Hoorcollege 6

Regressie

Het gaat in dit hoofdstuk over lineaire relaties. Aan de hand van het beschrijven van deze relaties kunnen er voorspellingen gemaakt worden. Bij een correlatiestudie wordt gekeken of er een relatie is tussen twee variabelen. Dit wordt gebruikt bij de verificatie van theorieën, betrouwbaarheid, validiteit en om te voorspellen. Validiteit voorspellen wordt regressie genoemd. De vraagstelling die hierbij hoort is: kan de waarde van iemand op een bepaald kenmerk voorspeld worden met kennis over andere kenmerken? We gaan dus kijken of we de relatie tussen twee variabelen kunnen beschrijven, zodat we voorspellingen kunnen maken over een bepaalde variabelen.

Een regressieanalyse heeft duidelijke rollen wat betreft de variabelen. De afhankelijke variabele (Y), dat onderzoekt de regressieanalyse, met een minimaal meetniveau interval. De onafhankelijke variabele (X) heeft een gebruikelijk meetniveau van interval. In regressie wordt de onafhankelijke variabele vaak de predictor genoemd.

Stap 1 het spreidingsdiagram

De eerste stap is net als bij correlatie, eerst een spreidingsdiagram maken. Het verschil tussen het gebruik van een correlatieanalyse of een regressieanalyse is dat bij een correlatieanalyse er een relatie is tussen de twee variabelen en dat de variabelen geen specifieke rol spelen. Bij een regressieanalyse doe je een voorspelling van een variabele (afhankelijk) aan de hand van de andere variabele (onafhankelijk). Er moet altijd gekeken worden naar de vraagstelling, want daar ligt het aan voor welke analyse je gebruikt. Je moet dus goed weten wat de afhankelijke, en wat de onafhankelijke variabele is. De onafhankelijke variabele ligt altijd op de x-as. De afhankelijke variabele altijd op de y-as.
De lijn die je ziet in het spreidingsdiagram zorgt ervoor dat het lineaire verband beter zichtbaar is. Ook geeft deze lijn het gemiddelde aan en kunnen we er voorspellingen mee doen.
Voordat je een voorspelling kunt doen, moet er eerst een vergelijking opgesteld worden.
De lineaire relatie die tussen twee variabelen wordt beschreven als $Y = bX + a$. b noemen we de richtingscoëfficiënt, of de regressiecoëfficiënt. De a is de Y-intercept, waar snijdt de lijn de y-as als x nul is? In SPSS wordt dit de constante genoemd.
In de statistiek bepalen we de vergelijking die de ‘gemiddelde’ relatie tussen twee variabelen X en Y beschrijft. We gaan dan op zoek naar de vergelijking die het beste bij de data past. We moeten kunnen meten hoe goed de lijn bij zo’n puntenwolk past. De techniek die hiervoor gebruikt wordt is de Least Squares Regression.
Least betekent het minste en Squares betekent kwadratensom. We gaan de regressie opstellen door de kleinste kwadratensom.

Vervolgens moet er een bepaalde definitie gemaakt worden welke lijn in de puntenwolk het beste past bij wat je wilt weten. Dan kan de afstand gemeten worden tussen de puntenwolk en de lijn. De afstand tussen een observatie en de lijn heet residu. Het gaat om de verticale afstand. Deze afstand wordt ook wel de schattingsfout genoemd. Als een punt onder de lijn ligt is het residu negatief. Als een punt echter boven de lijn ligt, dan is het verschil positief. Als je een lijn hebt die niet goed bij de residuen past, is de som van alle residuen groot. Als er sprake is van positieve en negatieve residuen, worden deze gekwadrateerd en vervolgens bij elkaar opgeteld. De vergelijking die de kleinste kwadratensom van residuen oplevert past het beste bij de data, de standaardfout is daar het kleinst. Die vergelijking die de kleinste kwadratensom van de residuen oplevert is de “winnaar”.
Dezelfde formule wordt genoemd: $\hat{Y} = bX + a$. Dit wordt uitgesproken als ‘Y-hat’ of Y-dakje. Een dakje geeft een schatting aan. Als je een formule ziet waar in plaats van een a en een b een alfa en een bèta staat, gaat het over de populatie. Dit wordt in deze cursus niet meer gebruikt, hier wordt slechts gekeken naar de steekproef.

