Onderzoekspracticum 2, aantekeningen college 11

www.joho.org

Multipele regressie
Multipele regressie is de uitbreiding van enkelvoudige lineaire regressie (zie college 10) naar regressie met meerdere continue onafhankelijke variabelen. Bij enkelvoudige regressie wordt de afhankelijke variabele y voorspeld door slechts één onafhankelijke variabele x. Bij multipele regressie wordt de afhankelijke variabele y voorspeld door p onafhankelijke variabelen: x_1, x_2, …, x_p. De populatieregressievergelijking is dan als volgt:

$$m_y = b_0 + b_1 x_1 + b_2 x_2 + ... + b_p x_p$$

Hieruit blijkt dat de gemiddelde respons (μ_y) een lineaire functie is van de onafhankelijke variabelen. Het model doet de aanname dat de standaarddeviatie van de populatie (σ) gelijk is voor alle x-waarden. In een dataset (in bijvoorbeeld SPSS) worden de gegevens als volgt weergegeven: de waarden van x_1 staan onder elkaar in de eerste kolom, daarnaast staan alle waarden van x_2 in een kolom en dit gaat zo verder tot en met x_p. In de laatste kolom komen de waarden van de afhankelijke variabele y te staan.

Uit de populatieregressievergelijking wordt het volgende statistische model voor multipele lineaire regressie afgeleid:

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + ... + b_p x_{ip} + e_i$$

In dit model staat i voor de i'de persoon. In totaal zijn er n aantal personen, dus i = 1, 2, …, n. De afwijkingen ε_i (gemiddelde afwijking per persoon) zijn onafhankelijk en normaal verdeeld met gemiddelde nul en standaarddeviatie σ. De modelparameters zijn β_0, β_1, β_2 , …, β_p en σ.

Schatten van multipele regressieparameters
β_0, β_1, β_2 , …, β_p worden geschat door de steekproefschattingen b_0, b_1, b_2, …, b_p. De voorspelde respons van persoon i wordt berekend door:

$$\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + ... + b_p x_{ip}$$

Het residu van persoon i is het verschil tussen zijn/haar geobserveerde respons en zijn/haar voorspelde respons. Dit residu geven we aan met e_i. Dus: e_i = geobserveerde respons – voorspelde respons $= y_i - \hat{y}_i = y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - ... - b_p x_{ip}$.

Evenals bij enkelvoudige lineaire regressie wordt gebruik gemaakt van het kleinste-kwadratenprincipe. Hierbij kies je de waarden van de b's zodanig dat de som van alle gekwadrateerde e_i's minimaal is. Het berekenen van de b's wordt verder niet behandeld, maar de waarde van onderstaande grootheid moet dus zo klein mogelijk zijn:

$$\sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - ... - b_p x_{ip})^2$$

De schatter voor σ^2 is net als bij enkelvoudige regressie het gemiddelde van de gekwadrateerde residuen:

$$s^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-p-1} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-p-1}$$

De schatter van σ is de wortel uit s^2. Het aantal vrijheidsgraden van s^2 is de steekproefgrootte min het aantal predictoren (onafhankelijke variabelen) min 1, oftewel n – p – 1. Bij enkelvoudige regressie is het aantal vrijheidsgraden n – 2, omdat het aantal predictoren dan 1 is.

Een voorbeeld
Een docent wil het cijfer van studenten voor onderzoekspracticum 2 (OP2) voorspellen uit hun behaalde cijfers voor onderzoekspracticum 1 (OP1) en hun behaalde cijfers voor SPPS. In SPSS worden alle benodigde waarden hiervoor uitgerekend door middel van Analyze **à** Regression **à** Linear. In de uitvoer zijn de waarden van de intercept (b_0), het effect van het cijfer voor SPSS (b_1) en het effect van het cijfer voor OP1 (b_2) te vinden. Als voorbeeld nemen we de waarden $b_0 = 0.594$, $b_1 = 0.366$ en $b_2 = 0.517$. De multipele regressievergelijking wordt dan: cijferOP2 = 0.594 + 0.366×cijferSPSS + 0.517×cijferOP1 + e_i.

