

# College 1: Field Hoofdstuk 6

## Hoofdstuk 6 Semi-partiele correlatie

Semi partiele correlatie is belangrijk voor regressie analyse. Semi partiele correlatie is de correlatie met een andere variabele, waaruit de correlatie met een andere variabele is weggehaald. Een variabele voegt weinig verklaarde variantie toe wanneer deze hoog correleert met een al gebruikte predictor. In een regressie model is:

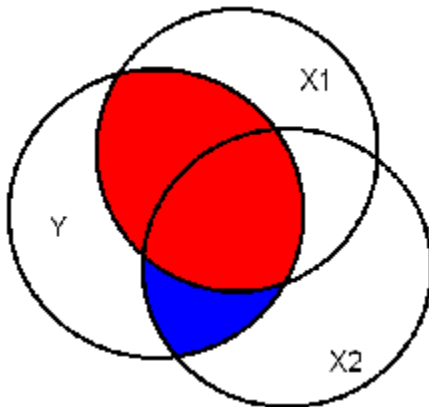
Y=wat je wilt voorspellen (notatie: Y')

X=predictor om Y te voorspellen (X<sub>1</sub>,X<sub>2</sub> etc.)

De beste predictor is een predictor die het meeste verklaring kan toevoegen aan de verklaarde variantie. Dus niet de X die het hoogst correleert. De verklaarde variantie is de hoeveelheid variantie die X kan verklaren. De semi partiele correlatie kun je berekenen door:

$$r_{y12} = \frac{r_{y2} - r_{y1} \cdot r_{12}}{\sqrt{1 - r_{12}^2}}$$

De uitkomst hiervan moet je kwadrateren om de verklaarde variantie te krijgen. Een uitkomst van bijvoorbeeld 0,0358 betekent dat het toevoegen van een tweede predictor 3,6% extra variantie van Y kan verklaren. De totale verklaarde variantie bepaal je door de verklaarde variantie van predictor 1 en predictor 2 bij elkaar op te tellen.



In de figuur is te zien dat X1 een groot gedeelte van de variantie van Y kan verklaren (rood). Toevoegen van X2 zorgt ervoor dat ook het blauwe gedeelte verklaard kan worden. Samen verklaren ze meer dan de helft van de variantie in Y.

## College 2: Field Hoofdstuk 7 Deel 1

Door middel van Regressie kun je met X voorspellingen doen over Y, gebruik makend van de samenhang van X en Y. hoe hoger de samenhang, hoe beter de voorspelling. Multiple regressie houdt in dat er meerdere predictoren gebruikt worden. Een voorbeeld van een regressielijn is

$$Y' = 0,8x_1 + 0,5x_2 + 0,6$$

Met x<sub>1</sub>,x<sub>2</sub> is predictoren, het cijfer ervoor geeft de regressiecoefficient aan, ook wel slope genoemd. Deze geeft namelijk aan hoe stijl de lijn is. Het laatste cijfer is de intercept. Dit geeft de beginwaarde weer. De formule voor de regressielijn is gebaseerd op de methode van de 'least squares' dit houdt in dat er zo min mogelijk voorspellingsfouten mogen zijn. Door deze formule krijg je de lijn met de minste voorspellingsfouten:

$$Y_j' = \bar{Y} + r_{XY} \frac{S_Y}{S_X} (X_j - \bar{X})$$

De lijn die ontstaat uit deze formule is:

$$Y = b_0 + b_1X_i + \epsilon_i$$

Met:

b<sub>1</sub>=regressiecoefficient voor de predictor, gradient (slope) voor de regressielijn

b<sub>0</sub>=intercept, waarde van Y als X=0. Punt waar de regressielijn kruist met de Y-as

Om te bekijken of een model goed is wordt gekeken naar de voorspelling die het model kan doen. Wanneer deze voorspelling beter is dan de voorspelling op basis van het gemiddelde van de data is het een goed model. Er wordt hierbij gekeken naar deviatie score: dit is de afwijking die de ware score heeft van het gemiddelde. Wanneer nu een voorspellingsmodel wordt toegevoegd als informatie wordt de deviatie kleiner en de voorspelling beter. Die verbetering ( $Y'-Y$ ) is het verschil tussen ware score en residu. Hoe groter het residu (voorspellingsfout) hoe slechter het model.

Wanneer je de bovenstaande formule toepast krijg je automatisch het beste model voor die lijn, maar om te checken of het model ook beter is dan het gemiddelde zou voorspellen kun je gebruik maken van:

1. Een betrouwbaarheidsinterval voor de voorspelling maken
2. Toetsen of het model goed is met ANOVA

## College 3:

### Field hoofdstuk 7 deel 2.

Met ANOVA (variantie analyse) kan worden nagegaan welk deel van de variantie te verklaren is door het model en welk deel daardoor niet te verklaren is. Daartoe worden de afwijkingen van  $Y$  met de waarnemingen van  $Y$  gemiddeld opgesplitst:

$$\begin{aligned} \text{Totaal: } (Y_j - \bar{Y}) &\hat{=} \sum (Y_j - \bar{Y})^2 : SS_{total} \\ \text{Model: } (Y'_j - \bar{Y}) &\hat{=} \sum (Y'_j - \bar{Y})^2 : SS_{model} \\ \text{Residu: } (Y_j - Y') &\hat{=} \sum (Y_j - Y')^2 : SS_{residual} \end{aligned}$$

Als de voorspelling van het model beter is dan de voorspelling van het gemiddelde wordt verwacht dat  $SS_m$  veel groter is dan  $SS_r$

Om te bepalen of een model goed is moeten de sums of squares worden omgezet in mean squares: MS. Sums of squares zijn totale waarden die je als een gemiddelde weer kunt geven door te delen door df.

$$\begin{aligned} MS_{residual} &= \frac{SS_{residual}}{n - \text{aantal predictoren} - 1} \\ MS_{model} &= \frac{SS_{model}}{\text{aantal predictoren}} \\ MS_{total} &= \frac{SS_{total}}{n - 1} \end{aligned}$$

De toetsingsgrootte  $F$  wordt berekend door de  $MS_m$  te delen door  $MS_r$  is de verhouding van wat door het model verklaard kan worden en wat niet door het model verklaard kan worden.

$$F = \frac{MS_M}{MS_R}$$

Wanneer de  $F$ -waarde gelijk is aan 1 is het zo dat het signaal gelijk is aan de ruis, de voorspellingsfout is dan gelijk aan het model. Dit is niet goed, een model is bevredigend wanneer  $F > 1$ . De 0-hypothese die altijd wordt gesteld bij het testen van een regressiemodel is: Het model dat is ontworpen doet het niet beter dan voorspellen op basis van het gemiddelde.

$R^2$ =hoeveelheid variantie die verklaard kan worden door het model.  $R$  is het symbool voor multiple correlatie. De correlatie van alle predictoren met  $Y$ . Als een parameter voldoende

van 0 afwijkt, dan is er een werkelijke bijdrage van de predictor aan de voorspelling. De intercept hoeft nooit getoetst te worden, dit is een constante.  $b_1$  moet wel getoetst worden  $\hat{a}$  regressie coefficient, die wordt vermenigvuldigd met X dus heeft veel invloed. Je kunt  $b_1$  toetsen door de gevonden waarde te vergelijken met de verwachte waarde onder de 0-hypothese.

$$t = \frac{b_{observed} - b_{expected}}{SEb} = \frac{b_{observed}}{SEb}$$

Bij meerdere predictoren zijn ook meerdere parameters nodig. De situatie wordt complexer. Nu moet ook gekeken worden naar de kwaliteit van de predictoren (welke is het best? Welke voorspelt het beste?). Ook moet gekeken worden naar de correlatie tussen de predictoren. Een te hoge correlatie veroorzaakt multicolineariteit. Het is dan niet zinvol om twee predictoren te gebruiken. Multicollineariteit is wanneer twee predictoren te hoog met elkaar correleren, dit wordt later nog verder uitgelegd.

$R^2$  (R square) = multiple correlatie tussen  $X_1$ ,  $X_2$  etc. en Y. een hoge  $R^2$  is een goed model. Ook de ANOVA is weer te berekenen. De fout die wordt gemaakt zonder toepassing van het model (total) is te splitsen in het verschil tussen de voorspelling en de meting (-residu) en het verschil tussen de voorspelling met gebruik van het model en zonder het model (-model).

