X_i$ natuurlijk weer een unieke factor $U_i$, het residu van $X_i$ na de regressie op $F_1$ en $F_2$. (We gebruiken hier hetzelfde symbool $U$ als boven, maar de unieke factoren hier zijn natuurlijk andere dan die bij één gemeenschappelijke factor.) Omdat zowel de geobserveerde variabelen $X_i$ als de gemeenschappelijke factoren $F_1$ en $F_2$ gestandaardiseerde variabelen zijn, geldt voor de unieke factoren $U_i$ ook hier dat $\overline{U}_i = 0$ en $\text{VAR}(U_i) \leq 1$.

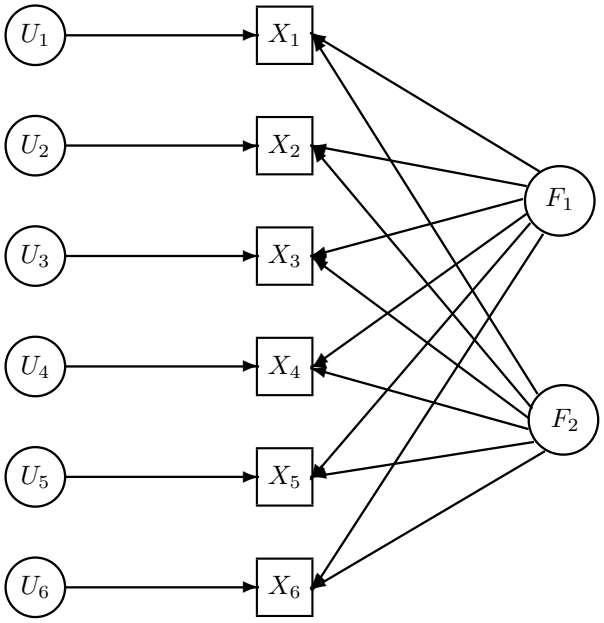
In de lineaire regressie zijn de residuen ongecorreleerd met alle predictoren. Analoog geldt hier dat de $U_i$ ongecorreleerd zijn met $F_1$ en $F_2$:

$$\boxed{r(U_i, F_j) = 0 \qquad \text{voor} \quad i = 1, \ldots, 6 \qquad \text{en} \qquad j = 1, 2.}$$

Net zoals in het model met één gemeenschappelijke factor, veronderstellen we ook hier dat de correlatie tussen de geobserveerde variabelen geheel kan worden toegeschreven aan de gemeenschappelijke factoren. Daarom moet de correlatie na uitzuivering van de gemeenschappelijke factoren (i.e., de correlatie tussen de unieke factoren onderling) gelijk zijn aan nul:

$$\boxed{r(U_i, U_j) = 0 \quad \text{voor} \quad i \neq j \quad .}$$

De grafische voorstelling van het model met twee gemeenschappelijke factoren wordt getoond in Figuur 5.2. Bij elke geobserveerde variabele eindigen nu drie in plaats van twee pijlen; twee komen van de gemeenschappelijke factoren $F_1$ en $F_2$, en één komt van de unieke factor $U_i$.




Figuur 5.2: Grafische voorstelling van het factormodel met twee gemeenschappelijke factoren.

Deze eis van onderling ongecorreleerde residuen of unieke factoren legt weer beperkingen op aan de nog onbekende factorladingen $a_{ij}$. Laten we bij wijze van voorbeeld nogmaals de vergelijkingen voor de variabelen $X_1$ en $X_2$ bekijken. Uit het feit dat de latente variabelen $F_1$, $F_2$, $U_1$ en $U_2$ alle onderling ongecorreleerd zijn volgt dat de correlatie tussen $X_1$ en $X_2$ gelijk is aan het product van de ladingen van deze twee variabelen op de eerste gemeenschappelijke factor plús het product van de ladingen op de tweede gemeenschappelijke factor:

$$\boxed{r(X_1, X_2) = a_{11} \cdot a_{21} + a_{12} \cdot a_{22} \quad .}$$

Volgens het model moet nu dus gelden $a_{11} \cdot a_{21} + a_{12} \cdot a_{22} = 0.72$. Er zijn natuurlijk soortgelijke uitdrukkingen voor de andere variabelenparen. Bijvoorbeeld, er moet gelden $r(X_3, X_5) = a_{31} \cdot a_{51} + a_{32} \cdot a_{52} = 0.35$.

Het is niet zo moeilijk om te laten zien dat de bovenstaande formule voor $r(X_1, X_2)$ klopt. Dit kunnen we op dezelfde manier laten zien als bij het model met één gemeenschappelijke factor. We maken opnieuw gebruik van het feit dat een covariantie tussen twee sommen van variabelen gelijk is aan de som van de covarianties tussen de samenstellende variabelen. In het model met twee gemeenschappelijke factoren zijn $X_1$ en $X_2$ elk de som van drie variabelen: $X_1$ is gelijk aan $a_{11}F_1 + a_{12}F_2 + U_1$ en $X_2$ is gelijk aan $a_{21}F_1 + a_{22}F_2 + U_2$. Het resultaat is dat $r(X_1, X_2) = \mathrm{COV}(X_1, X_2)$ geschreven kan worden als een som van negen covarianties. Hiervan zijn er zeven gelijk aan 0, en de resterende twee zijn gelijk aan $a_{11} \cdot a_{21}$, respectievelijk, $a_{12} \cdot a_{22}$.

### 5.2.2 Schattingen van de Factorladingen

Om de 12 onbekende factorladingen $a_{ij}$ te bepalen hebben we 15 vergelijkingen waarin een correlatiecoëfficiënt gelijkgesteld wordt aan een som van twee producten van factorladingen. De $a_{ij}$ zijn dus weer overgedetermineerd: we kunnen in principe geen waarden vinden die tegelijkertijd aan alle 15 vergelijkingen voldoen. Dit model met twee gemeenschappelijke factoren legt dus nog steeds beperkingen op aan de data en gaat dus niet op voor willekeurig welke correlatiematrix. We moeten weer, gegeven onze data, schattingen vinden voor de factorladingen $a_{ij}$ en dan nagaan hoe goed deze de geobserveerde correlaties reproduceren. Het vinden van deze schattingen kan opnieuw via het ML- of het ULS-criterium, die in dit geval (bij deze data en twee gemeenschappelijke factoren) precies gelijk zijn aan elkaar[2]. De oplossing wordt getoond in Tabel 5.3.

| | $F_1$ | $F_2$ |
|---|---|---|
| $X_1$ | 0.89 | -0.14 |
| $X_2$ | 0.79 | -0.12 |
| $X_3$ | 0.50 | 0.49 |
| $X_4$ | 0.43 | 0.42 |
| $X_5$ | 0.36 | 0.35 |
| $X_6$ | 0.30 | -0.04 |

Tabel 5.3: De twee-factoroplossing voor de correlatiematrix in Tabel 5.1.

Volgens deze *factorladingenmatrix* (in het Engels, *factor pattern matrix*) geldt dus bijvoorbeeld dat (de geschatte) $a_{11} = 0.89$, $a_{32} = 0.49$, enzovoort. Voor de vergelijkingen in termen van de latente factoren betekent het bijvoorbeeld dat

$$\begin{aligned} X_1 &= 0.89F_1 + (-.14)F_2 + U_1, \\ X_2 &= 0.79F_1 + (-.12)F_2 + U_2 \quad , \end{aligned}$$

enzovoort.

[2]Dit is altijd het geval als de residuele correlaties allemaal gelijk zijn aan 0 (zie Tabel 5.4).

Net zoals bij de één-factoroplossing gaan we nu onderzoeken hoe goed deze oplosssing met twee gemeenschappelijke factoren is. Met andere woorden, we gaan onderzoeken hoe goed de geobserveerde correlaties nu gereproduceerd worden. Voor de gereproduceerde correlatie tussen $X_1$ en $X_2$ geldt nu bijvoorbeeld:

$$\widehat{r}(X_1, X_2) = 0.89 \cdot 0.79 + (-0.14) \cdot (-0.12) = 0.72 \quad .$$

Aangezien $r(X_1, X_2) = 0.72$ betekent dit een perfect terugvinden van de geobserveerde correlatie. Maar $\widehat{r}(X_1, X_2)$ was ook al goed met slechts één factor. Laten we eens naar de correlatie voor de variabelen $X_3$ en $X_5$ kijken, die met één factor slecht werd gereproduceerd. Nu krijgen we

$$\widehat{r}(X_3, X_5) = 0.50 \cdot 0.36 + 0.49 \cdot 0.35 = 0.35 \quad ,$$

wat nu ook perfect overeenstemt met $r(X_3, X_5)$. Het totaaloverzicht van de gereproduceerde en residuele correlaties wordt getoond in Tabel 5.4. In dit geval, zijn alle residuele correlaties gelijk aan nul en hebben we dus een perfecte oplossing met twee gemeenschappelijke factoren. Voor deze correlatiematrix is dus geen derde factor nodig. Met nadruk moet erop gewezen worden dat dit heel uitzonderlijk is; de voorbeeldcorrelatiematrix was speciaal zo geconstrueerd.

Een perfecte factoroplossing is heel uitzonderlijk. Dit komt omdat we in de praktijk altijd te maken hebben met correlaties die op een eindig aantal observaties berekend zijn. Dit impliceert dat, indien het twee-factormodel opgaat voor de populatiecorrelaties (i.e., de correlaties berekend zijn op een oneindig aantal observaties), dit model vrijwel nooit perfect opgaat voor de geobserveerde correlaties. Dit komt omdat de geobserveerde correlaties vrijwel altijd afwijken van de populatiecorrelaties. Het is belangrijk om te zien dat het hier gaat om toevallige afwijkingen van het twee-factormodel, die afnemen naarmate het aantal observaties toeneemt.

In de praktijk, met echte data, kunnen we dus alleen hopen dat de residuele correlaties klein zijn, maar niet dat ze perfect gelijk zijn aan nul. Als ze klein zijn, dan spreken we van een goede oplossing met twee gemeenschappelijke factoren. De vraag is natuurlijk wat "klein" is. Hiervoor bestaat geen hard criterium, en daarom moeten we ons behelpen met een vuistregel. Een vaak gehanteerde vuistregel is dat alle residuele correlaties in absolute waarde kleiner moeten zijn dan 0.05.

**Gereproduceerd:**

| | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ |
|---|---|---|---|---|---|
| $X_1$ | | | | | |
| $X_2$ | 0.72 | | | | |
| $X_3$ | 0.38 | 0.34 | | | |
| $X_4$ | 0.32 | 0.29 | 0.42 | | |
| $X_5$ | 0.27 | 0.24 | 0.35 | 0.30 | |
| $X_6$ | 0.27 | 0.24 | 0.13 | 0.11 | 0.09 |

**Residueel:**

| | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ |
|---|---|---|---|---|---|
| $X_1$ | | | | | |
| $X_2$ | 0.00 | | | | |
| $X_3$ | 0.00 | 0.00 | | | |
| $X_4$ | 0.00 | 0.00 | 0.00 | | |
| $X_5$ | 0.00 | 0.00 | 0.00 | 0.00 | |
| $X_6$ | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |

Tabel 5.4: Matrix met gereproduceerde en residuele correlaties op basis van het twee-factormodel.

Bij toepassingen van de factoranalyse hoeft de onderzoeker nooit zélf de gereproduceerde en residuele correlaties te berekenen. Dit komt omdat een factoranalyse altijd uitgevoerd wordt met een computer en de meeste computerprogramma's berekenen, behalve de factorladingen, ook de gereproduceerde en residuele correlaties (als daarom gevraagd wordt). Oudere versies van SPSS drukken de gereproduceerde en residuele correlaties samen in één matrix af. Een voorbeeld van zo'n matrix wordt gegeven in Tabel 5.5. Op de diagonaal van deze matrix staan getallen met een "*". Dit zijn de zogenaamde *communaliteiten* die in de volgende paragraaf besproken worden. De getallen boven de diagonaal zijn de residuele correlaties en de getallen onder de diagonaal zijn de gereproduceerde correlaties. In de meest recente versies van SPSS worden de gereproduceerde en residuele correlaties in aparte matrices afgedrukt.

Reproduced Correlation Matrix:

| | Item 1 | Item 2 | Item 3 | Item 4 | Item 5 | Item 6 | Item 7 |
|---|---|---|---|---|---|---|---|
| Item 1 | .860* | .000 | .001 | .024 | -.019 | -.003 | -.012 |
| Item 2 | .783 | .719* | -.001 | -.001 | -.008 | .015 | -.006 |
| Item 3 | .845 | .771 | .830* | -.027 | .027 | -.005 | .019 |
| Item 4 | -.527 | -.436 | -.505 | .671* | .001 | .005 | -.029 |
| Item 5 | -.542 | -.450 | -.519 | .683 | .687* | -.006 | .037 |
| Item 6 | -.414 | -.338 | -.396 | .571 | .582 | .490* | -.021 |
| Item 7 | .017 | .030 | .011 | .076 | .054 | .057 | .242* |

Tabel 5.5: Voorbeeld van een matrix met gereproduceerde correlaties (onder de diagonaal), residuele correlaties (boven de diagonaal), en communaliteiten (op de diagonaal).

## 5.3 Ladingen, Communaliteit en Eigenwaarde

### 5.3.1 Ladingen als Correlaties

Zoals gezegd zijn de factorladingen regressiegewichten. De gevonden twee-factoroplossing resulteerde bijvoorbeeld in de volgende vergelijking voor $X_1$:

$$X_1 = 0.89F_1 - 0.14F_2 + U_1 \quad .$$

Hierin is dus $a_{11} = 0.89$ het regressiegewicht van de latente predictor $F_1$ en $a_{12} = -0.14$ is het regressiegewicht van de latente predictor $F_2$. Aangezien $X_1$, $F_1$ en $F_2$ gestandaardiseerde variabelen zijn, zijn de regressiegewichten hier de gestandaardiseerde regressiegewichten uit de regressieanalyse (ook bètagewichten genoemd). De twee latente predictoren $F_1$ en $F_2$ zijn ongecorreleerde variabelen; $r(F_1, F_2) = 0$. Uit de regressieanalyse weten we nog dat in zo'n geval het bètagewicht een zeer eenvoudige interpretatie heeft: het is een correlatie, en in dit geval de correlatie tussen een variabele en een gemeenschappelijke factor. Dus, $a_{11} = 0.89$ is de correlatie tussen $X_1$ en $F_1$, en $a_{12} = -0.14$ is de correlatie tussen $X_1$ en $F_2$. Evenzo geldt $r(X_4, F_2) = 0.42, r(X_6, F_1) = 0.30$,

enzovoort. In het algemeen geldt dus dat bij orthogonale factoren de factorladingenmatrix gelijk is aan de matrix van correlaties tussen geobserveerde variabelen en gemeenschappelijke factoren. Aangezien deze ladingen de enige informatie vormen die we over de latente factoren bezitten, kunnen we stellen dat de factoren hypothetische variabelen zijn die we alleen maar kennen via hun correlaties met de geobserveerde variabelen. We zullen later zien dat we op basis hiervan ook zullen proberen tot een interpretatie van de gevonden factoren te komen.

### 5.3.2 De Communaliteit van een Variabele

We vertrekken van het feit dat de variantie van een som van ongecorreleerde variabelen gelijk is aan de som van hun varianties. Omdat de twee gemeenschappelijke factoren ongecorreleerd zijn, kunnen we nu gemakkelijk de door de gemeenschappelijke factoren verklaarde variantie uitrekenen. Bijvoorbeeld, voor variabele $X_1$ leidt de gevonden vergelijking

$$X_1 = 0.89F_1 - 0.14F_2 + U_1 \quad ,$$

tot de volgende vergelijking in termen van varianties:

$$\text{VAR}(X_1) = \text{VAR}(0.89 \cdot F_1) + \text{VAR}(-0.14 \cdot F_2) + \text{VAR}(U_1) \quad .$$

Om tot dit resultaat te komen, wordt gebruikgemaakt van het feit dat $F_1$, $F_2$ en $U_1$ onderling ongecorreleerd zijn. En verder, omdat de variantie van een constante maal een variabele gelijk is aan het kwadraat van die constante maal de variantie van die variabele, kan men de bovenstaande formule als volgt schrijven:

$$\text{VAR}(X_1) = (0.89)^2 \cdot \text{VAR}(F_1) + (-0.14)^2 \cdot \text{VAR}(F_2) + \text{VAR}(U_1) \quad .$$

Omdat $X_1$, $F_1$ en $F_2$ gestandaardiseerde variabelen zijn, kunnen we dit als volgt uitwerken:

$$\begin{aligned} 1 &= (0.89)^2 + (-0.14)^2 + \text{VAR}(U_1) \\ 1 &= 0.79 + 0.02 + \text{VAR}(U_1) \\ 1 &= 0.81 + \text{VAR}(U_1) \quad . \end{aligned}$$

Via deze berekening zien we dus dat 0.79 van de variantie van $X_1$ wordt verklaard door de factor $F_1$, en 0.02 van deze variantie door $F_2$. In totaal, verklaren de gemeenschappelijke factoren dus 0.81 van de variantie van $X_1$. Het gedeelte variantie van een variabele dat wordt verklaard door de gemeenschappelijke factoren heet de *communaliteit* van de variabele. Het resterende deel, de onverklaarde variantie, ofwel de variantie van de unieke factor, heet de *uniciteit* van de variabele. Hier heeft $X_1$ dus een uniciteit van $1 - 0.81 = 0.19$. Omdat alle geobserveerde variabelen gestandaardiseerd zijn, komt het gedeelte verklaarde variantie rechtstreeks overeen met het percentage verklaarde variantie: 81% van de variantie van $X_1$ wordt verklaard door de gemeenschappelijke factoren, 19% van de variantie zit in de unieke factor $U_1$. In bovenstaande berekening hebben we gezien dat de communaliteit van een variabele gelijk is aan de som van de kwadraten van de ladingen van deze variabele op alle gemeenschappelijke factoren. Op deze manier zijn de communaliteiten dus gemakkelijk te berekenen

uit de factorladingenmatrix: kwadrateer alle ladingen in een bepaalde rij en tel deze kwadraten bij elkaar op. Deze communaliteiten zijn in Tabel 5.6 als een extra kolom aan de factorladingenmatrix toegevoegd. Het symbool $h_i^2$ is de, in de factoranalytische literatuur gebruikelijke, aanduiding voor de communaliteit van variabele $X_i$. We hebben bijvoorbeeld $h_4^2 = 0.43^2 + 0.42^2 = 0.36$, enzovoort.

| | $F_1$ | $F_2$ | $h_i^2$ |
|---|---|---|---|
| $X_1$ | 0.89 | -0.14 | 0.81 |
| $X_2$ | 0.79 | -0.12 | 0.64 |
| $X_3$ | 0.50 | 0.49 | 0.49 |
| $X_4$ | 0.43 | 0.42 | 0.36 |
| $X_5$ | 0.36 | 0.35 | 0.25 |
| $X_6$ | 0.30 | -0.04 | 0.09 |
| EW | 2.07 | 0.57 | 2.64 |
| % | 34.5% | 9.5% | 44% |

Tabel 5.6: Communaliteiten en eigenwaarden voor de twee-factoroplossing voor de correlatiematrix in Tabel 5.1.

Theoretisch kan de communaliteit van een variabele nooit groter zijn dan 1. De geobserveerde variantie van een variabele is immers gelijk aan 1 (want de variabelen zijn gestandaardiseerd) en men kan nooit méér variantie verklaren dan er geobserveerd is. In de praktijk komt het echter soms voor dat sommige variabelen een communaliteit groter dan 1 hebben. Men spreekt dan van een *Heywoodgeval*. Hiervoor zijn twee mogelijke oorzaken: (1) toevallige schommelingen in de data, en (2) men heeft te weinig factoren "getrokken[3]". Positief geformuleerd, als er geen toevallige schommelingen zitten in de data (i.e., als de correlaties berekend zijn op een groot aantal personen) en als het aantal getrokken factoren groter dan of gelijk is aan het aantal factoren dat de data gegenereerd heeft, dan komen Heywoodgevallen niet voor.

Het is goed om te weten dat het ML-criterium vaker tot Heywoodgevallen leidt dan het ULS-criterium. Indien, voor een bepaalde dataset, ook het ULS-criterium in een of meerdere Heywoodgevallen resulteert, dan is de enige mogelijke oplossing het vermeerderen van het aantal factoren.

### 5.3.3 De Eigenwaarde van een Factor

In het bovenstaande hebben we de verklaarde variantie per variabele bekeken. We kunnen ook bekijken hoeveel variantie in totaal verklaard wordt door iedere factor. Daarvoor tellen we gewoon van iedere factor bij elkaar op hoeveel variantie hij van iedere variabele verklaart. We berekenen dus weer de kwadraten van de ladingen, maar tellen deze nu per kolom bij elkaar op. Het resultaat hiervan staat ook in bovenstaande tabel, in de rij EW onder de variabelen. Bijvoorbeeld de totale hoeveelheid variantie verklaard door factor $F_1$ bedraagt

$$0.89^2 + 0.79^2 + 0.50^2 + 0.43^2 + 0.36^2 + 0.30^2 = 2.07 \quad .$$

[3]Het berekenen van de factorladingen van een aantal variabelen op een aantal factoren wordt in de literatuur vaak aangeduid als het "trekken" van factoren.

Deze totale hoeveelheid variantie verklaard door een gemeenschappelijke factor wordt de *eigenwaarde*[4] (in het Engels, *eigenvalue*) van de factor genoemd (vandaar de afkorting EW in Tabel 5.6). $F_1$ heeft dus een eigenwaarde van 2.07, terwijl de eigenwaarde van $F_2$ gelijk is aan $(-0.14)^2 + (-0.12)^2 + \ldots = 0.57$. Gezamenlijk verklaren de gemeenschappelijke factoren dus $2.07 + 0.57 = 2.64$ van de totale variantie van de geobserveerde variabelen. Omdat deze zes variabelen gestandaardiseerd zijn is de som van hun varianties gelijk aan 6 en door te delen door deze som vinden we de proporties verklaarde variantie. Voor $F_1$ wordt dit $2.07/6 = 0.345$, voor $F_2$ $0.57/6 = 0.095$. In percentages komt dit neer op 34.5% verklaard door $F_1$, 9.5% verklaard door $F_2$, en dus 44% verklaard door beide gemeenschappelijke factoren.

Contrasteer deze 44% verklaarde variantie met het feit dat er hier sprake is van een perfecte oplossing in twee gemeenschappelijke factoren. Dit illustreert dat het doel bij deze variant van de factoranalyse niet is om van de geobserveerde variabelen zoveel mogelijk variantie te verklaren, maar wel om zo goed mogelijk hun onderlinge correlaties te verklaren. Zoals we later zullen zien, ligt de zaak anders bij de hoofdcomponentenanalyse, de tweede variant van de factoranalyse. Daar worden de ladingen zó bepaald dat de bijbehorende factoren zoveel mogelijk variantie verklaren.

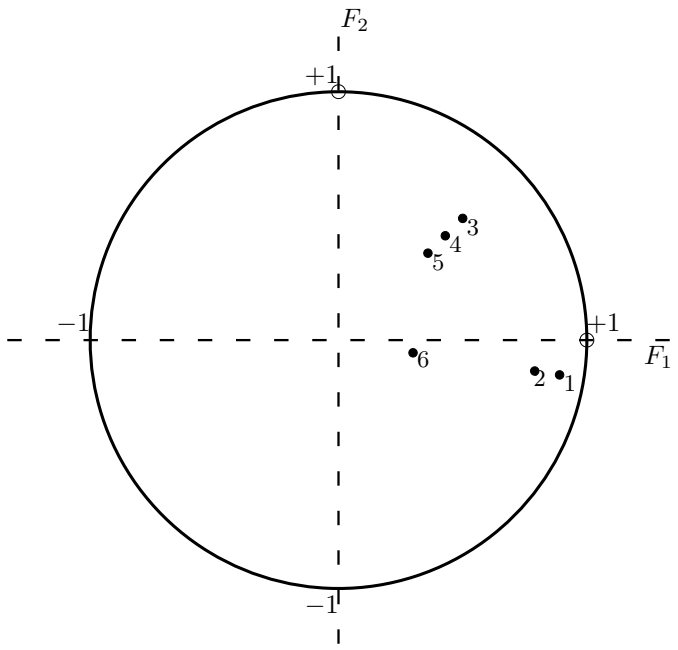
### 5.3.4 Grafische Weergave van de Factoroplossing

Nu we een oplossing hebben met slechts twee gemeenschappelijke factoren, kunnen we de gevonden oplossing mooi grafisch weergeven in een zogenaamde *factorladingenplot*. Dit doen we in Figuur 5.3. In deze figuur is een rechthoekig assenstelsel getekend, waarin de assen geassocieerd zijn met de gemeenschappelijke factoren; in de figuur is $F_1$ langs de horizontale en $F_2$ langs de verticale as afgezet. Het feit dat $F_1$ en $F_2$ ongecorreleerd zijn is in de tekening weergegeven door het loodrecht op elkaar staan van de twee assen. (Het Griekse woord voor rechthoekig of loodrecht is *orthogonaal*, en dit verklaart de benaming *orthogonale factoren* voor ongecorreleerde factoren.) In dit assenstelsel zijn de variabelen $X_1$ t/m $X_6$ getekend als dikke punten met volgnummers 1 t/m 6. Deze zijn zó geplaatst dat de coördinaten overeenkomen met de ladingen op de factoren. Zo heeft het punt 2 in de tekening de coördinaten $(0.79, -0.12)$ (dit wil zeggen, dit punt ligt horizontaal gezien 0.79 rechts van de oorsprong en verticaal gezien 0.12 onder de oorsprong). Deze getallen zijn gelijk aan de ladingen van $X_2$ op, respectievelijk, de factoren $F_1$ en $F_2$. Dit geldt ook voor de andere variabelen. Het feit dat de coördinaten van de variabelen in de tekening gelijk zijn aan de ladingen op de factoren betekent feitelijk dat volgnummer $i$ niet precies variabele $X_i$ representeert, maar eigenlijk de variabele

$$\widehat{X}_i = X_i - U_i = a_{i1}F_1 + a_{i2}F_2 \quad ,$$

het deel van $X_i$ dat wordt verklaard door $F_1$ en $F_2$. De factoren zelf komen ook in de tekening voor, namelijk als de twee open cirkels met coördinaten (1,0) voor $F_1$ en (0,1) voor $F_2$. Dit wil zeggen, iedere factor laadt 1 op zichzelf en 0 op de andere factor. Dit is nog steeds consistent met de interpretatie van ladingen als correlaties, aangezien inderdaad elke factor een correlatie van 1 met zichzelf heeft en van 0 met de andere factor (ze zijn immers orthogonaal).

