13. Combining Regression and ANOVA: Quantitative and Categorical Predictors

Multipele regressie kan ook gedaan worden met zowel een kwantitatieve als een categorische variabele tegelijk. Hierbij wordt gewone regressie-analyse (voor de kwantitatieve variabele) gecombineerd met variantie-analyse (voor de categorische variabele).

In veel studies is het zinvol om te controleren voor een kwantitatieve variabele. Bijvoorbeeld wanneer je het inkomen voor mannen en vrouwen wilt vergelijken, is het aantal jaren werkervaring een zinvolle controlevariabele. Zo'n kwantitatieve controle variabele noemen we dan een covariaat. Bijbehorende regressie noemen we dan 'covariantie-analyse'.

Vergelijken van gemiddelden en regressie lijnen

In dit hoofdstuk wordt een kwantitatieve predictor aangegeven met x, en een categorische predictor met z.

Het effect van x op y, gecontroleerd voor z

Stel we hebben een afhankelijke variabele y: de prijs van een huis. We hebben een kwantitatieve predictor x: de grootte van het huis. En een categorische predictor z: het huis is nieuw (1 = ja, 2 = 0). We kunnen het effect van x op y, gecontroleerd voor z, onderzoeken in een grafiek. Je krijgt dan twee lijnen. Er wordt een lijn getrokken van de invloed van x op y voor nieuwe huizen, en een lijn voor oude huizen. Deze grafiek zou er zo uit kunnen zien:

Wanneer er in dit model geen sprake is van interactie, betekent dat, dat een toename op x hetzelfde effect heeft op y wanneer een huis oud is (z = 0) als wanneer een huis nieuw is (z = 1). In de eerste grafiek is dat zo. Er is daar wel een verschil te zien in prijzen voor oude en nieuwe huizen, maar er is geen sprake van een interactie-effect. In de middelste grafiek is er ook geen sprake van interactie, en is er ook geen verschil tussen oude en nieuwe huizen. In de rechter grafiek is er wel sprake van een interactie-effect. Een toename op x leidt immers tot een andere waarde op y voor nieuwe huizen, dan voor oude huizen.

In dit model wordt het effect van een kwantitatieve variabele op y, gecontroleerd voor een categorische variabele getoetst.

Het effect van z op y, gecontroleerd voor x

Andersom kan ook. Het effect van z kan ook veranderen wanneer er gecontroleerd wordt voor x. Bijvoorbeeld wanneer je kijkt naar het jaarinkomen (y) van mannen en vrouwen (z), waarbij je controleert voor aantal jaren werkervaring (x). De grafieken kunnen er met betrekking tot interactie hetzelfde uitzien als bovenstaand. Alleen de middelste grafiek moet dan anders worden geïnterpreteerd: daarbij zullen vrouwen voornamelijk rondom het onderste deel van de lijn gecentreerd zijn, en de mannen voornamelijk rondom het bovenste deel van de lijn.

Regressie met kwantitatieve en categoriale verklarende variabelen

Eerst wordt er gekeken naar regressie zonder interactie. In dit model zitten twee verklarende variabelen (een kwantitatieve variabele x, en een categorische variabele met drie categorieën, dus twee dummyvariabelen z1 en z2). De regressieformule ziet er dan zo uit:

$E(y) = a + b(x) + b1(z1) + b2(z2)$. Hierbij is de 'b' van 'b(x)' het effect van x op het gemiddelde van y voor alle categorieën.