Deze vergelijking wordt als volgt geïnterpreteerd:

- Cijfer SPSS – Als het cijfer voor SPSS met één punt toeneemt, dan neemt het cijfer voor OP2 *gemiddeld* met 0.366 punten toe. Hierbij wordt het cijfer van OP1 constant gehouden.
- Cijfer OP1 – Als het cijfer voor OP1 met één punt toeneemt, dan neemt het cijfer voor OP2 *gemiddeld* met 0.517 punten toe. Hierbij wordt het cijfer van SPSS constant gehouden.
- Intercept – Als er het cijfer nul is behaald voor zowel OP1 als SPSS, dan is het verwachte cijfer voor OP2 gelijk aan 0.594.

Bij de interpretatie met betrekking tot de onafhankelijke variabelen zijn dus drie aspecten belangrijk:

- Er moet worden vermeld hoeveel de waarde van de afhankelijke variabele verandert als de onafhankelijke variabele met één punt toeneemt.
- Het moet duidelijk zijn dat het om een *gemiddelde* verandering (toename of afname) gaat.
- Er moet vermeld worden dat dit geldt onder constanthouding van alle overige onafhankelijke variabelen.

Betrouwbaarheidsinterval en significantietoets
Een coëfficiënt β_j heeft een betrouwbaarheidsinterval. Dit C%-betrouwbaarheidsinterval wordt berekend door:

$$b_j \pm t^* SE_{b_j}$$

SE_{bj} is de standaard error van b_j en t* is de waarde van een t-verdeling met n – p – 1 vrijheidsgraden waarvoor geldt dat de oppervlakte tussen –t* en t* C% is. De nulhypothese is $\beta_j = 0$. Voor het toetsen van deze nulhypothese wordt ook weer een t-verdeling gebruikt met n – p – 1 vrijheidsgraden. De alternatieve hypothese kan eenzijdig zijn (gevonden p-waarde niet vermenigvuldigen met 2), of tweezijdig (gevonden p-waarde wel vermenigvuldigen met

2). De t-waarde wordt berekend door:

$$t = \frac{b_j}{SE_{b_j}}$$

Evenals bij enkelvoudige regressie bestaan er bij multipele regressie ook voorspellingsintervallen en een betrouwbaarheidsinterval voor de gemiddelde respons. Het idee daarbij is hetzelfde als bij enkelvoudige regressie, maar details worden hier niet verder besproken.

ANOVA voor multipele regressie
Bij enkelvoudige regressie toetst de F-toets hetzelfde als de tweezijdige t-toets, namelijk dat de richtingscoëfficiënt β_1 gelijk is aan nul (nulhypothese). Bij multipele regressie toetst de F-toets de hypothese dat *alle* regressiecoëfficiënten gelijk zijn aan nul.

Bij multipele regressie geldt net als bij enkelvoudige regressie dat SST = SSM + SSE (de totale kwadratensom = de kwadratensom van het model + de kwadratensom van de error).

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 \text{ , } SSM = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \text{ en } SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Bij elke kwadratensom hoort een bepaald aantal vrijheidsgraden:

- DFM is het aantal predictoren (onafhankelijke variabelen), oftewel DFM = p.
- DFE is gelijk aan n – p – 1.
- DFT is gelijk aan n – 1 (DFM + DFE).