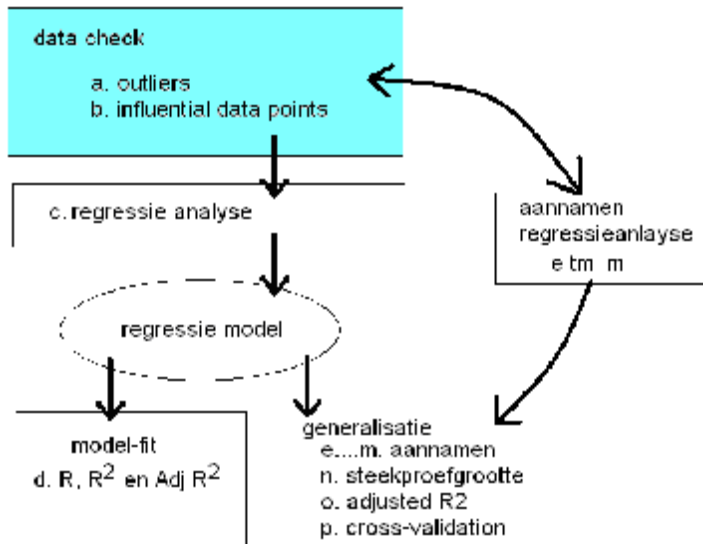
In de output van SPSS staat een B en er staat Beta. B staat op de oorspronkelijke meetschaal, de Beta is gestandaardiseerd en dus te interpreteren als een Z-score. Voorbeeld van SPSS output:

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-.490	2,735		-.179	,874
	X1	,796	,580	,548	1,372	,304
	X2	,816	,731	,446	1,117	,380

a. Dependent Variable: Y

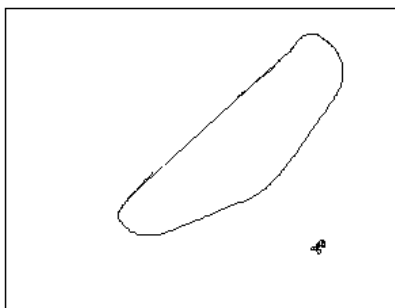
#### College 4: Field hoofdstuk 7, Deel 3

Een data check wordt uitgevoerd om te bekijken of er goede data is verzameld. Er kan namelijk sprake zijn van outliers of influential data points. Er is een schema dat hierop aansluit en wat gebruikt wordt bij de toetsing van een model.

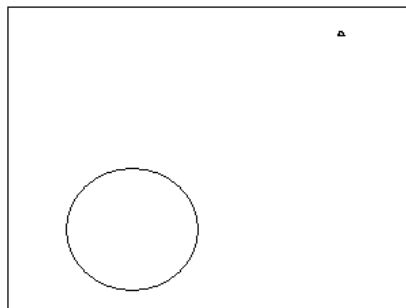


Aan de hand van dit schema wordt het bekijken van een model uitgelegd.

Een outlier en een influential data point zijn punten die buiten de normale range van getallen vallen. Ze hebben consequenties voor de hoogte van de correlatie en dus voor de kwaliteit van het regressie model. Het is daarom belangrijk dat de outliers worden opgespoord. Of deze verwijderd mogen worden hangt af van de situatie. Een outlier heeft een groot residu, een influential data point een klein residu. Een outlier verandert de regressielijn, een influential data point niet of nauwelijks. Een influential data point heeft wel invloed op de hoogte van de correlatie. Bij een outlier is er meestal iets misgegaan, bij een influential data point niet.



1. Outlier



2. Influential data point

Bij outliers is er sprake van:

1. Standardized residual: te interpreteren als een standaardwaarde - Z-score, bij een score  $>3$  en  $<-3$  is het residu te groot.
2. Studentized residual wordt berekend met de standaardafwijking, geschat voor de specifieke case. Dit zegt iets meer over de individuele case, dezelfde waarden worden gehanteerd.

Bij Influential data points is er sprake van:

1. Studentized deleted residual:  $sdr_{-}$  gestandaardiseerd. Verschil van meting met voorspelling zonder case ( $<-3$  of  $>3$ ). Wat is het verschil tussen de voorspelling met case en de voorspelling zonder case?
2. Cook's distance: maat voor overall influence op  $r$ ,  $b_1$ ,  $b_0$  enz. ( $>1$ )
3. Leverage ( $>3 (k=1)/n$ )
4. Mahalanobis distance ( $>11$ )

Het uitzoeken en elimineren van variabelen kan op verschillende manieren:

- Hierarchische methode (blockwise entry)
- Forced entry (SPSS) – alle data worden meegenomen – ENTER
- Stepwise Method (SPSS) : A Forward – de volgende predictor die wordt ingevoerd is de predictor die het meest verklaarde variantie kan toevoegen, aan wat is verklaard door eerdere predictoren.

B Backward – Eerst worden alle predictoren ingevoerd en daarna worden de predictoren die het minst extra kunnen verklaren verwijderd.

C Stepwise – Net als forward, maar na elke opname wordt gekeken of er misschien niet een predictor overbodig is geworden, zo ja, dan wordt deze verwijderd.

A, B, en C werken hier met semi-partiele correlatie

Model-fit wordt weergegeven in R, R<sup>2</sup> of adjusted R<sup>2</sup>. Adj R<sup>2</sup> = R<sup>2</sup> aangepast aan het aantal waarnemingen en predictoren, deze geeft aan in hoeverre je kunt generaliseren naar de populatie of naar een andere steekproef.

$$adj R^2 = 1 - \left[ \left( \frac{n-1}{n-k-1} \right) \left( \frac{n-2}{n-k-2} \right) \left( \frac{n+1}{n} \right) \right] (1 - R^2)$$

De shrinkage is R<sup>2</sup> – adj R<sup>2</sup> en geeft aan in hoeverre je kunt generaliseren. Bij een klein verschil kan er gegeneraliseerd worden naar de populatie of een andere steekproef. Adj R<sup>2</sup> geeft namelijk aan

R<sup>2</sup> je kunt in de populatie (of R<sup>2</sup> = 0 kan niet! Dan heel erg slecht goed model heeft waarnemingen en predictoren. Voorbeeld uit SPSS.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,905(a)	,818	,802	,78174

a Predictors: (Constant), x2, x1

hoeveel van verwachten steekproef). is het een model. Een veel weinig

Om te bepalen of een model goed is, moet er ook worden gekeken of er voldaan is aan de aannames die zijn gesteld. De aannames zijn:

E = Variabele type

F = Non-zero variance

G = Multicollineariteit

H = Geen correlatie met andere externe variabelen

I = Homoscedasticiteit

J = Errors onafhankelijk

K = Residuen normaal verdeeld

L = Uitkomstvariabele onafhankelijk voor elke waarneming

M = lineaire samenhang

E). Variabele type: de variabele mag niet norminaal zijn, maar hoort continu, numeriek, dichotoom of categoriaal te zijn. De uitkomstvariabele (criterium) moet continu zijn en niet aan een range gebonden.

F). Non-zero variance: de variantie moet groter zijn dan 0, er moet variantie zijn.

G). Multicollineariteit: er mag geen sprake zijn van een perfecte multicollineariteit, de predictoren mogen niet te hoog met elkaar correleren. Dit geeft namelijk statistische problemen, de uitkomsten berusten meer op toeval en zijn daarmee minder stabiel en er zijn dan 1 of meer predictoren die nauwelijks of geen extra variantie verklaren, dus overbodig zijn. Er zijn twee manieren om dit op te lossen:

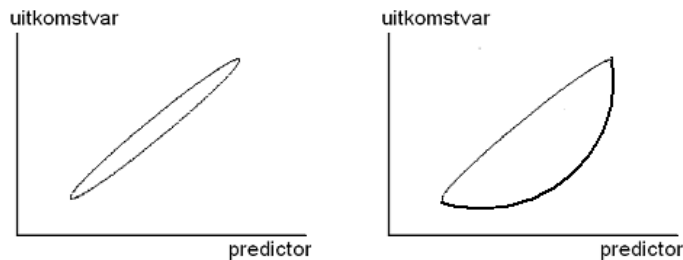
1. als je multicollineariteit vindt kun je een variabele laten vallen, want deze is overbodig. De variantie blijft vrijwel hetzelfde.

2. principale componenten analyse (PCA): set variabelen beschouwen en kijken welke er bij elkaar horen. Je kunt dan groepjes variabelen maken die samen hoog correleren en dus hetzelfde deel van de variantie verklaren. Van deze groepjes kunnen nieuwe variabelen worden gemaakt. Je hebt zo minder predictoren (minder toeval) en de variabelen correleren

onderling niet meer met elkaar. Je kunt multicollineariteit toetsen door de Tolerance methode of VIF (komen later aan bod).

H). Externe variabelen: de predictoren moeten ongecorrleerd zijn met externe variabelen (denk aan het third variable effect). Het is belangrijk dat bij een regressiemodel gebruik wordt gemaakt van een sluitende redenatie.

I). Homoscedasticiteit: hierbij gaat het om voorspellingsfouten (residuen). Wanneer deze overall ongeveer gelijk zijn is er sprake van homoscedasticiteit. Er moet sprake zijn van homoscedasticiteit: op elk niveau van de predictor moet de variantie van de residuen gelijk zijn.



Links: Homoscedastisch

Rechts: Heteroscedastisch

J). onafhankelijke errors: De errors (residuen, meetfouten) moeten onafhankelijk zijn voor elke twee observaties. Dit kan getoetst worden in SPSS met de toets van Durbin-Watson. Een waarde van 2 geeft perfecte onafhankelijkheid aan, kleiner dan 2 duidt op een positieve correlatie en groter dan 2 op een negatieve correlatie.

K). Normaal verdeelde errors: De residuen moeten normaal verdeeld zijn. Bij een normale verdeling liggen de errors rond de 0-lijn en deze ligt ook daadwerkelijk bij 0.

L). Onafhankelijke uitkomst variabele: Iedereen moet de test afzonderlijk invullen. Het criterium of de uitkomstvariabele is voor elke paar waarnemingen onafhankelijk.

M). Er is bij regressie sprake van een lineaire samenhang.

Bij het testen van een model moet ook worden gekeken naar steekproefgrootte, belangrijk is:

1. 10 tot 15 waarnemingen per predictor ongeveer
2. Testen overall model:  $N > 50 + 8k$
3. Testen individuele parameters:  $N > 104 + k$
4. Grafieken op blz 223 zijn nauwkeurig hierin

Onder shrinkage wordt verstaan het verschil tussen  $R^2$  en  $\text{adj } R^2$ . Er wordt onderscheid gemaakt tussen de twee verschillende methoden om dit te berekenen.