[4]Dit begrip heeft niets te maken met het gelijknamige psychologisch construct.

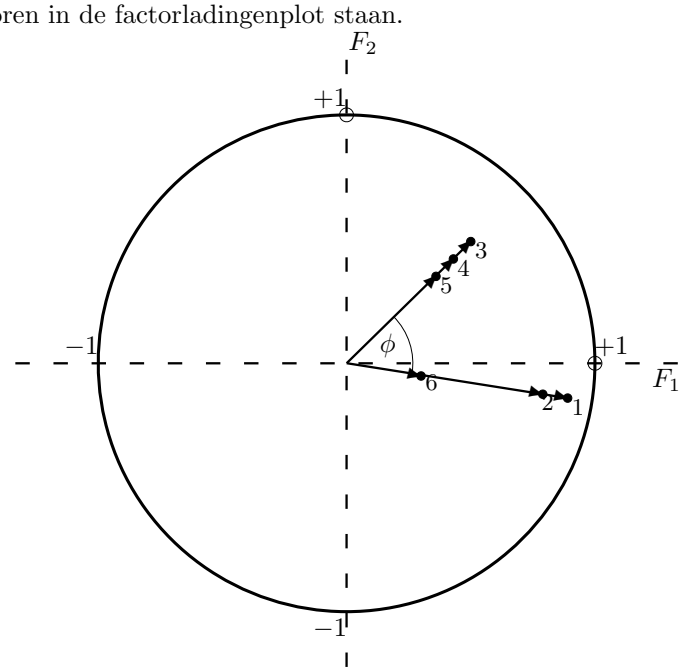



Figuur 5.3: Factorladingenplot van de twee-factoroplossing in Tabel 5.6.

Volgens de aloude stelling van Pythagoras is het kwadraat van de afstand van een punt tot de oorsprong gelijk aan de som van de kwadraten van de coördinaten van dat punt. Dus, aangezien coördinaten corresponderen met ladingen, is het kwadraat van de afstand van een punt in de factorladingenplot tot de oorsprong gelijk aan de som van de kwadraten van de ladingen van de met dat punt geassocieerde variabele. In het voorgaande hebben we gezien dat deze som aangeduid werd als de communaliteit van deze variabele. Kortweg, de communaliteit van een variabele wordt in bovenstaande figuur weerspiegeld door (het kwadraat van) de afstand van die variabele tot de oorsprong. Zo is direct zichtbaar dat bijvoorbeeld $X_1$ een aanmerkelijk hogere communaliteit heeft dan $X_5$ (controleer dit in de factorladingenmatrix). De grote cirkel in Figuur 5.3 bevat de punten met een afstand van 1 tot de oorsprong. Omdat we met gestandaardiseerde variabelen werken (variabelen met een variantie gelijk aan 1), is de communaliteit, het gedeelte van deze variantie verklaard door de gemeenschappelijke factoren, altijd kleiner dan of gelijk aan 1. Bijgevolg liggen alle variabelen bínnen deze eenheidscirkel. Naarmate variabelen een hogere communaliteit hebben liggen ze verder naar buiten, naarmate ze een hogere uniciteit hebben liggen ze verder naar binnen. Op de cirkel zelf komen variabelen terecht met een communaliteit van 1, dit wil zeggen variabelen die volledig door de gemeenschappelijke factoren worden verklaard. Zulke variabelen hebben we hier (zelfs in dit kunstmatige voorbeeld) niet waargenomen. De factoren zelf liggen natuurlijk altijd op de cirkel; zij worden uiteraard, op triviale wijze, volledig door zichzelf verklaard.

Men kan zich op een eenvoudige manier de gereproduceerde correlaties ruimtelijk voorstellen. Het vertrekpunt hierbij is dat men zich de punten in de factorladingenplot

als *vectoren* voorstelt. Een vector is een pijl die vertrekt van de oorsprong van het assenstelsel en eindigt in een bepaald punt (in dit geval, een punt met als coördinaten de factorladingen van een bepaalde test). Deze vectoren worden getoond in Figuur 5.4. Omdat de punten in groepjes van drie in elkaars verlengde liggen (3, 4, en 5 zijn één groepje en 1, 2, en 6 zijn een ander groepje) zijn er slechts twee lijnen te zien, terwijl er toch zes vectoren in de factorladingenplot staan.




Figuur 5.4: Factorladingenplot van de twee-factoroplossing in Tabel 5.6 met de variabelen als vectoren.

De gereproduceerde correlatie tussen twee variabelen hangt af van (1) de hoek tussen de vectoren die deze variabelen voorstellen en (2) de lengte van deze vectoren. In Figuur 5.4 wordt de grootte van de hoek aangeduid met de Griekse letter $\phi$ (spreek uit "fie"). We weten dat de communaliteit van een variabele $X_i$ (aangeduid met $h_i^2$) gelijk is aan het kwadraat van de lengte van de vector die deze variabele voorstelt. Hieruit volgt dat deze lengte eenvoudig te berekenen is als de vierkantswortel uit de communaliteit. De lengte van de vector voor de variabele $X_i$ duiden we aan met $h_i$ (dus, zonder het kwadraat). Nu, de gereproduceerde correlatie kan als volgt berekend worden:

$$\boxed{\widehat{r}(X_i, X_j) = h_i h_j \cos(\phi) \quad ,}$$

het product van de lengtes van de twee vectoren en de cosinus van de hoek tussen beide. Van de cosinus moet men weten dat hij (1) gelijk is aan 0 als de hoek precies gelijk is aan 90 graden, (2) toeneemt van 0 naar 1 als de hoek afneemt van 90 naar 0 graden, en (3) kleiner is dan 0 als de hoek groter is dan 90 graden. Uit het feit dat de hoek $\phi$ in Figuur 5.4 kleiner is dan 90 graden, volgt dat alle zes variabelen positief met elkaar correleren.

Met een gradenboog kun je controleren dat $\phi$ gelijk is aan 53.13 graden. De cosinus van 53.13 graden is 0.60. Gebruiken we nu de bovenstaande formule voor de berekening van de gereproduceerde correlatie tussen $X_2$ en $X_3$, dan halen we eerst $h_2 = \sqrt{0.64} = 0.8$ en $h_3 = \sqrt{0.49} = 0.7$ uit Tabel 5.6, en berekenen dan het volgende:

$$\begin{aligned} \widehat{r}(X_2, X_3) &= 0.8 \times 0.7 \times 0.6 \\ &= 0.336 \quad , \end{aligned}$$

hetgeen na afronding gelijk is aan 0.34, het getal dat we ook vinden in Tabel 5.4. Als we de gereproduceerde correlatie berekenen tussen variabelen waarvan de vectoren in elkaars verlengde liggen, dan hebben we te maken met een hoek van 0 graden, en de cosinus hiervan is gelijk aan 1. Voor de berekening van de gereproduceerde correlatie tussen $X_1$ en $X_6$ halen we eerst $h_1 = \sqrt{0.81} = 0.9$ en $h_6 = \sqrt{0.09} = 0.3$ uit Tabel 5.6, en berekenen dan het volgende:

$$\begin{aligned} \widehat{r}(X_1, X_6) &= 0.9 \times 0.3 \times 1.0 \\ &= 0.27 \quad , \end{aligned}$$

het getal dat we ook vinden in Tabel 5.4.

---

*Opgave 1*
Een onderzoekster heeft een factoranalyse uitgevoerd op vier tests. Ze heeft twee ongecorreleerde gemeenschappelijke factoren getrokken. De factorladingen staan in Tabel 5.7.

| | Factor 1 | Factor 2 |
|---|---|---|
| Test 1 | 0.70 | -0.30 |
| Test 2 | 0.50 | 0.40 |
| Test 3 | -0.80 | 0.60 |
| Test 4 | 0.40 | -0.70 |

Tabel 5.7: Tabel bij opgave 1.

1. Welke variabele heeft (in absolute waarde) de hoogste correlatie met de tweede factor?
2. Van welke variabele wordt de meeste variantie verklaard door de factoroplossing?
3. Welke factor heeft de kleinste eigenwaarde?
4. Welke factor verklaart de meeste variantie van de variabelen?
5. Wat is de gereproduceerde correlatie tussen Test 1 en Test 4?

*Opgave 2*
De correlaties tussen vier tests (Test 1 t/m Test 4) werden onderworpen aan een factoranalyse. De factorladingen worden getoond in Tabel 5.8.

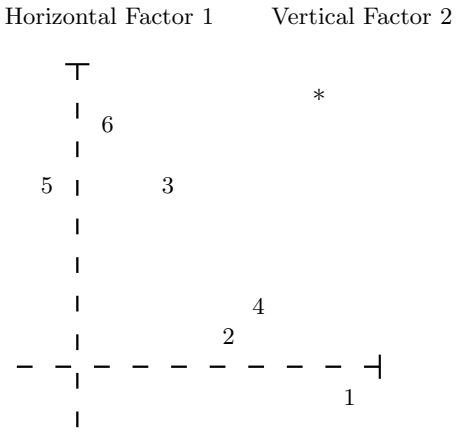
1. Wat is de gereproduceerde correlatie tussen Test 2 en Test 4?
2. De geobserveerde correlatie tussen Test 2 en Test 4 is 0.1. Wat is dan de residuele correlatie?
3. Welke test heeft de hoogste communaliteit? Laat zien hoe je aan je antwoord komt.
4. Wat verstaan we onder communaliteit? (Alleen de rekenformule geven is onvoldoende.)

| | Factor 1 | Factor 2 |
|---|---|---|
| Test 1 | -0.70 | 0.60 |
| Test 2 | -0.60 | 0.70 |
| Test 3 | 0.50 | 0.60 |
| Test 4 | 0.60 | 0.50 |

Tabel 5.8: Tabel bij opgave 2.

5. Maak een factorladingenplot van deze vier tests.

*Opgave 3*
In een factoranalyse op 6 variabelen worden twee gemeenschappelijke factoren getrokken. De factoroplossing wordt gegeven in Figuur 5.5. De uiteinden van de horizontale en de verticale as corresponderen met de punten $(1, 0)$ en $(0, 1)$.




Figuur 5.5: Tabel bij opgave 3.

1. Welke van de variabelen 2 of 3 correleert hoger met factor 1, of valt dit uit de figuur niet te beantwoorden? Leg uit.
2. Van welk variabelenpaar wordt de correlatie het best gereproduceerd, of valt dit uit de figuur niet te beantwoorden? Leg uit.
3. Had een van de variabelen terecht kunnen komen op de plaats van de $*$ (rechtsboven in de figuur)? Leg uit.
4. Had een van de variabelen in principe precies in de oorsprong van het assenstelsel terecht kunnen komen? Zo ja, welke waarden van de geobserveerde correlaties kunnen hiervoor verantwoordelijk zijn?

*Opgave 4*
De correlaties tussen zes tests (Test 1 t/m Test 6) werden onderworpen aan een factoranalyse. De factorladingen worden getoond in Tabel 5.9.

1. Wat is de gereproduceerde correlatie tussen Test 3 en Test 5?
2. Welke test heeft de hoogste communaliteit? Laat zien hoe je aan je antwoord komt.

| | Factor 1 | Factor 2 |
|---|---|---|
| Test 1 | 0.20 | 0.40 |
| Test 2 | 0.40 | 0.00 |
| Test 3 | 0.30 | 0.60 |
| Test 4 | 0.60 | 0.00 |
| Test 5 | 0.40 | 0.80 |
| Test 6 | 0.80 | 0.00 |

Tabel 5.9: Tabel bij opgave 4.

3. Kun je uit deze uitvoer afleiden of de twee-factoroplossing voldoende is voor de verklaring van de geobserveerde correlaties? Motiveer je antwoord.
4. Als je een test wilt die alleen de eerste gemeenschappelijke factor meet, welke zou je dan kiezen? Motiveer je antwoord.
5. Maak een factorladingenplot van deze zes tests.

*Opgave 5*
Een onderzoekster beschikt over twee groepen van telkens zes tests (Test 1 t/m Test 6 en Test 7 t/m Test 12). Zij voert twee aparte factoranalyses uit, één op elk van de twee groepen van zes tests. In Tabel 5.10 vind je een deel van de uitvoer van een factoranalyse op de eerste en de tweede groep van zes tests. Door één van de factoroplossingen (Analyse 1 of Analyse 2) wordt de *samenhang* tussen de variabelen beter beschreven dan door de andere. Door de andere factoroplossing wordt de *variantie* van de variabelen beter verklaard. Welke analyse doet wat? Beargumenteer je antwoord.

Analyse 1:
Reproduced Correlation Matrix:

| | Test 1 | Test 2 | Test 3 | Test 4 | Test 5 | Test 6 |
|---|---|---|---|---|---|---|
| Test 1 | .8045* | .0904 | .1254 | .0913 | .1654 | .2002 |
| Test 2 | .4563 | .7021* | .2434 | .1009 | .0786 | .2300 |
| Test 3 | .4521 | .3264 | .7021* | .1023 | .0547 | .1245 |
| Test 4 | .3214 | .2314 | -.2341 | .6843* | .1209 | .0234 |
| Test 5 | .2641 | .4213 | .1978 | .3251 | .6903* | .1232 |
| Test 6 | .2341 | .5653 | .1425 | .1213 | .4276 | .6532* |

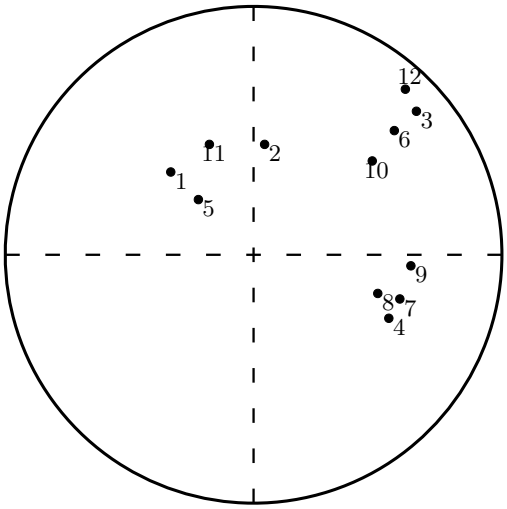
Analyse 2:
Reproduced Correlation Matrix:

| | Test 7 | Test 8 | Test 9 | Test 10 | Test 11 | Test 12 |
|---|---|---|---|---|---|---|
| Test 7 | .5437* | .0123 | .0912 | .0521 | .0286 | .1001 |
| Test 8 | .2356 | .6432* | .0254 | .0712 | .0743 | .0912 |
| Test 9 | .3212 | .2451 | .4987* | .0019 | .0231 | .0601 |
| Test 10 | .2123 | .3100 | .0981 | .4991* | .0101 | .0612 |
| Test 11 | .3111 | .2171 | .2215 | .1190 | .5009* | .0019 |
| Test 12 | .2126 | .3001 | .1983 | .1543 | .2513 | .4132* |

Tabel 5.10: Tabel bij opgave 5.

*Opgave 6*
In een factoranalyse op 12 variabelen worden twee gemeenschappelijke factoren getrokken. De factoroplossing wordt gegeven in Figuur 5.6.

1. Welke variabele heeft de laagste communaliteit, of valt dit uit de plot niet te beantwoorden? Leg uit.
2. Van welke paren van variabelen is de gereproduceerde correlatie (vrijwel) gelijk aan nul, of valt dit uit de plot niet te beantwoorden? Leg uit.
3. Waar of niet waar en leg uit: uit deze plot valt op te maken dat de twee-factoroplossing niet goed voldoet en dat er eigenlijk drie factoren in het spel zijn.

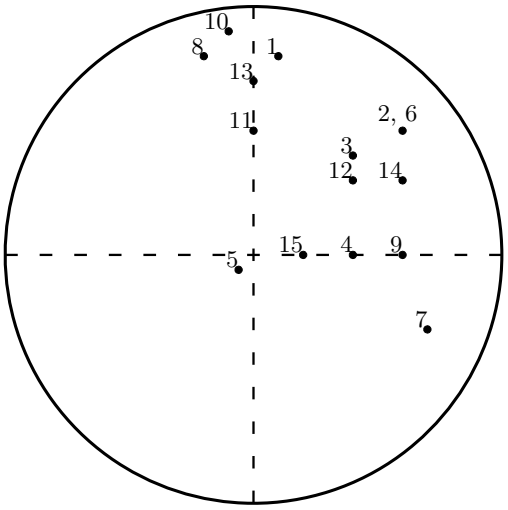



Figuur 5.6: Figuur bij opgave 6.

*Opgave 7*
In een factoranalyse op 15 variabelen werden twee gemeenschappelijke factoren getrokken. De ladingen van de variabelen op de factoren worden weergegeven in Figuur 5.7.

1. Welke variabele correleert het sterkst met de factor die op de horizontale as afgebeeld wordt? Leg uit.
2. Welke variabele heeft de grootste uniciteit? Leg uit.
3. Waar of niet waar en leg uit: de gereproduceerde correlatie tussen variabele 15 en variabele 4 is gelijk aan de gereproduceerde correlatie tussen variabele 15 en variabele 9.
4. Waar of niet waar en leg uit: de communaliteit van variabele 9 is gelijk aan de communaliteit van variabele 14.
5. Waar of niet waar en leg uit: de gereproduceerde correlatie tussen variabele 15 en variabele 11 is gelijk aan de gereproduceerde correlatie tussen variabele 15 en variabele 13?
6. Waar of niet waar en leg uit: als alle residuele correlaties precies gelijk zijn aan 0, kan je op basis van de gestandaardiseerde scores op variabele 6 de gestandaardiseerde scores op variabele 2 exact voorspellen.

*Opgave 8*
Van de volgende stellingen moet je aangeven of ze waar of onwaar zijn. Als je vindt dat een stelling onwaar is, leg dan ook uit waarom dat zo is.




Figuur 5.7: Figuur bij opgave 7.

1. De totale variantie die door een factor verklaard kan worden is maximaal gelijk aan 1.
2. De residuele correlatie van twee variabelen is dat gedeelte van de variantie dat door de unieke factoren wordt verklaard.
3. In absolute waarde is een residuele correlatie kleiner dan of gelijk aan de overeenkomstige geobserveerde correlatie.
4. Als de geobserveerde correlaties laag zijn, dan zijn de residuele correlaties hoog.

---

## 5.4 Hoofdcomponentenanalyse

Tot nu toe waren gemeenschappelijke factoren latente variabelen die de correlaties tussen de geobserveerde variabelen verklaarden. In de hoofdcomponentenanalyse wijken we hiervan af: nu zijn de gemeenschappelijke variabelen latente variabelen die de varianties van de geobserveerde variabelen verklaren. Het verklaren van varianties gebeurt met een ander model dan het verklaren van correlaties. Het verschil met het model dat we gebruiken om de correlaties te verklaren, is dat de residuen (de $U_i$'s) nu niet langer onderling ongecorreleerd zijn. Omdat we nu toelaten dat de residuen wél gecorreleerd zijn, vervangen we in de formules $U_i$ (de unieke factor van de i-de variabele) door $R_i$ (het residu van de i-de variabele). Begrijpelijkerwijs, noemen we $R_i$ geen unieke factor, want een variabele die correleert met andere variabelen is moeilijk uniek te noemen. In wat volgt, spreken we gewoon van residu. Het feit dat de residuen nu gecorreleerd mogen zijn, heeft implicaties voor de correlaties tussen de geobserveerde variabelen $X_i$. Dit wordt nu uitgelegd.

### 5.4.1 Het Model en zijn Implicaties voor de Correlaties

We keren terug naar het model met één gemeenschappelijke factor. Met $U_i$ vervangen door $R_i$, kan het model met één gemeenschappelijke factor als volgt geschreven worden:

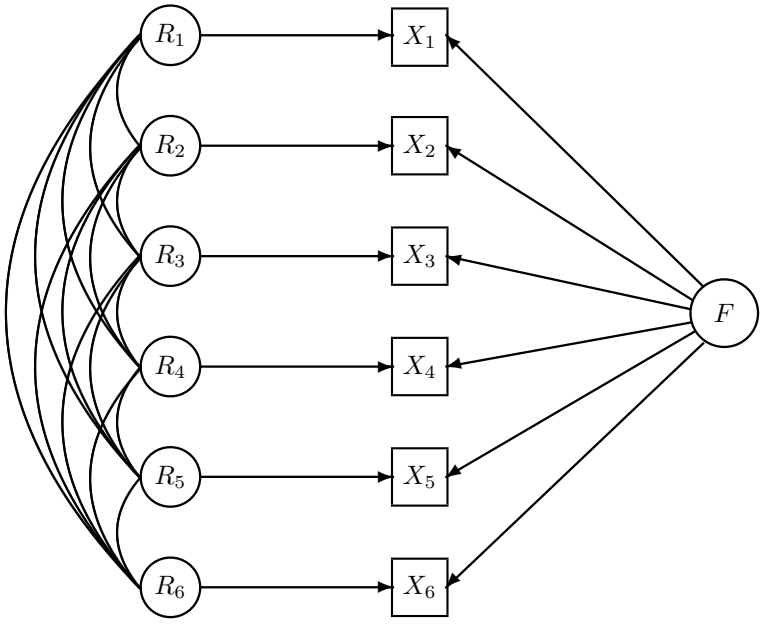
$$\boxed{X_i = a_i F + R_i \qquad \text{voor} \quad i = 1, \ldots, 6 \quad .}$$

Het essentiële kenmerk van een residu is dat het ongecorreleerd is met de gemeenschappelijke factoren. Dus, net zoals voor de unieke factoren, geldt het volgende:

$$\boxed{r(F, R_i) = 0 \qquad \text{voor} \quad i = 1, \ldots, 6 \quad .}$$

Het verschil met de unieke factoren is echter dat de residuen onderling gecorreleerd mogen zijn.

Men kan het model ook grafisch voorstellen en dit gebeurt in Figuur 5.8. Het nieuwe element in deze figuur zijn de verbindingen tussen de residuen. Deze verbindingen geven aan dat de residuen gecorreleerd zijn. In deze figuur is mooi te zien dat er twee bronnen zijn die bijdragen aan de correlaties tussen de geobserveerde variabelen: (1) het feit dat ze allemaal voor een deel opgebouwd zijn uit dezelfde gemeenschappelijke factor, en (2) het feit dat hun niet-gemeenschappelijke delen (de residuen) onderling gecorreleerd zijn.




Figuur 5.8: Grafische voorstelling van het factormodel met gecorreleerde residuen en één gemeenschappelijke factor.