Hierbij zijn een paar formules belangrijk:

- Richtingscoëfficiënt: $b = SP / SS_X$ of $b = r$ x (s_y/s_X). S is hierbij de standaarddeviatie.
- Residu: Het residu is $Y – \hat{Y}$ ofwel geobserveerde Y – voorspelde Y.
- $a = M_Y – bM_X$

Hoe kom je aan de richtingscoëfficiënt?

Dit wordt berekend aan de hand van de productensom gedeeld door de sum of squares van x. Als de Y-intercept uitgerekend moet worden gebruik je: M van y – bM van x. Dit is het resultaat van het punt (M van x, M van y) dat altijd precies op de regressielijn valt.
Er is echter ook een alternatieve formule die gebruikt maakt van de standaardafwijkingen. De ratio daarvan moet vermenigvuldigd worden met de gemiddelde standaardafwijking van Y en X. Ofwel: $b = r * (S_Y / S_X)$. Met deze formule kun je goed het nauwe verband tussen correlatie en regressie zien.

SPSS

In de eerste tabel staan alleen de variabelen die in de vergelijking komen en wat de onafhankelijke en afhankelijke variabele is.
Het eerste wordt gekeken naar de onderste tabel. Daar staat alle informatie over de coëfficiënten. Onder de tabel staat nog een keer de afhankelijke variabele. De onafhankelijke variabele staat in de tabel. In de kolom van B staan de waarde van *a* en *b* (richtingscoëfficiënt en het snijpunt met Y). De richtingscoëfficiënt komt dus voor de X waarde (de onderste rij). De bovenste rij is de constante en dus de waarde van *a*. Dit is het snijpunt met de Y as. De bovenste is dus *a* en de onderste is de waarde van *b* in de kolom B.

In het voorbeeld:
Voorspelde studiedruk = -0,7171 * gemiddeld cijfer + 7,834.

Voorspellen

De eerste stap van het voorspellen van de Y-score is om de X-waarde van die individu in de regressievergelijking te stoppen. Dit betekent dat je een voorspelling van de Y-waarde van een individu krijgt bij een bepaalde X. Het kan ook zijn dat het een schatting is van het gemiddelde van Y voor iedereen bij een bepaalde X.
Let op! De waarde van Y-dakje is dezelfde waarde maar de betekenis is anders. Let dus goed op dat deze waarde op twee manieren geïnterpreteerd kan worden. Bovendien moet je geen voorspellingen doen buiten het bereik van X (extrapoleren), dus niet buiten de waarden om die je geobserveerd hebt.

Er bestaat ook een gestandaardiseerde vorm van regressievergelijking, bij verschillende variabelen die gestandaardiseerd zijn. Hierbij moeten de z-scores gebruikt worden, waarbij alle eenheden vervallen. Alle variabelen krijgen dan dezelfde eenheden, waardoor de x-variabelen vergeleken kunnen worden. Hierbij moeten eerste de X en de Y gestandaardiseerd worden. Daarna standaardiseren (de z-scores uitrekenen). Hierbij geven we bèta als naam aan de gestandaardiseerde richtingscoëfficiënt. Bèta is dus de correlatiecoëfficiënt.
De gestandaardiseerde regressie is $z_{\hat{Y}}$ = beta * z_X. Bèta is gelijk aan *r*.

Residuen en standaardschattingsfout

Bij weinig spreiding rond de regressielijn zijn de residuen klein. De voorspellingen die we maken met betrekking tot de regressievergelijking zijn dan veel nauwkeuriger. Bij veel spreiding rond de regressielijn zijn de voorspellingen die we maken minder nauwkeurig. De spreiding rond de regressielijn wordt gemeten door een standaardafwijking. Deze standaardafwijking noemen we de standaardschattingsfout (standard error of the estimate). Dit kan gerapporteerd worden, waardoor je kan zien of de voorspelling nauwkeurig is. Aan de hand van de standaardschattingsfout kun je bijvoorbeeld kijken welk van twee modellen het meest nauwkeurig is. De modellen moeten dan wel hetzelfde meten. Deze standaardschattingsfout staat letterlijk in de SPSS-output, en wordt ook wel de standard error of the estimate genoemd. Het is in feite de standaardfout van de residuen.

