Hoofdstuk 8

8.1

We gaan nu twee concepten bespreken die te maken hebben met de regressie analyse. Namelijk de correlatie analyse, met de partiële en semi-partiële correlatie. De meervoudige regressie modellen bestaan uit twee of meer predictoren en een criterion variabele; dus er zijn op zijn minst drie variabelen betrokken bij dit model. De Pearson correlatie, die we eerst gebruikten, is gebaseerd op maar twee variabelen. Deze kunnen we dus niet gebruiken. De oplossing voor dit is de partiële- en semi-partiële correlatie.

Eerst bespreken we de partiële correlatie. De makkelijkste situatie is wanneer we drie variabelen hebben, die we X1, X2, en X3 noemen. Dus een voorbeeld van een partiële correlatie is, de correlatie tussen X1, en X2, waarbij X3 constant word gehouden. Dus de partiële correlatie laat de lineaire relatie tussen X1 en X2, zien onafhankelijke van de invloed van X3. Dit voorbeeld wordt genoteerd als r12.3. We berekenen dit als vogt:

$$r_{12.3} = r_{12} - r_{13}r_{23} / \sqrt{(1 - r_{13}^2)(1 - r_{23}^2)} \qquad (80)$$

Er kunnen extreme uitkomsten komen bij de partiële correlatie. Een voorbeeld is perfect collinearity, wat een groot probleem is. In dat geval is of r13, of r23 gelijk aan 1. Wanneer dit zo is, kan r12.3 niet worden berekend, omdat de noemer gelijk is aan 0. In deze situatie kunnen we de partiële correlatie niet berekenen.

We gaan nu kijken naar de semi-partiële correlatie. Opnieuw gebruiken we de makkelijkste situatie met drie variabelen, die weer X1, X2, en X3 worden genoemd. Een voorbeeld van een semi-partiële correlatie is de correlatie tussen X1 en X2, waarbij X3 is verwijderd van alleen X2. Dus deze semi-partiële correlatie is de lineaire relatie tussen X1 en X2 nadat een deel van X2 niet kan worden bepaald door X3 omdat deze is verwijderd uit X2. Dit voorbeeld wordt genoteerd als r1(2.3). We berekenen dit als volgt:

$$r_{1(2.3)} = r_{12} - r_{13}r_{23} / \sqrt{(1 - r_{23}^2)} \qquad (81)$$

8.2

In deze paragraaf zullen we de niet gestandaardiseerde en gestandaardiseerde meervoudige regressie modellen bespreken, de coëfficiënt van multipele determinatie, multipele correlatie (meervoudige correlatie), significantie toetsen en statistische aannames.

Het meervoudige lineaire regressie model gebaseerd op de steekproef om Y te voorspellen op basis van een aantal predictoren genaamd m X1,2,….m is

$$Y_i = b_1X_{1i} + b_2X_{2i} + \ldots + b_mX_{mi} + a + e_i \qquad (82)$$

Waarin:

- Y de criterion variabele is (afhankelijke variabele)
- X de predictor (onafhankelijke) variabelen zijn, waar k= 1,…,m
- bk de partiële helling van de regressie lijn, gebaseerd op de steekproef, waarbij X Y voorspelt.
- a de steekproef intercept van de regressie lijn is, voor Y voorspeld door de X'en
- ei de residuen of fouten van de predictor variabelen zijn.
- i de index voor een individu is. Deze kan waarden aannemen van 1,…,n.

De term partiële helling word gebruikt omdat dit de helling voorstelt van Y voor een bepaalde X. We hebben de invloed van de andere X'en uitgesloten.

