9. Lineaire Regressie en Correlatie

Lineaire verbanden

In dit hoofdstuk worden methoden gepresenteerd waarmee je kwantitatieve respons variabelen (afhankelijk) en verklarende variabelen (onafhankelijk) kunt analyseren. Dit gebeurt aan de hand van regressie analyse. Regressie analyse omvat drie analyses: 1) onderzoeken of er een verband bestaat tussen de variabelen, 2) de sterkte van dit verband bepalen, en 3) het maken van een regressieformule om zo de waarde van de response variabele te kunnen voorspellen aan de hand van de verklarende variabele.

Bij een lineair verband wordt de respons variabele weergegeven met 'y' en de verklarende variabele met 'x'. Met een lineaire functie kunnen we een rechte lijn trekken door de datapunten in een grafiek. Deze functie heeft deze vorm: y = a + b (x). Hierbij is ' α ' de intercept, en ' β ' de hellingscoëfficiënt.

Interpreteren van y-intercept en hellingscoëfficiënt

De y-intercept is de waarde van y wanneer x = 0. Want als x = 0, dan vervalt b(x), en hou je alleen y = β over. Daarbij geeft de y-intercept aan waar de lijn op de y-as begint.

De hellingscoëfficiënt geeft de verandering aan in y, bij een toename van 1 punt bij x. Wanneer x er 1 punt bij krijgt, verandert y met b.

Over het algemeen is het zo dat hoe groter β, hoe steiler de regressielijn. Als β positief is, betekent dat dat wanneer x hoger wordt, y ook hoger wordt. Dit kenmerkt een positief verband. Wanneer β negatief is, betekent het dat wanneer x hoger wordt, y lager wordt. Dit is een negatief verband. Wanneer β = 0, betekent het dat de waarde van y constant is en niet verandert wanneer x verandert. Dit kan betekenen dat de twee variabelen onafhankelijk van elkaar zijn.

De best passende regressielijn

Bij regressieanalyse beschouwen we a en b als onbekende parameters, die we gaan schatten aan de hand van de data. De eerste stap hierbij is het plotten van de data in een scatterplot. Zo kun je zien of het wel logisch om een lineaire formule te gaan maken. Wanneer de data immers een U vorm heeft, heeft het geen zin om daar een lineaire lijn door te trekken.

Omdat we y gaan benaderen met de regressie-analyse en het een schatting is, geven we dit aan met een dakje boven de y: $\hat{y}$. De regressie formule ziet er zo uit: $\hat{y}$ = a + b(x). Deze lijn zal de 'beste' lijn weergeven, in de zin dat deze het dichtste ligt bij alle datapunten. Dit wordt later toegelicht.

Effecten van outliers

Een regressie outlier (uitschieter) is een datapunt dat ver buiten de trend van de andere datapunten valt. Zo'n datapunt wordt invloedrijk genoemd wanneer het verwijderen ervan een grote verandering teweeg brengt in de regressieformule. Dit effect is kleiner bij een grote dataset. Het is soms beter om deze outliers te verwijderen.

Residuen

De regressieformule geeft een schatting van de y-waarden. Deze zullen niet helemaal overeenkomen met de daadwerkelijke (geobserveerde) y-waarden. Door het verschil tussen de geschatte waarden en de geobserveerde waarden te bekijken, kun je zien hoe 'goed' de regressielijn is. Het verschil tussen deze twee heet een residu. Het is het verschil tussen een geobserveerde waarde (y) en de voorspelde waarde ($\hat{y}$). Wanneer de geobserveerde waarde groter is dan de voorspelde waarde heb je een positief residu. Wanneer de geobserveerde waarde

kleiner is dan de voorspelde waarde heb je een negatief residu. Hoe kleiner de absolute waarde van het residu, hoe beter de voorspelling, en dus de regressielijn.