- In SPSS wordt gebruik gemaakt van de formule van Werry. Deze geeft aan hoe  $R^2$  verwacht wordt in de populatie
- De formule van Stein is met de hand te berekenen en geeft aan in hoeverre  $R^2$  verwacht wordt in een andere populatie.

Cross-validation is hoe goed een model het doet in een andere steekproef. Dit is te bekijken door data-splitting. De data wordt dan een twee helften gesplitst (A en B). Er wordt eerst gekeken naar helft A en hier wordt een regressie model mee gemaakt. De waarden van de data van B worden nu ingevuld en er kan nu gekeken worden in hoeverre dit klopt. Wanneer dit klopt, kunnen de data van A en B weer samengevoegd worden. Dit kan uiteraard ook omgekeerd toegepast worden.

Het vergelijken van modellen kan door te kijken naar de change van  $R^2$ . Is er een aanzienlijke verbetering van  $R^2$  als er een nieuwe variabele is toegevoegd? De  $R^2$  change

kan worden omgezet in een F-waarde, om te toetsen of het nieuwe model significant beter is. Bij een  $p < 0,05$  wordt de nieuwe variabele in het model opgenomen.

Multicollineariteit kun je toetsen door

- Tolerance:  $1 - R_j^2$  (multiple correlatie van een predictor met andere predictoren). Een tolerance lager dan 0,10 is slecht.
- VIF is ongeveer hetzelfde maar wordt berekend door:  $\frac{1}{\text{Tolerance}}$  bij een waarde  $> 10$  is het slecht voor het model.

## College 5

### Hoofdstuk 9 T-Test: Comparing two means

Meetschaal: Bij de analyses over de verschillen tussen de gemiddelden is de afhankelijke variabele gemeten op metrisch niveau (interval +ratio) en de onafhankelijke variabele op nominaal niveau.

Variatie in data kan zowel systematisch als onsystematisch zijn. Systematische variatie is dat wat bedoeld wordt door de onderzoeker en wordt veroorzaakt door de manipulatie. Onsystematische variatie komt door individuele verschillen of meetfouten. Wanneer groepen op voorhand al van elkaar verschillen is er bijvoorbeeld sprake van onsystematische variatie.

Bij een **dependent t-test** wordt één groep twee keer gemeten. Om de twee metingen zo gelijkwaardig mogelijk te houden wordt elke ppn random toegewezen aan een van de condities: 2-1 of 1-2. Het heet ook wel een paired sample t-test. De paired sample t-test is het testen van de verschilscore tussen de voor, -en nameting. Voor de 0-hypothese geldt  $H_0$ : voormeting=nameting, verschilscore ( $\Delta$ )=0. Wanneer Y de score op de nameting is en X de score op de voormeting geldt  $D=Y-X$ . Hierbij is t te berekenen door  $t = \frac{\bar{D} - \mu_D}{\frac{s}{\sqrt{n}}}$  Er wordt daarna verder gewerkt met de verschilscore.

De test kan over een neutraal of een controversieel onderwerp gaan. Dit wordt weergegeven door  $\bar{D}_{\text{controversieel}} = 1$  en  $\bar{D}_{\text{neutraal}} = 1$ . Er is dan een verschil in interpretatie: Neutraal:  $p < 0,01$ , Controversieel:  $p > 0,05$ . De aannamen van een dependent t-test zijn de volgende:

- De verschilscores moeten normaal verdeeld zijn
- De data zijn gemeten op interval niveau
- De populaties, waaruit de steekproeven zijn getrokken hebben ongeveer gelijke varianties.
- De scores zijn niet onafhankelijk, deze zijn per definitie afhankelijk

Het is goed om naast de nulhypothese nog een toets te hebben die niet afhankelijk is van steekproefgrootte, het kan namelijk nog wel eens zijn dat bij een zeer grote

$$r = \sqrt{\frac{t^2}{t^2 + df}} = \sqrt{\frac{t^2}{t^2 + n - 1}}$$

steekproef de nulhypothese verworpen kan worden waar dit bij een kleinere steekproef niet kan. Daarom wordt vaak ook de effectgrootte berekend:

Een effectgrootte moet als volgt geïnterpreteerd worden:  $r < 0,3$  : klein effect,  $0,3 < r < 0,5$ : middelmatig effect en  $r > 0,5$  : groot effect. Dit geldt voor zowel een neutraal als een controversieel onderwerp.

## College 6

### Field Hoofdstuk 9 deel 2

Onsystematische variantie kan bij een **independent t-test** voorkomen worden door een grote steekproef te nemen met random selectie. Bij een independent t-test wordt er gekeken naar het verschil tussen de gemiddelden van twee groepen proefpersonen. Hoe kleiner de varianties, hoe groter de t en hoe groter  $n_1$  en  $n_2$  en hoe kleiner de variantie van de steekproefverdeling is, hoe groter de t.

Bij een independent t-test gaat het erom dat twee groepen met elkaar worden vergeleken met allebei een gemiddelde. De nulhypothese is dan  $H_0: \bar{X} = \bar{Y}$ . Ook wel geschreven als  $\mu_x = \mu_y$  of  $\mu_x - \mu_y = 0$ . Hierbij moet onderscheid gemaakt worden tussen

gelijke steekproefgrootte:  $t = \frac{\bar{x} - \bar{y} - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$  en ongelijke steekproefgrootte:  $t = \frac{\bar{x} - \bar{y} - 0}{\sqrt{\frac{sp^2}{n_1} + \frac{sp^2}{n_2}}}$

met  $Sp^2 = \frac{(n_1 - 1)s_x^2 + (n_2 - 1)s_y^2}{n_1 + n_2 - 2}$ . Dit is standaardisatie. Belangrijk hierbij is dat aan de

aanname dat de factoren manipulatiefactoren zijn is voldaan.

Met classificatiefactoren kan de t-test niet uitgevoerd worden. Het verschil in gemiddelden en de variantie in de groepen spelen een grote rol in de grootte van de toestingsgrootte (t).

De assumpties voor de independent t-test zijn:

- De populaties, waaruit de steekproeven worden getrokken moeten normaal verdeeld zijn.
- De data moet zijn gemeten op interval niveau
- De populaties waaruit de steekproeven getrokken zijn moeten ongeveer dezelfde varianties hebben. Dit test je door middel van de Levene test. De Levene test laat zien of er voldaan wordt aan de aanname voor gelijke varianties. Als er niet aan voldaan wordt kan er geen dependent t-test worden uitgevoerd. Echter is er toch een uitweg. Er kan worden gekeken in SPSS naar de uitkomst bij equal variances not assumed en er kan een resultaat afgelezen worden. Er moet wel vermeld worden dat er niet voldaan is aan de assumptie.

$$t = \frac{\text{observed difference between sample means} - \text{expected difference between sample means if } H_0 \text{ is true}}{\text{estimate of standard error of differences between the two sample means}}$$

$$t = \frac{\text{effect door manipulatie}}{\text{error}}$$

$$t = \frac{\text{effect door manipulatie}}{\text{individuele verschillen + meetfouten}}$$



- De scores zijn onafhankelijk

De effectgrootte voor de independent t-test wordt als volgt gemeten:

$$r = \sqrt{\frac{t^2}{t^2 + df}} = \sqrt{\frac{t^2}{t^2 + n_1 + n_2 - 2}}$$

De interpretatie is hetzelfde als bij de dependent sample t-test.

Bij een **one-sample t-test** wordt gekeken of het gemiddelde afwijkt van een theoretisch gemiddelde. De nulhypothese is dan  $H_0: \mu_0 = 0$ . Bij het berekenen van de t kijk je naar het verschil tussen systematische en onsystematische variatie. Je kijkt

naar  $S_{\bar{x}} = \frac{S_x}{\sqrt{n}}$ . De verdeling wordt smaller wanneer n groter wordt. De gehele formule

voor t wordt dan:  $t = \frac{\bar{x} - \mu_0}{\frac{S_x}{\sqrt{n}}}$ .

De paired sample t-test is het testen van de verschillscore tussen de voor, -en nameting. Voor de 0-hypothese geldt  $H_0: \text{voormeting} = \text{nameting}$ , verschillscore ( $\Delta$ )=0. Wanneer Y de score op de nameting is en X de score op de voormeting geldt  $D = Y - X$ .

Hierbij is t te berekenen door  $t = \frac{\bar{D} - 0}{\frac{S_D}{\sqrt{n}}}$ .

Keuze maken voor independent of dependent t-test:

De dependent t-test geeft een grotere power, maar is niet altijd mogelijk. Het kan namelijk alleen wanneer de variabelen niet geclassificeerd zijn (man/vrouw) en wanneer de manipulatie geen langdurig effect met zich meebrengt. Er moet een manipulatie effect aanwezig zijn dit wil zeggen dat er een verschil is tussen de gemiddelden van de twee groepen door de manipulatie (groep met/groep zonder alcohol).