Het prediction model gebaseerd op de steekproef is als volgt:

$$Y_i' = b_1X_{1i} + b_2X_{2i} + \ldots + b_mX_{mi} + a \qquad (83)$$

Waar Yi' de voorspelde waarde van Y is. Het verschil tussen het prediction model en het regressie model is hetzelfde als in hoofdstuk 7. We berekenden de residuen als volgt:

$$e_i = Y_i - Y'_i \qquad (84)$$

Het is moeilijk om de steekproef partiële hellingen en intercept te bepalen. Om het makkelijk te houden gebruiken we een model met twee predictoren om het te laten zien. Over het algemeen wordt dit berekent met SPSS. In het geval van twee predictoren zijn de partiële hellingen gebaseerd op de steekproef:

$$b_1 = (r_{Y1} - r_{Y2}r_{12})s_Y/(1 - r_{12}^2)s_1$$
$$b_2 = (r_{Y2} - r_{Y1}r_{12})s_Y/(1 - r_{12}^2)s_2$$
$$a = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2 \qquad (85)$$

Een alternatieve methode om de partiële hellingen te berekenen is door gebruik te maken van de partiële correlaties:

$$b_1 = r_{Y1.2}\, s_Y\sqrt{1 - r_{Y2}^2} \Big/ s_1\sqrt{1 - r_{12}^2}$$
$$b_2 = r_{Y2.1}\, s_Y\sqrt{1 - r_{Y1}^2} \Big/ s_2\sqrt{1 - r_{12}^2} \qquad (86)$$

In het meervoudige lineaire regressie model gebruiken we het kleinste kwadraten criterium. Dus we moeten een regressie model vinden, met bepaalde partiële hellingen en een intercept dat de kleinste som van gekwadrateerde residuen heeft.

We zullen nu kijken naar het gestandaardiseerde regressie model. In dit model zijn de termen z-scores. Het gemiddelde en de variantie van de gestandaardiseerde variabelen zijn respectievelijk 0 en 1. Het gestandaardiseerde lineaire prediction model wordt:

$$z\,(Y_i') = b_1^* z_{1i} + b_2^* z_{2i} + \ldots + b_m^* z_{mi} \qquad (87)$$

Waarin bk* voor de gestandaardiseerde partiele helling staat. Deze gestandaardiseerde hellingen worden berekent met de volgende formule:

$$b_k^* = b_k \frac{s_k}{s_Y} \qquad (88)$$

Voor een model met twee predictoren worden de gestandaardiseerde partiele hellingen als volgt berekent:

$$b_1^* = b_1 \frac{s_1}{s_Y} \quad \text{or} \quad b_1^* = \frac{r_{Y1} - r_{Y2} r_{12}}{(1 - r_{12}^2)}$$
$$b_2^* = b_2 \frac{s_2}{s_Y} \quad \text{or} \quad b_2^* = \frac{r_{Y2} - r_{Y1} r_{12}}{(1 - r_{12}^2)} \qquad (89)$$

We willen nu weten wat de utiliteit is van de verschillende predictor variabelen. De makkelijkste manier om hier naar te kijken is door naar de verdeling van de totale som van de kwadraten te kijken. Dit wordt genoteerd als SStotal. In de meervoudige regressie analyse kunnen we dit als volgt schrijven:

$SS_{total} = [n\sum Y_i^2 - (\sum Y_i)^2]/n$ ..... or $SS_{total} = (n-1)s_Y^2$

$SS_{total} = SS_{reg} + SS_{res}$

$$\sum (Y_i - \bar{Y})^2 = \sum (Y_i' - \bar{Y})^2 + \sum (Y_i - Y_i')^2 \qquad (90)$$

Waarin:

- SSreg d de kwadratensomvan de voorspelling van Y door X is
- SSres de kwadratensom van de residuen is.