Sum of Squared Errors/Residual Sum of Squares

De beste regressielijn is die met de kleinste residuen. Om die te vinden, worden de residuen van de datapunten gekwadrateerd en opgeteld. Dit noemen we de SSE. De SSE staat voor 'sum of squared errors'. De formule is $SSE = (y - \hat{y})^2$. De beste regressielijn heeft de kleinste SSE van alle andere mogelijke lijnen. De SSE van de beste regressielijn heeft zowel negatieve als positieve residuen (die door het kwadrateren allemaal positief worden), waarvan samen de som en het gemiddelde 0 zijn. Daarbij loopt de regressielijn altijd door het punt van het gemiddelde van x en het gemiddelde van y, dus door het punt ($\bar{x}$,$\bar{y}$).

Het lineaire regressie model

Bij een regressieformule y = a + b(x) hoort bij elke x-waarde eenzelfde y-waarde. Dit heet een deterministisch model. Zo werkt het in de werkelijkheid niet. Stel bijvoorbeeld dat we inkomen (y) willen voorspellen aan de hand van opleidingsniveau (x), dan zien we dat niet iedereen met dezelfde opleiding ook hetzelfde inkomen heeft. In plaats van een deterministisch model kun je dan beter gebruik maken van een probabilistisch model. Deze ‘conditionele distributie’ refereert naar de variabiliteit in de y-waarden op een vaste waarde voor x.

Daarom veranderen we de formule nu naar E(y) = a + b(x). Hierbij geeft E(y) aan dat we kijken naar de conditionele distributie van y, en we proberen het gemiddelde van y te voorspellen.

Variatie op de regressielijn

Het lineaire regressiemodel kent nog een parameter, namelijk σ. Deze beschrijft de standaard afwijking van elke conditionele distributie. Het meet de variabiliteit van de y-waarden voor alle personen met die bepaalde x-waarde. We noemen σ de conditionele standaarddeviatie.

Omdat we deze echte standaardafwijking niet weten, gebruiken we wat we weten uit de steekproef, namelijk ‘s’. De formule van ‘s’ is $s = \sqrt{\frac{SSE}{n-2}}$, waarbij $SSE = \sum (y - \hat{y})^2$.

Als je deze ‘s’ kwadrateert heb je de zogenaamde ‘Mean Square Error’ of MSE.

Gestandaardiseerde regressie coefficient / Pearson correlatie

Nu gaan we kijken hoe sterk het eventuele verband is tussen x en y. We gebruiken daarvoor een gestandaardiseerde versie van correlatie en geven deze met 'r' aan. Deze ‘r’ wordt ook wel de gestandaardiseerde regressie coëfficiënt, of Pearson correlatie genoemd. Deze wordt berekend als volgt:

$$r = (\frac{s_x}{s_y})\, b$$

Hierbij is de steekproef deviatie van ‘x’ en S_y de steekproef deviatie van ‘y’. De formules van S_x en S_y , zijn als volgt:

$$S_x = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}} \quad \text{en} \quad S_y = \sqrt{\frac{\sum (y - \bar{y})^2}{n-1}}$$

Deze correlatie heeft een aantal kenmerken:

- Je kunt de correlatie alleen gebruiken wanneer een lineair verband zinvol is.
- ‘r’ valt tussen 1 en -1.

- ‘r’ is positief/negatief gelijk aan ‘b’. Als ‘b’ positief is (en er een positief verband is) is ‘r’ ook positief en als ‘b’ negatief is (en er een negatief verband is) is ‘r’ ook negatief.
- Hoe groter ‘r’, hoe sterker het lineaire verband.

r-kwadraat

De 'coëfficiënt ofdetermination' of r^2 is gerelateerd aan ‘r’ en geeft aan hoe goed y voorspeld kan worden door x. r-kwadraat heeft een aantal kenmerken die sterk overeenkomen met r:

- Omdat r tussen 1 en -1 valt, moet r^2 wel tussen 0 en 1 liggen.
- Als SSE = 0, dan r^2 = 1. Alle punten moeten op de lijn vallen.
- Als b = 0, r = 0, r^2 = 0.
- Hoe groter r^2, hoe sterker het lineaire verband.

