Hoofdstuk 7. Interval- en ratio associatiematen

In dit onderdeel zijn de associatiematen op interval- en rationiveau aan de orde. Zowel de bivariate analyses (correlatie en enkelvoudige regressie) als de multivariate analyse (meervoudige regressie) komen naar voren. Er treedt rangordening op, waardoor je verschillende soorten samenhang kan hebben: perfecte positieve samenhang (als x toeneemt, neemt ook y toe), perfecte negatieve samenhang (de waarde van y daalt als x toeneemt) of geen samenhang (een verandering in x gaat niet gepaard met een verandering in y).

Correlatie
De volledige naam van deze associatiemaat is *Pearson productmoment correlatiecoëfficiënt*, aangeduid met de letter *r*. De correlatiecoëfficiënt is een associatiemaat die alleen op interval- en rationiveau voorkomt. Er hoeft hierbij geen sprake te zijn van een onafhankelijke of afhankelijke variabele, dat betekent dat het om een symmetrische relatie gaat. Wel moeten **beide** variabelen interval of ratio zijn.

Een correlatie of samenhang is vaak te zien in een spreidingsdiagram. Dit is een x-as en een y-as met allemaal losse punten verspreid. Hieruit kun je vaak al opmaken of er een positieve samenhang (de losse punten beginnen laag en naarmate x toeneemt, neemt y ook toe), negatieve samenhang (de losse punten beginnen hoog en naarmate x toeneemt, daalt y) of geen samenhang (alle punten zijn willekeurig verspreidt) is.

Om een spreidingsdiagram in SPSS te krijgen, ga in je menubalk naar Graphs, daarna naar Legacy Dialogs, vervolgens kies je in het Scatter/Dot scherm voor het eenvoudige spreidingsdiagram: Simple Scatter. Bij Simple Scatter kun je vervolgens aangeven welke variabelen je wilt gebruiken voor de x- en y-as. Het spreidingsdiagram geeft een indruk van de samenhang, maar het berekenen van de correlatie geeft betere informatie. De correlatiecoëfficiënt kan een waarde aannemen tussen de -1 en +1.

- Correlatiecoëfficiënt is 1 = perfecte positieve samenhang, stijgende lijn
- Correlatiecoëfficiënt is 0 = geen samenhang
- Correlatiecoëfficiënt is -1 = perfecte negatieve samenhang, dalende lijn

Bij nominale en ordinale variabelen wordt vaak gekeken naar een kruistabel. Bij interval- en ratiovariabelen is het van belang om naar een spreidingsdiagram te kijken. De correlatiecoëfficiënt *r* laat namelijk alleen zien in hoeverre er een lineaire samenhang is. Als er bijvoorbeeld sprake is van een kromlijnig verband tussen de variabelen, zal de waarde van *r* 0 zijn. Er is echter wel degelijk een samenhang.

De correlatiecoëfficiënt kun je ook met de hand uitrekenen. De associatiemaat is gebaseerd op de standaarddeviatie en de variantie. Standaarddeviatie is de gemiddelde afstand ten opzichte van het gemiddelde. De variantie is de standaarddeviatie in het kwadraat. Hierbij kijken we naar de variantie tussen twee variabelen, zogenaamd **covariantie**. Dit is de mate waarin twee variabelen tegelijk variëren. Er zijn verschillende formules betrokken bij het berekenen van correlatiecoëfficiënt.

JoHo Samenvatting – Beschrijvende Statistiek

Formule variantie
Hierbij wordt gekeken naar alle individuele waarden van x (x_i) ten opzichte van het gemiddelde van x ($\dot{x}$, oftewel 'x streep'). Deze doe je elk in het kwadraat en vervolgens tel je alle losse uitkomsten bij elkaar op. Nadat je deze uitkomst hebt gedeeld door het aantal onderzoekseenheden – 1, heb je de variantie.

Formule van de variantie:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \dot{x})^2}{n-1}$$

Formule covariantie
Hierbij bereken je alle individuele waarden van x (x_i) ten opzichte van x-gemiddeld ($\dot{x}$) en ook alle individuele waarden van y (y_i) ten opzichte van y-gemiddeld ($\bar{y}$). Deze waarden vermenigvuldig je met elkaar. Vervolgens tel je alle uitkomsten van deze vermenigvuldiging bij elkaar op en deel je door het aantal onderzoekseenheden – 1.