We zullen kijken naar de coëfficiënt van de meervoudige determinaties, genoteerd als $R^2_{Y.1,\ldots,m}$. De subscript verteld ons dat Y is de criterion (afhankelijke variabele) en dat $X_{1,\ldots,m}$ is de predictor (onafhankelijke) variabele. D makkelijkste manier om R^2 te bereken is als volgt:

$$R^2_{Y.1,\ldots,m} = b_1^* r_{Y1} + b_2^* r_{Y2} + \ldots + b_m^* r_{Ym} \qquad (91)$$

De coëfficiënt van de meervoudige determinatie verteld ons het deel van de totale variatie in de afhankelijke variabele Y, dat is voorspeld door de predictor variabelen. We zien deze coëfficiënt ook vaak met de SS termen als: $R^2_{Y.1,\ldots,m} = SS_{reg}/SS_{total}$. We kunnen deze formule herschrijven als volgt:

$SS_{reg} = R^2 SS_{total}$ and $SS_{res} = (1-R^2)SS_{total} = SS_{total} - SS_{reg}$ (92)

De coëfficiënt wordt niet alleen bepaald door de kwaliteit van de predictor variabelen, maar ook door de kwaliteit van de relevante predictor variabelen die niet in het model zijn meegenomen. Ook wordt het bepaald door de totale variantie in de afhankelijke variabele, Y. De coëfficiënt van determinatie kan ook worden gebruikt om de effect grootte te bepalen (Klein effect: R2 = 0.10; gemiddeld effect: R2 = 0.30; groot effect R2 = 0.50).

R2 is erg gevoelig voor de steekproefgrootte en ook voor het aantal predictor variabelen in het model. Wanneer de steekproefgrootte en/of het aantal predictor variabelen groter wordt, zal R2 ook groter worden. R is een gebiased meervoudige populatiecorrelatie. Over het algemeen overschat R de meervoudige populatiecorrelatie. Daarom hebben we ook een aangepaste R2:

$$R^2_{adj} = 1 - (1- R^2)\left(\frac{n-1}{n-m-1}\right) \qquad (93)$$

Deze aangepaste waarde, past de waarde van R aan, aan de steekproefgrootte en het aantal predictoren. Deze kunnen we dus gebruiken om modellen te vergelijken met een verschillend aantal predictor variabelen. Het verschil tussen R2 en de aangepaste R wordt shrinkage genoemd.

Om te kijken of het model genoeg power heeft, kunnen we G*power gebruiken. Maar we moeten zeker weten dat de verhouding tussen n en m groot genoeg is. Dit zorgt er namelijk voor dat de bias zo klein mogelijk blijft en dat de resultaten beter te generaliseren zijn naar de populatie.

We gaan nu kijken naar de significantie toetsen. We kijken naar twee methoden die worden gebruikt in de meervoudige regressie analyse. De eerste is om de significantie van het hele regressie model te toetsen. De tweede is om de significantie van elke partiële helling te bepalen:

Significantie toets voor het hele regressie model

De hypothesen van deze toets worden geschreven met de coëfficiënt van de meervoudige determinatie. Ze zijn als volgt:

H_0: $\beta_1 = \beta_2 = \ldots = \beta_K = 0$
H_1: not all the $\beta_k = 0$

Wanneer H0 wordt verworpen, dan is een of meer van de regressie coëfficiënten niet significant verschillend van 0. Deze toets is gebaseerd op de volgende toetsingsgrootheid:

$$F = \frac{R^2/m}{(1-R^2)/(n-m-1)} \qquad (94)$$

Waarin F laat zien dat het om een F-toets gaat. M is het aantal predictor of onafhankelijke variabelen en n is de steekproefgrootte. Deze F toets wordt vergeleken met de kritische waarde van F, dit is altijd een eenzijdige toets, met een alpha level en met de vrijheidsgraden (n-m-1). De kritische waarde kan worden gevonden in tabel A.4. De toetsingsgrootheid kan ook als volgt worden geschreven:

$$F = \frac{SS_{reg}/df_{reg}}{SS_{res}/df_{res}} = \frac{MS_{reg}}{MS_{res}} \qquad (95)$$

Waarin df(reg) = m, en df(res) = (n-m-1).