We onderzoeken nu wat de implicaties zijn van dit model voor de correlaties tussen de geobserveerde variabelen $X_i$. Op dezelfde manier als voor het één-factormodel met

ongecorreleerde residuen, kan men de volgende formule afleiden voor de correlatie tussen twee willekeurige variabelen $X_i$ en $X_j$ $(i \neq j)$.

$$\boxed{r(X_i, X_j) = a_i \cdot a_j + \mathrm{COV}(R_i, R_j) \quad .}$$

Deze formule is anders dan de formule voor het model met ongecorreleerde residuen omdat $R_i$ en $R_j$ gecorreleerd kunnen zijn; hierdoor verschijnt de term $\mathrm{COV}(R_i, R_j)$ in de formule. Vertrekkende van het feit dat $X_i$ gelijk is aan $a_iF + R_i$ en $X_j$ aan $a_jF + R_j$, kunnen we deze formule afleiden:

$$\begin{aligned} r(X_i, X_j) &= \mathrm{COV}(X_i, X_j) \\ &= \mathrm{COV}(a_iF, a_jF) + \mathrm{COV}(a_iF, R_j) + \mathrm{COV}(R_i, a_jF) + \mathrm{COV}(R_i, R_j) \\ &= \mathrm{COV}(a_iF, a_jF) + \mathrm{COV}(R_i, R_j) \\ &= a_i \cdot a_j + \mathrm{COV}(R_i, R_j) \quad . \end{aligned}$$

In deze afleiding vallen de termen $\mathrm{COV}(a_iF, R_j)$ en $\mathrm{COV}(R_i, a_jF)$ weg omdat de residuen niet correleren met de gemeenschappelijke factor. De term $\mathrm{COV}(R_i, R_j)$ blijft staan omdat de residuen onderling wel mogen correleren.

Het verhaal is volledig analoog voor het model met twee gemeenschappelijke factoren. Dit model wordt als volgt geschreven:

$$\boxed{X_i = a_{i1}F_1 + a_{i2}F_2 + R_i \qquad \text{voor} \quad i = 1, \ldots, 6 \quad .}$$

De formule voor de correlatie tussen $X_i$ en $X_j$ $(i \neq j)$ is ook volledig analoog aan de formule voor het model met één gemeenschappelijke factor, en hij wordt op een analoge wijze afgeleid. Deze formule is de volgende:

$$\boxed{r(X_i, X_j) = a_{i1} \cdot a_{j1} + a_{i2} \cdot a_{j2} + \mathrm{COV}(R_i, R_j) \quad .}$$

### 5.4.2 Schattingen van de Factorladingen

De factorladingen kunnen niet op dezelfde manier geschat worden als bij het factormodel met ongecorreleerde residuen. Daar werden de factorladingen op zo'n manier bepaald dat de residuele correlaties zo klein mogelijk waren. De reden waarom we deze methode niet kunnen gebruiken bij het factormodel met gecorreleerde residuen is dat het aantal onbekenden groter is dan het aantal correlaties, zelfs in het model met slechts één gemeenschappelijke factor. Concreet, met zes geobserveerde variabelen, bevat het éénfactormodel de volgende onbekenden: zes onbekende factorladingen $a_i$ en 15 onbekende covariantietermen $\mathrm{COV}(R_i, R_j)$. Tegenover deze 21 onbekenden staan 15 geobserveerde correlaties. Door elk van de 15 covariantietermen $\mathrm{COV}(R_i, R_j)$ gelijk te stellen aan de corresponderende correlatie $r(X_i, X_j)$, verkrijgen we op een triviale manier dat de geobserveerde en de gereproduceerde correlaties gelijk zijn aan elkaar. Omdat zelfs het model met slechts één gemeenschappelijk factor geen restricties oplegt aan de gegevens (een

perfecte fit is altijd mogelijk), kan men zich met recht afvragen of er hier wel sprake is van een model. Immers, van een model dat opgaat voor elke denkbare correlatiematrix, hoe gek ook, kan men zich afvragen of het wel iets zegt over de gegevens.

We vinden een uitweg uit deze probleemsituatie als we het factormodel met gecorreleerde residuen niet langer beschouwen als een model dat restricties oplegt aan de data, maar wel als een middel om de data op een spaarzame manier te beschrijven. Met "spaarzaam" bedoelen we dat met een klein aantal gemeenschappelijke factoren zoveel mogelijk variantie in de geobserveerde variabelen verklaard wordt. De verklaarde variantie van een variabele wordt op dezelfde manier berekend als bij het factormodel met ongecorreleerde residuen. Concreet, voor het model met één gemeenschappelijke factor is de verklaarde variantie (communaliteit) van $X_i$ gelijk aan $a_i^2$, en voor het model met twee gemeenschappelijke factoren is hij gelijk aan $a_{i1}^2 + a_{i2}^2$. De reden waarom deze formules ook geldig zijn onder het model met gecorreleerde residuen, is dat hun afleiding alleen afhangt van de correlaties tussen de componenten bínnen een variabele (i.c., de gemeenschappelijke factoren en het residu) en niet van de correlaties tússen de variabelen.

We beschikken nu over een criterium op basis waarvan we de factorladingen kunnen schatten: bepaal de gemeenschappelijke factoren zodat zij zoveel mogelijk variantie van de geobserveerde variabelen verklaren. Voor een gegeven set gemeenschappelijke factoren zijn de geschatte factorladingen de gewone regressiegewichten van een analyse waarin de geobserveerde variabelen $X_i$ voorspeld worden op basis van deze gemeenschappelijke factoren. Concreet, voor het model met één gemeenschappelijke factor is dit criterium het volgende:

> Kies $F$ (met gemiddelde 0 en variantie 1) zodanig dat voor de bijbehorende factorladingen geldt dat
>
> $$\sum_{i=1}^{6} a_i^2 \qquad \text{maximaal is.}$$

De factorlading $a_i$ is het regressiegewicht van de predictor $F$ in de enkelvoudige regressieanalyse op de criteriumvariabele $X_i$. Omdat $X_i$ en $F$ gestandaardiseerd zijn, hebben we hier te maken met een gestandaardiseerde regressievergelijking, en is het regressiegewicht $a_i$ gelijk aan de correlatie tussen $X_i$ en $F$. De proportie verklaarde variantie voor de criteriumvariabele $X_i$ is in dit geval gelijk aan $a_i^2$ en de som van de proporties verklaarde variantie is gelijk aan $\sum_{i=1}^{6} a_i^2$. Alles draait hier om de keuze van de gemeenschappelijke factor $F$; als $F$ eenmaal bekend is, zijn de factorladingen uit te rekenen als eenvoudige correlaties. Hóe deze gemeenschappelijke factor, die de maximale variantie verklaart, bepaald kan worden, is voor de praktijk niet van belang. Het is voldoende om te weten dat dit gebeurt met een techniek die hoofdcomponentenanalyse genoemd wordt. De gemeenschappelijke factor die gevonden wordt met behulp van de hoofdcomponentenanalyse wordt ook vaak een hoofdcomponent genoemd.

In een hoofdcomponentenanalyse met twee gemeenschappelijke factoren ($F_1$ en $F_2$) worden deze factoren als volgt bepaald:

Kies $F_1$ en $F_2$ (beide met gemiddelde 0 en variantie 1, en onderling ongecorreleerd) zodanig dat voor de bijbehorende factorladingen geldt dat

$$\sum_{i=1}^{6}(a_{i1}^2 + a_{i2}^2) \qquad \text{maximaal is.}$$

De factorlading $a_{i1}$ en $a_{i2}$ zijn de regressiegewichten van de predictoren $F_1$ en $F_2$ in de meervoudige regressieanalyse op de criteriumvariabele $X_i$. Omdat $X_i$, $F_1$ en $F_2$ gestandaardiseerd zijn, hebben we hier te maken met een gestandaardiseerde regressievergelijking. En omdat $F_1$ en $F_2$ ongecorreleerd zijn, zijn de regressiegewichten $a_{i1}$ en $a_{i2}$ gelijk aan de correlatie tussen $X_i$ en $F_1$, respectievelijk, $X_i$ en $F_2$. De proportie verklaarde variantie voor de criteriumvariabele $X_i$ is in dit geval gelijk aan $a_{i1}^2 + a_{i2}^2$ en de som van de proporties verklaarde variantie is gelijk aan $\sum_{i=1}^{6}(a_{i1}^2 + a_{i2}^2)$.

Bij een hoofdcomponentenanalyse is het altijd zo dat de eerste hoofdcomponent de meeste variantie verklaart. De tweede hoofdcomponent, die ongecorreleerd is met de eerste, verklaart op zijn beurt het maximum van de nog resterende variantie (i.e., de variantie die niet door de eerste hoofdcomponent verklaard werd). En de derde hoofdcomponent, die ongecorreleerd is met de eerste twee, verklaart op zijn beurt het maximum van de daarna nog resterende variantie (i.e., de variantie die niet door de eerste twee hoofdcomponenten verklaard werd). Voor alle volgende hoofdcomponenten geldt analoog dat zij telkens het maximum van de nog resterende variantie verklaren. Deze eigenschap van de hoofdcomponentenanalyse is van weinig belang. Dit is zo omdat alleen de totale hoeveelheid verklaarde variantie belangrijk is, en niet de wijze waarop deze verklaarde variantie verdeeld is over de factoren. In de paragraaf over de *onbepaaldheid* van een factoroplossing zullen we zien dat er meerdere factoroplossingen bestaan die allemaal de maximale hoeveelheid variantie verklaren, maar die van elkaar verschillen in de wijze waarop deze verklaarde variantie verdeeld is over de factoren. In sommige factoroplossingen verklaren beide factoren elk ongeveer evenveel variantie, en in andere factoroplossingen verklaart een van de twee gemeenschappelijke factoren veel meer variantie dan de andere.

In Tabel 5.11 staan de factorladingen die geproduceerd werden door een hoofdcomponentenanalyse op de correlaties uit ons voorbeeld. In deze tabel staan ook de communaliteiten en de eigenwaarden van deze oplossing. De ladingen van de eerste factor zijn dezelfde als de ladingen van de één-factoroplossing. Dit komt omdat de eerste factor (hoofdcomponent) van een hoofdcomponentenanalyse zoveel mogelijk variantie verklaart, de tweede zoveel mogelijk van de nog resterende variantie, etcetera.

### 5.4.3 Een Vergelijking met het Factormodel met Ongecorreleerde Residuen

De twee-factoroplossing voor het model met ongecorreleerde residuen, verkregen op basis van het ML-criterium (die ook de ULS-oplossing is), en de bijbehorende communaliteiten en eigenwaarden, worden gegeven in Tabel 5.6. We vergelijken nu Tabel 5.6 met Tabel 5.11, de corresponderende tabel voor het model met gecorreleerde residuen, verkregen op basis van een hoofdcomponentenanalyse. Een eerste vaststelling is dat het percent-

| | $F_1$ | $F_2$ | $h_i^2$ |
|---|---|---|---|
| $X_1$ | 0.81 | -0.31 | 0.75 |
| $X_2$ | 0.78 | -0.34 | 0.72 |
| $X_3$ | 0.69 | 0.35 | 0.60 |
| $X_4$ | 0.62 | 0.40 | 0.54 |
| $X_5$ | 0.55 | 0.46 | 0.51 |
| $X_6$ | 0.39 | -0.59 | 0.50 |
| EW | 2.58 | 1.05 | 3.63 |
| % | 43.0% | 17.5% | 60.5% |

Tabel 5.11: Factorladingen, communaliteiten en eigenwaarden van de hoofdcomponentenanalyse met twee hoofdcomponenten voor de correlatiematrix in Tabel 5.1.

age verklaarde variantie (de som van de eigenwaarden/communaliteiten gedeeld door het aantal geobserveerde variabelen) groter is voor de hoofdcomponentenoplossing (60.5 procent) dan voor de ML/ULS-oplossing (44 procent). Dit is altijd zo, omdat de hoofdcomponentenanalyse, bij een gelijk aantal factoren, per definitie de meeste variantie verklaart. De afzonderlijke eigenwaarden en communaliteiten zijn meestal (maar niet altijd) groter voor de hoofdcomponentenoplossing dan voor de ML/ULS-oplossing. Een tweede vaststelling is dat het patroon van de ladingen op de eerste factor ongeveer hetzelfde is voor de hoofdcomponenten- en de ML/ULS-oplossing. Concreet, de volgorde van de ladingen is hetzelfde in de twee oplossingen. Dit is niet het geval voor de tweede factor. Omdat de interpretatie van een factor gebaseerd is op het patroon van hoge en lage ladingen op die factor (zie de paragraaf *Roteren naar Interpreteerbare Factoren*), zal de eerste factor van beide oplossingen wellicht op dezelfde manier geïnterpreteerd worden. Voor de tweede factor is dit niet zo zeker.

Behalve de ladingen, communaliteiten, en eigenwaarden, kan men ook de residuele correlaties van de hoofdcomponenten- en de ML/ULS-oplossing met elkaar vergelijken. Voor ons getallenvoorbeeld weten we al op voorhand dat de hoofdcomponentenoplossing het hier slechter gaat doen; de ML/ULS oplossing was immers perfect (alle residuele correlaties waren gelijk aan 0). Voor andere datasets zullen we vrijwel altijd hetzelfde observeren: de residuele correlaties zijn kleiner voor de ML/ULS-dan voor de hoofdcomponentenoplossing. De reden hiervoor is dat het ML/ULS-criterium precies deze residuele correlaties minimaliseert.

De hoofdcomponenten- en de ML/ULS-oplossing zijn bijna altijd verschillend. Voor de praktijk is zo'n verschil natuurlijk alleen van belang als het ook substantieel is. Welnu, het verschil kán substantieel zijn als het percentage verklaarde variantie (voor alle geobserveerde variabelen samen) klein is. Andersom geformuleerd, naarmate het percentage verklaarde variantie groter is, lijken de hoofdcomponenten- en de ML/ULS-oplossing meer op elkaar. Dit komt omdat, naarmate de geobserveerde variabelen voor een groter gedeelte bepaald worden door de gemeenschappelijke factoren (i.e., naarmate hun communaliteiten groter zijn), ook de correlaties tussen deze geobserveerde variabelen voor een groter gedeelte bepaald worden door de gemeenschappelijke factoren. Hieruit volgt dat, naarmate het percentage verklaarde variantie stijgt, de residuele correlaties

afnemen. Dus, naarmate de hoofdcomponentenanalyse meer variantie verklaart, wordt ook het ML/ULS-criterium (zo klein mogelijke residuele correlaties) beter benaderd.

We kunnen het bovenstaande ook op een meer precieze manier aantonen. We doen dit aan de hand van het model met één gemeenschappelijke factor. Op pagina 486 hebben we laten zien dat de correlatie tussen $X_1$ en $X_2$ als volgt geschreven kan worden:

$$r(X_i, X_j) = a_i \cdot a_j + \mathrm{COV}(R_i, R_j) \quad .$$

De eerste term in het rechterlid van deze formule, $a_i \cdot a_j$, is de gereproduceerde correlatie. Deze gereproduceerde correlatie wordt op dezelfde manier berekend als in de factoranalyse met ongecorreleerde residuen: als het product van de geschatte factorladingen (met meer dan één gemeenschappelijke factor, als een *som* van producten van geschatte factorladingen).

Het meest interessante deel van de bovenstaande formule is $\mathrm{COV}(R_i, R_j)$, de covariantie tussen de residuen $R_i$ en $R_j$. Deze covariantie kan men eenvoudig berekenen als men over de residuen $R_i$ en $R_j$ beschikt. Een andere manier om $\mathrm{COV}(R_i, R_j)$ te berekenen, zonder gebruik te maken van de residuen $R_i$ en $R_j$, is als het verschil tussen $r(X_i, X_j)$, de geobserveerde correlatie, en $a_i \cdot a_j$, de gereproduceerde correlatie (zie de bovenstaande formule). Met andere woorden, $\mathrm{COV}(R_i, R_j)$ is de residuele correlatie. Men mag zich hier niet in verwarring laten brengen door de term residuele *correlatie*, want het gaat hier over een covariantie en niet over een correlatie. De term residuele correlatie is ongelukkig maar het is wel de standaardterm voor het verschil tussen een geobserveerde en een gereproduceerde correlatie.

We laten nu zien dat $\mathrm{COV}(R_i, R_j)$ kleiner wordt (in absolute waarde) naarmate het percentage verklaarde variantie toeneemt. We beginnen met het herschrijven van $\mathrm{COV}(R_i, R_j)$ als een functie van de correlatie $r(R_i, R_j)$:

$$\begin{aligned} \mathrm{COV}(R_i, R_j) &= \mathrm{SD}(R_i) \times \mathrm{SD}(R_j) \times r(R_i, R_j) \\ &= \sqrt{\mathrm{VAR}(R_i)} \times \sqrt{\mathrm{VAR}(R_j)} \times r(R_i, R_j) \quad . \end{aligned}$$

In het één-factormodel zijn de onverklaarde varianties $\mathrm{VAR}(R_i)$ en $\mathrm{VAR}(R_j)$ gelijk aan, respectievelijk, $1 - a_i^2$ en $1 - a_j^2$ (1 min de communaliteit). Een toename van het percentage verklaarde variantie is natuurlijk equivalent met een afname van het percentage ónverklaarde variantie. Bijgevolg, als de onverklaarde varianties $\mathrm{VAR}(R_i)$ en $\mathrm{VAR}(R_j)$ in de bovenstaande formule naar 0 gaan, dan gaat ook $\mathrm{COV}(R_i, R_j)$ naar 0, eenvoudig omdat de correlatie $r(R_i, R_j)$ begrensd is door $-1$ en $+1$. Dus, naarmate de hoofdcomponentenoplossing meer variantie verklaart, zijn ook de residuele correlaties $\mathrm{COV}(R_i, R_j)$ kleiner. En omdat in de ML/ULS-oplossing de residuele correlaties geminimaliseerd worden, volgt hieruit dat, naarmate de hoofdcomponentenoplossing meer variantie verklaart, deze oplossing meer op de ML/ULS-oplossing gaat lijken.

De belangrijkste vraag voor de praktijk is wanneer men de ene of de andere variant van de factoranalyse (factoranalyse met ongecorreleerde residuelen of hoofdcomponentenanalyse) moet gebruiken. Jammer genoeg, bestaat er geen bruikbare richtlijn voor deze keuze. Dit komt omdat het in de praktijk vrijwel nooit duidelijk is of het nu belangrijker is om de correlaties zo goed mogelijk te verklaren of om zoveel mogelijk variantie te verklaren. De reden waarom we in dit boek beide varianten van de factoranalyse

behandelen, is dat ze beide gebruikt worden in de praktijk en dat hun oplossingen soms verschillen.

Voor wat er in de volgende twee paragrafen behandeld zal worden (de onbepaaldheid van de factoroplossing en het roteren naar interpreteerbare factoren), is het niet van belang om te weten op basis van welk criterium men de factorladingen bepaald heeft; het is voldoende dat er een set van factorladingen ligt die aan een of ander optimaliteitscriterium voldoet (ML, ULS, maximale verklaarde variantie, ...).

---

*Opgave 9*

Geef, van de volgende stellingen, aan of ze waar of onwaar zijn. Als je vindt dat een stelling onwaar is, leg dan ook uit waarom dat zo is.

1. Een hoofdcomponentenanalyse verklaart altijd minstens evenveel variantie als een factoranalyse met ongecorreleerde residuen, gegeven dat het aantal factoren gelijk is.
2. Naarmate een hoofdcomponentenoplossing meer variantie verklaart, zijn ook de residuele correlaties kleiner.
3. Een ULS-factoroplossing die kleine residuele correlaties levert, verklaart ook veel variantie.
4. In een hoofdcomponentenanalyse kunnen de residuen positief gecorreleerd zijn als de residuele correlaties negatief zijn.
5. Als bij een hoofdcomponentenanalyse alle residuele correlaties gelijk zijn aan nul, dan geldt hetzelfde voor een factoroplossing met ongecorreleerde residuen (met hetzelfde aantal factoren).
6. In een factoranalyse die gebruik maakt van het ULS criterium zijn de residuen alleen ongecorreleerd als de residuele correlaties gelijk zijn aan nul.

---

## 5.5 De Onbepaaldheid van een Factoroplossing

### 5.5.1 Illustratie en Omschrijving

Het zou mooi zijn als we hiermee de beschrijving van de factoranalyse konden afsluiten. Er is echter een belangrijke complicatie die we tot nog toe verzwegen hebben, maar nu onder ogen moeten zien. De kwestie kan geïllustreerd worden aan de hand van de factorladingenmatrix in Tabel 5.12. De factorladingen in deze tabel voldoen aan het ULS-criterium, en horen dus bij het factormodel met ongecorreleerde residuen. Het punt dat we hier zullen maken, staat echter los van het criterium op basis waarvan men de factorladingen bepaald heeft. De factorladingen in Tabel 5.12 zijn de ladingen op twee andere orthogonale gemeenschappelijke factoren (andere factoren dan $F_1$ en $F_2$ in Tabel 5.6 op pagina 476). Deze nieuwe factoren duiden we aan met $F_1^*$ en $F_2^*$ (dus, met een $^*$-tje). De ladingen van dezelfde variabelen $X_1$ tot en met $X_6$ op deze nieuwe factoren zijn duidelijk verschillend van die op $F_1$ en $F_2$ (zie Tabel 5.6). Dus, $F_1^*$ en $F_2^*$ zijn duidelijk verschillend van $F_1$ en $F_2$. Niettemin, als we de door $F_1^*$ en $F_2^*$ gereproduceerde correlaties bekijken, vinden we bijvoorbeeld

$$\widehat{r}(X_1, X_2) = 0.90 \cdot 0.80 + 0.00 \cdot 0.00 = 0.72 \quad ,$$

en ook

$$\widehat{r}(X_3, X_5) = 0.42 \cdot 0.30 + 0.56 \cdot 0.40 = 0.35 \quad .$$

In beide gevallen is de gereproduceerde waarde gelijk aan de geobserveerde. Dit blijkt te gelden voor alle door $F_1^*$ en $F_2^*$ gereproduceerde correlaties. Met andere woorden, deze factoren $F_1^*$ en $F_2^*$ geven ook een perfecte oplossing van $X_1$ tot en met $X_6$ in twee factoren. Blijkbaar is de eerst gevonden oplossing in termen van $F_1$ en $F_2$ niet uniek.

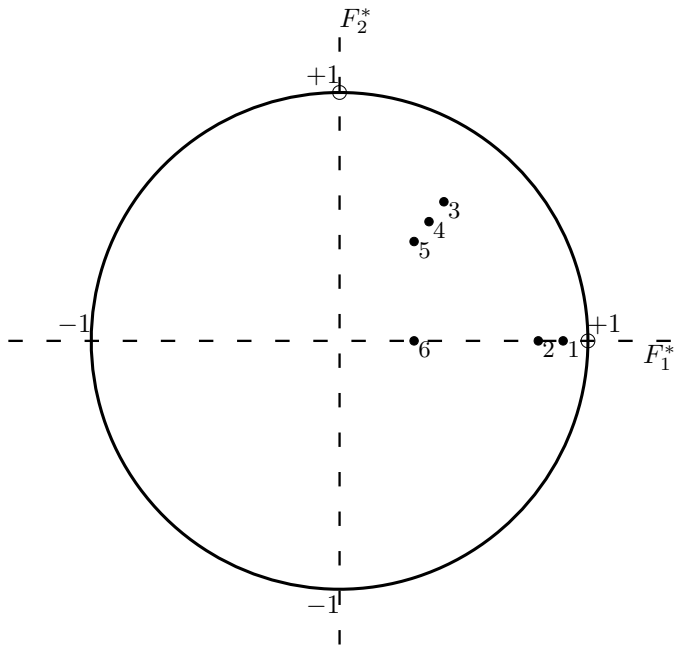
| | $F_1^*$ | $F_2^*$ | $h_i^2$ |
|---|---|---|---|
| $X_1$ | 0.90 | 0.00 | 0.81 |
| $X_2$ | 0.80 | 0.00 | 0.64 |
| $X_3$ | 0.42 | 0.56 | 0.49 |
| $X_4$ | 0.36 | 0.48 | 0.36 |
| $X_5$ | 0.30 | 0.40 | 0.25 |
| $X_6$ | 0.30 | 0.00 | 0.09 |
| EW | 1.94 | 0.70 | 2.64 |
| % | 32.3% | 11.7% | 44% |

Tabel 5.12: Factorladingenmatrix voor een alternatieve twee-factoroplossing voor de correlatiematrix in Tabel 5.1.

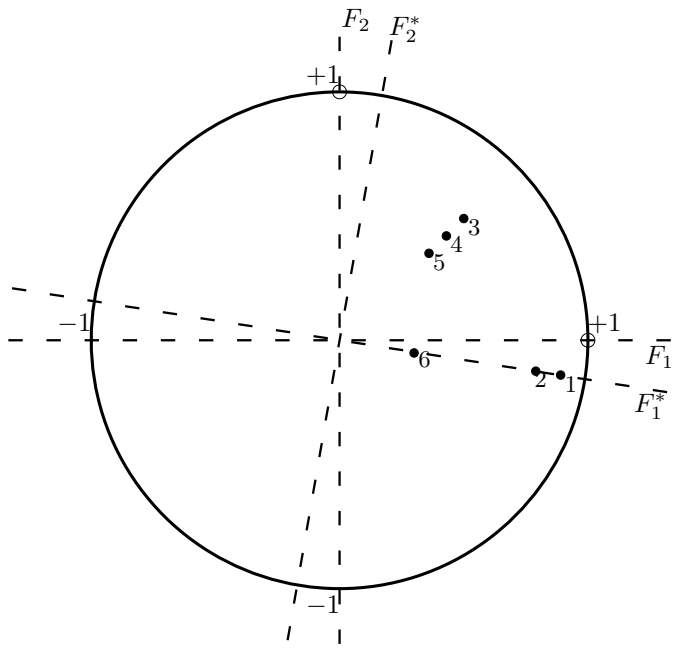
Zoals de tweede oplossing nu gepresenteerd is, zonder uit te leggen waar ze vandaan komt, lijkt het alsof de twee oplossingen niets met elkaar te maken hebben. Zo wordt misschien de suggestie gewekt dat de hele zaak triviaal is, dat wellicht willekeurig welke oplossing met twee gemeenschappelijke factoren voldoet. Niets is echter minder waar, er blijkt wel degelijk een nauw verband te bestaan tussen de factoren $F_1$ en $F_2$ enerzijds en $F_1^*$ en $F_2^*$ anderzijds. Dit is het duidelijkst meetkundig te zien, als we de factorladingenplot voor de alternatieve oplossing bekijken. Deze factorladingenplot wordt getoond in Figuur 5.9.