Formule van de covariantie:

$$Cov(x, y) = \frac{\sum_{i=1}^{n} (x_i - \dot{x})(y_i - \bar{y})}{n-1}$$

Formule voor correlatie
Het lijkt moeilijker dan het in werkelijkheid is. Stap 1 bij het berekenen van de correlatie is de datamatrix erbij pakken en het gemiddelde van x en y berekenen (alle waarden optellen, gedeeld door *n*). Stap 2 is het berekenen van de standaarddeviatie voor x en y (dat is de wortel uit de variantie, formule is eerder weergegeven). Daarna bereken je de covariantie (hierbij doe je hetzelfde als bij de variantie alleen nu neem je alle waarden van y ten opzichte van y-gemiddeld ook mee). Stap 3, de laatste stap, is het invullen van de formule van r, alle gegevens heb je immers al berekend in de voorgaande stappen.

Formule van de correlatie:

$$r_{xy} = \frac{Cov(x,y)}{s_x s_y}$$

Er is nog een makkelijker weer te geven formule voor *r*, hierin is namelijk de covariantie al meegenomen:

$$r = \frac{\sum_{i=1}^{n} (x_i - \dot{x})(y_i - \bar{y})}{(n-1)\, s_x s_y}$$

Berekening in SPSS
Natuurlijk kan SPSS ook correlatie (r) berekenen. Eerst ga je naar Analyze → Correlate → Bivariate. In het vakje Variables kun je de variabelen aangeven waarbij de berekening van correlatie nodig is. Pearson van Pearsons productmoment correlatiecoëfficiënt staat automatisch al aangevinkt.

Kendalls tau-b en Spearmans rho kunnen eveneens uitgerekend worden via Bivariate Correlations. Het enige wat je hiervoor hoeft te doen, is dezelfde stappen, zoals hierboven beschreven, te doorlopen en vervolgens het juiste hokje aan te vinken.

Enkelvoudige regressie
Bij het vorige onderdeel van correlatie ging het om een symmetrische relatie. Bij enkelvoudige regressie gaat het om een asymmetrische relatie, dus een relatie tussen één onafhankelijke en één afhankelijke variabele. Bij enkelvoudige regressie wil je een regressieanalyse uitvoeren en hierbij is het de bedoeling om een patroon te achterhalen in de samenhang tussen x en y, zodat je voorspellingen of schattingen kunt doen over y, indien x is gegeven.

Regressielijn
Een regressielijn is een lineaire lijn die je trekt in het spreidingsdiagram zo dicht mogelijk langs alle punten. Deze lijn $\hat{y}$ (uitspraak y dakje) kun je ook berekenen met de formule:

$\hat{y} = a + b(x)$.

Het dakje op de y geeft aan dat het om een schatting gaat. De *a* uit de formule wordt **intercept** of **constante** genoemd, hierbij snijdt de lijn de y-as. Anders gezegd: dit is de voorspelde waarde van y als x = 0. De *b* in de formule heet de **ongestandaardiseerde regressie coëfficiënt**. Deze bepaalt met hoeveel eenheden de waarde van y verandert als x toeneemt met één eenheid, dus hoe stijl je stijgende of dalende regressielijn wordt in het spreidingsdiagram.

Bij de berekening van enkelvoudige regressie pak je weer de datamatrix erbij. Vervolgens bereken je de gemiddelden voor x en y, omdat je deze allebei nodig hebt in de formules. Daarna begin je met de berekening van *b*, deze doe je eerst omdat je *b* nodig hebt om de intercept *a* uit te rekenen. Het beste is om alle informatie voor jezelf in een tabel te zetten. Zodoende kun je alle gegevens voor *b* gemakkelijk invullen en daarna *a* berekenen.

JoHo Samenvatting – Beschrijvende Statistiek

Formule van *b*:

$$b = \frac{\sum (x - \dot{x})(y - \bar{y})}{\sum (x - \dot{x})^2}$$

Formule van *a:*

$$a = \bar{y} - b\dot{x}$$

Proportie verklaarde varia(n)tie
Uiteindelijk zal blijken als de lijn getrokken is in het spreidingsdiagram op basis van het berekenen van de regressielijn, dat de scores niet precies op de geschatte lijn liggen. Dat betekent dat er altijd een mogelijkheid is op variatie rond de regressielijn. Nu is het mogelijk om de mate waarin de regressielijn de varia(n)tie verklaart, te berekenen. Dit doen we met de proportie verklaarde variantie (R^2) die gebaseerd is op de proportie voorspellingsverbetering. Hij komt daarmee deels overeen met Goodman & Kruskals tau en lambda. De proportie verklaarde variantie kan echter alleen voorkomen op interval- en rationiveau.