Toets van significantie bk

Deze toets bepaalt of alle aparte niet gestandaardiseerde regressie coëfficiënten significant verschillend zijn van 0. Deze test is hetzelfde voor . De hypothesen zijn als volgt:

$H_0: \beta_k = 0$
$H_1: \beta_k \neq 0$

In het meervoudige regressie model is het nodig om de standaardfouten voor elke regressie coëfficiënt te bepalen. De standaardfout van de geschatte waarden wordt als volgt berekend:

$$s^2_{res} = SS_{res}/df_{res} = MS_{res} \quad (96)$$

Finally we need to compute a standard error for each b_k:

$$s(b_k) = \frac{s_{res}}{\sqrt{(n-1)s^2_k(1-R^2_k)}} \quad (97)$$

Deze toets wordt vergeleken met de kritische waarde van t, een tweezijdige toets, voor een onzijdige hypothese, met een alpha level, en vrijheidsgraden (n-m-1). Deze kan worden gevonden in tabel A.2. We kunnen het betrouwbaarheidsinterval berekenen met de formule:

Where
s^2_k is the sample variance for predictor X
R^2_k is the squared multiple correlation between X_k and the remaining X_k's

The test statistic is as follows:

$$t = \frac{b_k}{s(b_k)} \quad (98)$$

We gaan nu kijken naar de verschillende aannames waaraan moet worden voldaan behorende bij het meervoudige regressie model. De aannames zijn (a) onafhankelijkheid, (b) homogeniteit, (c) normaliteit, (d) lineariteit, (e) fixed X, en (f) non-collineariteit

Onafhankelijkheid

De makkelijkste manier om te bepalen of aan deze aanname voldaan, is om een residuenplot maken van e tegen de verwachte waarden van de afhankelijke variabele, Y. Of een grafiek van e tegen elke onafhankelijke variabele X. Wanneer aan deze aanname is voldaan, zullen de residuen op een willekeurige manier verdeeld zijn. Wanneer deze aanname geschonden is, kan dit de bepaalde standaardfouten beïnvloeden.

Homogeniteit

Hierbij moeten de conditionele verdelingen dezelfde constante variantie hebben voor alle waarden van X. Dit kan je ook in een residuenplot bekijken.. Wanneer niet aan deze aanname is voldaan, zullen de berekende standaardfouten groter zijn, en zal ook de conditionele verdeling niet normaal zijn.

Normaliteit

De conditionele verdelingen van de scores van Y, of van de predictieve fouten (prediction errors) volgen een normale verdeling. Wanneer niet aan deze aanname voldaan wordt, kan dit komen door een uitbijter. Je kan een frequency distribution, Q-Q plots, en de kurtosis en skewness waarden gebruiken om dit te bekijken.

Lineariteit

Er moet een lineaire relatie zijn tussen de geobserveerde scores van de afhankelijke variabele Y en de waarden van de onafhankelijke variabele X. Wanneer aan deze aanname wordt voldaan, zullen de steekproef partiële hellingen en intercept niet gebiased zijn. Wanneer deze relatie niet lineair is, betekend dit dat de verwachte toename van Y afhangt van de waarde van X. Dus de verwachte toename is niet constant. Wanneer niet aan de aanname voldaan is, kan je dit zien in een residuenplot De residuen moeten rondom de lineaire lijn vallen (standaardfouten).

Fixed X

Als de onafhankelijke variabele X gefixeerde waarden heeft (dus niet random), dan is het regressie model alleen valide voor de waarden van X die zijn geobserveerd in het model. Over het algemeen willen we geen voorspellingen doen over individuen die een combinatie van X scores hebben anders dan de waarden die wij hebben gebruikt in het model (extrapolating). Daarnaast willen we ook geen voorspellingen doen op basis van individuen die een combinatie van X-scores hebben binnen de waarden die we hebben gebruikt om het model te voorspellen (interpolating). Het is bewezen dat wanneer aan alle andere aannames is voldaan, het niet uit maakt of X fixed of random is.