We vergelijken de plot voor de eerste oplossing, getoond in Figuur 5.3 op pagina 478, met de plot voor de alternatieve oplossing, getoond in Figuur 5.9. Het valt onmiddellijk op dat ze erg op elkaar lijken. De relatieve posities van de 6 variabelen zijn bijvoorbeeld dezelfde. Als je de twee figuren op elkaar legt en je draait vervolgens de tweede plot over een kleine hoek rechtsom, dan blijken de variabelen in de plots zelfs samen te vallen. In Figuur 5.10 wordt deze operatie uitgevoerd.

De tweede oplossing is dus meetkundig uit de eerste verkregen door een draaiing, een *rotatie*, uit te voeren. Hieraan zien we dat de factoren $F_1^*$ en $F_2^*$ inderdaad ook orthogonaal zijn: na een rotatie van de orthogonale $F_1$ en $F_2$ over eenzelfde hoek staan de nieuwe assen nog steeds loodrecht op elkaar. Ook verklaart dit waarom de communaliteiten in de twee oplossingen gelijk zijn (vergelijk de toegevoegde kolom van de twee factorladingenmatrices in Tabel 5.6 en Tabel 5.12). Immers, zoals we boven uiteengezet hebben, is deze communaliteit een eenvoudige functie van de afstand van de variabele tot de oorsprong. Een rotatie verandert deze afstand niet. Een direct gevolg van dit gelijkblijven van de communaliteiten is dat de twee geroteerde factoren $F_1^*$ en $F_2^*$ samen precies evenveel variantie verklaren als $F_1$ en $F_2$. Dit ziet men als men de nieuwe factorladingenmatrix in Tabel 5.12 vergelijkt met de oude factorladingenmatrix in Tabel 5.6.




Figuur 5.9: Factorladingenplot van de twee-factoroplossing in Tabel 5.12.




Figuur 5.10: Rotatie van de factoren.

De eigenwaarden van de afzonderlijke factoren zijn wél veranderd na rotatie (zie weer de factorladingenmatrices), hetgeen weerspiegelt dat de rolverdeling van de twee geroteerde factoren wat anders is geworden. We zien bijvoorbeeld dat, waar de variabelen $X_1$, $X_2$ en $X_6$ licht negatief, en de variabelen $X_3$, $X_4$ en $X_5$ eerst gematigd positief laadden op $F_2$, de eerste set variabelen nu in het geheel niet meer laadt op $F_2^*$, terwijl de ladingen van de andere set op $F_2^*$ juist groter zijn geworden. De ladingen van de eerste set, $X_1$, $X_2$ en $X_6$, op $F_1^*$ nemen zeer licht toe of blijven gelijk ten opzichte van $F_1$, terwijl de ladingen van de tweede set, $X_3$, $X_4$ en $X_5$, afnemen.

Onder rotatie veranderen de gereproduceerde correlaties niet. Dit mag niet als een verrassing klinken. Eerder hebben we immers de volgende formule voor de gereproduceerde correlatie voorgesteld:

$$\boxed{\widehat{r}(X_i, X_j) = h_i h_j \cos(\phi) \quad .}$$

Volgens deze formule is de gereproduceerde correlatie een functie van (1) de lengte van de vectoren die de variabelen voorstellen en (2) de hoek tussen deze vectoren. Welnu, deze vectoren (i.c., hun lengte en hun positie ten opzichte van elkaar) veranderen niet onder rotatie, en daarom veranderen ook de gereproduceerde correlaties niet.

Wat hier aan de hand van één specifiek voorbeeld, één bepaalde rotatie, is geïllustreerd blijkt algemeen te gelden: wanneer we een bepaalde factoroplossing hebben gevonden, levert een oplossing verkregen door rotatie van de gevonden factoren dezelfde gereproduceerde correlaties en verklaart zo'n geroteerde oplossing evenveel variantie. Met één oplossing in gemeenschappelijke factoren hebben we dus meteen oneindig veel verschillende oplossingen; we kunnen immers over oneindig veel hoeken draaien.

### 5.5.2 Het Belang van de Onbepaaldheid

Afhankelijk van de hoek waarover geroteerd wordt, kunnen de ladingen van de verschillende equivalente factoroplossingen sterk van elkaar verschillen. Dit heeft belangrijke implicaties voor de interpretatie van de factoren. In de paragraaf *Roteren naar Interpreteerbare Factoren* gaan we meer uitgebreid in op het interpreteren van factoren. Op dit ogenblik is het voldoende om te weten dat interpretatie van factoren neerkomt op naamgeving. Deze naamgeving gebeurt op basis van de factorladingen. De basis voor de naamgeving van een bepaalde factor (bijv., "rekenvaardigheid", "extraversie", "gewetensvolheid") zijn de variabelen die substantieel op deze factor laden. Een speciale plaats wordt hier ingenomen door de variabelen die slechts op één van de factoren substantieel laden, de zogenaamde *factorieel zuivere variabelen*. Factorieel zuivere variabelen kunnen gezien worden als een zuivere representant van deze factor. De naam die men kiest voor een bepaalde factor moet overeenstemmen met de inhoud van de deze factorieel zuivere variabelen; "rekenvaardigheid" als het rekenitems zijn, "extraversie" als de items peilen naar initiatief-nemend sociaal gedrag, en "gewetensvolheid" als de items peilen naar gewetensvol gedrag. Nu, omdat er meerdere equivalente sets van factorladingen bestaan, waarin telkens weer andere variabelen als factorieel zuiver naar voren komen, stelt zich hier de vraag op welke set men de interpretatie moet baseren.

De onbepaaldheid van het factormodel heeft de bijdrage van dit model aan de psychologie sterk beperkt. In de jaren 1930 tot 1960 had men grote verwachtingen van dit

model en de bijbehorende methoden om de factorladingen te bepalen. Dit was de periode waarin een groot aantal psychologische tests gepubliceerd werden. Hierdoor ontstond er een keuzeprobleem: men wist niet welke tests men moest gebruiken om iemands kenmerken binnen een bepaald domein (dat van de vaardigheden, de persoonlijkheid, de interesses, ... ) te beschrijven. Het idee ontstond dat er slechts een beperkt aantal *basisdimensies* bestaan waarop mensen van elkaar verschillen. In termen van het factormodel, zijn deze basisdimensies de factoren. Eysenck heeft het standpunt verdedigd dat er in het domein van de persoonlijkheid slechts drie basisdimensies zijn: neuroticisme, extraversie en psychoticisme. De scores op het grote aantal beschikbare tests worden bepaald door de posities op dit kleine aantal basisdimensies. De tests verschillen van elkaar door de mate waarin ze bepaald worden door de verschillende basisdimensies. Zo kunnen er persoonlijkheidstests zijn die zowel bepaald worden door neuroticisme als door extraversie, maar helemaal niet door psychoticisme. Andere persoonlijkheidstests kunnen vooral bepaald worden door neuroticisme en psychoticisme, en slechts in beperkte mate door extraversie. De mate waarin een testscore bepaald wordt door de verschillende basisdimensies wordt uitgedrukt door de factorladingen: een test met een sterk van nul afwijkende lading op een bepaalde basisdimensie wordt in sterke mate bepaald door die basisdimensie.

Het idee is toen ontstaan om tests te construeren die slechts door één van de basisdimensies bepaald worden. In termen van de factorladingen betekent dit dat men tests wilde construeren die slechts op één van de factoren een lading heeft die substantieel van 0 afwijkt. De score op een dergelijke factorieel zuivere test kan men beschouwen als een vervanger voor de factor. (Een factorieel zuivere test is natuurlijk geen perfecte vervanger voor de factor, want de testscore wordt ook door het residu bepaald.) Beschikt men eenmaal over factorieel zuivere tests, dan kan men iemands kenmerken binnen een bepaald domein volledig beschrijven met behulp van één test per factor.

Binnen het intelligentiedomein hebben Thurstone en Guilford een groot aantal factorieel zuivere tests ontwikkeld. De tests van Thurstone, de *Primary Mental Abilities* (PMA), hebben aan de basis gestaan van de in Nederland veelgebruikte *Groninger Intelligentie Test* (GIT). Binnen het domein van de persoonlijkheid hebben vooral Eysenck en Cattell een belangrijke bijdrage geleverd aan de ontwikkeling van factorieel zuivere tests. De tests van Eysenck hebben aan de basis gestaan van de *Amsterdamse Biografische Vragenlijst* (ABV).

De ontwikkeling van factorieel zuivere tests zou een stuk eenvoudiger zijn indien men, voor een bepaalde dataset, de bijbehorende factorladingen op unieke wijze kon achterhalen. In dit geval, zou men kunnen spreken van dé factoren, en zou het niet gaan om één van de vele equivalente factoroplossingen die elk tot een andere factorinterpretatie aanleiding geven. Hetzelfde geldt voor de factorieel zuivere tests: als men de factorladingen op unieke wijze kon achterhalen, zou men kunnen spreken van dé factorieel zuivere tests. Omdat er echter meerdere equivalente factoroplossingen zijn, tot elkaar te herleiden door middel van rotatie, zullen sommigen een bepaalde test als factorieel zuiver aanduiden en anderen niet. En omdat het gaat om equivalente oplossingen, is de keuze eerder een subjectieve aangelegenheid.

## 5.6 Roteren naar Interpreteerbare Factoren

Zoals boven aangegeven, is de onbepaaldheid van een factoroplossing een groot probleem voor de factoranalyse. Deze procedure kan blijkbaar om principiële redenen niet de onderliggende factoren opsporen. Wat men echter gedaan heeft is van de nood een deugd maken. Men is de bestaande vrijheid tot roteren gaan benutten om onder de oplossingen die equivalent zijn wat betreft gereproduceerde correlaties dat stelsel van gemeenschappelijke factoren te zoeken dat het best interpreteerbaar is.

### 5.6.1 Enkele Richtlijnen bij de Interpretatie

Een interpretatie van een factor moet gebaseerd zijn op het patroon van ladingen van de variabelen op die factor. Dit is immers de enige informatie die we over de factor hebben. Met name zijn we geïnteresseerd in de variabelen die hoog (dit wil zeggen, ofwel zeer positief, dicht bij $+1$, ofwel zeer negatief, dicht bij $-1$) op de factor laden. Bij orthogonale factoren zijn de ladingen op een factor gelijk aan de correlatiecoëfficiënten tussen variabelen en factor. Dit geeft een duidelijke indicatie dat een factor iets meet wat gemeenschappelijk is aan de variabelen die er hoog op laden. Een speciale plaats wordt ingenomen door de factorieel zuivere variabelen, de variabelen die slechts op één van de factoren hoog laden. Factorieel zuivere variabelen kunnen gezien worden als een zuivere representant van deze factor.

Als een inhoudelijke inspectie van de variabelen ons leert dat ze een bepaalde psychologische grootheid gemeenschappelijk hebben die niet of in veel mindere mate aanwezig is in de overige variabelen, dan kunnen we verdedigen dat de factor in kwestie deze onderliggende, meer abstracte psychologische grootheid meet. Variabelen met een hoge negatieve lading moeten dan natuurlijk ook inhoudelijk in de tegenovergestelde richting wijzen.

Het voorgaande geeft aan dat de inhoudelijke interpretatie van een factor een subjectieve aangelegenheid is zonder correcte en incorrecte antwoorden. Een inhoudelijke analyse van de variabelen moet het argument leveren voor het verkiezen van de ene oplossing boven een andere. Ter illustratie nemen we even aan dat de variabelen $X_1$ tot en met $X_6$ uit het voorbeeld testscores zijn op 6 verschillende types redactiesommen. Het is een nuttige werkhypothese om ervan uit te gaan dat redactiesommen in meer of mindere mate een beroep doen op rekenvaardigheid en taalvaardigheid. Dan ligt het voor de hand om bij de onderliggende gemeenschappelijke factoren te zoeken naar zoiets als rekenvaardigheid en taalvaardigheid. Met deze mogelijke duiding van de factoren in het achterhoofd is de eerste door ons gevonden oplossing (met de factoren $F_1$ en $F_2$) sowieso al weinig aantrekkelijk. Immers, de testscores $X_1$, $X_2$ en $X_6$ laden licht *negatief* op $F_2$. Welke vaardigheid we ook met $F_2$ willen associëren (taal- of rekenvaardigheid), dit impliceert dat hoge scores op deze drie tests over het algemeen samengaan met een *gebrek* aan die vaardigheid. Dit is niet aannemelijk. Wat dit betreft, levert de geroteerde oplossing in $F_1^*$ en $F_2^*$ betere vooruitzichten, mits we na een inspectie van de tests $X_1$, $X_2$ en $X_6$ hard kunnen maken dat deze feitelijk slechts een beroep doen op één bepaalde vaardigheid (bijvoorbeeld rekenvaardigheid), terwijl de overige drie tests, $X_3$, $X_4$ en $X_5$ duidelijk beide vaardigheden vereisen. In zo'n situatie is deze geroteerde oplossing heel goed te interpreteren door $F_1^*$ met rekenvaardigheid en $F_2^*$ met taalvaardigheid te

identificeren. Mocht bij de inhoudelijke analyse van de tests echter blijken dat juist de tests $X_3$, $X_4$ en $X_5$ een beroep doen op slechts één vaardigheid (zeg, weer rekenvaardigheid), en $X_1$, $X_2$ en $X_6$ op beide, dan is de oplossing met de factoren $F_1^*$ en $F_2^*$ niet goed interpreteerbaar, maar dan dient zich wel een alternatief aan. In de laatste plot hebben we gezien dat de factoren $F_1^*$ en $F_2^*$ verkregen zijn door, uitgaande van de oplossing in $F_1$ en $F_2$, de assen rechtsom te draaien totdat $X_1$, $X_2$ en $X_6$ op de geroteerde horizontale as terechtkomen. Maar we kunnen natuurlijk nog verder draaien, namelijk totdat $X_3$, $X_4$ en $X_5$ op de geroteerde verticale as vallen. Deze oplossing zou juist te verkiezen zijn in het laatst beschreven geval, waarbij nu de verder geroteerde verticale factor geïnterpreteerd kan worden als rekenvaardigheid en de verder geroteerde horizontale factor als taalvaardigheid.

Het moet benadrukt worden dat onze voorbeelddata kunstmatig zijn. In de praktijk zullen de variabelen nooit zo mooi in de factorladingenplot liggen als hier, in twee groepjes van drie, op rechte lijnen door de oorsprong. Kort samengevat zijn dit de belangrijkste richtlijnen bij de interpratie van de factoroplossing:

1. Het vinden van een goed interpreteerbare oplossing is vaak meer een kunst dan een kunde. Er zijn hier geen objectieve, formele criteria en dus ook geen objectief juiste of onjuiste antwoorden. Succes is ook niet gegarandeerd, soms is er gewoon geen bevredigende interpretatie te vinden. Een inhoudelijke analyse van de geobserveerde variabelen speelt een essentiële rol, maar die levert niet altijd het gewenste aanknopingspunt voor de interpretatie.

2. Een factoroplossing is het beste te interpreteren als er voor elke factor een aantal factorieel zuivere variabelen zijn. De factor valt dan inhoudelijk samen met deze factorieel zuivere variabelen.

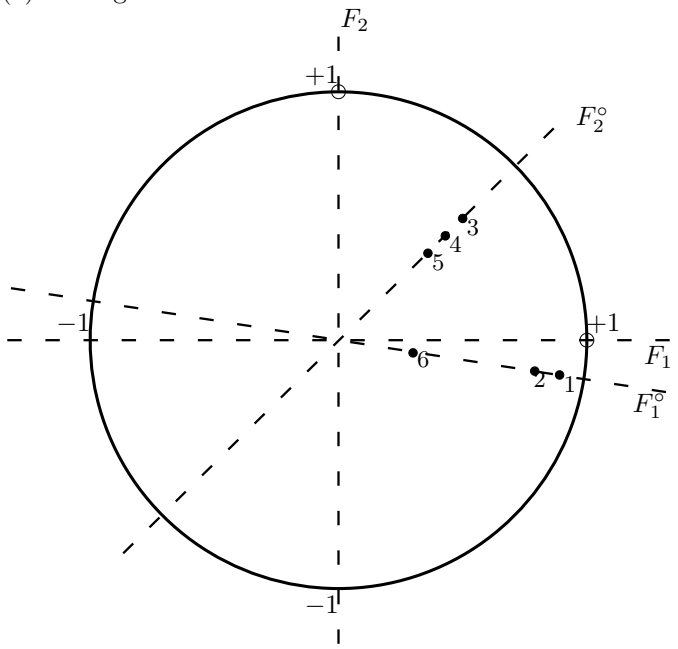
3. Als er weinig of geen factorieel zuivere variabelen zijn, dan moet men zich baseren op die variabelen die hoog op de factor laden (maar die ook op andere factoren kunnen laden). De factor is goed te interpreteren als er een duidelijke scheiding is tussen variabelen die hoog (positief of negatief) op deze factor laden en die welke er niet of nauwelijks op laden. Zo'n factor is dan redelijk goed bepaald: hij moet een psychologische grootheid representeren die gedeeld wordt door de hoog ladende variabelen, maar waarop geen beroep gedaan wordt door de niet-ladende variabelen.

4. Een factor met allemaal lage of middelmatige ladingen (zeg, niet hoger dan $\pm 0.30$ of $\pm 0.40$) is bijna nooit goed te interpreteren.

Bij test- en vragenlijstconstructie is men bijzonder geïnteresseerd in factorieel zuivere items. Bij deze toepassing van de factoranalyse vertrekt men van een inhoudelijk heterogene itemset die men aan een groep personen voorlegt. Op deze data wordt dan een factoranalyse uitgevoerd. Als uit de factoroplossing blijkt dat sommige items factorieel zuiver zijn, dan voegt men deze items samen in een zogenaamde *schaal*. Items die tot dezelfde schaal behoren meten allemaal dezelfde gemeenschappelijke factor.

### 5.6.2 Oblieke Rotatie

In het voorgaande hebben we ons beperkt tot echte rotaties, waarbij een orthogonaal stelsel van factoren overgaat in een nieuw orthogonaal stelsel. Dat het gaat om een orthogonaal stelsel van factoren kan men zien aan het feit dat de twee assen een rechte hoek maken. Niets belet ons echter om de assen ook schuin ten op zichte van elkaar te kiezen. De rotatie naar een schuin assenstelsel wordt een *oblieke rotatie* genoemd alhoewel het, strikt genomen, geen rotatie is. In Figuur 5.11 wordt zo'n oblieke rotatie getoond. Net zoals een orthogonale rotatie, heeft ook een oblieke rotatie geen invloed op de gereproduceerde correlaties. Dit volgt onmiddellijk uit het feit dat de gereproduceerde correlaties alleen afhangen van (1) de hoek tussen de vectoren die de variabelen voorstellen en (2) de lengtes van deze vectoren.




Figuur 5.11: Oblieke rotatie van de factoren zodanig dat elke variabele op slechts één gemeenschappelijke factor laadt.

#### De Interpretatie van een Oblieke Factoroplossing

Een oblieke rotatie biedt ons meer vrijheid, en die vrijheid kunnen we gebruiken om een factoroplossing te zoeken die beter interpreteerbaar is dan het resultaat van een orthogonale rotatie. Door obliek te roteren is het gemakkelijker om factorieel zuivere variabelen te vinden, variabelen die op slechts één van de factoren laden en er daarom inhoudelijk mee samenvallen (het residu buiten beschouwing gelaten). Dit is precies wat er gebeurt in Figuur 5.11. Stel dat een inhoudelijke analyse van de tests $X_1$ t/m $X_6$ uitwijst dat $X_1$, $X_2$ en $X_6$ eigenlijk puur rekenkundige items bevatten, terwijl $X_3$, $X_4$ en $X_5$ juist uit puur taalkundige items bestaan. Zoals blijkt uit Figuur 5.11, is dit met

oblieke factoren uitstekend te representeren: de rekenkundige factor zetten we af langs de lijn door de punten 1, 2 en 6, en de taalkundige factor langs de lijn door de punten 3, 4 en 5.

Een consequentie van een oblieke rotatie is dat de factoren nu gecorreleerd zijn in plaats van ongecorreleerd. De correlatie tussen de factoren is gelijk aan de cosinus van de hoek tussen de assen die de factoren voorstellen. In ons voorbeeld (zie Figuur 5.11), is de hoek tussen de assen 53.13 graden, en de cosinus hiervan is 0.60. Dus, als we de factoren interpreteren als rekenvaardigheid en taalvaardigheid, dan impliceert onze oblieke factoroplossing dat er een correlatie van 0.60 bestaat tussen deze vaardigheden. Dit lijkt een zeer realistische aanname.

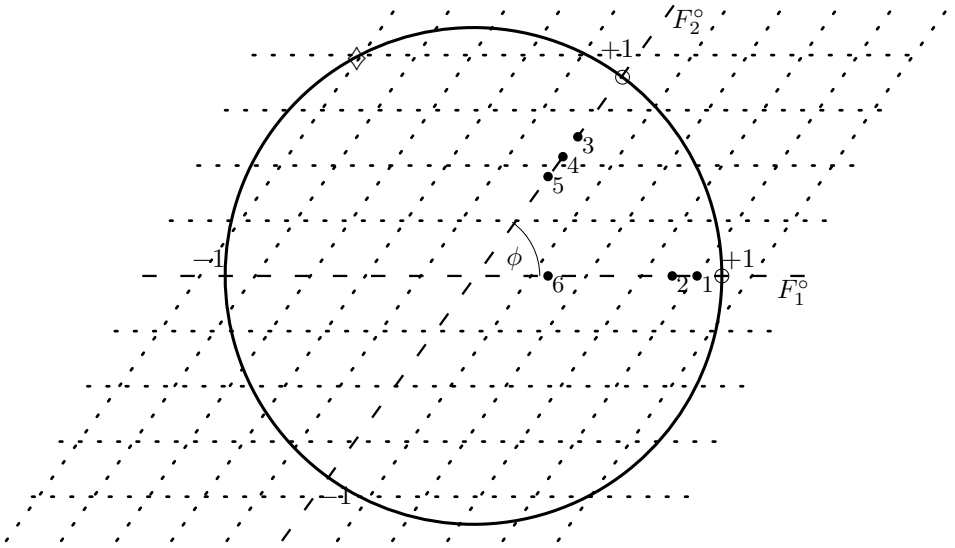
De interpretatie van een oblieke factoroplossing dient te gebeuren op basis van de factorladingen én de correlaties tussen de factoren. Als vertrekpunt neemt men de variabelen die op slechts één van de factoren substantieel laden. Het residu buiten beschouwing gelaten, vallen deze factorieel zuivere variabelen inhoudelijk gezien samen met de factoren. De interpretatie op basis van deze factorieel zuivere variabelen moet in overeenstemming zijn met de correlaties tussen de factoren. Bijvoorbeeld, het kan niet dat twee negatief correlerende factoren geïnterpreteerd worden als technische leesvaardigheid en begrijpende leesvaardigheid.

### Oblieke Factoranalyse als Regressieanalyse met Gecorreleerde Latente Predictoren

Het is goed om hier weer gebruik te maken van de analogie tussen factoranalyse en regressieanalyse: factoranalyse is een set van regressieanalyses, één per geobserveerde variabele, waarin de predictoren niet geobserveerd (dus, latent) zijn. Een orthogonale factoroplossing is vergelijkbaar met regressieanalyse met ongecorreleerde predictoren. De factorladingen zijn de bètagewichten van de latente predictoren (factoren). Het feit dat deze latente predictoren ongecorreleerd zijn heeft drie belangrijke gevolgen. Ten eerste, de factorladingen zijn correlaties tussen de geobserveerde variabelen en de latente predictoren. Ten tweede, de proportie verklaarde variantie van een geobserveerde variabele is te berekenen als de som van de gekwadrateerde factorladingen van deze variabele. En ten derde, de totale hoeveelheid verklaarde variantie kan opgedeeld worden in de afzonderlijke bijdragen van de twee gemeenschappelijke factoren, zoals gemeten door de som van de gekwadrateerde factorladingen per factor, de eigenwaarde. Deze eigenschap is volledig analoog aan het feit dat, in een regressieanalyse met twee ongecorreleerde predictoren, de proportie verklaarde variantie gelijk is aan de som van de afzonderlijke bijdragen van de twee predictoren (zoals gemeten door het kwadraat van de correlatie met het criterium).