Berekening
R^2 kan een waarde aannemen tussen 0 en 1, dat betekent minimaal 0% verklaring tot maximaal 100% verklaring. De totale variatie is de kwadratensom van de verschillen met het gemiddelde. De onverklaarde variatie, ook wel **residu** genoemd, is het kwadraat van het verschil tussen y en de voorspelde waarde van y (ŷ). Hierbij kun je al behoorlijk wat gegevens van de vorige berekende regressielijn overnemen: $\hat{y} = a + b(x)$. Het beste hierbij is wederom om alles uit te schrijven in een tabel.

De formule voor R^2:

$$R^2 = \frac{E_1 - E_2}{E_1}$$

De uitgebreide formule voor R^2:

$$R^2 = \frac{\sum(y - \bar{y})^2 - \sum (y - \hat{y})^2}{\sum (y - \bar{y})^2} = \frac{\text{totale variatie} - \text{onverklaarde variatie}}{\text{totale variatie}}$$

JoHo Samenvatting – Beschrijvende Statistiek

Berekening in SPSS
Het uitvoeren van een regressieanalyse in SPSS gaat als volgt: Analyze → Regression → Linear. Bij het vakje Dependent voer je de afhankelijke variabele in en bij Independent de onafhankelijke variabele(n).

Uitleg output
Bij de output van SPSS krijg je drie tabelletjes. De eerste tabel, genaamd Coefficients, geeft de intercept, de ongestandaardiseerde- en de gestandaardiseerde regressiecoëfficiënt weer. **Intercept** staat in de tabel gelijk onder B en na (Constant). Dit houdt in dat wanneer x = 0, y het getal is dat je hier vindt. De **ongestandaardiseerde regressiecoëfficiënt** (b) wordt onder intercept en achter de onafhankelijke variabele weergegeven. De **gestandaardiseerde regressiecoëfficiënt** kan worden gevonden onder Bèta, die geeft het zuivere effect van de onafhankelijke variabele op de afhankelijke variabele.

De tweede tabel genaamd Model Summary, geeft de **proportie verklaarde variantie** weer, onder R square. Dit getal geeft aan of de onafhankelijke variabele een goede verklaring is voor de afhankelijke variabele.

De derde tabel heet ANOVA en deze geeft de berekening van de proportie verklaarde variantie weer. Bij deze laatste tabel kun je de **totale variantie** vinden onder Total (ook wel SST; Sum of Squares Total) en de **onverklaarde variantie** onder Residual (ook wel SSE; Sum of Squares Error).

Meervoudige regressie
Bij een meervoudige regressieanalyse wordt er gebruikt gemaakt van meerdere onafhankelijke variabelen in tegenstelling tot de enkelvoudige regressie. Er wordt nog wel steeds gekeken naar het voorspelde effect van deze variabelen. De formule kan verveelvoudigd worden afhankelijk van de hoeveelheid onafhankelijke variabelen die meespelen. De formule bij drie onafhankelijke variabelen is:

$$\hat{y} = a + b_1x_1 + b_2x_2 + b_3x_3$$

De intercept blijft enkelvoudig, omdat dit het punt is waarop de y-as gesneden wordt als alle x-en nul zijn. De ongestandaardiseerde regressiecoëfficiënt van de onafhankelijke variabele geeft het effect van die variabele op y weer als de andere onafhankelijke variabelen niet veranderen en dus constant worden gehouden.

Berekening in SPSS
Bij SPSS in de tabel Coefficients zijn er nu drie bèta's te zien oftewel drie gestandaardiseerde regressiecoëfficiënten. Dat wil zeggen dat deze drie de partiële zuivere effecten aangeven Omdat die regressiecoëfficiënten gestandaardiseerd zijn, mag je de drie onafhankelijke variabelen met elkaar vergelijken. De bèta's variëren altijd van -1 tot 1. Als er meer dan drie bèta's zijn, is *R* de multipele correlatiecoëfficiënt niet meer gelijk aan een correlatiecoëfficiënt |r|. Na de standaardisering mag je de formule invullen voor meervoudige regressie, hiermee kun je voorspellingen doen.

JoHo Samenvatting – Beschrijvende Statistiek

Samenvattend overzicht

In dit hoofdstuk zijn de associatiematen op interval- en rationiveau uitgelegd, waarbij de correlatie (r) op basis van een symmetrische relatie wordt berekend en de proportie verklaarde variantie (R^2) en de gestandaardiseerde regressiecoëfficiënt (bèta) op basis van een asymmetrische relatie worden berekend.

Je kijkt bij een meervoudige regressieanalyse naar het partiële effect van een onafhankelijke variabele, onder constant houding van de andere onafhankelijke variabelen.