Geen collineariteit

Deze aanname wordt alleen gebruikt voor de meervoudige lineaire regressie. Wanneer niet aan deze aanname voldaan is, dan betekent dit dat er collineariteit aanwezig is. Dit betekent dat er een hele sterke lineaire relatie is tussen twee of meer van de predictor variabelen. Dit is een probleem in verschillende opzichten. Ten eerste, het zorgt voor meer instabiliteit in de regressie coëfficiënten. Ook kunnen de geschatte waarden veranderen van grootte. Dit komt doordat de standaardfouten groter zijn, wat het moeilijker maakt om een significant model te krijgen. Ten tweede, het kan ook zijn dat het volledige regressiemodel significant is, maar dat geen van de individuele predictoren significant is.

Collineariteit treedt op wanneer er grote veranderingen zijn in de voorspelde coëfficiënten doordat (a) een variabele wordt toegevoegd of verwijderd en/of (b) een observatie wordt toegevoegd of verwijderd.

We kunnen kijken of aan deze aanname voldaan is door speciale regressie analyses uit te voeren. Bijvoorbeeld een regressievergelijking op stellen voor elke X waarin deze predictor wordt voorspeld door alle andere X'en. Wanneer een van de resultaten een waarde rondom 1 heeft (groter dan 0.9) dan zal collineariteit een probleem zijn. Een grote R2 kan ook veroorzaakt worden door een kleine steekproef. Wanneer het aantal predictoren groter of gelijk is aan n, dan kan er perfecte collinearteit zijn (zie 8.1). Een andere manier om collineariteit te vinden is door de variance inflation factor (VIF) te bepalen. Deze is gelijk aan 1/(1- R2). De VIF word gedefinieerd als de toename die optreed voor per regressie coëfficiënt als de predictoren correleren.

De grootste VIF waarde moet kleiner zijn dan 10 om aan de aanname te voldoen.

Er zijn ook andere methodes die met collineariteit werken. De eerste is dat je een of meer gecorreleerde predictoren kan verwijderen. De tweede methode bestaat uit ridge regressie technieken. De derde methode gebruikt principal component scores die worden gevonden door de principal component analysis (PCA). De vierde methode is het transformeren van variabelen.

Een samenvatting van de aannames en wat er gebeurt als de data niet aan deze aannames voldoet.

Aanname	Effect wanneer niet aan de aanname voldaan is
Onafhankelijkheid	Beïnvloedt de standaardfouten
Homogeniteit	Biases in de varianties van de residuen Kan de standaardfouten vergroten, en dus de kans op een Type II fout vergroten Kan zorgen voor een niet normale conditionele verdeling
Normaliteit	Minder precieze hellingen, intercept en R2
Lineair	Vooroordeel in de helling en intercept Verwachte verandering in Y is niet constant en hangt af van de waarde van X
Vastgestelde X-waarden	Extrapolating buiten de waarden van X: predictieve fouten worden groter, kan ook leiden tot biases in de helling en intercept. Interpolating: Binnen de waarden van X: kleinere effect dan voorheen. Wanneer aan alle andere aannames is voldaan dan is dit een verwaarloosbaar effect
Non-collinearity van de X'en	Regressie coëfficiënten kunnen onstabiel zijn over de steekproeven (omdat standaardfouten groter zijn) R2 kan significant zijn, terwijl geen van de predictoren significant is Minder generalisatie van het model.

8.3

Het meervoudige predictor model kan worden gezien als een simultane regressie (simulteneous regression). Dat betekent, alle predictoren die worden geberuikt zijn gelijktijdig ingevoerd, zodat alle regressie parameters gelijktijdig kunnen worden geschat. Er zijn drie andere methoden om deze onafhankelijke variabelen in te voeren, namelijk systematisch. Dit wordt sequential regression of sequentiële regressie genoemd. We bespreken drie van deze methoden:

Backward elimination

In deze regressie worden de variabelen geëlimineerd gebaseerd op de hoeveelheid die ze bijdragen aan het voorspellen van de criterion variabele. In de eerste fase van de analyse worden alle potentiële predictoren ingevoerd. In de tweede fase wordt de predictor verwijderd die het minst bijdraagt aan het voorspellen van de criterion variabele.