Een oblieke factoroplossing is vergelijkbaar met een regressieanalyse met gecorreleerde predictoren. Net zoals in de regressieanalyse met gecorreleerde predictoren, zijn de bètagewichten, hier dus de ladingen van de variabelen op de factoren (latente predictoren), niet meer gelijk aan de correlaties tussen variabelen en factoren. Hun grootte is ook moeilijker te interpreteren, want ze hoeven nu zelfs niet meer tussen $-1$ en $+1$ te liggen. Dit kunnen we illustreren aan de hand van Figuur 5.12. In deze figuur wordt, behalve het oblieke assenstelsel, ook aangegeven hoe de ladingen op de factoren bepaald moeten worden. Dit gebeurt met de fijne stippellijntjes. Om voor een bepaalde vari-

abele (een punt binnen de cirkel) de factorladingen te bepalen, moet men deze variabele parallel aan de stippellijntjes projecteren op de assen. Het is eenvoudig te zien dat er punten zijn binnen de cirkel waarvan de projecties op de assen buiten het interval $(-1, +1)$ liggen.




Figuur 5.12: Grafische weergave van een oblieke factoroplossing met aanduiding van de richting waarin de coördinaten (factorladingen) bepaald worden.

Net zoals in een regressieanalyse met gecorreleerde predictoren, is het ook bij een oblieke factoroplossing niet mogelijk om de proporties verklaarde variantie van de geobserveerde variabelen uit te drukken in termen van de bètagewichten van de (hier, latente) predictoren (de factorladingen). Dit kunnen we illustreren aan de hand van Figuur 5.12. We willen de communaliteit berekenen van een variabele (zijn proportie verklaarde variantie) die zich bevindt op de plaats van het ruitje. De ladingen van deze variabele op de twee gemeenschappelijke factoren zijn beide groter dan 1 in absolute waarde. Daarom: als we de oude formule voor de communaliteit zouden gebruiken (de som van de gekwadrateerde factorladingen), dan zouden we een verklaarde variantie groter dan 2 vinden, hetgeen meer is dan het dubbele van de geobserveerde variantie. Dat het in het algemeen niet mogelijk is om in een oblieke factoroplossing de communaliteit te berekenen als de som van de gekwadrateerde factorladingen, betekent niet dat het nooít mogelijk is. Bijvoorbeeld, voor variabelen die op slechts één van de factoren laden (en dus precies op één van de assen vallen), zoals in Figuur 5.12, kan men de communaliteit wél berekenen als de som van de gekwadrateerde factorladingen (die in dit geval maar uit één term bestaat).

In plaats van te kijken naar de proportie verklaarde variantie per variabele (de com-

munaliteit), kunnen we ons ook afvragen hoe de totale hoeveelheid verklaarde variantie (de som van de communaliteiten) verdeeld wordt over de factoren. Nu, in de regressieanalyse met twee gecorreleerde predictoren kunnen we de proportie verklaarde variantie niet opsplitsen in afzonderlijke bijdragen van de twee predictoren. Analoog hieraan, geldt voor een oblieke factoroplossing (waarin de factoren gecorreleerd zijn) dat de totale hoeveelheid verklaarde variantie niet opgedeeld kan worden in de afzonderlijke bijdragen van de twee gemeenschappelijke factoren. Dit is de reden waarom men in een oblieke factoroplossing niet spreekt van de hoeveelheid verklaarde variantie door een factor (in een orthogonale factoroplossing, de eigenwaarde genoemd).

### De Correlaties tussen de Variabelen en de Factoren

Bij de interpretatie van een oblieke factoroplossing verliest men gemakkelijk uit het oog dat de factoren gecorreleerd zijn. Als men zich laat leiden door de matrix van factorladingen, dan krijgt men door het grote aantal bijna-nul ladingen snel de indruk dat de variabelen weinig met elkaar gemeenschappelijk hebben. Om deze indruk tegen te gaan, gebruikt men voor de interpretatie ook vaak de matrix van correlaties tussen variabelen en factoren, de zogenaamde *factorstructuurmatrix*. De factorstructuurmatrix kan men berekenen op basis van de factorladingenmatrix en de matrix van correlaties tussen de factoren. In Tabel 5.13 worden deze drie matrices getoond voor de gegevens uit ons voorbeeld. Uit de factorstructuurmatrix blijkt dat de meeste variabelen een substantiële correlatie hebben met beide factoren, ook al blijkt uit de factorladingenmatrix dat ze allemaal op slechts één van beide factoren laden. Deze correlaties zijn er natuurlijk omdat de factoren zelf gecorreleerd zijn.

| | $F_1^\circ$ | $F_2^\circ$ |
|---|---|---|
| $X_1$ | 0.90 | 0.00 |
| $X_2$ | 0.80 | 0.00 |
| $X_3$ | 0.00 | 0.70 |
| $X_4$ | 0.00 | 0.60 |
| $X_5$ | 0.00 | 0.50 |
| $X_6$ | 0.30 | 0.00 |

| | $F_1^\circ$ | $F_2^\circ$ |
|---|---|---|
| $F_1^\circ$ | 1.00 | 0.60 |
| $F_2^\circ$ | 0.60 | 1.00 |

| | $F_1^\circ$ | $F_2^\circ$ |
|---|---|---|
| $X_1$ | 0.90 | 0.54 |
| $X_2$ | 0.80 | 0.48 |
| $X_3$ | 0.42 | 0.70 |
| $X_4$ | 0.36 | 0.60 |
| $X_5$ | 0.30 | 0.50 |
| $X_6$ | 0.30 | 0.18 |

Tabel 5.13: Factorladingenmatrix (links), matrix van correlaties tussen de factoren (midden), en factorstructuurmatrix (rechts) voor de oblieke factoroplossing in Figuur 5.12.

Elk element van de factorstructuurmatrix wordt berekend door een rij getallen uit de factorladingenmatrix te vermenigvuldigen met een kolom getallen uit de matrix met correlaties tussen de factoren. Dit illustreren we aan de hand van de berekening van de correlatie tussen de derde variabele en de eerste factor in Tabel 5.13 (0.42). We verkrijgen deze correlatie door de derde rij van de factorladingenmatrix te vermenigvuldigen met de eerste kolom van de matrix met de correlaties tussen de factoren. Het vermenigvuldigen van een rij met een kolom gebeurt als volgt: neem het product van het eerste element van de rij en de kolom ($0.00 \cdot 1.00$), vermeerder dit met het product van het tweede element van de rij en de kolom ($0.70 \cdot 0.60$), enzovoort (in het geval dat de rij uit meer

dan twee elementen bestaat). De andere correlaties in de factorstructuurmatrix worden op dezelfde manier berekend. Bijvoorbeeld, de correlatie tussen de vijfde variabele en de tweede factor wordt als volgt berekend:

$$0.00 \cdot 0.60 \ + 0.50 \cdot 1.00 = 0.50 \quad .$$

In termen van de factorladingenplot (zie Figuur 5.12), hangt de correlatie tussen een variabele en een factor af van (1) de hoek tussen de vector die de variabele voorstelt en de lijn die de factor voorstelt (aangeduid met de Griekse letter $\phi$) en (2) de lengte van de vector die de variabele voorstelt. De lengte van deze vector is eenvoudig te berekenen als de vierkantswortel uit de communaliteit. Nu, de correlatie tussen de variabele $X_i$ en de factor $F^\circ$ kan als volgt berekend worden:

$$\boxed{\widehat{r}(X_i, F^\circ) = h_i \cos(\phi) \quad .}$$

Deze formule is identiek aan de formule voor een gereproduceerde correlatie tussen twee variabelen, $\widehat{r}(X_i, X_j) = h_i h_j \cos(\phi)$, voor het bijzondere geval waarin $X_j$ een factorieel zuivere meting is van de factor $F^\circ$ met een communaliteit gelijk aan 1.

---

*Opgave 10*
Een onderzoeker construeert een vragenlijst die probeert vast te leggen in welke mate een persoon op kan schieten met zijn collega's. Hij heeft hiertoe een aantal vragen geformuleerd die op een schaal van 1 tot 7 worden gescoord. Deze vragenlijst is afgenomen bij een grote steekproef en op de resulterende data is een hoofdcomponentenanalyse uitgevoerd. De vragen zijn:

1. Ik kan goed met mijn collega's opschieten. (OPSCHIET)
2. Ik heb dezelfde opvattingen over beleid als mijn collega's. (OPVATTIN)
3. Koffiedrinken en lunchen zijn gezellige activiteiten op mijn afdeling. (KOF&LUN)
4. Ik heb zelden een conflict met mijn collega's. (ZELDCONFL)
5. Op mijn afdeling weten we problemen op de juiste manier op te lossen. (PROBLOPL)
6. Mijn afdeling vormt een goed team omdat we elkaar persoonlijk goed liggen. (PERSLIG)
7. Er worden op mijn afdeling duidelijke afspraken gemaakt waar iedereen zich ook aan houdt. (AFSPRAAK)
8. Goede onderlinge verhoudingen zorgen ervoor dat we goed werk afleveren. (ONDVERH)
9. Ik heb veel plezier in mijn werk. (PLEZIER)

De factorladingen van de twee-factoroplossing wordt gegeven in Tabel 5.14.

1. Wat is de gereproduceerde correlatie tussen OPSCHIET en ZELDCONFL?
2. Welk item heeft de laagste communaliteit, en hoe kun je dat inhoudelijk verklaren?
3. Maak een factorladingenplot en interpreteer de factoroplossing. Als je denkt dat een rotatie de oplossing beter interpreteerbaar maakt, voer deze rotatie dan uit en geef de bijbehorende interpretatie.

*Opgave 11*
Deze opgave gaat over een onderzoek naar interesse in de sociaal-politieke actualiteit. Aan een groep personen werden 20 uitspraken voorgelegd, waaronder de volgende:

| | Factor 1 | Factor 2 |
|---|---|---|
| OPSCHIET | 0.04 | 0.81 |
| OPVATTIN | 0.58 | 0.68 |
| KOF&LUN | -0.65 | 0.60 |
| ZELDCONFL | 0.12 | 0.65 |
| PROBOPL | 0.61 | 0.56 |
| PERSLIG | -0.59 | 0.66 |
| AFSPRAAK | 0.55 | 0.65 |
| ONDVERH | -0.45 | 0.72 |
| PLEZIER | 0.14 | 0.25 |

Tabel 5.14: Tabel bij opgave 10.

- De uitslag van de verkiezingen in Israël interesseert me niet veel. (ISRAEL)
- Als er ergens in de wereld hongersnood is, dan zoek ik in de krant altijd naar berichten over hoe het met die mensen gesteld is. (HONGER)
- De sociaal-politieke hervormingen in het voormalige Oostblok zijn een fascinerend gebeuren. (OOSTBLOK)
- Ze moeten mij niet lastig vallen met de gevolgen van de AIDS-epidemie in Afrika. (AIDS)

Deze uitspraken werden beoordeeld op een schaal van 1 (volledig mee eens) tot 7 (volledig mee oneens). Vervolgens werd een factoranalyse uitgevoerd, waarbij twee gemeenschappelijke factoren werden getrokken. Deze twee factoren werden orthogonaal geroteerd. In Tabel 5.15 geven we de factorladingen van de vier bovenstaande vragen. Van de andere 16 vragen worden de factorladingen niet gegeven.

| | Factor 1 | Factor 2 |
|---|---|---|
| ISRAEL | -0.83 | 0.06 |
| HONGER | 0.04 | 0.90 |
| OOSTBLOK | 0.81 | -0.02 |
| AIDS | -0.05 | -0.87 |

Tabel 5.15: Tabel bij opgave 11.

1. Wat is de gereproduceerde correlatie tussen ISRAEL en OOSTBLOK? Laat zien hoe je dit berekent.
2. Wat is de communaliteit van HONGER? Laat zien hoe je dit berekent.
3. Stel dat je de communaliteiten van deze vier variabelen opnieuw zou berekenen, maar nu op basis van de ongeroteerde factorladingen. Wat zou je dan vinden? Leg dit uit, vertrekkende van de factorladingenplot.
4. Zijn de factorladingen van ISRAEL, HONGER, OOSTBLOK en AIDS informatief voor de interpretatie van de factoren? Motiveer je antwoord en, indien mogelijk, geef een interpretatie.

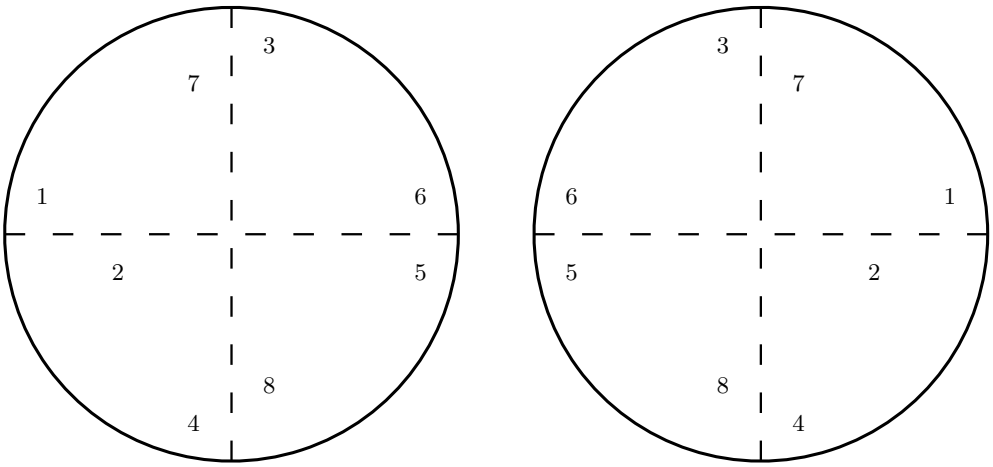
*Opgave 12*
Geef, voor elk van de volgende stellingen, aan of ze waar of onwaar zijn. Indien ze onwaar zijn, leg dan uit waarom:

1. Met het uitvoeren van een rotatie verandert de totale hoeveelheid door de factoren verklaarde variantie.
2. Met het uitvoeren van een rotatie verandert in de regel de hoeveelheid variantie die door elk van de factoren verklaard wordt.
3. Met het uitvoeren van een rotatie verandert in de regel de hoeveelheid variantie van een variabele die verklaard wordt door alle factoren samen.
4. Met het uitvoeren van een rotatie veranderen in de regel de residuele correlaties.
5. Met het uitvoeren van een rotatie veranderen in de regel de uniciteiten.

*Opgave 13*
Deze opgave is een vervolg van Opgave 4.

Is er een rotatie mogelijk zodat alle zes tests elk slechts één factor meten? Zo ja, hoe ziet die rotatie eruit?

*Opgave 14*
Een studente voert binnen het kader van haar scriptieonderzoek een factoranalyse uit. Zij wil de factorladingen spiegelen rond de verticale as (zie Figuur 5.13) omdat dit tot een eenvoudiger interpretatie leidt. Haar begeleider denkt echter dat dit niet tot een correcte factoroplossing leidt. Geef aan waarom dit wel of niet mag.




Gevonden oplossing Gespiegelde oplossing

Figuur 5.13: Figuur bij opgave 14.

*Opgave 15*
Een onderzoeker beschikt over een interessante dataset: van 532 personen zijn 23 tests afgenomen die verband houden met affectieve variabelen zoals angst, depressie, etcetera. De onderzoeker weet weinig af van methodologie en multivariate methoden, maar wil toch deze gegevens analyseren. Een bevriende collega is bereid om de klus voor hem op te knappen. Er wordt afgesproken dat de collega zal proberen om de onderliggende affectieve dimensies van de 23 tests te achterhalen. De onderzoeker is wantrouwig van aard en vraagt een tweede collega om hetzelfde te

doen zodat hij de resultaten met elkaar kan vergelijken. Bij aflevering van de resultaten blijkt dat beide collega's factoranalyse hebben uitgevoerd. Beide collega's hadden de factorladingen bepaald op basis van het ULS-criterium en beide vonden een bevredigende oplossing met twee gemeenschappelijke factoren. Maar al vlug merkt de onderzoeker dat er een belangrijk verschil is tussen de twee analyses. Zo blijkt dat beide collega's tot de conclusie gekomen zijn dat test 22 in zeer belangrijke mate Factor I meet. Maar, collega 1 zegt dat test 22 Factor II niet meet en collega 2 beweert het tegendeel. Bovendien zegt collega 1 dat er wél een correlatie is tussen test 22 en Factor II. Ook voor andere tests doen er zich gelijkaardige verschillen voor. De onderzoeker vermoedt dat collega 1 fouten heeft gemaakt in de analyse. Is zijn vermoeden correct? Hoe kan je de vaststellingen van de twee collega's verklaren?

*Opgave 16*
In een factoroplossing met twee oblieke factoren heeft een test de volgende factorladingen: 0.7 en 0.0. En een tweede test heeft de volgende factorladingen: 0.7 en 0.8. Geef aan, voor elk van de onderstaande beweringen, of ze waar of onwaar zijn. Motiveer je antwoord, als je vindt dat de bewering onwaar is.

1. Deze twee tests zijn factorieel zuiver.
2. Omdat de factorladingen van deze twee tests positief of nul zijn, correleert geen van beide tests negatief met de twee factoren.
3. De tweede test is een Heywoodgeval.

*Opgave 17*
Geef, voor elk van de volgende stellingen, aan of ze waar of onwaar zijn en, indien ze onwaar zijn, leg uit waarom:

1. Bij een orthogonale factoroplossing zijn de correlaties tussen de factoren gelijk aan 0.
2. Na een oblieke rotatie zijn de correlaties tussen de factoren veranderd.
3. In de regel, veranderen de communaliteiten door een oblieke rotatie.
4. Omdat de factoren gecorreleerd zijn, verandert de totale hoeveelheid verklaarde variantie door een oblieke rotatie.
5. Met een oblieke rotatie (in vergelijking met een orthogonale rotatie) is het eenvoudiger om een factoroplossing te vinden waarin elke variabele slechts op één van de factoren substantieel laadt.
6. Een variabele die negatief laadt op een factor is ook altijd negatief gecorreleerd met deze factor.
7. Door een oblieke rotatie is het mogelijk dat variabelen die eerst bínnen de eenheidscirkel liggen, nu buíten de eenheidscirkel liggen.
8. Na een oblieke rotatie is het mogelijk dat variabelen met een positieve lading op een bepaalde factor toch negatief gecorreleerd zijn met deze factor.
9. Na een factoranalyse met gecorreleerde residuen kan men beter een oblieke dan een orthogonale rotatie uitvoeren. Voor een factoranalyse met ongecorreleerde residuen maakt dit niets uit.

*Opgave 18*
Geef, voor elk van de onderstaande uitspraken, aan of ze waar of onwaar zijn.

Een voordeel van oblieke boven orthogonale rotatie is dat men

1. meer variantie kan verklaren.

2. de correlaties beter kan reproduceren.
3. beter factorieel zuivere tests kan verkrijgen.
4. de totale hoeveelheid verklaarde variantie kan opsplitsen per factor.

*Opgave 19*
Na een oblieke rotatie vindt een onderzoeker de factorladingen- en factorstructuurmatrices in Tabel 5.16. Is de correlatie tussen de factoren positief, ongeveer nul, of negatief? Leg uit.

| | Ladingen | | Structuur | |
|---|---|---|---|---|
| | Factor 1 | Factor 2 | Factor 1 | Factor 2 |
| Test 1 | 0.90 | 0.05 | 0.8650 | -0.5800 |
| Test 2 | 0.70 | 0.06 | 0.6580 | -0.4300 |
| Test 3 | 0.80 | 0.09 | 0.7370 | -0.4700 |
| Test 4 | 0.06 | 0.80 | -0.5000 | 0.7580 |
| Test 5 | 0.07 | 0.85 | -0.5250 | 0.8010 |
| Test 6 | 0.10 | 0.95 | -0.5650 | 0.8800 |

Tabel 5.16: Tabel bij opgave 19.

---

## 5.7 Het Model met Meer dan Twee Gemeenschappelijke Factoren

### 5.7.1 Veralgemening van de Formules

In de praktijk zullen we vaak meer dan twee gemeenschappelijke factoren nodig hebben voor we een bevredigende oplossing vinden. Maar met drie, vier of nog meer factoren gaat het in principe op dezelfde manier als met twee. Bijvoorbeeld, bij vier gemeenschappelijke orthogonale factoren en ongecorreleerde residuen geldt het volgende voor de correlatie tussen $X_i$ en $X_j$:

$$r(X_i, X_j) = a_{i1} \cdot a_{j1} + a_{i2} \cdot a_{j2} + a_{i3} \cdot a_{j3} + a_{i4} \cdot a_{j4} \quad .$$

Indien, daarentegen, de residuen gecorreleerd zijn, dan moet men in het rechterlid de term $\mathrm{COV}(R_i, R_j)$ toevoegen:

$$r(X_i, X_j) = a_{i1} \cdot a_{j1} + a_{i2} \cdot a_{j2} + a_{i3} \cdot a_{j3} + a_{i4} \cdot a_{j4} + \mathrm{COV}(R_i, R_j) \quad .$$

De gereproduceerde correlatie $\widehat{r}(X_i, X_j)$ wordt opnieuw berekend met de bovenstaande formule voor $r(X_i, X_j)$ in het geval van ongecorreleerde residuen. In deze formule worden de geschatte factorladingen ingevuld. Deze geschatte factorladingen worden berekend op basis van het ULS- of ML-criterium (in het geval van ongecorreleerde residuen) of op basis van het criterium van maximale verklaarde variantie (in het geval van gecorreleerde residuen). Ook de communaliteiten van de variabelen worden op dezelfde wijze berekend, namelijk als de som van de gekwadrateerde ladingen van de variabele op de verschillende

factoren. Bijvoorbeeld, bij vier gemeenschappelijke factoren wordt de communaliteit als volgt berekend:

$$h_i^2 = a_{i1}^2 + a_{i2}^2 + a_{i3}^2 + a_{i4}^2 \quad .$$

Natuurlijk hebben alle factoren weer hun eigen eigenwaarde. Deze is te berekenen door de gekwadrateerde ladingen per kolom op te tellen. Dit alles geldt natuurlijk alleen indien de factoren orthogonaal zijn.

### 5.7.2 Grafische Weergave van de Factoroplossing

Met meer gemeenschappelijke factoren wordt de grafische weergave van een factoroplossing door een factorladingenplot wat moeilijker. Met drie factoren bijvoorbeeld worden de variabelen in een driedimensionaal assenstelsel getekend, waarin de factoren afgezet zijn langs drie onderling loodrechte assen. (Vergeleken met de boven gegeven tweedimensionale plots kan men zich voorstellen dat er nu nog een as corresponderend met de derde factor is die loodrecht op het papier staat.) De coördinaten van de variabelen in deze ruimte worden weer gegeven door hun ladingen op de drie factoren en omdat nog steeds de communaliteit van een variabele ten hoogste gelijk kan zijn aan 1, liggen de variabelen binnen de bol met straal 1 om de oorsprong. Variabelen met een hoge communaliteit liggen weer dichter bij de buitenkant van de bol, variabelen met een lage communaliteit dicht bij de oorsprong. Wij kunnen ons dit ruimtelijk wel voorstellen maar het is moeilijk om dit in een tweedimensionale figuur te tekenen. Voor vier en meer factoren hebben we zelfs een vier- of hogerdimensionale ruimte, met daarin vier of meer onderling loodrechte assen en met alle variabelen binnen de hyper-bol met straal 1 om de oorsprong. Dit gaat ons voorstellingsvermogen teboven. Wat we dan wel kunnen doen is tweedimensionale projecties van zulke hoogdimensionale ruimtes bekijken. Dit wil zeggen, we kiezen een tweetal factoren en tekenen dan de plot die bij deze twee gemeenschappelijke factoren hoort, waarin de variabelen als coördinaten hun ladingen op deze twee gemeenschappelijke factoren hebben. Voor elk tweetal factoren is er in principe zo'n plot te construeren, en op deze manier werpen we vanuit alle gezichtshoeken een blik in de hoogdimensionale ruimte.

### 5.7.3 Roteren met Meer dan Twee Factoren

Ook het rotatieprobleem wordt, praktisch gezien, ingewikkelder met meer factoren. Met twee gemeenschappelijke factoren valt vaak met het oog wel te zien welke rotatie van een gevonden oplossing veelbelovend is, met drie factoren lukt dit soms ook nog, maar met meer factoren missen we het totaaloverzicht. De tweedimensionale plots voor paren van factoren kunnen hierbij wel helpen, maar men moet er steeds op verdacht zijn dat een bepaalde rotatie in zo'n tweedimensionale deelruimte (in het vlak van de plot) ook effecten heeft op de situatie in sommige andere plots. Het wordt vaak moeilijk de operaties in de verschillende vlakken (plots) goed te integreren tot de meest geschikte rotatie van de gehele hoogdimensionale ruimte.