De coëfficiënten en dergelijke zijn gegeven bij dit voorbeeld, en ingevuld:

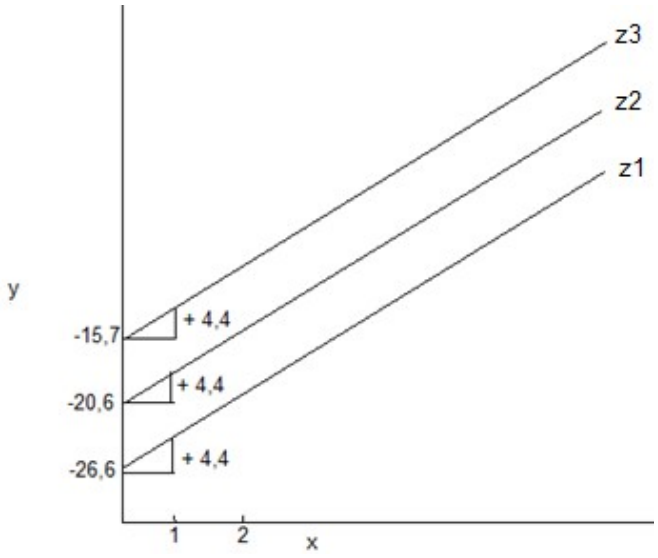
$E(y) = -15,7 + 4,4(x) – 10,9(z1) – 4,9(z2)$

Hiermee kunnen drie lijnen in een grafiek worden getrokken, namelijk voor elke categorie één.

De lijn van z1: E(y) = -15,7 + 4,4(x) – 10,9 = -26,6 + 4,4(x)

De lijn van z2: E(y) = -15,7 + 4,4(x) – 4,9 = -20,6 + 4,4(x)

De lijn van z3: E(y) = -15,7 + 4,4(x)



Coëfficiënten interpretatie

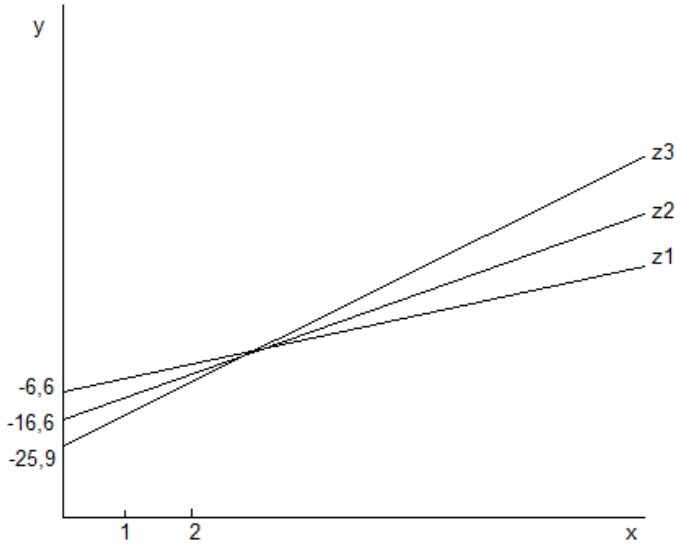
Wanneer x met 1 punt omhoog gaat, gaat de y met ‘b’ (=4,4) omhoog. Deze helling is voor elke categorie gelijk, daarom lopen de lijnen ook parallel.

In de grafiek wordt ook goed duidelijk hoe de coëfficiënten de verschillen in de gemiddelden weergeven. Want tussen z1 en z3 zit precies 10,9 (=b1) verschil. Tussen z2 en z3 zit precies 4,9 (=b2) verschil. Tussen z1 en z2 zit precies 6 (=b1-b2) verschil. Ofwel:

B1 is het verschil tussen z1 en z3 (de referentiecategorie). B2 is het verschil tussen z2 en z3 (de referentiecategorie). De coëfficiënten geven dus het verschil aan in gemiddelde tussen de bijbehorende categorie en de referentiecategorie.

Toevoegen van interactie bij regressie met kwantitatieve en categoriale variabelen
Meestal is er echter wel sprake van interactie in een model. Als we dat toevoegen aan ons model (zonder getallen), dan ziet het regressiemodel er zo uit:
E(y) = a + b(x) + b1(z1) + b2(z2) + b3(z1*x) + b4(z2*x).

De bijbehorende coëfficiënten zijn gegeven:
E(y) = -25,9 + 5,2(x) + 19,3(z1) + 9,3(z2) – 2,4(z1*x) – 1,1(z2*x).