Het doen van een puntschatting is vrij nauwkeurig, maar de kans dat deze werkelijk gelijk is aan de ware (verschil)score is zeer klein. De betrouwbaarheid is dus laag. Bij een intervalschatting is de betrouwbaarheid een stuk hoger, maar de nauwkeurigheid lager. Wanneer de standaardafwijking kleiner wordt wordt de intervalschatting nauwkeuriger. Hoe betrouwbaarder de schatting, hoe minder nauwkeurig en andersom. Een schatting maak je van het effect door toeval. De steekproefverdeling geeft weer wat er op basis van toeval verwacht kan worden gegeven dat de nulhypothese waar is. deze is theoretisch af te leiden, maar blijft een schatting:

$$S_{mean} = \frac{S}{\sqrt{N}}$$

Bij het vergelijken van gemiddelden kun je de nulhypothese en de alternatieve hypothesen weergeven in een grafiek. Bij de nulhypothese zijn de gemiddelden en de rest ook gelijk en bij de alternatieve hypothese verschilt alleen het gemiddelde.

## College 7

### Hoofdstuk 10 Independent ANOVA

De hypothesen die je kunt stellen bij een ANOVA zijn als volgt:

$H_0: \mu_1 = \mu_2 = \mu_3$

$H_a$ : de gemiddelden zijn niet allemaal gelijk

Wanneer we gemiddelden met elkaar willen vergelijken gebruiken we normaal gesproken een t-test. Gebruik van een t-test bij het vergelijken van meerdere gemiddelden heeft echter nadelen:

- Vaak zijn we geïnteresseerd in het vergelijken van gemiddelden van meer dan twee groepen
- Een t-test kan alleen gebruikt worden bij één predictor en één onafhankelijke variabele
- Een ANOVA is een uitgebreide regressie
- Door het gebruiken van een t-test bij meerdere gemiddelden kunnen er fouten ontstaan. Je hebt dan een grotere kans om de nulhypothese te verwerpen terwijl deze in werkelijkheid wel waar is (Type I fout) – Kanskapitalisme.

ANOVA is een omnibus test, het toetst of er verschillen zijn, maar niet welke en in welke richting. De theorie achter ANOVA is als volgt: Totale variantie = systematische variantie (variantie door manipulatie) + onsystematische variantie (error).

Bij het uitvoeren van een ANOVA begin je met het berekenen van de total sum of squares (=SSt).

$$SSt = \sum (x_i - \bar{x}_{grand})^2 \quad S^2 = \frac{SS}{n-1}$$
$$SSt = S^2_{grand}(n-1) \quad SS = S^2(n-1)$$

Het aantal vrijheidsgraden is het aantal waarden dat vrij kan variëren. Het wordt meestal berekend door 1 af te halen van het aantal waarden waarmee SS is berekend  $df_i = (n-1)$ .

Na de berekening van de total sum of squares ga je de model sum of squares berekenen. Dit doe je door  $SSm = \sum n_i (\bar{x}_i - \bar{x}_{grand})^2$ . De vrijheidsgraden bereken je door te kijken hoeveel waarden er zijn gebruikt (k).  $df_m = (k-1)$ .

De SSr is de sum of squares van het residu.  $SSr = \sum (x_i - \bar{x}_i)^2$  of  $SSr = \sum S_i^2(n_i - 1)$ .  $S^2$  en SS bereken je hetzelfde als bij de total sum of squares.  $SSr = S^2_{group1}(n_1 - 1) + S^2_{group2}(n_2 - 1) + S^2_{group3}(n_3 - 1)$ . De vrijheidsgraden bereken je door:  $df_r = df_{group1} + df_{group2} + df_{group3}$ . De gevonden waarden zijn te controleren, SSt moet namelijk uitkomen op SSm + SSr.

Na het berekenen van de Sum of squares ga je over tot het berekenen van de mean squares. De  $MS_{model} = \frac{SSm}{df_m}$ ,  $MS_{residu} = \frac{SSr}{df_r}$ . Nu als laatste bereken je dan de F-ratio.

$$F = \frac{MSm}{MSr}$$

Ook bij het doen van een ANOVA zijn er aannames. Dit zijn de volgende:

- De populaties waaruit de steekproeven zijn getrokken zijn normaal verdeeld, dit kun je controleren door middel van de descriptives in SPSS. De populaties waaruit de steekproeven zijn getrokken hebben ongeveer dezelfde varianties. Je kunt dit testen met de test van Levene. Hierbij heb je een nulhypothese (varianties  $\sigma$  zijn gelijk) en een nulhypothese (niet alle varianties  $\sigma$  zijn gelijk).
- De data zijn onafhankelijk van elkaar. Wanneer proefpersonen aselekt zijn toegewezen aan de condities is hieraan voldaan.
- De data zijn gemeten op intervalratio

## College 8

### Field hoofdstuk 10 Vervolanalyses.

Na het doen van een ANOVA wordt er altijd gebruik gemaakt van een vervolganalyse. De F-toets geeft namelijk alleen aan dat de gemiddelden verschillen (of niet) en niet in welke richting. Ook weet je dan nog niet welke gemiddelden van elkaar verschillen.

Vervolanalyses zijn:

Contrasten: gebaseerd op een specifieke hypothese en is a priori geformuleerd op basis van de theorie. Het idee is dat de variatie die wordt uitgelegd door het model (SS<sub>m</sub>) veroorzaakt wordt doordat de deelnemers zijn toegewezen aan verschillende groepen. De variatie kan verder gespecificeerd worden door hypothesen te stellen waar de groepen verschillen en welke groepen. De hypothese hierover wordt gesteld vóór de start van het onderzoek. Er moet één unieke hypothese getoetst worden, er moeten ten minste twee 'chunks' gebruikt worden (K-1) en er moet één contrast minder zijn dan het aantal groepen. Het eerste contrast is vaak de vergelijking met de controle groep en de experimentele groep (2 chunks). Daarna kan nog een hypothese gesteld worden door verschil te bekijken tussen de twee experimentele groepen bijvoorbeeld. Geplande contrasten: groepen met een positieve lading moeten vergeleken worden met groepen met een negatieve lading. De som moet gelijk zijn aan 0 en als een groep niet in de vergelijking zit wordt een gewicht van 0 toegekend. Ook is belangrijk dat de gewichten voor 1 chunk gelijk zijn aan het aantal groepen in de andere chunk. De hypothesen hierbij zijn als volgt:  $m_2 + m_3$

$$\text{contrast 1: } H_0 : m_1 = \frac{m_2 + m_3}{2} \quad H_A : \text{het is niet zo}$$

$$m_1 - 1/2 m_2 - 1/2 m_3 = 0$$

$$2 * m_1 - 1 * m_2 - 1 * m_3 = 0 \rightarrow 2, -1, -1 : \text{gewichten}$$

$$\text{contrast 2: } H_0 : 0 * m_1 - 1 * m_2 + 1 * m_3 = 0$$

$$H_A : \text{het is niet zo} \quad \text{gewichten dus : } 0, -1, 1$$

Er bestaat nog een soort contrasten: Orthogonale contrasten. Dit zijn contrasten die niet correleren. Hiervoor geldt dat de systematische variantie geheel op te delen is in de variantie door de contrasten. De formule is  $SS_m = SS_{c1} + SS_{c2} (+ \dots)$ . Niet orthogonale contrasten correleren juist wel. De formule is  $SS_m \neq SS_{c1} + SS_{c2} (+ \dots)$ . de systematische variantie is niet gelijk aan de totale variantie van de contrasten.

Post-hoc tests: zijn exploratief en vergelijkt alle paren met gemiddelden. Er moet wel rekening gehouden worden met kanskapitalisatie. Er is een controle ingebouwd, een voorbeeld is die van Bonferroni:  $\alpha = \frac{\alpha}{\text{number of tests}}$ . Bij een post-hoc test zijn varianties en steekproeven even groot – REGWQ/Turkey HSD. Steekproefgrootte klein verschil – Gabriel's. Steekproefgrootte groot verschil - Hochberg's GT2. Varianties niet gelijk – Games Howell.

Trend analysis is een speciaal contrast, ook wel een polynomiaal contrast. Deze is theoretisch vaak heel waardevol en kan worden toegepast als de niveaus van de onafhankelijke variabele op interval niveau zitten.

Effectgrootte: eta squared is de makkelijkste maat, is goed te interpreteren, maar is biased dus geen zuivere schatter. De formule is  $R^2 = \eta^2 = \frac{SS_m}{SS^*}$ . Een betere maat is dan omega squared. Deze schatting is wel zuiver, maar alleen geldig als de steekproeven even groot zijn. De formule is:

$$\omega^2 = \frac{SSm - (dfm)MSr}{SSr + MSr}$$

## College 9

### Hoofdstuk 13: Herhaalde metingen (dependent ANOVA)

Bij de repeated measures of dependent ANOVA wordt voor elke conditie dezelfde set proefpersonen gebruikt. Ze worden dus meerdere keren getest. De verschillen tussen de proefpersonen kunnen worden geschat en van de non systematische variantie worden afgetrokken. Omdat er dan minder non systematische variantie is, wordt  $F(\text{syst}/\text{unsyst})$  groter. De power neemt toe. Er zijn voor dit design minder proefpersonen nodig, een nadeel is dat proefpersoonuitval grote gevolgen kan hebben.

Het is niet altijd mogelijk herhaalde metingen te doen. Dit kan alleen als er aselekt aan de condities kan worden toegewezen. Bij classificatiefactoren zoals man/vrouw kan niet herhaaldelijk worden gemeten en als de manipulatie een langdurig effect heeft ook niet.