Dit is een tijdrovende, veel rekenwerk vereisende klus en daarom bij uitstek iets wat we graag aan een computer zouden willen uitbesteden. In software die factoranalyses uitvoert is dan ook altijd een procedure opgenomen die op een gevonden oplossing een

rotatie uitvoert, in de hoop dat de geroteerde oplossing beter interpreteerbaar is voor de gebruiker van het programma. Zo'n geautomatiseerde rotatieprocedure kan natuurlijk helemaal geen oplossing zien of interpreteren en moet haar keuze voor een rotatie baseren op een formeel criterium. Er zijn verscheidene van zulke criteria in omloop. Het meest gebruikte criterium is het zogenaamde VARIMAX-criterium, dat onder andere beschikbaar is binnen SPSS. VARIMAX voert een echte rotatie uit; de geroteerde factoren zijn nog steeds orthogonaal. Er bestaan ook geautomatiseerde procedures om oblieke rotaties uit te voeren, zoals OBLIMIN, dat ook beschikbaar is binnen SPSS. In wat volgt, concentreren we ons op het VARIMAX-criterium.

We hebben hierboven al opgemerkt dat een factor vaak makkelijker te interpreteren is als de ladingen van de verschillende variabelen op de factor in absolute waarde ver uit elkaar liggen. Dit wil zeggen, als er een duidelijke opsplitsing is in een groep van hoge en een groep van lage factorladingen. Dit is precies wat men met het VARIMAX-criterium probeert te bereiken, maar dan gemiddeld over de factoren. (We kunnen het criterium namelijk niet zomaar op elke afzonderlijke factor toepassen, want de uiteindelijke oplossing moet door rotatie te verkrijgen zijn uit de oorspronkelijke oplossing. Daarom heeft wat we met één factor doen consequenties voor wat nog met de andere factoren mogelijk is.)

De naam VARIMAX roept vaak ten onrechte de associatie op dat de hoeveelheid verklaarde variantie in de variabelen gemaximaliseerd wordt. Een rotatie, het doet er niet toe welke, verandert nooit de hoeveelheid verklaarde variantie van de variabelen. Wat VARIMAX wél maximaliseert, is de variantie (over de variabelen) van de gekwadrateerde factorladingen, gemiddeld over de factoren. Hierbij dient het kwadrateren van de factorladingen alleen om de factorladingen positief te maken; men had dus net zo goed de absolute waarde kunnen nemen. Door de variantie van de gekwadrateerde factorladingen te maximaliseren wordt verkregen dat, gemiddeld over de factoren, de factorladingen zoveel als mogelijk in absolute waarde ver uit elkaar liggen. Het ideaal dat men op deze manier probeert te bereiken is dat, voor elke factor, de ladingen uiteenvallen in een groep met (in absolute zin) zeer hoge waarden en een andere groep met waarden gelijk aan nul.

## 5.8 De Keuze van het Aantal Gemeenschappelijke Factoren

We zijn nog niet expliciet ingegaan op de vraag hoe te beslissen hoeveel gemeenschappelijke factoren er getrokken moeten worden. De regel hierbij is dat we niet meer gemeenschappelijke factoren willen trekken dan nodig is. De vraag hoeveel er nodig zijn wordt verder in deze paragraaf behandeld. We blijven eerst even stilstaan bij het feit dat het onverstandig is om zoveel mogelijk gemeenschappelijke factoren te trekken. Stel dat het doel is om de correlaties te verklaren. Als het aantal factoren te groot wordt, dan legt het model geen enkele beperking meer op aan de correlatiematrix. Dan bestaat er altijd een perfecte oplossing, welke waarden er ook in de correlatiematrix voorkomen. Zo'n oplossing in veel gemeenschappelijke factoren is daarom triviaal, zonder verklarende waarde. Als we bijvoorbeeld evenveel latente factoren toelaten als er geobserveerde variabelen

zijn kunnen we, in de geest van de klassieke testtheorie, als latente factoren de ware-scorecomponenten van alle geobserveerde variabelen nemen. Concreet, als $X_i = T_i + E_i$ de opsplitsing in ware-scoredeel en meetfoutgedeelte is van de $i$-de geobserveerde variabele, dan nemen we alle ware scores $T$ als onze latente factoren. Het is dan zeker waar dat deze factoren de correlaties tussen de geobserveerde variabelen volledig verklaren: na uitzuivering van de ware-scoredelen, blijft er van de correlatie tussen twee geobserveerde variabelen slechts de correlatie tussen de meetfoutgedeelten over, en die is gelijk aan nul. Het is echter ook duidelijk dat we op deze manier weinig verklaard hebben. Er heeft geen reductie plaatsgevonden, het probleem is slechts verplaatst en de vraag luidt nu: hoe verklaren we de correlaties tussen deze ware-scorevariabelen? Kunnen deze correlaties zijn teweeggebracht door een veel kleiner aantal echt gemeenschappelijke factoren? Men kan een gelijkaardig verhaal vertellen als het doel is om variantie te verklaren: door meer gemeenschappelijke factoren te trekken, verklaart men ook meer variantie, maar dit gaat ten koste van de mate van reductie die we bewerkstelligen met de factoranalyse.

We behandelen nu de vraag hoeveel gemeenschappelijke factoren er nodig zijn voor een bepaalde dataset. De criteria die men hierbij hanteert verschillen naargelang men correlaties of variantie wil verklaren. Als men correlaties wil verklaren, dan zijn de residuele correlaties van belang, en als men variantie wil verklaren dan zijn de eigenwaarden van belang. Daarnaast is er ook een criterium dat voor beide analyses van belang is: de interpreteerbaarheid van de gevonden oplossing. Tenslotte, voor de factoranalyse waarin correlaties verklaard worden, bestaat er ook een statistisch criterium. In de praktijk wordt de waarde van dit statistisch criterium vaak overschat, en daarom besteden we er hier wat aandacht aan.

### 5.8.1 Inspectie van de Residuele Correlaties.

Dit criterium is van belang voor de factoranalyse met ongecorreleerde residuen. We pasten het reeds toe bij het voorbeeld. In dit voorbeeld werden de correlaties tussen de variabelen $X_3$, $X_4$ en $X_5$ duidelijk niet goed gereproduceerd door de beste één-factoroplossing, terwijl alle correlaties perfect werden gereproduceerd door twee gemeenschappelijke factoren. Hier was het dus duidelijk dat voor deze data twee gemeenschappelijke factoren getrokken moesten worden. Met echte data in de praktijk zal het natuurlijk nooit zo duidelijk liggen.

Het vertrekpunt is altijd een factoroplossing met een bepaald aantal factoren. Als de residuele correlaties over het algemeen klein zijn, en de afwijkingen van nul toevallige schommelingen lijken, dan heeft men wat dit betreft geen reden om nog een factor toe te voegen. Indien er nog wél vrij grote residuele correlaties zijn, en vooral wanneer deze systematisch voorkomen (zoals in ons voorbeeld met de één-factoroplossing, geconcentreerd rond een drietal variabelen), kan het toevoegen van een factor nuttig zijn. Dit zijn geen objectieve criteria, maar eerder subjectieve oordelen. De onderzoeker zelf moet bijvoorbeeld beslissen hoe *klein* klein genoeg en hoe *algemeen* algemeen genoeg is. Omdat er voor "klein" en "algemeen" geen hard richtgetal bestaat, moeten we ons behelpen met een vuistregel. Een vaak gehanteerde vuistregel is dat alle residuele correlaties in absolute waarde kleiner moeten zijn dan 0.05.

### 5.8.2 Inspectie van de Eigenwaarden

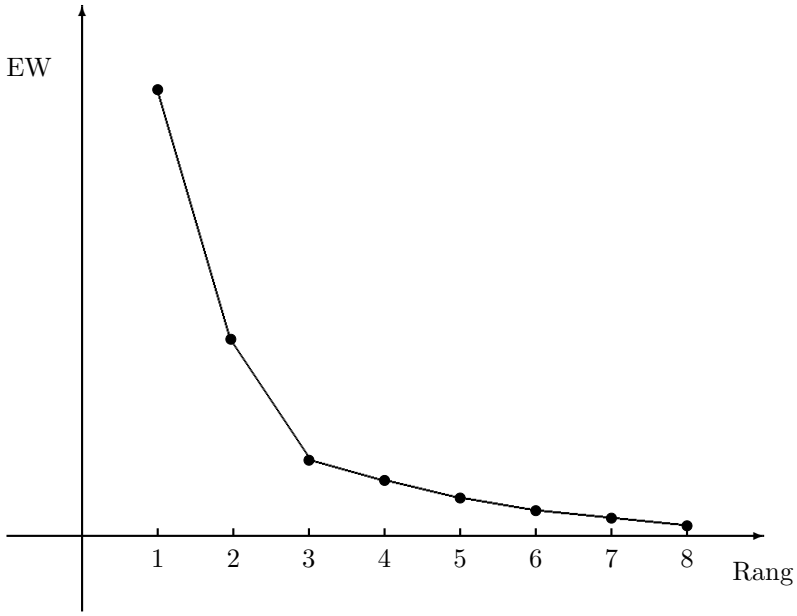
Dit criterium is van belang voor de hoofdcomponentenanalyse. Met de hoofdcomponentenanalyse wordt gezocht naar de factoren die zoveel mogelijk variantie verklaren. De hoeveelheid variantie verklaard door één gemeenschappelijke factor, wordt de eigenwaarde van die factor genoemd. In de hoofdcomponentenanalyse worden de factoren getrokken in de volgorde van de eigenwaarden (de factor met de grootste eigenwaarde als eerste, die met de tweede grootste eigenwaarde als tweede, ...). Het ligt dan voor de hand om factoren te blijven trekken zolang hun eigenwaarden nog voldoende groot zijn. Er bestaat echter geen objectief criterium voor wat "voldoende groot" is.

De grootte van de eigenwaarden kan op twee manieren geëvalueerd worden: op een absolute of een relatieve manier. De absolute manier houdt in dat alleen de factoren met een eigenwaarde groter dan 1 behouden worden. De rationale hierachter is dat de variantie die door een factor verklaard moet worden minstens gelijk moet zijn aan de variantie van één variabele. En die variantie is gelijk aan 1, omdat alle variabelen gestandaardiseerd zijn.

De relatieve manier om de grootte van de eigenwaarden te evalueren vertrekt van een grafiek, de zogenaamde *eigenwaardenplot*, waarin de eigenwaarden uitgezet worden tegen hun rangnummer naar grootte. In Figuur 5.14 wordt zo'n eigenwaardenplot getoond. Om het aantal factoren te bepalen op basis van een eigenwaardenplot, zoekt men naar een knik in de curve. In Figuur 5.14 ligt deze knik bij de derde eigenwaarde. Het aantal factoren dat behouden wordt is gelijk aan het aantal factoren vóór de knik (twee, in Figuur 5.14). De verantwoording van deze procedure is gebaseerd op de idee dat de grote en de kleine eigenwaarden twee duidelijk te onderscheiden groepen vormen, met binnen de groep van de kleine eigenwaarden veel kleinere onderlinge verschillen dan binnen de groep van de grote eigenwaarden. Het is niet gegarandeerd dat er in elke eigenwaardenplot ook zo'n mooie knik zit als in Figuur 5.14; soms zit er geen duidelijk knik in de eigenwaardenplot en soms zitten er meerdere in. Deze procedure om het aantal factoren te bepalen is dus niet altijd bruikbaar. Indien hij wel bruikbaar is, dan zegt men dat het aantal factoren bepaald werd op basis van een *scree-test*. Deze naam komt van het Engelse *scree*, wat puinhelling betekent, een naam die begrijpelijk is als men weet dat, in de regel, alleen kleine stukjes puin van een helling afrollen.

### 5.8.3 De Interpreteerbaarheid van de Gevonden factoroplossing.

Dit criterium is van belang voor beide vormen van factoranalyse. Als een gevonden oplossing goed interpreteerbaar is, hebben we wat dit betreft geen reden om nog een factor toe te voegen. Vinden we geen geschikte interpretatie voor onze factoren, dan kunnen we altijd proberen of het toevoegen van een factor nieuwe gezichtspunten oplevert. Soms heeft men voor alle factoren op één na een goede interpretatie, maar die ene factor lijkt niets specifieks te doen (heeft bijvoorbeeld op alle variabelen lage tot middelmatige ladingen). Dan kan men overwegen een nieuwe analyse te doen met één gemeenschappelijke factor minder. Deze eenvoudigere oplossing met minder factoren is te verkiezen indien deze oplossing (1) wél goed te interpreteren is, (2) (voor de factoranalyse met ongecorreleerde residuen) niet leidt tot een al te grote toename van de residuele correlaties, en (3) (voor de hoofdcomponentenanalyse) niet leidt tot een al




Figuur 5.14: Eigenwaardenplot.

te grote afname van de totale hoeveelheid verklaarde variantie. Is dit laatste niet het geval, dan vervulde deze niet-interpreteerbare factor blijkbaar toch een nuttige functie en moeten we hem in onze oplossing behouden.

### 5.8.4 De Uitslag van een Statistische Toets

Men kan deze toets uitvoeren indien de factoroplossing bepaald wordt op basis van het ML-criterium. In oudere versies van SPSS vind je deze toets onder het kopje *Chi-square Statistic* en in nieuwe versies onder *Goodness-of-fit Test*. De nulhypothese die getoetst wordt heeft betrekking op het factormodel met ongecorreleerde residuen. Deze nulhypothese houdt in dat, in de populatie, de correlatiematrix met het opgegeven aantal gemeenschappelijke factoren te verklaren is.

Er zijn twee problemen met deze toets. Ten eerste, deze toets is alleen zinvol indien men ook echt gelooft in de nulhypothese die men ermee toetst, en veel gebruikers van de factoranalyse doen dat niet. Concreet, weinigen geloven dat het factormodel met een specifiek aantal gemeenschappelijke factoren ook echt kan opgaan in de populatie. Of we op zo'n toets een significant resultaat verkrijgen of niet is dan (namelijk, als het model, de nulhypothese van de toets, niet perfect klopt) een functie van het onderscheidingsvermogen (in het Engels, de *power*) van de toets. Dit onderscheidingsvermogen hangt in sterke mate af van de grootte van de steekproef waarop de correlaties zijn berekend. Is deze steekproef groot, dan zullen we snel een significant resultaat vinden, leidend tot een verwerping van het model met het gegeven aantal factoren, en dus mogelijk tot het toevoegen van een factor.

De meeste gebruikers van de factoranalyse zien het factormodel met een klein aantal gemeenschappelijke factoren als een *benadering* van de werkelijkheid. Als deze gebruikers blind zouden varen op de uitslag van de genoemde statistische toets, dan zouden zij vaak hun doel voorbijschieten. Immers, een factoroplossing met een klein aantal factoren kan best een heel bruikbare benadering geven van de werkelijkheid, met bijvoorbeeld interpreteerbare factoren en kleine residuen, maar de statistische toets kan toch significant zijn omwille van de grootte van de steekproef.

Ten tweede, indien de steekproef kleín is, kan men weinig vertrouwen hebben in de p-waarde van toetsstatistiek die in de uitvoer afgedrukt wordt. De reden hiervoor is dat deze p-waarde slechts een *asymptotische* p-waarde is. De exacte verdeling van de toetsstatistiek onder de genoemde nulhypothese is immers niet gekend, en daarom moet men gebruik maken van de asymptotische verdeling, die wel gekend is. Nu, die asymptotische verdeling komt pas redelijk in de buurt van de exacte verdeling als de steekproef rond de 500 personen bevat.

Deze bedenkingen bij de statistische toets die SPSS voor ons berekent wekken misschien de indruk dat hij helemaal niet nuttig zou zijn. Deze indruk klopt niet; een niet-significante p-waarde bij een grote steekproef heeft wel degelijk een nuttige interpretatie. Deze interpretatie is dat het opgegeven aantal gemeenschappelijke factoren voldoende is om de geobserveerde correlaties te verklaren. Het interpretatieprobleem doet zich zowel voor bij significante p-waarden als bij kleine steekproeven: een significante p-waarde betekent niet automatisch dat de factoroplossing met het opgegeven aantal gemeenschappelijke factoren onbruikbaar is als benadering van de werkelijkheid, en een kleine steekproef betekent juist dat we geen vertrouwen kunnen hebben in de p-waarde van de toetsstatistiek (wat die p-waarde ook is).

## 5.9 Het Bepalen van de Factorscores

Soms is het nuttig om, behalve de factorladingen ($a_{i1}$, $a_{i2}$, ...), ook de factorscores ($F_1$, $F_2$, ...) te kennen. Het is bijvoorbeeld nuttig in een onderzoek waarin men een criterium wil voorspellen op basis van een zeer groot aantal predictoren. Meestal zijn de predictoren items van een vragenlijst. Als het doel alleen is om het criterium zo goed mogelijk te voorspellen, dan kan men dit groot aantal predictoren best allemaal tegelijk in een meervoudige regressieanalyse opnemen. Vaak wil men echter ook weten wat de unieke bijdragen zijn van de individuele predictoren en daar een samenhangende inhoudelijke interpretatie aan geven. Dit is onmogelijk als het aantal predictoren erg groot is. Een oplossing voor dit probleem is dat men eerst een factoranalyse uitvoert op het grote aantal variabelen dat men eerst als predictoren wou gebruiken en dan de factorscores van deze factoroplossing gebruikt als de feitelijke predictoren.

Het bepalen van de factorscores verloopt anders in de factoranalyse met gecorreleerde en ongecorreleerde residuen. De situatie is het eenvoudigst in de factoranalyse met gecorreleerde residuen (de hoofdcomponentenanalyse). In een hoofdcomponentenanalyse worden de factorscores gewoon samen met de factorladingen berekend. Het blijkt dat deze factorscores eenvoudige gewogen sommen zijn van de geobserveerde variabelen. In de factoranalyse met ongecorreleerde residuen worden de factorladingen eerst bepaald, en daarna pas de factorscores. In de literatuur vindt men verschillende methoden die hier-

voor gebruikt kunnen worden. De meest bekende zijn de regressiemethode, de methode van Anderson en Rubin, en de methode van Bartlett. Het blijkt niet mogelijk te zijn om met deze methoden factorscores te bepalen die altijd de correcte varianties en correlaties hebben, zoals wel het geval is bij de hoofdcomponentenanalyse. Met "correcte varianties" bedoelen we dat zij allemaal gelijk zijn aan 1, zoals verondersteld wordt door het factoranalysemodel. En met "correcte correlaties" bedoelen we dat deze correlaties in overeenstemming zijn met de gekozen rotatie. In de praktijk, produceren deze methoden meestal factorscores die redelijk in de buurt komen van de correcte varianties en correlaties.

## 5.10 Uitvoeren van een Factoranalyse met de SPSS-procedure Factor Analysis

Men vindt de procedure Factor Analysis in het menu *Analyze* bij de categorie *Data Reduction*. In het dialoogvenster van Factor Analysis moet men eerst aangeven welke variabelen betrokken moeten worden in de analyse. Dit doet men door de variabelenamen te selecteren en naar het vakje onder *Variables* over te brengen.

Onderaan het dialoogvenster staan vijf knoppen die nu één voor één besproken worden.

*Descriptives* Via deze knop kan je vragen om in de uitvoer een aantal optionele statistieken af te drukken. De opties *Univariate Statistics* en *Initial Solution* zijn zelden nuttig. Om de hoeveelheid uitvoer te beperken worden ze niet aangekruist.

Onder *Correlation Matrix* is voor ons alleen de optie *Reproduced* van belang. Deze optie geeft niet alleen de gereproduceerde maar ook de residuele correlaties.

*Extraction* Via deze knop kan men onder andere aangeven hoe de factorladingen berekend moeten worden: met *maximum likelihood*, *unweighted least-squares* of *principal components analysis*. Naast deze drie opties, kan men ook nog andere kiezen, maar de corresponderende technieken worden in dit boek niet besproken.

In het dialoogvenster *Extraction* kan men ook aangeven of men een analyse wil uitvoeren op de correlaties of op de covarianties; in de meeste toepassingen is alleen een analyse op de correlaties zinvol.

Onder *Display* kun je aangeven of je een *scree plot* wil, en als je voor *principal components* gekozen hebt, moet je dat ook doen. De optie *Unrotated factor solution* onder *Display* is vrijwel nooit interessant.

Onder *Extract* specificeert men de wijze waarop het aantal factoren bepaald wordt; meestal gebeurt dat door een getal in te vullen achter *Number of factors*.

Enkele methoden om factorladingen te bepalen, zoals maximum likelihood, doen beroep op zogenaamde *iteratieve algoritmen*[5]. Het maximum aantal iteraties kan

[5]Iteratief betekent *herhalend*. Het best gekende voorbeeld van een iteratief algoritme is de staartdeling die we geleerd hebben op de basisschool. Bij het uitvoeren van een staartdeling wordt een elementaire rekenoperatie (i.c., het bepalen van een veelvoud van de deler en het aftrekken van dit veelvoud van een restwaarde die overblijft na een vorige aftrekoperatie) een aantal keer na elkaar uitgevoerd op het

men binnen SPSS zelf bepalen en dit gebeurt achter het label *Maximum iterations for convergence*. Schrijf in het vakje achter dit label het getal 500. Dit aantal is voldoende voor alle analyses die je in de hierna volgende opgaven moet uitvoeren.

*Rotation* Via deze knop kan men aangeven wat voor soort rotatie men wil uitvoeren. Bij sommige van deze rotaties (i.c., *Direct Oblimin* en *Promax*) kan men met een extra parameter (i.c., *Delta* en *Kappa*) het rotatiecriterium wat aanpassen. In dit boek wordt hier geen aandacht aan geschonken.

Onder *Display* kan men aangeven of de geroteerde factoroplossing afgedrukt moet worden. Deze optie wordt bijna altijd aangekruist. Met de optie *Loading plot(s)* kan men aangeven of er een factorladingenplot afgedrukt moet worden. Deze optie is nuttig als er slechts een klein aantal factoren getrokken wordt (twee of drie).

Enkele rotaties, zoals *Varimax* en *Direct Oblimin*, doen beroep op iteratieve algoritmen. Het maximum aantal iteraties schrijft men achter het label *Maximum iterations for convergence*. Schrijf in dit vakje het getal 500.

*Scores* Via deze knop kan men aangeven óf en, zo ja, op welke manier de factorscores berekend moeten worden. Men kan hier kiezen uit de drie methoden die in dit hoofdstuk kort besproken werden.

*Options* Via deze knop kan men onder andere bepalen hoe de factorladingen afgedrukt worden. Als men *Sorted by size* aankruist, dan worden de factorladingen in volgorde van grootte afgedrukt. Hierdoor springen de variabelen die voor de interpretatie belangrijk zijn sneller in het oog. De optie *Suppress absolute value less than* beoogt hetzelfde doel: het vereenvoudigen van de interpretatie. Door deze optie aan te kruisen en een bepaalde minimumwaarde in te vullen worden alleen die factorladingen getoond die in absolute waarde groter zijn dan deze minimumwaarde. Hierdoor voorkom je dat er een heleboel getallen afgedrukt worden waarvan er slechts een klein aantal betekenisvol is. Meestal kiest men voor een minimumwaarde tussen 0.25 en 0.35.

---

*Opgave 20*

In een factoroplossing met drie orthogonale factoren heeft de eerste variabele de volgende factorladingen: 0.7, 0.2, en 0.3. En de tweede variabele heeft de volgende factorladingen: 0.5, 0.8, en 0.4.

1. Wat is de gereproduceerde correlatie tussen de eerste en de tweede variabele?
2. Is de tweede variabele een Heywoodgeval? Leg uit waarom wel of waarom niet.
3. Waar of onwaar en leg uit (ook als het waar is): Omdat alle factorladingen van deze twee variabelen positief zijn, correleren beide variabelen positief met elk van de drie factoren.
4. Wat gebeurt er met de factorladingen als de tweede factor gespiegeld wordt?

---

resultaat van de vorige elementaire rekenoperatie. Het resultaat van de elementaire rekenoperatie is telkens (1) het veelvoud van de deler dat afgetrokken wordt en (2) de restwaarde die overblijft na de aftrekoperatie (de invoer voor de volgende uitvoering van de elementaire rekenoperatie). Dit is kenmerkend voor een iteratief algoritme: een elementaire rekenoperatie wordt herhaald (iteratief) toegepast op het resultaat van de vorige herhaling (iteratie).

*Opgave 21*
Op de correlaties tussen zes tests wordt een factoranalyse uitgevoerd waarbij drie orthogonale factoren getrokken worden. De factorladingen worden gegeven in Tabel 5.17.

| | Factor 1 | Factor 2 | Factor 3 |
|---|---|---|---|
| TEST 1 | .70 | .00 | -.50 |
| TEST 2 | .10 | .70 | -.10 |
| TEST 3 | .00 | -.50 | .80 |
| TEST 4 | .50 | .80 | .00 |
| TEST 5 | .70 | .00 | -.70 |
| TEST 6 | -.20 | .00 | .90 |

Tabel 5.17: Tabel bij opgave 21.