Nu kunnen er weer drie lijnen in een grafiek getrokken worden. Nu zullen ze echter niet parallel lopen, omdat de hellingen niet meer hetzelfde zijn. De formules voor elke lijn (van elke categorie) zijn:
De lijn van z1: E(y) = -25,9 + 5,2(x) + 19,3 – 2,4(x) = -6,6 + 2,8(x)
De lijn van z2: E(y) = -25,9 + 5,2(x) + 9,3 – 1,1(x) = -16,6 + 4,1(x)
De lijn van z3: E(y) = -25,9 + 5,2(x)

Coëfficiënten interpretatie
Nu is de interactie goed te zien: een toename op x heeft voor alle drie de groepen een ander effect op y. Daarbij kunnen we ook niet zo makkelijk meer iets zeggen over de coëfficiënten die het verschil zijn tussen de geschatte gemiddelden van y. Want in de grafiek is te zien, dat hoe groter x wordt, hoe groter het verschil tussen de gemiddelden wordt.

Om te kijken of het model met interactie nu beter is dan het originele model, kan je kijken naar de toegenomen waarde van R^2.

Inferentie voor regressie met kwantitatieve en categorische variabelen
Nu gaan we significantie testen doen voor regressie met kwantitatieve en categorische variabelen, en daarna worden er schattingsmethoden voor covariantie modellen besproken. Eerst moet er getest worden of er sprake is van interactie.
De hypothesen hebben betrekking op de complete en gereduceerde modellen, en worden gedaan met de F-statistiek. Deze kan op twee manieren worden berekend:

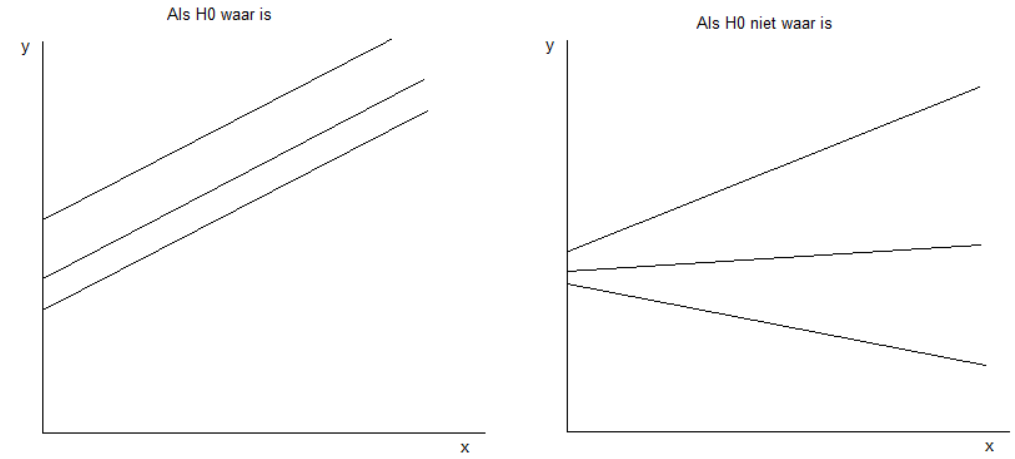
$F=SSEr-SSEc/df1SSEc/df2$ of $Rc2-Rr2/df1-Rc2/df2$.

Hierbij zijn de Residual Sum of Squares van het complete en het gereduceerde model, df1 is het verschil in aantal termen tussen het complete en gereduceerde model, en df2 is de waarde van de ‘mean square’ van het complete model. Deze statistieken worden gewoon gegeven in SPSS, in de tabel ANOVA.

Het toetsen van interactie
Hierbij is de nulhypothese dat er geen interactie is, dus dat het complete model niet beter is dan het gereduceerde model zonder interactietermen.