Significantie toetsen bij regressie

t-toets voor onafhankelijkheid

Het principe bij deze test is hetzelfde als die bij de Chi-kwadraattoets, namelijk kijken of de variabelen onafhankelijk zijn. Je gaat ervan uit dat het zo is, dus je H0: $\beta = 0$ en Ha: $\beta \neq 0$. Bij het doen van deze test wordt er van uitgegaan dat er aan een aantal assumpties is voldaan:

- Randomisatie
- Het gemiddelde van y wordt benaderd door de formule E(y) = a + b (x)
- De conditionele standaard deviatie is gelijk voor elke x-waarde
- De conditionele distributie van y voor elke x-waarde is normaal verdeeld

Manier 1:

De t-score wordt berekend door b te delen door de standaardfout van b. De formule voor t is $t = \frac{b}{se}$.

De opbouw van deze formule is niet belangrijk. De vorm van de formule is gelijk aan die van elke t-score. Namelijk de schatting minus de nulhypothese (die hier 0 is, en dus gewoon verdwijnt), gedeeld door de standaardfout van de schatting.

Voor het opzoeken van de p-waarde gebruik je df = n – 2.

Manier 2:

Je kunt de t-score ook berekenen met deze formule:

$$t = \frac{r}{\sqrt{(1 - r^2)/(n - 2)}}.$$

Er zal hetzelfde getal uitkomen als bij manier 1. Hiermee test je of de correlatie die is gevonden significant is, en daarmee test je dus ook of de variabelen onafhankelijk van elkaar zijn.

Betrouwbaarheidsinterval

Je kunt ook een betrouwbaarheidsinterval voor de hellingscoëfficiënt maken. Je doet dan de hellingscoëfficiënt plus en min de waarde van t, vermenigvuldigd met se.

$$\text{B.I. voor } \beta = b \pm t\,(se)$$

Wat je invult voor t is afhankelijk van de precisie en zekerheid die je wilt gebruiken. Je moet dan wel opletten dat je voor df = n – 2 gebruikt.

Assumpties

Het gemiddelde van y kan worden benaderd met een lineair model

Het is belangrijk dat je altijd eerst een scatterplot maakt om te kijken of het wel zinvol is om een lineair model te maken voor jouw onderzoek. Als je dit niet doet, kun je het gevaar lopen een lineair verband te ontdekken in data die helemaal niet lineair is. Als je data bijvoorbeeld een U vorm heeft, kan SPSS alsnog een lineair verband ontdekken: de y verandert immers als de x verandert.

Niet extrapoleren

Het is niet verstandig om je regressieformule te gebruiken voor zelfbedachte datapunten buiten je dataset. Dit omdat het verband wellicht bij die datapunten helemaal niet meer lineair is, of omdat de schattingen van y dan niet realistisch zijn.

Outliers

Sommige outliers kunnen grote effecten hebben op de regressielijnen en de correlaties. Het is soms nodig dat je bepaalde outliers eruit haalt om nog een keer je analyses te lopen. Één punt kan al veel invloed hebben, in het bijzonder bij een kleine steekproef.

Steekproefgrootte

Bij een kleine steekproef heb je minder variatie in je x-waarden dan bij een grote steekproef. Je loopt dan het risico dat je niet een realistische dataset hebt verkregen, en dat je analyses niet veel waard zijn. Je kunt correlaties het beste gebruiken bij datasets die willekeurig verkregen zijn, en representatief zijn wat betreft de variatie in de x- en y-waarden.

Tabel: Onafhankelijkheidstesten en associatiematen

Meetniveau			
	Nominaal	**Ordinaal**	**Interval**
Nulhypothese	H0: onafhankelijkheid	H0: onafhankelijkheid	H0: onafhankelijkheid (b = 0)
Test statistiek	$X^2 = \sum \frac{(f_o - f_e)^2}{f_e}$	$z = \frac{\hat{y}}{se}$	$t = \frac{b}{se},\ df = n - 2$
Meting samenhang	$\hat{\pi}_2 - \hat{\pi}_1$ Odds ratio	$\hat{\gamma} = \frac{C - D}{C + D}$	$r = b(\frac{s_x}{s_y})$ $r^2 = \frac{TSS - SSE}{TSS}$