1. Wat is de gereproduceerde correlatie tussen TEST 2 en TEST 4?
2. Wat is de communaliteit van TEST 2? En wat is zijn uniciteit?
3. Stel dat de geobserveerde correlatie tussen TEST 1 en TEST 6 gelijk is aan $-0.45$. Wat is dan de residuele correlatie?
4. Hoeveel procent van de totale variantie wordt verklaard door de tweede factor alleen?
5. Stel dat we een oblieke rotatie uitvoeren op de gevonden factoroplossing. Welke verandering kunnen we dan verwachten in de totale hoeveelheid verklaarde variantie door de drie factoren?

*Opgave 22*
Geef, voor elk van de volgende stellingen, aan of ze waar of onwaar zijn en, indien ze onwaar zijn, leg uit waarom:

1. De VARIMAX-rotatie heeft tot doel de hoeveelheid door de factoren verklaarde variantie te maximaliseren.
2. Met het uitvoeren van een VARIMAX-rotatie is, gemiddeld over alle variabelen, de proportie door de factoren verklaarde variantie toegenomen.
3. Na een VARIMAX-rotatie zijn de factorladingen en de gereproduceerde correlaties veranderd, maar de communaliteiten zijn gelijk gebleven.
4. Met het uitvoeren van een VARIMAX-rotatie zijn, gemiddeld over alle paren van variabelen, de residuele correlaties kleiner geworden.
5. Wanneer alle variabelen dicht bij elkaar liggen in een factorladingenplot, dan kan men met een VARIMAX-rotatie ervoor zorgen dat ze meer gespreid binnen de eenheidscirkel komen te liggen.
6. Met een VARIMAX-rotatie kan men ervoor zorgen dat, gemiddeld over de factoren, de ladingen per factor zoveel als mogelijk in absolute waarde uit elkaar liggen.
7. Na een OBLIMIN-rotatie zijn de ladingen veranderd maar de correlaties tussen de factoren zijn gelijk gebleven.
8. Indien de statistische toets die uitgevoerd wordt bij de ML-methode significant is, dan moet men alleen dán een groter aantal factoren trekken als het aantal proefpersonen groot is.

*Opgave 23*
Negen items uit een vragenlijst naar huwelijkstevredenheid worden aan een factoranalyse onderworpen. Drie van deze items, VLD1, VLD2 en VLD3, meten hoe voldaan men is met de eigen situatie, drie andere, AFH1, AFH2 en AFH3, hoe afhankelijk men zich voelt van de partner, en de laatste drie, OPN1, OPN2 en OPN3, in hoeverre men vrijelijk spreekt over intieme zaken. Er worden drie factoren getrokken met de ML-methode. De uitvoer begint met de melding:

ML Extracted 3 factors. 19 Iterations required.
Chi-square Statistic: 10.9505, D.F.: 12, Significance: .5332

Op de factoren wordt een VARIMAX-rotatie uitgevoerd. De factorladingenmatrix zoals die verschijnt in de SPSS-uitvoer wordt gegeven in Tabel 5.18. In de SPSS-uitvoer wordt ook de matrix met gereproduceerde en residuele correlaties gegeven. Deze matrix wordt gegeven in Tabel 5.19.

| | Factor 1 | Factor 2 | Factor 3 |
|---|---|---|---|
| VLD1 | .86 | -.35 | .00 |
| VLD2 | .81 | -.25 | .02 |
| VLD3 | .85 | -.32 | -.01 |
| AFH1 | -.30 | .75 | .10 |
| AFH2 | -.32 | .77 | .05 |
| AFH3 | -.21 | .66 | .06 |
| OPN1 | .04 | .05 | .49 |
| OPN2 | -.09 | .03 | .95 |
| OPN3 | .01 | .04 | .54 |

Tabel 5.18: Tabel bij opgave 23.

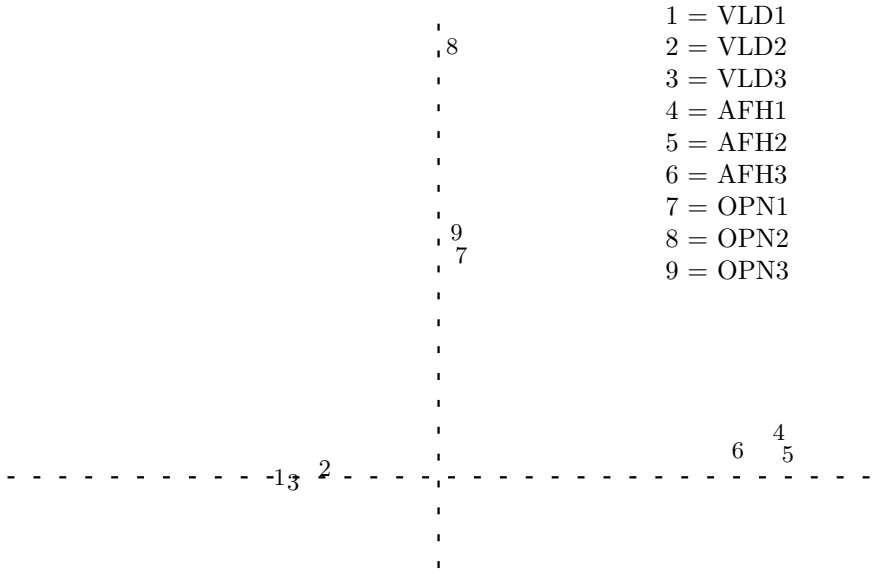
Reproduced Correlation Matrix:

| | VLD1 | VLD2 | VLD3 | AFH1 | AFH2 | AFH3 | OPN1 | OPN2 | OPN3 |
|---|---|---|---|---|---|---|---|---|---|
| VLD1 | .860* | .000 | .001 | .024 | -.019 | -.003 | -.012 | .001 | -.015 |
| VLD2 | .783 | .719* | -.001 | -.001 | -.008 | .015 | -.006 | -.003 | .042 |
| VLD3 | .845 | .771 | .830* | -.027 | .027 | -.005 | .019 | -.000 | -.008 |
| AFH1 | -.527 | -.436 | -.505 | .671* | .001 | .005 | -.029 | .003 | -.014 |
| AFH2 | -.542 | -.450 | -.519 | .683 | .687* | -.006 | .037 | -.001 | -.007 |
| AFH3 | -.414 | -.338 | -.396 | .571 | .582 | .490* | -.021 | -.001 | .029 |
| OPN1 | .017 | .030 | .011 | .076 | .054 | .057 | .242* | .000 | .006 |
| OPN2 | -.084 | -.058 | -.097 | .144 | .102 | .098 | .461 | .909* | -.001 |
| OPN3 | .000 | .014 | -.007 | .078 | .053 | .056 | .267 | .514 | .295* |

Tabel 5.19: Tabel bij opgave 23.

1. Van welk variabelenpaar wordt de correlatie het minst goed gereproduceerd? Leg uit.
2. Wat is de gereproduceerde correlatie tussen AFH1 en OPN2?
3. Vind je de gevonden factoroplossing bevredigend? Leg uit.
4. De ladingen op de factoren Factor 1 en Factor 2 zijn over het algemeen tegengesteld gericht. Is dit een aanwijzing dat deze twee factoren negatief gecorreleerd zijn? Leg uit.

5. Denk je dat een oblieke rotatie van de eerste en de tweede factor tot een beter interpreteerbare oplossing leidt? Motiveer je antwoord.
6. In Figuur 5.15 worden de ladingen op twee van de drie factoren weergegeven in een plot. Welke van de drie factoren Factor 1, Factor 2 en Factor 3, staat op de horizontale en welke op de verticale as in deze plot?
7. In Figuur 5.15 ligt de variabele VLD2 (2) veel dichter bij de oorsprong dan AFH3 (6), terwijl uit de Reproduced Correlation Matrix in Tabel 5.19 blijkt dat VLD2 een duidelijk hogere communaliteit heeft dan AFH3 ($h_2^2 = .719$, $h_6^2 = .490$). Hoe kan men dit verklaren?




Figuur 5.15: Figuur bij opgave 23.

*Opgave 24*
Deze vraag gaat over een onderzoek naar de houding ten opzichte van bedreigingen van het milieu. Aan een groep personen werden 20 uitspraken voorgelegd, waaronder de volgende:

- Men zou een plan moeten opstellen om alle kerncentrales op de wereld te ontmantelen (KERNCENTRALES).
- Dat de Siberische tijger met uitsterven bedreigd is, is een zaak van minder belang (TIJGER).
- In de media wordt teveel aandacht besteed aan het broeikaseffect (BROEIKAS).
- Het verlies van genetische diversiteit door de vernietiging van het Amazone-woud zal de mensheid later nog zuur opbreken (GENDIVERS).

Deze uitspraken werden beoordeeld op een schaal van 1 (volledig mee eens) tot 7 (volledig mee oneens). Vervolgens werd een factoranalyse uitgevoerd, waarbij twee gemeenschappelijke factoren werden getrokken. Bovendien werd er een VARIMAX-rotatie uitgevoerd. In de Tabel 5.20 geven we de factorladingen voor de vier bovenstaande uitspraken.

1. Hoe bereken je de gereproduceerde correlatie tussen KERNCENTRALES en BROEIKAS?

| | Factor 1 | Factor 2 |
|---|---|---|
| KERNCENTRALES | 0.85 | 0.04 |
| TIJGER | 0.07 | -0.92 |
| BROEIKAS | -0.82 | -0.05 |
| GENDIVERS | -0.03 | 0.88 |

Tabel 5.20: Tabel bij opgave 24.

2. Hoe bereken je de communaliteit van TIJGER?
3. Uit de berekening van de residuele correlaties blijkt dat de twee-factoroplossing goed past bij de gegevens. Een collega van de onderzoeker beweert dat die goede fit vooral te danken is aan de VARIMAX-rotatie, en dat het er wellicht slechter zal uitzien als er voor een andere rotatie gekozen wordt. Ben je het hier mee eens? Leg uit waarom wel of waarom niet.
4. Zijn de factorladingen van KERNCENTRALES, TIJGER, BROEIKAS en GENDIVERS informatief voor de interpretatie van de factoren? Indien ja, geef een interpretatie, en indien nee, leg uit waarom ze niet informatief zijn.

*Opgave 25*
Deze vraag gaat over een onderzoek naar interesse in sociaal-politieke actualiteit. Aan een groep personen werden 20 uitspraken voorgelegd, waaronder de volgende:

- De politieke situatie in Afghanistan kan me weinig schelen. (AFGHANISTAN)
- Als er hier in de gemeente over sportaccomodatie e.d. gediscussieerd wordt dan heb ik daar meestal geen mening over. (SPORT)
- De buitenlandse berichtgeving in de krant lees ik bijna altijd. (KRANT)
- In geval van een conflict tussen de woningcorporatie en het gemeentebestuur wil ik best als bewonersafgevaardigde optreden. (WONING)

Deze uitspraken werden beoordeeld op een schaal van 1 (volledig mee oneens) tot 7 (volledig mee eens). Vervolgens werd een factoranalyse uitgevoerd, waarbij twee gemeenschappelijke factoren werden getrokken. Op de factoren werd een VARIMAX-rotatie uitgevoerd. In Tabel 5.21 geven we de factorladingen van de vier bovenstaande items. Van de andere 16 items worden de factorladingen niet gegeven.

| | Factor 1 | Factor 2 |
|---|---|---|
| AFGHANISTAN | -0.85 | 0.04 |
| SPORT | 0.07 | -0.92 |
| KRANT | 0.82 | -0.05 |
| WONING | -0.03 | 0.88 |

Tabel 5.21: Tabel bij opgave 25.

1. Wat is de gereproduceerde correlatie tussen AFGHANISTAN en KRANT?
2. Zijn de bovenstaande gegevens voldoende om de eigenwaarde van Factor 1 te berekenen? Zo ja, laat zien hoe je haar berekent.
3. Kun je op basis van deze gegevens iets zeggen over de kwaliteit van de twee-factoroplossing? (Ga er hierbij vanuit dat de factorladingen van de andere 16 vragen vergelijkbaar zijn met die van de vier gegeven vragen.)

4. Zijn de factorladingen van AFGHANISTAN, SPORT, KRANT en WONING informatief voor de interpretatie van de factoren? Motiveer je antwoord en, indien mogelijk, geef een interpretatie.

*Opgave 26*
Een onderzoekster deed een onderzoek naar risicoperceptie. Ze legde daartoe de volgende stellingen voor aan haar proefpersonen:

> ... is een ernstig risico voor mijn welzijn.

Op de lege plek werden verschillende bedreigingen ingevuld. Op deze manier werd telkens een ander item geconstrueerd. De bedreigingen zijn de volgende: (1) inbraak (INBR), (2) een kernramp (KERN), (3) een giframp (GIFR), (4) verkeersdeelname (VERK), (5) oorlog (OORL), (6) genetische manipulatie (GENM), (7) terrorisme (TERR), en (8) een roofoverval (ROOF). De proefpersonen moesten op een schaal van 1 tot 7 aangeven in hoeverre ze deze bedreigingen als een risico voor hun welzijn ervaren. Er werd een factoranalyse uitgevoerd waarbij er twee factoren werden getrokken op basis van het ML-criterium. Op de factoren werd een VARIMAX-rotatie uitgevoerd. Het deel van de SPSS-uitvoer dat relevant is voor de onderstaande vragen wordt gegeven in Tabel 5.22, Tabel 5.23 en Tabel 5.24.

Correlation Matrix:

| | INBR | KERN | GIFR | VERK | OORL | GENM | TERR | ROOF |
|---|---|---|---|---|---|---|---|---|
| INBR | 1.00 | | | | | | | |
| KERN | -.11 | 1.00 | | | | | | |
| GIFR | .08 | .79 | 1.00 | | | | | |
| VERK | .53 | .39 | .36 | 1.00 | | | | |
| OORL | .24 | .53 | .71 | .46 | 1.00 | | | |
| GENM | -.13 | .71 | .54 | .32 | .56 | 1.00 | | |
| TERR | .62 | .41 | .54 | .65 | .50 | .37 | 1.00 | |
| ROOF | .71 | .11 | .32 | .53 | .35 | .06 | .64 | 1.00 |

Tabel 5.22: Tabel bij opgave 26.

Rotated Factor Matrix:

| | Factor 1 | Factor 2 |
|---|---|---|
| INBR | -.16 | .92 |
| KERN | .92 | .04 |
| GIFR | .84 | .24 |
| VERK | .36 | .63 |
| OORL | .64 | .36 |
| GENM | .76 | .01 |
| TERR | .43 | .75 |
| ROOF | .11 | .79 |

Tabel 5.23: Tabel bij opgave 26.

1. Een gedeelte van de matrix met gereproduceerde correlaties werd onleesbaar gemaakt doordat de onderzoekster koffie knoeide op de uitvoer (aangeduid met $\diamondsuit$). Vul de ontbrekende gegevens aan.

Reproduced Correlation Matrix:

| | INBR | KERN | GIFR | VERK | OORL | GENM | TERR | ROOF |
|---|---|---|---|---|---|---|---|---|
| INBR | .87* | .00 | -.01 | .01 | .01 | -.01 | .00 | .00 |
| KERN | -.11 | .84* | .02 | .04 | -.06 | .01 | .00 | -.01 |
| GIFR | .08 | .77 | .75* | -.09 | .10 | -.09 | .01 | .05 |
| VERK | .52 | .35 | .45 | .53* | .01 | .04 | .03 | .00 |
| OORL | .23 | .59 | .61 | .45 | .54* | .07 | -.04 | .00 |
| GENM | -.11 | .69 | .63 | .27 | .48 | .57* | .05 | -.02 |
| TERR | .62 | .41 | .53 | .62 | .54 | .32 | .74* | ♢ |
| ROOF | .71 | .12 | .27 | .53 | .35 | .08 | ♢ | ♢ |

Tabel 5.24: Tabel bij opgave 26.

2. Welke variabele heeft de hoogste uniciteit en wat betekent dit?
3. Zou je nog een derde factor trekken? Motiveer je antwoord.
4. Geef een gemotiveerde interpretatie van de factoren en geef aan hoe je hierbij te werk gaat.

*Opgave 27*
Een arbeids- en organisatiepsycholoog doet onderzoek naar arbeidstevredenheid met behulp van een korte vragenlijst, bestaande uit de volgende items:

- Er wordt vaak druk op mij uitgeoefend. (DRUK)
- Ik vind mijn werk uitdagend. (UITDAGING)
- Ik kan mijn werk in alle rust doen. (RUST)
- Mijn werk deprimeert me. (GEDEPRIMEERD)

Elk item wordt beantwoord met een cijfer tussen 1 (volledig oneens) en 7 (volledig eens).

Er worden gegevens verzameld van 100 werknemers van een bedrijf. Op deze gegevens wordt een factoranalyse uitgevoerd waarin twee gemeenschappelijke factoren getrokken worden. Deze factoren zijn orthogonaal. De onderzoeker vindt twee tabellen met factorladingen in de computeruitvoer. Deze worden hier getoond in Tabel 5.25 en Tabel 5.26.

| | Factor 1 | Factor 2 |
|---|---|---|
| DRUK | -0.60 | -0.45 |
| UITDAGING | -0.45 | 0.60 |
| RUST | 0.60 | 0.45 |
| GEDEPRIMEERD | 0.45 | -0.60 |

Tabel 5.25: Tabel bij opgave 27.

1. Bereken de gereproduceerde correlatie tussen DRUK en RUST op basis van Tabel 5.25 en daarna op basis van Tabel 5.26.
2. Leg uit wat de relatie is tussen de factoroplossing in Tabel 5.25 en die in Tabel 5.26?
3. Interpreteer de factoren. Op welke tabel baseer je je antwoord en waarom? Wat is je interpretatie?
4. Een collega van de onderzoeker beweert dat men zijn keuze voor een factoroplossing niet alleen mag baseren op de inhoudelijke interpretatie maar dat men ook rekening moet houden met de uniciteiten. Wat vind jij van dit standpunt?

| | Factor 1 | Factor 2 |
|---|---|---|
| DRUK | 0.00 | 0.75 |
| UITDAGING | 0.75 | 0.00 |
| RUST | 0.00 | -0.75 |
| GEDEPRIMEERD | -0.75 | 0.00 |

Tabel 5.26: Tabel bij opgave 27.

*Opgave 28*
Een klinisch psychologe was geïnteresseerd in de mate waarin men de tevredenheid met de partnerrelatie kan voorspellen op basis van de verwachtingen die men heeft ten aanzien van het samenleven. Om deze verwachtingen te inventariseren, formuleerde zij 60 items en maakte daarmee een vragenlijst. Zij beschikte over (1) de antwoorden van een grote groep personen die deze vragenlijst invulden kort voordat zij effectief met hun partner gingen samenleven en (2) een beoordeling van hun tevredenheid met de partnerrelatie na effectief drie à vijf jaar samengeleefd te hebben. De psychologe wilde niet alleen een zo goed mogelijke voorspelling maken van de tevredenheid; zij wilde ook begrijpen waarom bepaalde predictoren een grotere bijdrage leveren aan de voorspelling dan andere. Om te voorkomen dat zij door de bomen het bos niet meer zag, besliste zij om eerst een factoranalyse uit te voeren op de items van de vragenlijst, en de factorscores te gebruiken als de predictoren in de regressieanalyse.

1. De psychologe wilde dat de factorscores de itemscores zo goed mogelijk voorspellen. Zou jij de psychologe hebben aangeraden om een hoofdcomponentenanalyse uit te voeren of een factoranalyse met ongecorreleerde residuen? Motiveer je antwoord.
2. Zou jij de psychologe hebben aangeraden om orthogonaal of obliek te roteren? Wat zijn de voor- en nadelen van deze twee opties?

*Opgave 29* COMPUTEROPDRACHT
Een sociaal psycholoog was geïnteresseerd in het verband tussen zelfbeoordelingen en beoordelingen door anderen. Wat betreft de zelfbeoordelingen, vroeg hij aan de proefpersonen om zichzelf te beschrijven aan de hand van 12 uitspraken. Op elk van deze uitspraken moest geantwoord worden met een getal tussen 0 en 9 (inclusief), waarbij 0 aangeeft dat de uitspraak helemaal niet van toepassing is op zichzelf, en 9 dat deze uitspraak volledig van toepassing is op zichzelf. Wat betreft de beoordelingen door anderen, maakte hij gebruik van video-opnames van de proefpersonen. Het betrof een video-opname van de uitvoering van een groepstaak. Elke proefpersoon werd beoordeeld aan de hand van drie uitspraken waarvan vijf beoordelaars dienden aan te geven in welke mate ze van toepassing waren op deze proefpersoon (eveneens met een getal tussen 0 en 9). De oordelen van deze vijf beoordelaars werden gecombineerd door er het gemiddelde van te nemen. Aan het onderzoek namen 500 proefpersonen deel. Alle gegevens zijn opgeslagen in het gegevensbestand "zelfandr.sav". Hieronder staan de 12 uitspraken, elk gevolgd door de variabelennaam waaronder het antwoord op de betreffende uitspraak in het genoemde bestand is opgeslagen.

1. Ik hou ervan om nieuwe mensen te leren kennen (NIME).
2. Ik neem meestal het initiatief bij het maken van nieuwe vrienden (INITIA).
3. Ik heb haast altijd een antwoord klaar wanneer mensen zich tot mij wenden (ANTW).
4. Ik hou ervan om dingen te doen waarbij ik snel moet handelen (SNELHAND).
5. Ik maak me vaak zorgen over zaken die ik niet had moeten doen of zeggen (VAZORG).
6. Ik ben vlug geïrriteerd (IRRITE).

7. Ik voel me vaak eenzaam (EENZAAM).
8. Ik ben soms erg energiek en op andere momenten dan weer erg sloom (ENERSLO).
9. Ik heb vijanden die mij proberen te kwetsen (VIJAND).
10. Het gebeurt vaak dat andere mensen hun relatie met mij verbreken zonder dat het mijn fout is (RELATIE).
11. Ik vind het fijn als anderen bevreesd zijn voor mij (VREES).
12. Andere mensen vertellen mij een hoop leugens (LEUGENS).

De drie uitspraken aan de hand waarvan de proefpersonen beoordeeld werden (wederom gevolgd door de variabelennaam), zijn de volgende:

1. Deze persoon is goed aangepast (AANPAS).
2. Het gedrag van deze persoon is goed voorspelbaar (VOORSPEL).
3. Deze persoon is aangenaam in de omgang (AANGENA).

De opdrachten

1. Voer een factoranalyse uit op de 15 variabelen in dit onderzoek. Gebruik het ML-criterium en stel het aantal te trekken factoren op 2. Vraag om de gereproduceerde en de residuele correlaties. Zou je, op basis van de resultaten, nog een derde factor trekken? Waarom?
2. Laat nu drie factoren trekken en vraag opnieuw naar de gereproduceerde en de residuele correlaties. Is deze nieuwe factoroplossing beter dan de vorige? Waarom of waarom niet?
3. Roteer de gevonden factoroplossing volgens de VARIMAX-methode. Vraag om de geroteerde factoroplossing en druk deze af. Geef een interpretatie van de geroteerde factorladingenmatrix. Je kunt hierbij steunen op het feit dat de 12 zelfbeoordelingsitems afkomstig zijn uit de *Eysenck Personality Inventory*, een vragenlijst die gebaseerd is op factoranalytisch onderzoek waaruit blijkt (volgens Eysenck en Eysenck, 1968) dat er drie ongecorreleerde persoonlijkheidsdimensies bestaan: extraversie/introversie, neuroticisme/emotionele stabiliteit en psychoticisme (de eerste twee zijn zogenaamde *bipolaire* factoren). De eerste vier items (NIME t/m SNELHAND) behoren tot de extraversie/introversieschaal, de volgende vier (VAZORG t/m ENERSLO) tot de neuroticisme/emotionele stabiliteitsschaal, en de laatste vier (VIJAND t/m LEUGENS) tot de psychoticisme schaal. Wat kunnen we nu leren van de positie van de drie beoordelingen door anderen in de geroteerde factoroplossing? Of, wat explicieter geformuleerd, op welke manier worden de beoordelingen door anderen bepaald door de drie persoonlijkheidsdimensies?