Het complete model ziet er zo uit : E(y) = a + b(x) + b1(z1) + b2(z2) + b3(x*z1) + b4(x*z2).
Wanneer we er van uit gaan dat er geen interactie is, gaan we er van uit dat b3 en b4 allebei 0 zijn. We moeten testen of deze b3 en b4 significant zijn.

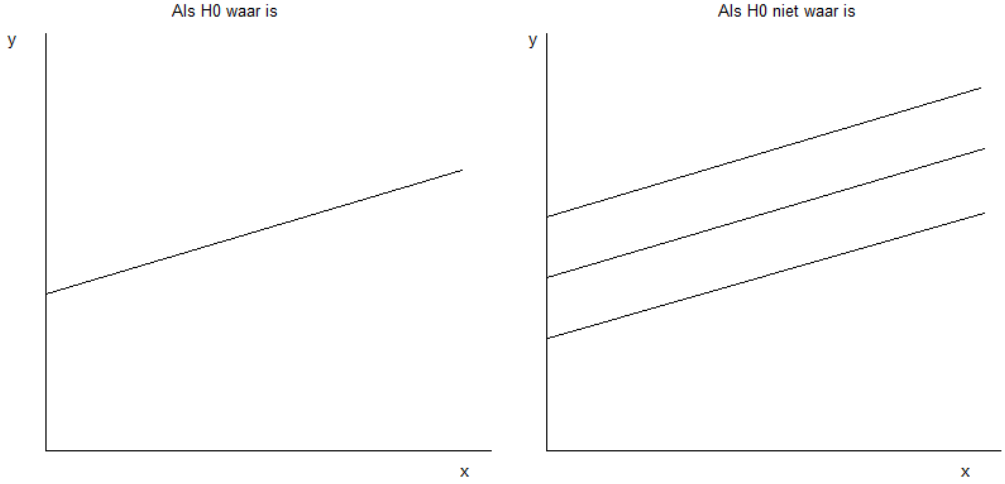
Wanneer er geen sprake is van interactie (zoals H0 voorspelt), lopen de regressielijnen parallel aan elkaar. Als je dat grafisch weergeeft:



Je ziet hier dat er als H0 waar is, geen interactie-effect is. Dat betekent dat een toename op x, hetzelfde effect op y heeft voor elke categorie. Als er wel een interactie-effect is, dan heeft een toename op x voor elke categorie een ander effect op y.

Het toetsen van het effect van de categorische variabele, gecontroleerd voor x
X is hier een kwantitatieve variabele, en de z1 en z2 zijn hier twee (van de drie) dummies voor de categorische variabele. Als we willen testen wat het effect van deze categorische variabele is, moeten we dus testen of de coëfficiënten van z1 en z2 significant zijn.

Als de categorische variabele geen effect heeft op y, dan zullen de regressielijnen van de categorieën niet (significant) van elkaar verschillen. Ze zouden in feite op elkaar kunnen liggen, of in ieder geval heel dicht bij elkaar. Als je dat grafisch weergeeft:



Je ziet hier dat als H0 waar is, de categorische variabele geen invloed heeft op y. Dit betekent dat alle lijnen op elkaar liggen, want de groepen verschillen niet van elkaar. Als de groepen wel van elkaar verschillen, dan is de categorische variabele wel een belangrijke voorspeller voor y, want blijkbaar heeft niet elke waarde op x dan eenzelfde waarde op y voor elke groep.

Het toetsen van het effect van x, gecontroleerd voor de categorische variabele
Voor deze toets stelt H0 dat de coëfficiënt van x 0 moet zijn. Dat betekent dus dat x geen invloed heeft op de waarde van y. Als x geen invloed heeft op y, dan zal de regressielijn geen helling hebben, want als x hoger wordt, zal dat geen effect hebben op y. Als x wel invloed heeft op y, heeft de lijn dus wel een bepaalde helling (dan wel positief of negatief).

Het is belangrijk dat je de grafieken leert interpreteren, om zo een duidelijker beeld te krijgen van wat er gebeurt in je regressiemodel en welke invloed de variabelen hebben.