Nadelen van het gebruik van herhaalde metingen is dat alle proefpersonen de condities in een andere volgorde moeten ondergaan. Dit is omdat een eventueel volgorde of leer effect tegen te gaan. Als het onderzoek om een aantal metingen in de tijd gaat (bijvoorbeeld vooruitgang in een cursus of therapie) is het niet mogelijk herhaald te meten. Een ander nadeel is dat de aanname van onafhankelijk bij herhaalde metingen vrijwel altijd geschonden wordt. Om de gevolgen daarvan te voorkomen moet aan een nieuwe aanname worden voldaan: de aanname van sphericity. Dit houdt in dat de varianties van de verschillen scores tussen de condities ongeveer gelijk moet zijn. Wanneer je gebruik maakt van drie condities (A, B en C) krijg je:

$$\begin{aligned} \sigma_{A-B}^2 &= \sigma_{A-C}^2 = \sigma_{B-C}^2 \\ s_{A-B}^2 &= s_A^2 + s_B^2 - 2s_{AB} \\ s_{A-C}^2 &= s_A^2 + s_C^2 - 2s_{AC} \\ s_{B-C}^2 &= s_B^2 + s_C^2 - 2s_{BC} \end{aligned}$$

**Sphericiteit      homogeniteit      compound symmetry**

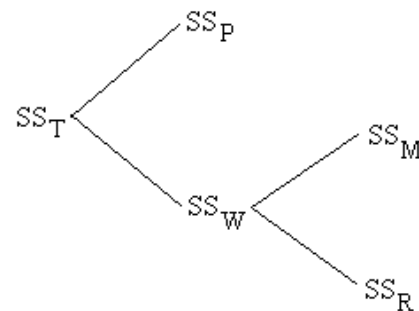
De aanname van Sphericiteit is makkelijker aan te voldoen dan die van homogeniteit en compound symmetry. Hier wordt namelijk alleen geëist dat het totaal gelijk is en niet de onderdelen apart. Om te kijken of aan de aanname is voldaan voer je de test van Mauchly uit. Deze toets laat zien op de sphericiteit afwijkt van de ideale waarde van 1 ( $\epsilon=1$ ). In plaats van te werken met sphericiteit kan ook gewerkt worden met non-sphericiteit. De schatters die je dan gebruikt zijn die van Greenhouse-Geisser( $\epsilon$ ) en Huynh-Feldt( $\epsilon$ ).

Wanneer de varianties van de verschillen scores gelijk zijn dan is de sphericity assumed. De waarde is dan gelijk aan 1. Dit is de ideale sphericiteit. De ondergrens die je stelt aan sphericity is de zo slecht mogelijke situatie =  $1/(\text{het aantal condities}-1)$ . Je kunt in plaats van sphericiteit ook een multivariate toets doen. Het is mogelijk om de verschillen scores (A-B) en (B-C) te beschouwen als twee afhankelijke variabelen. In dat geval is een multivariate toets geschikt. Hier geldt de aanname van sphericiteit namelijk niet.

## College 10

### Interpretatie Hoofdstuk 13

Je kunt de beste strategie aflezen uit de output van de repeated measures. Als de schatting van Huynh-Feldt groter is dan 0,7 kies je voor de multivariate analyse. Als de schatting kleiner is dan 0,7 een univariate analyse. Als Mauchly's test niet significant is kies je Sphericity assumed en bij wel significant kies je het gemiddelde van de overschrijdingskansen (Sig) van Greenhouse-Geisser en Huynh-Feldt.



Bij de herhaalde metingen wordt er op een andere manier opgesplitst dan bij de independent ANOVA. De verschillen tussen deelnemers gaan dan namelijk meespelen. In een plaatje ziet dat er zo uit: De totale variantie (SS<sub>T</sub>) is gesplitst in variatie tussen personen (SS<sub>P</sub>) en individuele verschillen (SS<sub>W</sub>). Van SS<sub>W</sub> is de variatie deels het gevold van de manipulatie en deels meetfouten. SS<sub>W</sub> wordt daarom opgesplitst in de variatie door de manipulatie (SS<sub>M</sub>) en de variatie van het residu (SS<sub>R</sub>). Hierbij is  $k$  het aantal metingen per persoon en  $n$  het aantal personen dat  $k$  keer wordt gemeten.

De dependent ANOVA test gaat ook uit van een aantal assumpties:

- De data moeten normaal verdeeld zijn
- De data moeten gemeten zijn op interval niveau
- De verschillen tussen de metingen moeten gelijke varianties hebben (=aanname van sphericiteit)

## College 11

### Hoofdstuk 11 Covariantie analyse

Bij het uitvoeren van een ANCOVA wordt een covariaat toegevoegd. Dit is een continue variabele die samenhang heeft met de afhankelijke variabele. Deze kan een groot deel van de variantie in de afhankelijke variabele verklaren. Door deze variatie te verwijderen wordt de variantie door de manipulatie duidelijker. Het beste voorbeeld is een voor- en na-meting. De voormeting kan worden ingevoerd als covariaat. Het is namelijk zeer waarschijnlijk dat de scores op de voormeting samenhangen met scores op de na-meting. Een eis die wordt gesteld is dat de scores op de covariaat ongeveer gelijk moeten zijn voor alle groepen. Wanneer dit niet zo is kan de variantie verklaard door de covariaat niet uit de afhankelijke variabele worden verwijderd en weet je nog niets. Als de proefpersonen aselekt worden toegewezen aan condities zijn de groepen ongeveer gelijk op elke variabele en dus ook op de covariaat. Bij kleine groepen en klassificatie gaat dit echter niet op. Je kunt dit zien aan de t-test. Wanneer de groepen nauwelijks verschillen op de voortest is de t-test ook niet significant. Dan zie je dat er gelijke groepen zijn door aselekt toewijzing. De covariaat is dan gelijk voor de groepen.

Berekeningen hiervan worden alleen uitgevoerd door SPSS. De interpretatie is wel belangrijk. De aannamen zijn als volgt:

- De data moeten onafhankelijk zijn van elkaar.

- De populatie varianties zijn ongeveer gelijk, getest met Levine (intercept + voortest +groep).
- De populaties waaruit de steekproeven zijn getrokken zijn ongeveer gelijk verdeeld
- De regressie coëfficiënten van de samenhang tussen de afhankelijke variabele en de covariaat zijn voor elke groep gelijk.

De regressie coëfficiënten van de samenhang tussen de covariaat en de afhankelijke variabele zijn voor elke groep gelijk. Je kunt voor elke groep apart een model maken. De regressie coëfficiënten moeten aan elkaar gelijk zijn.

## College 12

### ANCNOVA, interactie in SPSS Hoofdstuk 11

Als de regressielijnen evenwijdig lopen is er geen sprake van een interactie tussen de groep en de covariaat. Wanneer de lijnen niet evenwijdig lopen is wel sprake van interactie. De manier om te toetsen of de lijnen evenwijdig lopen is dus dezelfde als de manier om te toetsen of er interactie bestaat. Volgens de aanname van een ANCNOVA mag er geen interactie zijn, de  $H_0$  van geen interactie mag dus niet worden verworpen. Er worden bij de ANCNOVA geen gebruikelijke post-hoc toetsen gebruikt, omdat de gemiddelden worden gecorrigeerd. Wel kan gekeken worden of deze gecorrigeerde gemiddelden van elkaar verschillen. In SPSS kun je dan naar Options en 'display means for' groep (algemeen de factor). Dan klik je compare main effects. Er zijn hierbij drie adjustment levels die je kunt gebruiken:

- Tukey LSD: geen correctie, niet aanbevolen
- Bonferroni correctie: aanbevolen
- Sidak correctie: minder conservatief als Bonferroni, dus beter voor de power. Deze wordt het sterkst aanbevolen.

Wanneer in de SPSS output te zien is dat de interactie tussen de covariaat en de onafhankelijke variabele niet significant is is er dus geen interactie tussen groep en onafhankelijke variabele. Er is dan voldaan aan de aanname van gelijke regressie slopes. Het is te lezen in de test van between subjects effects. Deze mag niet gebruikt worden om het hoofdeffect af te lezen. **HIERUIT KUN JE DUS NIET CONCLUDEREN OF ER WEL OF GEEN EFFEC T IS!** De p mag niet gebruikt worden.

De Interpretatie van een ANCNOVA gaat namelijk als volgt: eerst kijk je of de covariaat voor alle groepen ongeveer gelijk is (aanname). Dit kun je doen met een t-test of ANOVA. Hierbij moet dus voldaan worden aan de nulhypothese van geen verschil! Als tweede ga je na of het toepassen van een ANCNOVA zinvol is: is de correlatie tussen covariaat en afhankelijke variabele groot genoeg? Als derde kijk je of er voldaan is aan de aanname van gelijke varianties. Daarna is het belangrijk om met gebruik van het model van interactie tussen de covariaat en de groep na te gaan of er sprake is van gelijke regressie coëfficiënten. Als laatste lees je pas het resultaat van de ANCNOVA af. Hierna kun je eventueel kijken naar vervolganalyse. Een voorbeeld test is een IQ-test. Het resultaat van de ANCNOVA is uiteindelijk af te lezen bij de test van between subjects effects bij dependent variable test en

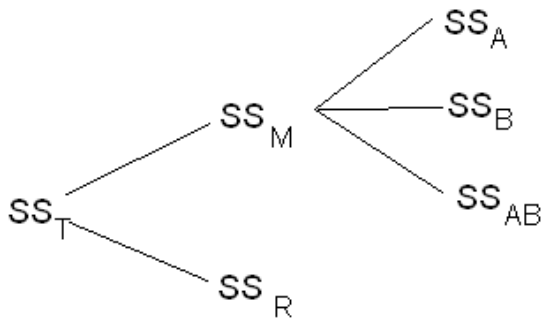
dan bij groep significantie. De gecorrigeerde gemiddelden zijn te verkrijgen door estimated marginal means te bekijken.