Dit kan je zien door de variabele te verwijderen met de kleinste F of t-waarde. In de fases daarna zal steeds de predictor met de kleinste bijdrage worden verwijderd. Dit gaat door totdat elke predictor die er nog is een significante bijdrage levert aan het voorspellen van Y. Dit kan je bekijken door de t- of F-waarde te vergelijken met de kritische waarden.

Forward selection

Bij deze methode worden de variabelen toegevoegd of geselecteerd op basis van de maximale bijdrage aan het voorspellen van Y. In het begin wordt geen van de predictor variabelen toegevoegd aan het model. In de eerste fase wordt de predictor toegevoegd die de grootste bijdrage levert (grootste t of F-waarde). De fases daarna zal steeds een nieuwe predictor worden geselecteerd die daarna de grootste bijdrage levert. Dit gaat door totdat alle geselecteerde predictor variabelen een significante bijdrage leveren aan het voorspellen van Y (vergelijk de F- of t-waarde met de kritische waarde).

Stepwise selection

Dit is een aanpaste vorm van het forward selection model. Er is een belangrijk verschil, namelijk dat de predictoren die zijn geselecteerd later ook weer kunnen worden verwijderd uit het model. Dit kan gebeuren wanneer een predictor in het begin een significante bijdrage leverde, maar naarmate er meer predictoren worden toegevoegd, deze bijdrage niet meer significant is. Ook in dit model is er in het begin nog geen enkele predictor toegevoegd. In de eerste fase, wordt de predictor toegevoegd die de grootste bijdrage levert (grootste F- of t-waarde). De fases daarna wordt steeds de predictor geselecteerd die daarna de grootste bijdrage levert. Daarnaast wordt er elke keer wanneer een nieuwe predictor wordt toegevoegd gekeken of de bestaande predictoren nog significant zijn. Wanneer dit niet zo is, dan worden ze verwijderd. Dit gaat zo door totdat alle predictoren een significante bijdrage leveren (vergelijk de F-waarde of t-waarde met de kritische waarde).

All possible subsets regression

Stel er zijn 5 potentiele predictoren. In deze methode worden alle mogelijke een-, twee-, drie-, en vier-variabelen modellen geanalyseerd. Dus er zullen 5 een-predictor modellen, 10 twee-predictor modellen en 10 drie-predictor modellen, en 5 four-predictor modellen zijn. Het beste model met k (aantal) predictoren zal worden gekozen. Dit model heeft dan de hoogste R2.

Deze methode wordt niet geadviseerd, eigenlijk geen van deze methoden, wanneer het aantal potentiele predictoren groot is. Het aantal modellen dat met deze methode kan worden gemaakt is gelijk aan 2m.

Hierarchial regression

In dit model beschrijft de onderzoeker van tevoren de volgorde van de predictor variabelen. Deze analyse gaat te werk als een forward selectie, backward selectie of stepwise selectie methode. Deze methode is anders omdat de onderzoeker voorzichtig zal bepalen welke volgorde voor hem het beste is gebaseerd op theorie en eerder onderzoek. Een type van een hiërarchische regressie is een setwise regressie (block-wise, chunck-wise, of forces stepwise regressie). Hierbij beschrijft de onderzoeker van tevoren de volgorde. Deze methode is hetzelfde als de hiërarchische methode waarbij de onderzoeker de volgorde bepaald. Het verschil is dat de setwise methode de reeksen van predictor variabelen gebruikt per fase in plaats van één predictor variabele per fase.

Er zijn een paar opmerkingen over de sequentiële regressie procedures. Het eerst is dat verschillende statistici problemen hebben gevonden in de step-wise methode namelijk (a) er wordt vaak noise (ruis) geselecteerd in plaats van belangrijke predictoren; (b) De waarden van R2 en de aangepaste R2 nemen toet; (c) betrouwbaarheidsintervallen voor de partiële hellingen zijn te smal; (d) p-waarden zijn niet betrouwbaar; (e) belangrijke predictoren worden bijna nooit uit het model gehaald, wat het mogelijk maakt dat het echte model niet gevonden wordt; (f) er kan per ongelijk sterke kanskapitalisatie plaatsvinden, door het aantal modellen dat wordt geanalyseerd.

