11. Multipele Regressie en Correlatie

Meervoudig regressie model

Nu gaan we kijken naar een relatie tussen een responsvariabele en meerdere verklarende variabelen. Een bivariate regressielijn ziet er in formule zo uit: E(y) = a + b(x). Deze regressielijn geeft informatie over één verklarende variabele. Maar omdat wij kijken naar meerdere verklarende variabelen, moeten we deze toevoegen aan de formule. Om de verschillende variabelen te onderscheiden geven we een nummer aan de 'x'-waarden: X1, X2, X3, etc. Om de verschillende coëfficiënten aan te geven, voegen we daar ook nummers aan toe. Een multipele regressieformule ziet er dan zo uit: E(y) = a + b1(X1) + b2(X2).. (etc.)

Stel dat we kijken naar een relatie tussen misdaadpercentage en opleidingsniveau. Bij deze relatie hoort een bivariate formule : E(y) = -51,3 + 1,5(opleiding/X1). Dit is een positieve relatie: wanneer opleiding omhoog gaat, gaat ook het misdaadpercentage omhoog.

We gaan nu ook urbanisatie toevoegen aan de formule, want we weten dat dit sterk gecorreleerd is aan het misdaadpercentage. Nu krijgen we de multipele formule: E(y) = 58,9 – 0,6(opleiding/X1) + 0,7(urbanisatie/X2). Nu zien we een negatieve relatie tussen opleiding en misdaadpercentage. Er is sprake van Simpson's paradox: de bivariate formule geeft een andere richting aan dan de multipele formule.

In de multipele formule zien we dat het effect van opleiding -0,6 is, wanneer gecontroleerd voor urbanisatie. Controleren voor urbanisatie betekent dat we de waarde van urbanisatie constant houden. Stel dat we voor urbanisatie het gemiddelde invoeren, namelijk 50. De formule wordt dan E(y) = 59,9 – 0,6(X1) + 0,7(50) = 58,9 – 0,6(X1) + 35. Door te controleren voor urbanisatie, verandert de invloed van opleiding niet, alleen de helling verandert. Andersom, wanneer gecontroleerd voor opleiding, geldt hetzelfde voor urbanisatie.

We noemen deze formule (E(y) = 58,9 – 0,6(X1) + 35) een partiële regressie formule, omdat deze formule maar naar een deel van de mogelijke observaties kijkt (namelijk alleen naar die gevallen die een urbanisatieniveau van 50 hebben).

Dit is het basisverschil tussen multipele en bivariate regressie:
In multipele regressie geeft een coëfficiënt het effect van een verklarende variabele op een responsvariabele, wanneer gecontroleerd voor andere variabelen in het model.
Bij bivariate regressie geeft een coëfficiënt het effect van een verklarende variabele op een responsvariabele, terwijl alle andere mogelijke verklarende variabelen genegeerd worden.

De coëfficiënt van een predictor geeft aan wat de verandering is in het gemiddelde van y wanneer de predictor met een punt omhoog gaat, en gecontroleerd voor alle andere variabelen in het model. Deze coëfficiënten noemen we 'partiële regressie coëfficiënten'. De parameter 'a' geeft aan wat het gemiddelde is van y, wanneer alle andere verklarende variabelen 0 zijn.

Residuen

Net zoals bij het bivariate model, gebruikt het meervoudige regressie model residuen om de voorspellingsfouten te meten. Voor iemand met een voorspelde response 'ŷ' en een gemeten respons 'y', is het residu het verschil tussen deze twee: y – ŷ. Wanneer we deze voor iedereen in de sample optellen, hebben we het 'SSE' (Sum of Squared Errors/Residual Sum of Squares). Deze vat samen hoe goed de regressielijn 'past' bij de data. De formule hiervoor is:
$\text{SSE} = \sum (y - \hat{y})^2$.

Het verschil wordt gekwadrateerd, zodat er geen negatieve waarden ontstaan. SPSS zal de lijn kiezen met de kleinste opgetelde residuen.

Multipele correlatie en R^2

R

Een ‘r’ en ‘r²’ beschrijft de sterkte van het lineaire verband van een bivariate relatie. Voor multipele regressie hebben we daarvoor ‘R’ en ‘R²’. Deze beschrijven de sterkte van het verband tussen y en de set van verklarende variabelen. Deze multipele correlatie voor een regressie model is de correlatie tussen de geobserveerde waarden en de voorspelde waarden. Deze correlatie beschrijven we met ‘R’. R valt altijd tussen 0 en 1. Hoe groter R, hoe beter de voorspelde waarden ($\hat{y}$) correleren met de geobserveerde waarden (y).