## College 13

### Hoofdstuk 12 Factorial ANOVA

Er is bij een factoriele ANOVA sprake van meerdere onafhankelijke variabelen, twee of meer factoren. Meerder one-way ANOVA's doe je niet hierbij omdat de kanskapitalisatie vergroot wordt en er sprake kan zijn van interactie. Welke analyse gedaan wordt is te zien aan de naam. Eerst zeg je het aantal factoren (onafhankelijke variabelen): "two-way, four-way etc." en daarna de toets ANOVA, mixed design ANOVA etc. de gemiddelden van de cellen waarbij je deze toets uitvoert zullen regelmatig van elkaar verschillen. Dit zegt echter niks over de effecten of interactie.

$SS_T$  = variantie tussen alle scores en bestaat uit  $SS_R$  = error variantie en  $SS_M$  = variantie door de experimentele manipulaties, de laatste is onder te verdelen in effect A ( $SS_A$ ), effect B ( $SS_B$ ) en het interactie effect tussen A en B ( $SS_{A \times B}$ ). Uitgelegd in een tekening is dit:



De aannamen van een factoriele ANOVA

zijn als volgt:

1. De populaties waaruit de steekproeven zijn getrokken zijn normaal verdeeld
2. De populaties waaruit de steekproeven zijn getrokken hebben gelijke varianties
3. De data zijn onafhankelijk van elkaar
4. De data zijn gemeten op interval niveau

Bij een twee-weg independent ANOVA met drie niveaus van factor A en 4 niveaus van factor B zijn er 12 condities (en 12 cellen). De berekening voor het aantal cellen is dus: niveau's factor A \* niveaus factor B = aantal cellen. De aannames over de normale verdeling en de gelijke varianties hebben betrekking op de 12 condities. Deze moeten bij benadering allemaal aan elkaar gelijk zijn.

Vervolganalyse: De contrasten zijn te vergelijken met de eenweg ANOVA die al besproken is. Het is weer de keuze tussen de standaardcontrasten op SPSS. Ook is er de mogelijkheid van een simpel contrast. Als je graag een contrast over alle condities wilt doen, dan kan dit met behulp van de syntax. De keuze gaat op dezelfde gronden als bij de eenweg ANOVA. Bij het kiezen voor de post-hoc helmerts test kan dit alleen bij factoren met meer dan twee categorieën.

Een voorbeeld waarmee de factoriele ANOVA kan worden uitgelegd is een experiment met de invloed van vermoeidheid en alcohol. Er zijn twee factoren die voor deze studie van belang zijn: de eerste is alcoholgebruik en vermoeidheid en de invloed op reactietijd. De vermoeidheidsfactor (B) heeft twee factoren namelijk fit (geen manipulatie) en vermoeid (1 nacht geen slaap + 20 km lopen). Aan de hand van de gemiddelden kan nu worden onderzocht of de gemiddelden van de 6 condities verschillen en er kan worden gekeken naar hoofd –en interactie effecten. Gekeken kan worden naar:

- Hoofdeffect van factor A: heeft alcohol effect op de reactietijd?
- Hoofdeffect van factor B: heeft vermoeidheid effect op reactietijd?
- Interactie-effect: heeft er een interactie tussen A en B plaatsgevonden?

Wanneer er een significant verschil gevonden wordt tussen de gemiddelden van de 6 condities kan  $H_0$  verworpen worden. Het resultaat van de toets in SPSS is af te lezen bij 'corrected model'.

## College 14

### Hoofdstuk 14&16 Mixed ANOVA

Verschillende effecten die je kunt vinden bij het uitvoeren van een analyse zijn:

- Hoofdeffecten: elke onafhankelijke variabele kan een hoofdeffect vertonen. Voor elk effect kun je de F-waarde berekenen.
- Tweeweg interacties bestaan uit twee variabelen die een interactie vertonen. Je kunt in één onderzoek testen op meerdere interacties en daar f waarden voor berekenen.
- Drieweg interacties kun je ook krijgen door analyse. Deze zijn meestal echter erg complex en zeer moeilijk te interpreteren. Waar je bij een tweeweg interactie drie keer kunt testen op twee variabelen die interacteren kun je dit met een drieweg interactie samenvakken en kijken hoe ze alle drie met elkaar interacteren.

een mixed ANOVA is vaak te complex om handmatig te berekenen. Dit gebeurt daarom meestal met SPSS. De interpretatie is dan erg belangrijk. Bij het gebruik van een mixed design moet op zowel de aannamen en beperkingen van de herhaalde meting worden gelet als op de beperkingen van de independent ANOVA omdat het gemixed is. De aannamen van de mixed ANOVA zijn:

- De populaties waaruit de steekproeven getrokken zijn moeten normaal verdeeld zijn
- De data moet zijn gemeten op interval niveau
- De data moet onafhankelijk zijn
- De populatievarianties zijn gelijk. Dus voor alle condities is de variantie van de between factor gelijk. Bijvoorbeeld mannen en vrouwen moeten in elke conditie dezelfde variantie hebben. Dit wordt getest met de Levine test.
- Ook de aanname van Sphericity is weer belangrijk (dependent). De variantie van de verschil scores moeten aan elkaar gelijk zijn, dit wordt getoetst met de Mauchly's test.

Vervolganalyses zijn dezelfde als bij de independent en dependent ANOVA en de post-hoc toets gaat weer volgens hetzelfde principe. Hier nogmaals uitgelegd hoe dit werkt:

- Gelijke varianties en even grote steekproeven – REGWQ of Tukey HSD
- Steekproefgrootten klein verschil – Gabriel's
- Steekproefgrootten groot verschil – Hochberg;s GT2
- Varianties niet gelijk – Games Howell



# Hoofdstuk 16

## MANOVA

MANOVA is ook een analyse om verschillen in gemiddelden tussen groepen te toetsen. Hierbij is het het geval dat er meer dan een afhankelijke variabele aanwezig is. Het is opnieuw beter dan aparte ANOVA's uit te voeren want:

- Kanskapitalisatie
- Houdt rekening met relatie tussen de afhankelijke variabelen
- Is meer in overeenstemming met de werkelijkheid

De power in een MANOVA analyse is afhankelijk van de steekproefgrootte, effectgrootte, grootte van de alpha en van de correlaties tussen de afhankelijke variabelen. Het is echter nog niet duidelijk hoe dit precies in elkaar zit. De F is afhankelijk van de verhouding  $SS_M/SS_R$ . Bij de MANOVA test zijn er vier test statistics die je ook te zien krijgt in SPSS. Dit zijn de:

1. Pillai's trace
2. Hotelling's trace
3. Wilk's lambda
4. Roy's largest root

Al deze statistics zijn gebaseerd op eigenschappen van een correlatie matrix. Allemaal hebben ze een andere formule. Bij het kiezen van de juiste uit deze vier moet er rekening gehouden worden met een aantal factoren:

- Power
- Robuustheid
- Gelijkheid van de grootten van de steekproeven

Pillai's trace is robuust als de steekproeven even groot zijn, Roy's largest root heeft de beste power (Pillai's trace ook). Bij Pillai's trace zijn de steekproeven eventueel even groot door delete cases.

Ook bij de MANOVA zijn er weer aannamen waaraan voldaan moet worden:

- Er moet voldaan zijn aan multivariate normaliteit
- Er is een aanname van gelijke variantie-covariantie matrices. Deze moet getoetst worden met een box-test. De toets is zeer gevoelig voor normaliteit.

Wanneer de test niet significant is is het moeilijk te bepalen of de waarden niet normaal verdeeld zijn of dat de matrices ongelijk zijn.

Gelijke variantie-covariantie matrices is uit te leggen met dit plaatje:

$$\begin{array}{|c|c|} \hline S_1^2 & S_{12} \\ \hline S_{12} & S_2^2 \\ \hline \end{array} = \begin{array}{|c|c|} \hline S_1^2 & S_{12} \\ \hline S_{12} & S_2^2 \\ \hline \end{array} = \begin{array}{|c|c|} \hline S_1^2 & S_{12} \\ \hline S_{12} & S_2^2 \\ \hline \end{array}$$

**Groep 1**                      **Groep 2**                      **Groep 3**

De aanname gelijke variantie-covariantie matrix is te toetsen met de box test. Een nadeel is dus dat deze zeer gevoelig is voor normaliteit. Een strategie om hiermee om te gaan bestaat uit:

1. Toets met Levene of de varianties in de groepen gelijk zijn. Als deze gelijk mogen worden beschouwd zijn deze diagonalen aan elkaar gelijk.

2. Ten tweede moet je nagaan of er sprake is van een multivariate normale verdeling. Dit is alleen niet te toetsen. Univariate normaliteit gelukkig wel. Hiervoor gebruik je de Kolmagorov-Smirnov of de Shapiro-Wilk toets (onder descriptives-explore). Je toetst hiermee of variabele 1 normaal verdeeld is en je toetst of variabele 2 normaal verdeeld is in elke groep. Dit geeft alleen geen garantie voor een normale multivariate verdeling in beide groepen.
3. De box test voer je vervolgens uit. Samen met de test van Levene kan deze aangeven of de diagonalen gelijk zijn. We hebben dan zoveel mogelijk gegevens over de gelijkheid van de matrices.