---

## 5.11 Appendix bij Hoofdstuk 5

### De Formule van de Covariantie tussen twee Sommen van Variabelen

We bekijken hier twee variabelen ($Y$ en $Z$) die elk een som zijn van twee andere variabelen ($Q$ en $R$, respectievelijk, $V$ en $W$).

$$\begin{aligned} Y &= Q + R \\ Z &= V + W \quad . \end{aligned}$$

De covariantie tussen $Y$ en $Z$ is als volgt gedefinieerd:

$$\mathrm{COV}(Y,Z) = \frac{1}{n}\sum_{p=1}^{n}[Y_p - \overline{Y}][Z_p - \overline{Z}]$$

Omdat $Y$ en $Z$ elk een som zijn van twee andere variabelen, kunnen we de formule voor $\mathrm{COV}(Y,Z)$ als volgt uitwerken:

$$\begin{aligned}
\mathrm{COV}(Y,Z) &= \frac{1}{n}\sum_{p=1}^{n}[Q_p + R_p - (\overline{Q} + \overline{R})][V_p + W_p - (\overline{V} + \overline{W})] \\
&= \frac{1}{n}\sum_{p=1}^{n}[(Q_p - \overline{Q}) + (R_p - \overline{R})][(V_p - \overline{V}) + (W_p - \overline{W})] \\
&= \frac{1}{n}\sum_{p=1}^{n}[Q_p - \overline{Q}][V_p - \overline{V}] + \frac{1}{n}\sum_{p=1}^{n}[Q_p - \overline{Q}][W_p - \overline{W}] + \\
&\quad \frac{1}{n}\sum_{p=1}^{n}[R_p - \overline{R}][V_p - \overline{V}] + \frac{1}{n}\sum_{p=1}^{n}[R_p - \overline{R}][W_p - \overline{W}] \\
&= \mathrm{COV}(Q,V) + \mathrm{COV}(Q,W) + \mathrm{COV}(R,V) + \mathrm{COV}(R,W) \qquad ,
\end{aligned}$$

de som van de covarianties tussen de samenstellende variabelen.

## De Formule van de Covariantie tussen Variabelen die met een Constante Vermenigvuldigd worden

We zijn geïnteresseerd in de covariantie tussen de variabelen $aY$ en $bZ$. Hierin zijn $Y$ en $Z$ variabelen en $a$ en $b$ constanten. We vertrekken van de definitie van de covariantie die we toepassen op $aY$ en $bZ$:

$$\mathrm{COV}(aY,bZ) = \frac{1}{n}\sum_{p=1}^{n}[aY_p - \overline{aY}][bZ_p - \overline{bZ}] \quad .$$

We maken nu gebruik van het feit dat het gemiddelde van een variabele die met een constante vermenigvuldigd wordt (hier, $aY$ en $bZ$) gelijk is aan die constante maal het gemiddelde van de variabele (hier, $a\overline{Y}$ en $b\overline{Z}$). We krijgen dan:

$$\begin{aligned}
\mathrm{COV}(aY,bZ) &= \frac{1}{n}\sum_{p=1}^{n}[aY_p - a\overline{Y}][bZ_p - b\overline{Z}] \\
&= (ab)\frac{1}{n}\sum_{p=1}^{n}[Y_p - \overline{Y}][Z_p - \overline{Z}] \\
&= ab\ \mathrm{COV}(Y,Z) \quad .
\end{aligned}$$

Dus, bij een covariantie tussen variabelen die met een constante vermenigvuldigd worden mag men de constanten buiten de haakjes brengen. In het bijzondere geval dat $Y$ gelijk is aan $Z$, is $\mathrm{COV}(Y,Z)$ gelijk aan $\mathrm{VAR}(Y)$.

# Literatuur


Aiken, L. S. & West, S. G. (1991). *Multiple regression: testing and interpreting interactions.* Newbury Park: Sage.

Anastasi, A. (1982). *Psychological testing.* New York: MacMillan.

Babbie, E. (1998). *The practice of social research.* Belmont: Wadsworth.

Bales, R. F. (1950). *Interaction process analysis; a method for the study of small groups.* Cambridge: Addison-Wesley.

Bales, R. F. (1952). A set of categories for the analysis of small group interaction. *American Sociological Review*, *15*, 257-263.

Barash, D. P. (1977). Human ethology: Exchanging cheetahs for chevrolets? *Environment and Behavior*, *9*, 487-490.

Barber, T. X. & Silver, M. J. (1968a). Fact, fiction, and the experimenter bias effect. *Psychological Bulletin*, *70*, 1-29.

Barber, T. X. & Silver, M. J. (1968b). Pitfalls in data analysis and interpretation: A reply to Rosenthal. *Psychological Bulletin*, *70*, 48-62.

Bowers, K. S. (1987). Revisioning the unconscious. *Canadian Psychology*, *28*, 93-104.

Brehm, S. S. (1981). Psychological reactance and the attractiviness of unobtainable objects: Sex differences in children's responses to an elimination of freedom. *Sex Roles*, *7*, 937-949.

Brinkman, J. (1994). *De vragenlijst.* Groningen: Wolters-Noordhoff.

Cepeda-Benito, A. & Short, P. (1998). Self-concealment, avoidence of psychological services, and perceived likelihood of seeking professional help. *Journal of Consulting Psychology*, 58-64.

Cronbach, L. J. & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions.* New York: Irvington.

Crusco, A. H. & Wetzel, C. G. (1984). The Midas touch: The effects of interpersonal touch on restaurant tipping. *Personality and Social Psychology Bulletin*, *10*, 512-517.

Darley, J. M. & Latané, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology*, *8*, 377-383.

Devlin, B., Fienberg, S. E., Resnick, D. P. & Roeder, K. (1995). Galton redux: Eugenetics, intelligence, race, and society: A review of "The bell curve: Intelligence and class structure in American lifes. *Journal of the American Statistical Association*, 1483-1488.

Eysenck, H. J. & Eysenck, S. B. (1968). A factorial study of psychoticism as a dimension of personality. *Multivariate Behavioral Research*, *spec. iss.*, 15-31.

Festinger, L., Riecken, H. & Schachter, S. (1956). *When prophecy fails.* Minneanapolis: University of Minnesota Press.

Griffin, J. H. (1960). *Black like me.* New York: The American Library.

Hartup, W. W. (1974). Agression in childhood: Development perspectives. *American Psychologist*, *29*, 336-341.

Herrnstein, R. J. & Murray, C. (1994). *The bell curve: Intelligence and class structure in American life.* New York: The Free Press.

Hite, S. (1987). *Woman and love: A cultural revolution in progress.* New York: Knopf.

Jones, L. V. (1984). White-black achievement differences: The narrowing gap. *American Psychologist*, *39*, 1207-1213.

Jöreskog, K. G. & Sörbom, D. (1979). *Advances in factor analysis and structural equation models.* Cambridge: Abt Books.

Latané, B., Williams, K. & Harkin, S. (1979). Many hands make light the work: The causes and consequences of social loafing. *Journal of Personality and Social Psychology*, *37*, 822-832.

Lewin, K. (1951). *Field theory in social science: Selected theoretical papers.* New York: Harpers.

Lorch, R. F. & Myers, J. L. (1990). Regression analyses of repeated measures data in cognitive research. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 149-157.

Manuck, S. B., Kaplan, J. R., Adams, M. R. & Clarckson, T. B. (1988). Studies of psychosocial influences on coronary artery atherogenesis in cynomolgus monkeys. *Health Psychology*, *7*, 113-124.

Maxwell, S. E. & Delaney, H. D. (1990). *Designing experiments and analyzing data: A model comparison perspective.* Pacific Grove: Brooks/Cole.

Milgram, S., Liberty, H. J., Toledo, R. & Wackenhut, J. (1986). Response to intrusion into waiting lines. *Journal of Personality and Social Psychology*, *51*, 683-689.

Miller, G. A. (1964). *Mathematics and psychology.* New York: Wiley.

Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, *17*, 776-783.

Parry, H. J. & Crossley, H. M. (1950). Validity of responses to survey questions. *Public Opinion Quarterly*, *14*, 61-80.

Piaget, J. (1952). *The child's conception of number.* (C. Gattegno & F. M. Hodgson, Vert.). London: Routledge & Kegan Paul. (Oorspronkelijk werk uitgegeven in 1941)

Pishkin, V. & Shurley, J. T. (1983). Electrophysiological parameters in anxiety and failure: Evaluation of doxepin and hydroxine. *Bulletin of the Psychonomic Society*, *21*, 21-23.

Poulton, E. C. (1973). Unwanted range effects from using within-subjects designs. *Psychological Bulletin*, *80*, 113-121.

Poulton, E. C. (1975). Range effects in experiments on people. *American Journal of Psychology*, *88*, 3-32.

Poulton, E. C. (1982). Influential companions. Effects of one strategy on another in the within-subjects designs of cognitive psychology. *Psychological Bulletin*, *91*, 673-690.

Roethlisberger, F. J. (1977). *The elusive phenomena: An autobiographical account of my work in the field of organized behavior at the Harvard Business School.* Cambridge: Division of Research, Graduate School of Business Administration (distributed by Harvard University Press).

Rosenhan, D. L. (1973). On being sane in insane places. *Science*, *179*, 250-258.

Rosenthal, R. & Fode, K. (1963). The effects of experimenter bias on the performance of the albino rat. *Behavioral Science*, *8*, 183-189.

Scheffé, H. (1959). *The analysis of variance.* New York: Wiley.

Sharma, S. & Moskowitz, H. (1972). Effect of marihuana on the visual autokinetic phenomenon. *Perceptual and Motor Skills*, *35*, 891-894.

Shaughnessy, J. J., Zechmeister, E. B. & Zechmeister, J. S. (2000). *Research methods in psychology.* New York: McGraw-Hill.

Stalmeier, P. F. M., Unic, I. J., Verhoef, L. C. G. & van Daal, W. A. J. (1999). Evaluation of a shared decision making program for women suspected to have a genetic predisposition to breast cancer: Prelininary results. *Medical Decision Making*, 230-241.

Taylor, H. C. & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection. Discussion and tables. *Journal of Applied Psychology*, *23*, 565-578.

Taylor, R. M. (1994). "The bell curve". *Contemporary Sociology*, *24*, 153-158.

Williams, K., Harkins, S. & Latané, B. (1981). Identifiability as a deterrent to social loafing: Two cheering experiments. *Journal of Personality and Social Psychology*, *40*, 303-311.

# Index


Afhankelijke variabele, *zie* Variabelen
Alternatieve-vormmethode, 84

Base-rate, 114
Beoordelingsschaal, 41
Betrouwbaarheid, 70–101
Binnen-subjectenexperiment, 27–30

Causaal verband, 20, 132
Codering, 41–42
Collineariteit, 350–353
  definitie, 351
  instabiele bétagewichten, 352
  instabiele bètagewichten, 351
  voorbeeld en omschrijving, 350–351
Communaliteit, 477–478
Condities versus groepen, 132
Constructvaliditeit, 105–109
Contrabalancering, 28–30
Convergente validiteit, 105
Correctie voor attenuatie, 98–100, 111–112
Covariantieanalyse, 135, 254–277
  ANOVA-tabel, 273–274
  effectkwadratensommen, 264–269
    kwadratensom van de covariaat, 268–269
    kwadratensom van de factor uitgezuiverd voor de covariaat, 265–266
    kwadratensom van de factor zonder uitzuivering voor de covariaat, 268
  errorkwadratensom, 269–270
  F-toetsen, 270–271
  nulhypothese, 262–263
  overzichtstabellen, 271
  relativering van de kracht van statistische uitzuivering, 277
  twee redenen om uit te zuiveren voor een covariaat
    1. uitzuiveren voor een storende variabele, 274–275
    2. gevoeliger toetsen, 275–276
  uitgezuiverde gemiddelden, 260–262
  uitgezuiverde verschil, 256–258, 260–262
  wanneer stel je de covariaatscores vast?, 276–277
Criteriumvaliditeit, 109–113
Criteriumvaliditeit en beslissingen, 113–117
Cronbachs alpha, 86–91

Designs met een voor- en een nameting, 135, 287–292
  design met één groep, 287–288
    causale interpretatie, 288
    voorbeeld en notatie, 287–288
  design met twee groepen
    causale interpretatie, 289–290
    evaluatie vanuit pragmatisch oogpunt, 291–292
    voorbeeld en notatie, 289
Diepte-interview, 50
Differentiële transfer, 28–29
Discriminantanalyse, 111
Divergente validiteit, 106
Dubbel-blind experiment, 47
Dummycodering, 400–406
  ∼ voor regressieanalyse met een categorische predictor, 401–403
  ∼ voor regressieanalyse met groepsspecifieke regressiegewichten, 403–406
Duurzaam effect, 28

Effectonderzoek, 20–30, 131

Effectparameters, 146–148
Eigenwaarde, 478–479
Enkelvoudige regressieanalyse, 308–318
∼ voor gestandaardiseerde variabelen, 308–310
gemiddelde kwadratische fout, 309–310
∼ voor ongestandaardiseerde variabelen, 310–311
gemiddelde kwadratische fout, 310–311
proporties verklaarde en onverklaarde variantie, 312
standaardschattingsfout van de regressie, 313
verklaarde en onverklaarde variantie, 311–312
voorbeeld, 314–318
zuivere schatter van de gemiddelde kwadratische fout, 313–314
Experimentele controle van de condities, 25–27, 131, 135–136

F-change statistiek
∼ in hiërarchische meervoudige regressieanalyse, 377–378
∼ in regressieanalyse met categorische predictoren, 385–386, 388–389
∼ in regressieanalyse met groepsspecifieke regressiegewichten, 394–395
overzicht van de verschillende types F-toetsen binnen het algemene lineaire model, 396–400
F-statistiek
in regressieanalyse, *zie* F-change statistiek
in variantieanalyse, *zie* F-toets(en)
$\mathcal{F}$-verdeling
in regressieanalyse, *zie* F-change statistiek
in variantieanalyse, *zie* F-toets(en)
Factoranalyse, 463–525
∼ met gecorreleerde residuen, *zie* Hoofdcomponentenanalyse, 465–476
bepalen van de factorscores, 514–515
binnen het onderzoek naar construct-validiteit, 107
keuze van het aantal gemeenschappelijke factoren, 510–514
inspectie van de eigenwaarden, 511–512
inspectie van de residuele correlaties, 511
interpreteerbaarheid, 512–513
statistische toets, 513–514
model met meer dan twee gemeenschappelijke factoren, 508–510
onbepaaldheid van een factoroplossing, 493–497
belang van de onbepaaldheid, 496–497
illustratie en omschrijving, 493–496
roteren naar interpreteerbare factoren, 497–504
oblieke rotatie, 499–504
richtlijnen bij de interpretatie, 498–499
Factoranalyse met ongecorreleerde residuen, 300, 465–476
model met één gemeenschappelijke factor, 465–471
implicaties voor de correlaties, 467–469
model, 465–467
schattingen van de factorladingen, 469–471
model met twee gemeenschappelijke factoren, 471–476
implicaties voor de correlaties, 473–474
model, 472–473
schattingen van de factorladingen, 474–476
Factorladingen
interpretatie van ∼ (bij oblieke factoren), 500–504
interpretatie van ∼ (bij orthogonale factoren), 476–477

Gelijktijdige validiteit, 110
Generaliseerbaarheidstheorie, 74
Gerandomiseerd experiment, *zie* Randomisering

Gereproduceerde correlaties, 469, 481
Gestandaardiseerd interview, *zie* Vragenlijst
Gestandaardiseerde observatie, 36–40
Gestructureerde observatie, 36–37
Gewogen gemiddelden, 151, 167–170, 172–175

Halvering, 85–86
Hertesting, 84–85
Heywoodgeval, 478
Hiërarchische meervoudige regressieanalyse (HMRA), 374–380
  ∼ binnen het onderzoek naar de rol van mediërende variabelen, 378–380
  ∼ binnen het onderzoek naar de rol van storende variabelen, 374–378
Hoofdcomponentenanalyse, 300, 486–493
  implicaties voor de correlaties, 487–488
  model, 487
  schattingen van de factorladingen, 488–490
  vergelijking met het factormodel met ongecorreleerde residuen, 490–493

Inhoudsvaliditeit, 104–105
Interactie
  ∼ tussen een kwantitatieve en een categorische variabele, 392–396
  ∼ tussen twee categorische variabelen (factoren), 210–212
  ∼ tussen twee kwantitatieve variabelen, 413–425
Interbeoordelaarsovereenstemming, 43
Interceptvariabele, 401
  groepsspecifieke interceptvariabelen, 401
Itemanalyse, 93–95

Klassieke testtheorie (basisformule), 74–79
Kleinste-kwadratencriterium, *zie* Kleinste-kwadratenschatters
Kleinste-kwadratenschatters
  ∼ in regressieanalyse, 304–305, 308–310, 321, 323–324
  ∼ in variantieanalyse, 149–151, 194, 196, 227–228, 292–296

Laboratoriumexperiment, 38–40, 131
Leeftijdsdifferentiatie, 108
Likelihood-ratio (LR) statistiek (in logistische regressieanalyse), 442–443
Lineaire regressie, 301
Logistische regressieanalyse, 426–447
  berekening van de logistische regressiecoëfficiënten, 430–431
  eigenschappen van de logistische regressiecoëfficiënten, 431–432
  evalueren van de effectgrootte, 435–440
    ∼ voor categorische predictoren, 436–438
    ∼ voor kwantitatieve predictoren, 438–440
  kwaliteit van de voorspelling op basis van een ∼, 432–435
  statistische toetsen, 440–443
    likelihood-ratio (LR) statistiek, 442–443
    scorestatistiek, 443
    Waldstatistiek, 441–442

Matching, 23, 133
Meervoudige correlatiecoëfficiënt
  definitie en gebruik, *zie* Meervoudige regressieanalyse (twee predictoren en willekeurig veel predictoren)
  gecorrigeerde ∼, 332–333
Meervoudige regressieanalyse (twee predictoren), 320–330
  ∼ voor gestandaardiseerde variabelen, 321–323
  ∼ voor ongestandaardiseerde variabelen, 323–324
  meervoudige correlatiecoëfficiënt, 324–327
  standaardschattingsfout van de regressie, 326–327
  verklaarde en onverklaarde variantie, 325–326
  voorbeeld, 327–330

zuivere schatter van de gemiddelde kwadratische fout, 327
Meervoudige regressieanalyse (willekeurig veel predictoren), 330–335
standaardschattingsfout van de regressie, 332
zuivere schatter van de gemiddelde kwadratische fout, 332
theorie, 330–333
voorbeeld, 333–335
ML-criterium
in factoranalyse, 470
in logistische regressieanalyse, 431
Moderatorvariabele, 393–394
Momentaan effect, 29
Multicollineariteit, *zie* Collineariteit
Multitrait-multimethod, 107–108

Niet-specifieke factoren, 46

Observatiemethoden, 33–48
Observatieverslag, 40–41
Observationeel onderzoek, 21, 134
Onafhankelijke variabele, *zie* Variabelen
Onechte correlatie, 343–346
definitie, 344–346
voorbeeld en omschrijving, 343–344
Ongebalanceerde designs, 161–164
Ongewogen gemiddelden, 166–170, 172, 174–175
Open vraag, 51
Operationele definities, 66–68

Parallelle tests, 80–81
Partcorrelatiecoëfficiënt, 335–339
theorie, 335–338
voorbeeld, 338–339
Participerende observatie, 35–36
Placebo-effect, 45–47
Placeboconditie, 46
Populatie, 52
Predictieve validiteit, 109
Predictor, 300
Psychologische constructen, 65–66
Psychologische test, 57–58

Quasi-experiment, 135

Randomisering, 22, 24, 29, 131, 133
Reactiviteit van de proefpersoon, 44–45
Regressieanalyse, 299–462
Regressieanalyse met één predictor, *zie* Enkelvoudige regressieanalyse
Regressieanalyse met twee predictoren, *zie* Meervoudige regressieanalyse (twee predictoren)
Regressieanalyse met willekeurig veel predictoren, *zie* Meervoudige regressieanalyse (willekeurig veel predictoren)
Regressieanalyse met categorische predictoren, 380–409
algemene lineaire model, 390–392
groepsspecifieke regressiegewichten, 392–396
regressievergelijking, 382–384
statistische significantie, 384–390
∼ van een categorische predictor, 385–388
∼ van een kwantitatieve predictor, 388–390
voorbeeld, 381–382
Regressieanalyse met interactie tussen twee kwantitatieve predictoren, 413–425
belang van ∼, 424–425
inhoudelijke interpretatie, 421–423
regressievergelijking met een interactievariabele, 417–421
regressievergelijking met niet-gecentreerde predictoren, 423–424
voorbeeld, 413–417
Regressiegewichten
omschrijving en functie, *zie* Enkelvoudige regressieanalyse, Meervoudige regressieanalyse
standaardfout van ∼, 354
statistische significantie van ∼, 353–357
Representatieve steekproef, 52–53
Residuele correlaties, 469
Respons, 55–56
Responsbias, 55–56
Restriction-of-range

invloed op de betrouwbaarheid, 95–97
invloed op de validiteitscoëfficiënt, 112–113

Scorestatistiek, 443
Selecteren van predictoren, 365–367
automatische selectie van predictoren, 366–367
achterwaartse methode, 366
stapsgewijze methode, 366
voorwaartse methode, 366
Selectiebias, 53–55
Selectienorm, 113
Selectieratio, 115
Selectiewinst, 115
Semi-partiële correlatiecoëfficiënt, *zie* Partcorrelatiecoëfficiënt
Slaagnorm, 113
Spearman-Brown formule, 82
Standaardfout van meting, 98
Standaardisering, 33
Statistische controle, *zie* Statistische uitzuivering
Statistische uitzuivering, 27, 134
∼ voor een categorische variabele, *zie* Variantieanalyse met twee factoren
∼ voor een kwantitatieve variabele, *zie* Covariantieanalyse
Storende variabelen
categorische ∼, 163–164, 169–170
kwantitatieve ∼, 255–256
storende persoonsvariabelen, 22–24, 132, 134, 163
storende situatievariabelen, 24–27, 135, 163–164
Succesratio, 115
Suppressie, 346–350
definitie, 348–350
voorbeeld en omschrijving, 346–348

T-statistiek
in regressieanalyse , *zie* statistische significantie van regressiegewichten
$\mathcal{T}$-verdeling
in regressieanalyse , *zie* statistische significantie van regressiegewichten
Taylor-Russell-tabellen, 115–116
Testverlenging en betrouwbaarheid, 82–84
Transfer, 28

ULS-criterium (in factoranalyse), 470

Validiteit, 104–117
Validiteitscoëfficiënt, 111
Variabelen, 17–20
afhankelijke ∼, 20
categorische ∼, 18, 134, 138, 380
experimentele ∼, 20–22
kwantitatieve ∼, 135, 138, 380
natuurlijke ∼, 20–22, 133
onafhankelijke ∼, 20
persoons∼, 18–20
situatie∼, 18–20
storende ∼, *zie* Storende variabelen
Variantieanalyse met één factor, 137–158
ANOVA-tabel, 156
effectkwadratensom, 148–152
errorkwadratensom, 152–153
F-statistiek
berekening en interpretatie, 155–157
F-toets, 143–145
kwadratensom-ontbinding, 153–154
notatie voor de gegevens, 139
proportie verklaarde variantie, 157–158
rekenvoorbeeld, 140–143
totale kwadratensom, 153–155
variantieanalysemodel, 145–148
Variantieanalyse met meer dan twee factoren, 242
Variantieanalyse met twee factoren, 161–244
designs met lege cellen, 241
met interactie-effect, 210–244
ANOVA-tabel, 243
errorkwadratensom, 229–230
F-toetsen, 231–234
kwadratensom van het interactie-effect, 227–228
kwadratensom-ontbinding, 230
kwadratensommen van de hoofdeffecten, 228–229
model, 225–226

overzichtstabellen, 235
proportie verklaarde variantie, 234–235
rekenvoorbeeld, 213–223
notatie voor de gegevens, 170–175
nut van toevoegen van factoren, 239–240
relativering van de kracht van statistische uitzuivering, 240
zonder interactie-effect, 175–204
F-toetsen, 200–204
rekenvoorbeeld, 179–191
effectkwadratensommen, 193–197
errorkwadratensommen, 197–198
F-toetsen, 175–179
kwadratensom-ontbindingen, 198–200
overzichtstabellen, 204
totale kwadratensom, 198–200
Variantieanalyse voor ongebalanceerde designs, 134
VARIMAX-rotatie, 510
Veldexperiment, 37–38, 131
Verwachte waarde, 143–144
Verwachtingen van de onderzoeker, 47–48
Verwachtingen van de proefpersoon, 45–47
Voorspellen zonder predictoren, 304–307
gemiddelde kwadratische fout, 305–307
kleinste-kwadratenschatting, 304–305
voorbeeld, 307
Vraagkenmerken, 45
Vragenlijst, 50–57
Vrij (niet-gestandaardiseerd) interview, 50
Vrije (niet-gestandaardiseerde) observatie, 34–35
Vrijheidsgraden
in regressieanalyse, *zie* F-change statistiek, T-statistiek
in variantieanalyse, *zie* F-toets(en)

Waaierhypothese, 290
Waldstatistiek
multivariate ∼, 441–442
univariate ∼, 441

Zuiver toevallige steekproeftrekking, 53
Zuiver toevallige toewijzing, *zie* Randomisering