R²

Met R² kun je meten wat de relatieve ‘verbetering’ is van de predictoren aan het gemiddelde. Er zijn hierbij twee standaard regels: 1) Wanneer je y wilt voorspellen zonder de predictoren, dan is het gemiddelde y– de beste voorspeller; 2) Wanneer je y wilt voorspellen door middel van de predictoren, dan is de regressieformule de beste voorspeller.

De formule voor het berekenen van R² is:

$$R^2 = \frac{TSS - SSE}{TSS}$$

Hierbij is TSS de ‘total sum of squares’. De formule hier voor is :

$$TSS = \sum (y - \bar{y})^2$$

Dit is het verschil tussen het gemiddelde en een geobserveerde y-waarde. Dit wordt gekwadrateerd (om negatieve verschillen te voorkomen), en vervolgens opgeteld.
SSE is 'sum of squared error' of de 'residual sum of squares'. De formule hier voor is:

$$SSE = \sum (y - \hat{y})^2$$

Dit is het verschil tussen de voorspelde y en de geobserveerde y. Dit wordt gekwadrateerd (om negatieve verschillen te voorkomen), en vervolgens opgeteld.

R² meet de proportie van de totale variatie in y, die verklaard wordt door de verklarende variabelen (via het regressiemodel).

In SPSS: wanneer je dit zelf wilt berekenen in SPSS, moet je kijken in de tabel in de output ‘ANOVA’. De TSS is dan wat er achter ‘Total’ staat, en onder ‘Sum of Squares’. De SSE is dan wat er achter ‘Residual’ staat, en onder ‘Sum of Squares’.

Een aantal kenmerken van R²:

- Valt altijd tussen 0 en 1.
- Hoe groter de waarde van R², hoe beter de verklarende variabelen gezamenlijk y voorspellen.
- R²=1, alleen wanneer alle residuen 0 zijn. Dat is wanneer y = ŷ, waardoor SSE = 0. De regressielijn loopt dan precies door alle punten in de dataset.
- R²=0, betekent dat alle coëfficiënten 0 zijn. De formule heeft dus geen voorspellende waarde.
- R² kan niet verminderen wanneer er extra verklarende variabelen aan de formule worden toegevoegd.

Multicollineariteit met >1 verklarende variabele

Wanneer je veel verklarende variabelen hebt die sterk met elkaar gecorreleerd zijn, dan heeft R² minder toegevoegde waarde naar mate het aantal verklarende variabelen toeneemt. Dit betekent niet dat die variabelen y niet goed kunnen voorspellen, maar dat ze niet veel meer toevoegen gegeven de predictoren die al in de formule zitten. Dit noemen we multicollineariteit. Problemen met multicollineariteit zijn minder groot bij grotere steekproeven.

Ideaal zou zijn als de steekproef minstens tien keer zo groot is als het aantal predictoren (bijv. bij vier predictoren ten minste 40 mensen in je steekproef).

Significantie toetsen bij multipele regressie, F-toets

Bij multipele regressie worden er twee significantie toetsen uitgevoerd. De eerste heeft betrekking op het hele model, en kijkt of één van de predictoren statistisch gerelateerd is aan y. De tweede heeft betrekking op de afzonderlijke predictoren en bekijkt welke predictoren een significant effect geven. De tweede toets wordt niet besproken. Voor het uitvoeren van die analyses is het berekenen van een standaardfout vereist. Die berekening wordt niet handmatig uitgevoerd, maar met SPSS berekend. SPSS berekent op deze manier direct of de toets significant is.

Toets 1: het hele model

Om te toetsen of een van de variabelen een significant effect heeft op y, toets je de nulhypothese die stelt dat er geen effect is. Alle coëfficiënten zijn gelijk aan 0.

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \ ... \ = 0$$

De alternatieve hypothese stelt dat ten minste één verklarende variabele gerelateerd is aan y.

$$H_a : \text{Ten minste één } \beta_i \neq 0$$

Om deze toets uit te voeren, maken we gebruik van de F-statistiek. De formule voor het berekenen van deze F-waarde is:

$$F = \frac{R^2 / k}{(1 - R^2) / [n - (k + 1)]}$$

N is hierbij de steekproefgrootte. ‘k’ is het aantal predictoren.

Kenmerken van de F-distributie:

- De distributie kent alleen positieve waarden.
- Rechts scheef
- Gemiddelde van ca. 1
- Hoe groter de R^2, hoe groter de F-statistiek (zie formule).
- Hoe groter de F-waarde, hoe groter de kans dat je H0 moet verwerpen.