8.4

We gaan nu kijken hoe je om kan gaan met een niet lineair model. We zullen verschillende meervoudige regressie modellen laten zien die toegepast kunnen worden als er geen lineaire relatie is tussen de criterion variabelen en de predictor variabelen. Eerst de polynomiale regressie modellen. In deze modellen, worden de machten van de predictor variabelen gebruikt als volgt:

$$Y = b_1X + b_2X^2 + \ldots + b_mX^m + a + e \qquad (100)$$

Wanneer het model alleen bestaat uit X, dan hebben we een enkelvoudige lineaire regressie (eerstegraad polynomiaal). Een tweede-graad polynoom omvat een X tot de macht 2 (kwadratisch model). Een derdegraads polynoom omvat X tot de macht 3(kubisch model). Het is belangrijk dat wanneer je een polynoom van een hogere-graad hebt dat je ook de eerstegraads polynoom moet meenemen in het model.

8.5

In onderstaande regressievergelijking is ook een interactieterm opgenomen. Deze vergelijking kan worden gebruikt in alle typen regressie modellen. Onderstaand model is een enkelvoudig model met twee onafhankelijke variabelen en een interactieterm.

$$Y = b_1X_1 + b_2X_2 + b_3X_1X_2 + a + e \qquad (101)$$

X1X2 is de interactie tussen de predictor variabelen 1 en 2. Een interactie wordt gedefinieerd als de relatie tussen Y en X1 die afhangt van X2. Dus X2 is de moderator variabele. Maar wanneer de variabelen heel erg correleren kan er sprake zijn van collineariteit.

8.6

We hebben tot nu toe alleen maar gekeken naar continue predictoren (onafhankelijke variabelen die op een interval of ratio schaal gemeten zijn). Maar, het kan ook zijn dat je een predictor hebt die op een categorische schaal gemeten wordt. Deze variabelen moeten opnieuw gecodeerd worden, zodat ze op een schaal van 0 en 1 zitten. Dit wordt ‘dummy coding’ genoemd. Bijvoorbeeld 0 zijn de vrouwen en 1 zijn de mannen.

8.7

We gaan nu de stappen bespreken om een meervoudige lineaire regressie analyse uit te voeren in SPSS. We hebben data met een afhankelijke variabele en twee onafhankelijke variabelen:

- Ga naar “analyse” en selecteer “regression” en daarna “linear”
- Sleep de afhankelijke variabele in de “dependent” box. Sleep de onafhankelijke variabelen in de “independent(s)” box.
- Vanuit de “linear regression” box, klik op “statisticts”. Hier moet je de volgende dingen aanvinken (a) estimates, (b) Cis, (c) model fit, (d) R squared change, (e) descriptives, (f) part and partial correlations, (g) collinearity diagnostics, (h) Durbin-Watson en (i) case wise diagnostics. Klik op “continue”
- Vanuit de “linear regression” dialog box, klik op “plots”. Hier moet je de volgende dingen aanvinken, (a)histogram, (b) normal probabillity plot, 9c) produce all partial plots. Klik op “continue”.
- Vanuit de “linear regression” dialog box, klik op “save”. Hier moet je de volgende dingen aanvinken onder het kopje predicted values: unstandardized. Onder het kopje residuals vink (a) unstandardised en (b) studentized) aan. Onder het kopje distances vink (a) mahalanobis, (b) Cook’s en (c) leverage values aan. Onder het kopje influence statistics vink (a) DFBETA(s) aan. Klik op “continue” en klik op “OK”.

De resultaten staan op pagina 395-399

Een belangrijke interpretatie van deze resultaten:

De aangepaste R2 wordt geïnterpreteerd als het percentage verklaarde variantie in de afhankelijke variabele nadat er is gecorrigeerd voor de steekproefgrootte en het aantal predictoren.

