Hoofdstuk 11: Meervoudige regressie

Inleiding
In veel gevallen wordt variabele y beïnvloed door meerdere verklarende variabelen. Stel bijvoorbeeld dat je cijfers op een rekentoets wilt voorspellen. In dat geval kun je kijken naar verschillende verklarende variabelen: IQ, motivatie en werkhouding.

11.1 Multipele regressie

Meerdere voorspellers
Het simpele lineaire regressiemodel gaat ervan uit dat het gemiddelde van responsvariabele y afhangt van x. De bijbehorende formule is: $\mu_y = \beta_0 + \beta_1 x$. Als we echter gebruik maken van meerdere predictoren, dan gebruiken we een andere formule:

- $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$.

Bij simpele lineaire regressie is er maar één voorspeller, waardoor de observaties samengevat kunnen worden als (x_i, y_i). Als er meerdere voorspellers zijn, dan maken we gebruik van de notatie x_{ij}. In dit verband staat j voor de j-ste variabele en i voor het i-ste geval (case).

Regressielijn voor meerdere voorspellers
We combineren de regressielijn voor de populatie en de aannames over variantie om een meervoudig lineair regressiemodel te maken. De subpopulatie-gemiddelden gaan over het fit-gedeelte van het model. Het residu-gedeelte gaat over de variantie die niet verklaard kan worden aan de hand van het model. We gebruiken ook hier het symbool ε als we het hebben over in hoeverre een individuele observatie afwijkt van het subpopulatie-gemiddelde. Deze afwijkingen zijn normaal verdeeld met een gemiddelde van 0 en een onbekende standaarddeviatie die niet afhangt van de waarden van x. Dit zijn aannames die we kunnen verifiëren door de residuen te bestuderen.

- Het *statistische model voor multipele lineaire regressie* is: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i$.
- De *gemiddelde respons* μ_y is een lineaire functie van alle verklarende variabelen: $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$.
- De *afwijkingen* (ε_i) zijn normaalverdeeld met een gemiddelde van 0 en standaarddeviatie σ. We kunnen dit samenvatten als $N(0,\sigma)$. De parameters van het model zijn $\beta_0 + \beta_1$, $\beta_2, \ldots, \beta_p$ en σ.

Het schatten van parameters bij multipele regressie
Zoals bij simpele lineaire regressie maken we bij het schatten van parameters (β) gebruik van steekproefwaarden (b). De details zijn echter wat ingewikkelder.

- $b_0, b_1, b_2, \ldots, b_p$ worden gebruikt om $\beta_0, \beta_1, \beta_2, \ldots \beta_p$ te schatten.
- Voor de i-ste observatie is de voorspelde y ($\hat{y}_i$): $b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_p x_{ip}$.
- Het i-ste residu (e_i) is het verschil tussen de geobserveerde en de voorspelde respons= $y_i - \hat{y}_i$. Dit is hetzelfde als: $y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \ldots - b_p x_{ip}$.
- Vervolgens moet de volgende formule gebruikt worden: $\Sigma(y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \ldots - b_p x_{ip})^2$. Dit betekent dat alle residuen gekwadrateerd moeten worden om niet op 0 uit te komen.

- De parameter σ^2 wordt geschat aan de hand van s^2. We vinden s^2 als volgt: Σe^2_i /n-p-1. In deze formule staat n voor de steekproefgrootte en p voor het aantal predictoren. Om de standaarddeviatie (σ) te vinden trekken we de wortel uit s^2.

Betrouwbaarheidsintervallen voor multipele regressie
We kunnen betrouwbaarheidsintervallen berekenen en significantietoetsen uitvoeren voor de regressiecoëfficiënten van alle predictoren (β_j).

- Het betrouwbaarheidsinterval voor β_j is $b_j \pm t^* SE_{bj}$. In deze formule is SE_{bj} de standaardfout van b_j en t* is de waarde van t(n-p-1).
- Om de hypothese $\beta_j=0$ te toetsen berekenen we een *t-toets:*
- $t= b_j / SE_{bj}$. De alternatieve hypothese kan zowel eenzijdig als tweezijdig zijn.

ANOVA-tabel voor multipele regressie
Omdat er sprake is van meerdere predictoren bij multipele regressie, worden de vrijheidsgraden voor SSM en SSE op een andere manier berekend:

BRON (source)	Vrijheidsgraden (DF)	SS (sum of squares)	MS (mean square)	F
Model	p (aantal predictoren)	$\Sigma(\hat{y}_i$-y-streepje)^2	SSM/DFM	MSM/MSE
Error	n-p-1	$\Sigma(y_i - \hat{y}_i)^2$	SSE/DFE	
Totaal	n-1	$\Sigma(y_i$-y-streepje)^2	SST/DFT	

Significantietoetsen voor regressiecoëfficiënten bij multipele regressie

- Bij multipele regressie kunnen we de nulhypothese toetsen die stelt dat alle regressiecoëfficiënten 0 zijn: $\beta_1 = \beta_2 = \ldots = \beta_p = 0$. De nulhypothese zegt dus eigenlijk dat geen van de x-variabelen een voorspeller is van de y-variabele.
- De alternatieve hypothese stelt dat tenminste één van de regressiecoëfficiënten (β_j) niet 0 is. Deze hypothese zegt eigenlijk dat minstens één van de x-variabelen een voorspeller is van de y-variabele.
- De F-toets wordt als volgt gevonden: MSM/MSE. Als de nulhypothese waar is, dan heeft F de F(p, n-p-1) distributie.

Tot slot kunnen we berekenen hoeveel variantie in y wordt verklaard door alle verklarende variabelen tezamen: R^2=SSM/SST.

11.2 Een voorbeeld
Vanaf bladzijde 615 wordt ingegaan op een voorbeeld waarbij een y-variabele wordt verklaard door meerdere predictoren. Om te kijken hoe de informatie uit 11.1 in de praktijk toegepast wordt, kan deze paragraaf doorgelezen worden.