Voor de standaardschattingsfout gebruiken we weer een kwadratensom gedeeld door de vrijheidsgraden. De kwadratensom is $SS_{residual}$. De $df_{residual}$ zijn n-2. De standaardschattingsfout is dan de wortel van de $SS_{residual}$ gedeeld door $df_{residual}$.
Dan wordt in de SPSS-output in de middelste twee tabellen gekeken (slide 39). De standaardschattingsfout staat rechtsboven (in het voorbeeld is het 0,3884). Er staat een ANOVA tabel omdat we in principe hetzelfde doen. Daar staan de SS van de residual. In dit geval is het 0,603. Daarnaast staan de vrijheidsgraden en als we de SS delen door de *df* zien we de MS. De wortel van MS is de standaardschattingsfout. Wanneer je de standaardschattingsfout in het kwadraat doet krijg je de MS.

Effectgrootte

Ook bij de regressie kijken we naar de effectgrootte. De effectgrootte bepalen we aan de hand van de proportie verklaarde variantie. Dit wordt gemeten door r^2. r^2 is het kwadraat van de correlatie tussen X en Y. r^2 heet ook wel de 'coëfficiënt of determination' en meet de proportie van de totale spreiding van de Y die verklaard wordt door de lineaire relatie met X. ($1-r^2$) meet dan dus de proportie van de spreiding dat niet verklaard wordt door de regressie relatie. De totale spreiding van Y = SS_y, of SS_{total}. Het stuk spreiding dat verklaard wordt door het regressiemodel is $SS_{regression}$. Dan wordt het dus de volgende formule:
$r^2 = SS_{regression} / SS_Y = SS_{regression} / SS_{total}$.
Deze laatste formule staat niet goed in het boek en ook niet op het formuleblad. De twee formules van het onverklaarde stuk spreiding en het verklaarde stuk spreiding staan wel op het formuleblad. Van deze formules kun je de effectgrootte zelf afleiden. De gegevens die je met deze formule uitrekent staan echter ook vaak gewoon in de SPSS-output. Op slide 44 is de SPSS-output te zien. De R is de correlatie in absolute waarde.

Toetsing

Tot slot moet getoetst worden of de relatie wel significant is. Dit is eigenlijk hetzelfde als kijken of de X-variabele wel goed functioneert. Wordt er wel een significant deel van de spreiding verklaard? Dit kan op twee manieren getoetst worden. Een toets voor de richtingscoëfficiënt of een toets voor een significante relatie.

De eerste is kijken of de richtingscoëfficiënt in de populatie (bèta) gelijk is aan nul of niet en kan met een gewone t-toets. De p-waarde is ook hetzelfde als voor de correlatietoets.
H0: $\beta = 0$
H1: $\beta \neq 0$ (voor de richtingscoëfficiënt in de populatie).
t = b – bs / Sb = b / sb.
Alleen er zijn n – 2 vrijheidsgraden. Deze t-toets wordt altijd uit de SPSS-output gehaald en dit is te zien in slide 50.
De p waarde is 0,032 en het is dus significant.

De tweede manier van toetsen is de toets voor een significante relatie. Deze toetst of een significant deel van de spreiding in Y wordt verklaard door de lineaire relatie. Deze toets wordt ook wel de regressieanalyse genoemd en is eigenlijk hetzelfde als de variantieanalyse, de F-toets dus. Hierbij wordt de totale spreiding verdeeld onder de regressie (verklaard deel) en de residuen (onverklaard deel). De totale spreiding wordt dan gesplitst en een voorbeeld hiervan is te zien op slide 52. In een enkelvoudige regressie zijn de t-toets en de F-toets identiek.