De gemiddelde kwadratensommen worden berekend door (voorbeeld dia 23):

- MSM = SSM/DFM
- MSE = SSE/ DFE

De F-waarde wordt gevonden door F = MSM/MSE. Deze waarde geeft aan of alle regressiecoëfficiënten gelijk zijn aan nul. De intercept wordt niet meegenomen in de toetsing. De hypothesen worden op de volgende manier opgeschreven:

H_0: $\beta_1 = \beta_2 = \ldots = \beta_p = 0$
H_a: Tenminste één β_j is ongelijk aan 0 / niet alle *β's zijn gelijk aan 0*

Als je wilt weten *welke* β_j ongelijk is aan 0, dan kijk je in de uitvoer van SPSS bij de significantieniveaus van de t-waarden uit de tabel voor coëfficiënten. Bij een significant effect mag je concluderen dat β_j afwijkt van 0. Er is dan een effect van de onafhankelijke op de afhankelijke variabele.

Het percentage verklaarde variantie (R^2) geeft aan welk percentage van de variantie verklaard wordt door het model en welk percentage verklaard wordt door de error. Bij enkelvoudige regressie geeft R^2 aan in hoeverre de geobserveerde waarden op de regressielijn liggen. Bij multipele regressie geeft R^2 aan hoe goed de predictoren de waarden van y voorspellen. R^2 = SSM/SST. Wanneer R^2 gelijk is aan 1, dan worden de y-waarden perfect voorspeld door de predictoren en is er geen error.

In SPSS worden ook gestandaardiseerde gewichten (coëfficiënten) gegeven. Deze worden net als populatiegewichten weergegeven met *β*, maar de betekenis is anders. Een gestandaardiseerd gewicht is namelijk het regressiegewicht dat je zou krijgen wanneer je de regressie zou uitvoeren met *gestandaardiseerde* scores. De *β*'s die terug te zien zijn in artikelen zijn altijd gestandaardiseerde gewichten, nooit populatiewaarden. De interpretatie van gestandaardiseerde gewichten is als volgt: als een bepaalde onafhankelijke variabele met één *standaarddeviatie* toeneemt, dan neemt de waarde van de afhankelijke variabele met gemiddeld x *standaarddeviaties* toe, onder constanthouding van de overige onafhankelijke variabelen. Bijvoorbeeld: als het cijfer voor OP1 met één standaarddeviatie toeneemt, dan neemt het cijfer van OP2 met gemiddeld 0.53 standaarddeviaties toe, onder constanthouding van de rest.

Opdracht dia 33
Inhoudelijke interpretatie van de constante b0. En heeft deze een zinvolle betekenis?
B0 houdt in dat wanneer je voor alle vakken een 0 haalt je voor OP2 een -1.842 kan halen, dit is niet zinvol want je kan geen 0 en al helemaal niet onder de 0 halen.

Inhoudelijke interpretatie van de regressiegewichten van het cijfer van de SPSS toets, het cijfer van OP1 en het cijfer van grondslagen :
SPSS : wanneer het cijfer van SPSS met 1 punt toeneemt neemt het cijfer van OP2 toe met gemiddeld .499 punten onder constant houding van de rest.
OP1 : Wanneer het cijfer voor OP1 met 1 punt toeneemt neemt het cijfer van OP2 toe met gemiddeld 0.465 punten onder constante houding van de rest.
Grondslagen : idem, maar dan met 0.322.

Welk van de onafhankelijke variabelen heeft een significant effect op het tentamencijfer van OP2?
SPSS en OP1, beide p-waarde zijn lager dan 0.05.

Is het toevoegen van het cijfer van grondslagen aan het model zinvol geweest?
Cijfer van grondslagen is niet significant, SPSS is door toevoeging van grondslagen significant geworden. Dit komt waarschijnlijk doordat :

- Er een negatieve correlatie is tussen grondslagen en SPSS (hoe lager het cijfer van grondslagen , hoe hoger het cijfer van SPSS).
- SPSS een positieve bijdrage levert aan OP2. Wanneer gecorrigeerd wordt voor het cijfer van grondslagen dan wordt daardoor het effect van SPSS op OP2 sterker, en dus opeens wel significant.

Antwoord op de vraag :

- Niet als je de voorspelling van het cijfer van OP2 wilt verbeteren.
- Wel als je de verhoudingen wilt weten tussen predictoren onderling.