Korte samenvatting:

- Levene: alle  $p > 0,05$ . De diagonalen zijn niet significant verschillend en dus gelijk. Er wordt dan voldaan aan de homogeniteits aanname.
- Normaliteitstoets: alle verdelingen univariaat normaal? Door naar de Box's test
- Box's test:  $p > 0,05$  er is voldaan aan de aanname van gelijke variantie covariantie matrices.

## College 15

### Hoofdstuk 17 PCA

PCA is principale component analyse. Als een aantal variabelen onderling een hoge correlatie hebben komt dat waarschijnlijk omdat de door eenzelfde dimensie worden beïnvloed. Deze worden componenten of factoren genoemd. Factoranalyse is op drie manieren te gebruiken:

1. Structuur ontdekken in een set variabelen, kijken welke variabelen dezelfde dimensie meten.
2. Het is ook handig om bij het maken van een vragenlijst de achterliggende latente variabele te ontdekken en een item-analyse uit te voeren. Elk groepje met variabelen die eenzelfde latente variabele meten vormen een schaal van de test.
3. Als derde is het goed om het aantal variabelen terug te brengen tot een beter onderzoek wat meer handelbaar is. Er kunnen bijvoorbeeld minder predictoren nodig zijn dan gedacht zonder dat er informatie verloren gaat. Het doen van factoranalyse kan een oplossing zijn voor multicollineariteit (wanneer twee predictoren te veel met elkaar correleren dat een predictor bijna geen variantie meer kan verklaren).

De correlatiematrix van de ingevoerde variabelen wordt geanalyseerd en het resultaat heet de factor ladingen matrix. De manier waarop de variabelen worden omgezet in factoren wordt de hoofdassen methode genoemd. Deze is niet mer de hand te berekenen. Door deze methode kunnen een aantal dingen bereikt worden:

- De factoren of componenten kunnen op volgorde van belang worden gepresenteerd.
- Er kan worden gekeken of het zo kan worden gemaakt dat componenten niet correleren.
- De factoren kunnen worden weergegeven in een assenstelsel. De variabelen kunnen hierin worden geplaatst.
- Hoe meer lading van variabelen op de factor (hoe meer invloed), hoe meer de factor de relatie tussen de variabelen verklaart.

Je kunt ook factorscores gebruiken. Dit is zeer handig voor:

- Het reduceren van een groot aantal tot een kleiner aantal. Hierbij geven de factor scores de score van het individu weer op een nieuwe combinatie meting. Elke verdere analyse kan met deze scores worden uitgevoerd.

- Factorscores kunnen dus het multicollineariteitsprobleem opheffen. De factorscores worden dan als predictoren gebruikt. Omdat de predictoren dan niet correleren heb je betere en minder predictoren.

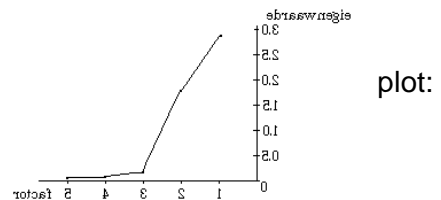
Het selecteren van componenten uit een analyse kan erg belangrijk zijn. Het is vaak zo dat er in eerste instantie te veel componenten worden ingevoerd. Er zijn dan evenveel componenten als ongevoerde variabelen. Je wilt echter minder variabelen, dus minder componenten. Daarom moet je componenten gaan selecteren. Dit doe je door te kijken naar eigenwaarde. De componenten zijn namelijk op volgorde van belang gerepresenteerd. Dit betekent dat de eerste component de meeste variantie kan verklaren, de tweede wat minder variantie en zo door. De verklaarde variantie per component = eigenwaarde van component  $k$  =

De formule van eigenwaarde = 
$$I_k = \sum_{i=1}^k a_{ik}^2$$

Selectie van componenten kan door te kijken naar een bepaald criterium. Zo is er het criterium van Kaiser. Je selecteerd hierbij de componenten met een eigenwaarde die groter is dan 1. 1 is de gemiddelde eigenwaarde, dus eigenwaarde groter dan het gemiddelde. Volgens dit criterium kunnen de eerste twee componenten worden geselecteerd.

Een andere manier is door het maken van een screen-

Hierbij zou je alle componenten boven de knik kunnen selecteren.



Als manier is er nog cumulatieve verklaarde variantie. De onderzoeker kiest een percentage dat de geselecteerde componenten gezamenlijk moeten verklaren. Dit kan uiteenlopen van 60% tot 99%. Dit is te berekenen door alle eigenwaarden te delen door de som van alle eigenwaarden en te vermenigvuldigen met 100. Bijvoorbeeld:

$$\text{perc var C1} = \frac{2,93}{5} * 100 = 58,6\%$$

Verder heb je percentage verklaarde variantie door een component. Een onderzoeker bepaalt dat een geselecteerde component ten minste 10% of 15% van de variantie van de variabelen moet verklaren.

De interpretatie van de componenten na de selectie gaat aan de hand van de geroteerde factorladingen matrix. Er moet aan de hand van de inhoud van de variabelen een naam worden gegeven aan de componenten. Deze naam is zeer subjectief, je kunt er niks uit ontlenu.

## College 16 hoofdstuk 17&18

$$h_i^2 = \sum_{k=1}^m a_{ik}^2$$

Communaliteit is de gemeenschappelijke variantie van de variabelen met de componenten. Dit wordt ook wel common variance genoemd. De formule voor het berekenen is:

De correlatie tussen de variabelen is uit de complete factorladingen matrix af te lezen.

De formule hiervoor is: 
$$r_{ig} = \sum_{k=1}^m a_{ik} a_{gk}$$

Eigenschappen van een goede correlatiematrix:

- De kwaliteit van de data bepaalt de kwaliteit van de analyse (GI  $\Rightarrow$  GO).
- De variabelen moeten goed correleren
  - $r > .3$ .
- Ze moeten niet te hoog correleren (= multicollineariteit)
  - $r > .80$ .

- Ze mogen zeker niet perfect correleren (= singulariteit):
  - $r = 1$ .
- Inspecteer de correlatiematrix!!

Eigenschappen van de correlatiematrix kun je bestuderen op bepaalde eigenschappen. Hieruit kun je afleiden of het een goede matrix is.

1. De determinant is zo'n eigenschap. Als deze te laag is duidt dit op multicollineariteit. De determinant moet namelijk groter zijn dan 0.00001. Te lage correlaties zijn slecht.
2. De tweede methode is de Kaiser-Meyer-Olkin (KMO) methode. Deze meet of de correlatie-matrix geschikt is voor de analyse. KMO moet namelijk groter zijn dan 0.5. Je kunt hiermee onderscheid maken tussen een compact patroon (goed) en een diffuus patroon (goed).
3. Als derde methode kun je de sphericiteit toetsen. Dit doe je door de Bartlett's test of Sphericity. Je toetst hiermee of de correlatie matrix een identiteitsmatrix is. Een voorbeeld van een identiteitsmatrix is dat hij significant moet zijn met een  $p < 0,05$ . Wanneer dit het geval is wijken de correlaties significant af van 0 en zijn dus groot genoeg.

Om factoren te vinden bestaan er ook verschillende methodes. Deze zijn allemaal anders toe te passen. Heel algemeen kun je een indeling maken in conirmatief en exploratief gebruik. Confrimatief wordt hier niet behandeld, maar betekent dat je van tevoren aangeeft welke factoren en welke variabelen daarop thuishoren. Dit wordt dan getoetst. Bij exploratieve factoranalyse is onderscheid te maken in een aantal verschillende technieken. De principale componenten analyse (factor analyse) wordt het meest gebruikt. Een andere manier om dit te doen is door de principale Axis Factoring (PAF).

Principale Axis Factoring gaat niet uit van een correlatie matrix, maar van een matrix waarbij naast de diagonaal wel de correlaties staan. Op de diagonaal staat echter de multiple correlatie met andere variabelen. Dit worden common variances of communaliteiten genoemd. Unieke variantie waaronder errorvariantie wordt dus niet in de analyse opgenomen. Dit is het grootste verschil met PCA. Deze gebruikt namelijk alle variantie in de analyse, dus ook de errorvariantie.

Een ander verschil is dat PCA uitgaat van een correlatiematrix waarbij componenten alle variantie van de variabelen verklaren. Een aanname hierbij is dat alle informatie uit de variabelen nog in de componenten aanwezig is. De communaliteiten van elke variabele is voorafgaand aan de selectie gelijk aan 1,00. Na de selectie wordt niet meer alle variantie verklaard door de overgebleven componenten: de communaliteit wordt dan kleiner dan 1.

Er bestaat ook een theoretische oplossing. Deze heeft geen variantie error  $\rightarrow$  FA. Als je een studie hebt ontworpen op basis van onderliggende construsten (dus niet heel erg exploratief) waarvan je verwacht dat ze scores zullen gaan produceren op je geobserveerde variabelen kies dan voor de A methode. Wanneer je een empirische samenvatting wilt van een dataset kan het beste PCA gebruikt worden.

We gaan er vanuit dat factoren algebraïsch zijn psychologische constructen weergeven. Hier kan een naam aan gegeven worden die aangeeft wat de inhoud is van deze constructen. Sommige wetenschappers zeggen dat componenten slechts statistische waarheden en psychologische bedenkzels zijn.

## Hoofdstuk 18

### Categorische data

In dit hoofdstuk gaat het eerste over categoriale data, dan wordt de chi-kwadraat toets besproken en vervolgens longlinaire modellen.