De F-distributie is afhankelijk van twee soorten van vrijheidsgraden. Df1 = k (het aantal predictoren). Df2 = n – (k + 1).

Voorbeeld: Stel we hebben n = 40, twee verklarende variabelen (k = 2), en R^2 = .339.

$$F = \frac{R^2 / k}{(1 - R^2) / [n - (k + 1)]}$$

De F-waarde is F = .339/2.661/[40-2+1] = 9,5.

De vrijheidsgraden zijn df1 = 2, en df2 = 37. Nu kun je in de tabel van de F-distributie opzoeken wat de bijbehorende p-waarde is. De p-waarde is < .001 en dus significant.

Als je de F-waarde wilt berekenen met de gegevens die je in de ANOVA tabel vindt in de SPSS output, dan kun je kijken onder de kolom 'Mean Square'. Je deelt wat er bij regressie staat door datgene wat bij residuen staat.

Interactie effecten

Een multipele regressie formule gaat er van uit dat de lijnen van de verklarende variabelen parallel lopen aan elkaar, en samen een betere formule leveren voor het voorspellen van y. Maar vaak is het zo dat er interactie is tussen verklarende variabelen. Er is sprake van interactie tussen verklarende variabelen en hun effect op y wanneer het effect van een variabele verandert als de waarde van andere verklarende variabele verandert. Dit kan worden opgelost door het gebruik van 'kruisproducten' (cross-product terms). Het model ziet er zo uit (bij interactie tussen x1 en x2):

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

Je kunt toetsen of er een interactie effect is. SPSS doet dit al wel voor je en deelt de coëfficiënt door de bijbehorende standaardafwijking. Als dat significant is, dan is er een interactie effect. Dit betekent dat het effect van de ene variabele onder meer afhangt van de waarde van de andere variabele.

Centreren

De coëfficiënten van verklarende variabelen zijn vaak niet zo heel erg nuttig, omdat ze alleen aangeven wat het effect is van die variabelen, gegeven dat de andere variabelen constant worden gehouden. Je kunt ze nuttiger maken door ze te centreren. Hierbij centreer je elke verklarende variabele rondom 0, door het gemiddelde er van af te trekken. We geven dit aan met een C in het symbool: $x_1^C = x_1 - \mu_{X_1}$, en $x_2^C = x_2 - \mu_{X_2}$. Als we dit invullen in de regressie formule krijg je :

$$E(y) = \alpha + \beta_1(x_1 - \mu_{x_1}) + \beta_2(x_2 - \mu_{x_2}) + \beta_3(x_1 - \mu_{x_1})(x_2 - \mu_{x_2})$$

Nu geeft de coëfficiënt van x1 (dus b1) aan wat het effect is van x1, wanneer x2 gemiddeld is.

Regressie modellen vergelijken

Het hoeft helemaal niet zo te zijn dat de hele uitgebreide modellen beter zijn dan de korte modellen. Je kunt met statistiek testen of een model significant beter is dan een ander model. We noemen een volledig model met alle predictoren een compleet model, en een model met alleen een paar van deze predictoren een gereduceerd model. Stel dat je bijvoorbeeld een compleet model hebt met drie verklarende variabelen en interactie-effecten:

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \beta_6 x_2 x_3$$

En je hebt een gereduceerd model:

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

De hypothese zal stellen dat het complete model niet beter is dan het gereduceerde model, en stelt dus $H_0: \beta_4 = \beta_5 = \beta_6 = 0$

De toetsingsgrootheid voor het vergelijken van twee regressie modellen, vergelijkt de SSE scores van beide formules. De SSE van het complete model wordt SSEc en de SSE van het gereduceerde model wordt SSEr. De SSE van het gereduceerde model zal groter zijn dan de SSE van het complete model, omdat die minder predictoren heeft en minder precies is en dus meer schattingsfouten maakt. De toetsingsgrootheid gebruikt dit verschil, dat ontstaat door het toevoegen van meerdere predictoren.

De toetsingsgrootheid is

$$F = \frac{(SSE_r - SSE_c) / df_1}{SSE_c / df_2} = \frac{(R_c^2 - R_r^2) / df_1}{(1 - R_c^2) / df_2}$$

Hier is df1 het aantal extra predictoren in het complete model, en df2 = n – (k + 1). Een groot verschil in de SSE zorgt voor een grotere F-waarde en een kleinere p-waarde, en dus meer bewijs tegen H0.