Categoriale data zijn gemeten op ordinaal of nominaal niveau. Elke persoon wordt toegewezen aan een ongeordende of geordende categorie. Dit kunnen categorieën zijn van 1 actor, maar ook van 2,3,4 enz factoren. Als je te maken hebt met twee factoren, krijg je een twee dimensionale kruistabel.

De Pearson's chi-kwadraat toets wordt gebruikt om te toetsen of er afhankelijkheid tussen twee variabelen bestaat. Dat doet de toets door de geobserveerde frequenties te vergelijken met de frequenties die worden verwacht wanneer de data onafhankelijk is. E is hierbij de verwachte waarde E. De formule:

$$X^2 = \sum_{\text{cellen}} \frac{(\text{Observed}_{ij} - \text{Model}_{ij})^2}{\text{Model}_{ij}} = \sum_{\text{cellen}} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad i=\text{rijen}, j=\text{kolommen}$$

de formule van het model is gebaseerd op de verwachte frequenties (expected frequencies).

Deze worden voor elke cel berekend:  

$$\text{Model}_{ij} = E_{ij} = \frac{\text{rij} - \text{totaal} * \text{kolom} - \text{totaal}}{n}$$

Je kunt van de verwachte waarden een tabel maken per variabele. In de rechter kolom kun je dan de geobserveerde waarden ernaast zetten. Let op: de som van de verwachte waarden is gelijk aan de marginales. Ook zijn de geobserveerde frequenties samen ongeveer gelijk aan de verwachte frequentie.

De Pearson's test statistic geeft de verhouding tussen verwachte frequentie en geobserveerde frequentie weer ( $X^2$ ). De toetsingsgrootte is  $X^2$ . Deze heeft  $(r-1)(c-1)$  vrijheidsgraden. Als de uitkomst significant is kan de nulhypothese van onafhankelijkheid worden verworpen. Dit houdt in dat er sprake is van afhankelijkheid (twee factoren hebben een correlatie met elkaar). Deze grootte is chi-kwadraat verdeeld.

Een andere grootte waarvoor hetzelfde geldt is de likelihood ratio statistic. Deze heeft in bepaalde situaties een voorkeur boven  $X^2$ . Het teken hiervan is  $G^2$ . De formule die hoort bij  $G^2$  is de volgende:

$$G^2 = L_{c^2} = 2 * \sum_{\text{cellen}} \text{Observed}_{ij} \ln \frac{\text{Observed}_{ij}}{\text{Model}_{ij}} = 2 * \sum_{\text{cellen}} O_{ij} \ln \frac{O_{ij}}{E_{ij}}$$

De interpretatie is hetzelfde als die bij  $X^2$  en de grootte is ongeveer hetzelfde. Toch kan het soms verschil uitmaken tussen wel en niet significant. De interpretatie van de Chi-kwadraat in het algemeen:

- De toetsingsgrootte heeft een overall resultaat
- De toetsingsgrootte kan worden afgebroken in 'standardized residuals'. Dit geeft aan welke invloed de frequentie in de betreffende cel heeft op de grootte van  $X^2$ . Twee dingen zijn hierbij belangrijk:
  1. Standardized residuals jennem een directe relatie tot de toetsingsgrootte
  2. Het zijn Z-scores en leveren geen significante bijdrage aan  $X^2$
- De effectgrootte: de Odds ratio kan gebruikt worden voor het aflezen van effect grootte bij een chi kwadraat toets. Het geeft de steekproef-onafhankelijke maat voor effectgrootte. Voorbeelden zijn jongen/meisje en eten/niet eten. Wanneer data onafhankelijk zijn is de odds ratio gelijk aan 1. Wanneer een waarde van 6 wordt gevonden moet deze geïnterpreteerd worden als 1/6. Het is dus een asymmetrische schaal. Er kan worden getoetst of de waarde van 1 awijkt.

Een chi-kwadraat test heeft ook aannames, dit zijn:

- De waarnemingen moeten onafhankelijk van elkaar zijn

- De verwachte waarden moeten in elke cel groter zijn dan 5 (in grote tabellen is deze regel wat soepeler). Eigenlijk is deze aanname meer een eis aan de steekproefgrootte.

## College 17

### hoofdstuk 18

Longlinear analyse wordt gebruikt als er meer dan twee factoren zijn. Een chi kwadraat toets is dan niet meer toe te passen. Het ziet er als volgt

Twee factoren:

factoren:

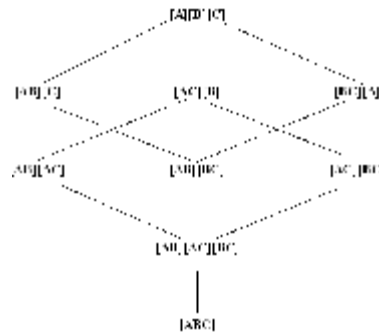
$H_0$ : onafhankelijkheid: [A][B]

$H_A$ : afhankelijkheid: [AB]

[A][B]

|  
[AB]

Drie



Bij drie factoren moet je gebruik gaan maken van modelpassing. Bij drie factoren geldt dus:

- Als model van volledige onafhankelijkheid wordt verworpen, blijven er nog 8 andere modellen over.

Elk van de 8 modellen wordt bestudeerd hoe ze passen bij de gegeven frequentie-tabel. Het model wat het beste past zal worden geïnterpreteerd. Wel wordt voor elk model de verwachte waarde berekend en vergeleken met de gevonden waarden. Voor elk model geldt een ander verschil tussen O en E. Elk model heeft dus een andere  $X^2$  en  $G^2$  en df. Aan de hand hiervan wordt voor elk model de p-value afgeleid. Je stelt een tabel op met de verwachte waarden gegeven dat de 0-hypothese van volledige onafhankelijkheid waar is. Je kunt bekijken wat  $G^2$  dan zou zijn. Vervolgens ga je kijken naar de verwachte waarde onder het model en de  $G^2$  daarbij. Wanneer die  $G^2$  gelijk is aan de df weet je al bijna zeker dat het model gaat passen. Je moet echter eerst nog kijken naar de significanties en dit in een tabel weergeven. Voorbeeld staat hieronder:

Model:	$G^2$	df	Sig = p
[D][E][F]			
[DE][F]			
[FE][G]			
[DE][DF]			

Alle mogelijke modellen:

Op het hoogste niveau het model van volledige onafhankelijkheid.

[A][B][C]

Een niveau lager volgt de modellen met de onafhankelijkheid van één factor

[AB][C] C is onafhankelijk van A en B, A en B hangen samen

[AC][B] B is onafhankelijk van A en C, A en C hangen samen

[BC][A] A is onafhankelijk van B en C, B en C hangen samen

Weer een niveau lager volgen de modellen de conditionele onafhankelijkheid

[AB][BC] A en C zijn onafhankelijk, gegeven B

[AB][AC] B en C zijn onafhankelijk, gegeven A

[AC][BC] A en B zijn onafhankelijk, gegeven C

Dan volgt het model van homogene associatie  
[AB][AC][BC] Drie tweeweg interacties tussen de variabelen.

En ten slotte: het verzadigde model:  
[ABC] Drieweg interactie

Om te kijken welk model het beste past worden eerst de modellen die significant zijn geselecteerd. Vanuit deze modellen wordt een selectie gemaakt van het beste model op basis van een aantal regels:

- Kies het eenvoudigste model. Het eenvoudigste model is het model met de minste onafhankelijkheidstermen.
- Kies alleen een model met meer onafhankelijkheidstermen als dat model significant beter is. Je kijkt hierbij niet naar de levels van significantie in de tabel, deze zeggen namelijk niets hierover. Je kijkt naar het verschil tussen de  $G^2$  en de d. het is namelijk zo dat  $\Delta G^2$  chi-kwadraat verdeeld is. Om te toetsen of het model met meer termen significant beter is bereken je de rechter kritieke grens van dat model en kijk je of het verschil groter is, zo ja? Dan is het model significant beter.

$$\Delta G^2 = G^2(\text{eenvoudiger model}) - G^2(\text{complexer model})$$

$$\text{en is } \chi^2 \text{ verdeeld met } \Delta df = df(\text{eenv. mod}) - df(\text{compl. mod})$$

In de vergelijking is af te lezen, dat een term meer invloed heeft op de uiteindelijke uitkomst, als de b-waarde groter is. Door het schatten van de b-waardes kun je schatten hoe groot de samenhang is. Hoe groter de b-waarde, hoe meer samenhang. Als de b-waarde heel klein is, is deze van geen belang voor de uitkomst. Dit kan worden getoetst met omzetting van de b-waarden in Z-scores. Bij een hoge significantie is er samenhang. Je mag hier echter nooit een definitieve conclusie uit trekken.

Assumpties zijn als volgt:

- Data moet onafhankelijk zijn
- De verwachte frequenties moeten groter zijn dan 5. De E's mogen voor 20% kleiner zijn dan 5, maar moeten groter zijn dan 1. Als deze aanname is geschonken neemt de power sterk af. Wanneer je hier niet aan kunt voldoen kun je:
  1. Samenklappen over twee factoren: je maakt dan geen onderscheid meer. Samenvoegen geeft meer waarnemingen per cel dus stijging van E.
  2. Klap samen over meer niveaus, zelfde effect als bij 1, echter kan alleen als dit theoretisch ook verantwoord is.
  3. Verzamel meer data
  4. Accepteer verlies van power.

Als laatste nog de hypothesen bij modeltoetsen:

$H_0$ : Complexe model past niet beter dan het simpele model

$H_A$ : Complexe model past wel beter dan het simpele model