We zullen nu kijken naar de waarden die we hebben opgeslagen van onze data:

- PRE_1 zijn de niet gestandaardiseerde voorspelde waarden
- RES_1 zijn de niet gestandaardiseerde residuen. Dit is het verschil tussen de geobserveerde en voorspelde waarden
- SRE_1 zijn de studentized residuen. Dit is een type van gestandaardiseerde residuen die meer gevoelig is voor uitbijters. Deze worden berekend door de niet gestandaardiseerde residuen te delen door een voorspelde waarde van de standaard deviatie. De studentized residuen met een absolute waarde groter dan 3 kunnen worden gezien als uitbijters.
- MAH_1 zijn Mahalanobis afstand waarden die kunnen helpen om uitbijters te herkennen. Gekwadrateerde mahalanobis afstand waarden gedeeld door het aantal variabelen die groter zijn dan 2.5 (kleine steekproeven) of 3-4 (grote steekproeven) kunnen uitbijters zijn.
- COO_1 zijn Cook’s afstand waarden en geven een indicatie van de invloed van aparte gevallen. Als regel, wanneer de Cook’s waarde groter is dan 1.0 geeft dit aan dat het problematisch kan zijn.
- LEV_1 staat voor de waarden van leverage, dit laat de afstand tussen een bepaalde waarde en het gemiddelde van de predictor zien.

- SDB0_1 en SDB1_1 zijn gestandaardiseerde DFBETA waarden. Deze kan je makkelijk interpreteren door ze te vergelijken met de niet gestandaardiseerde DFBETA waarden. Gestandaardiseerde waarden groter dan 2 geven aan dat dit geval onnodige invloed uitoefent op de parameters van het model.

Om te zien aan welke aannames is voldaan moeten we verschillende dingen doen. Voor de aanname van onafhankelijkheid moeten we de volgende grafieken maken (a) studentized residuen tegen de niet gestandaardiseerde geschatte waarden en (b) studentized residuen tegen elke onafhankelijke variabele. Wanneer aan de aanname is voldaan zullen de punten willekeurig in de grafiek liggen in een gebied van -2.0 en +2.0

We kunnen dezelfde grafieken gebruiken om te kijken naar homogeniteit. Wanneer aan de aanname is voldaan dan zal de verdeling van de residuen ongeveer constant zijn tegen de niet gestandaardiseerde geschatte waarden, en de geobserveerde waarden van de onafhankelijke variabele.

Deze grafieken kunnen we ook bekijken voor de lineaire relatie. Wanneer er een diagonale lijn te zien is dan is aan deze aanname voldaan.

Voor de normaliteitsaanname kan je de methoden gebruiken die eerder zijn besproken zoals de waarden van skewness en kurtosis, Q-Q plots of een boxplot.

Wanneer er multicollineariteit is, is er een sterke correlatie tussen twee onafhankelijke variabelen. Dit kan je bekijken door te kijken naar de VIF en tolerance waarden. Wanneer de waarde van tolerance (1-R2) dicht bij 0 licht (0.10 of minder) kan er een probleem zijn met multicollineariteit. Een tolerance van 0.10 betekend dat 90% van de variantie in een van de onafhankelijke variabelen kan worden uitgelegd door een andere onafhankelijke variabele. VIF word berekend door 1/tolerance. Waarden groter dan 10 suggereren multicollineariteit.

8.8

We zullen ook hier G*power gebruiken om de post hoc en a priori power te berekenen. Voor de post hoc analyse moeten we de goede testsoort selecteren. Dit doe je door “tests” te selecteren, daarna “correlation and regression” en vervolgens “linear multiple regression: fixed model, R2, deviation from zero”. Daarna zal de test soort automatisch veranderen in een F-toets. De input parameters zijn nu: (1) effect size, (2) alpha level, (3) total sample size, en (4) number of predictoren. We kunnen het pop-up schermpje gebruiken om de effect size te berekenen. Klik op “calculate” om de effect grootte te berekenen, klik daarna op “calculate and transfer to main window” om de berekende waarde in het model te plaatsen.

Voor de a priori power analyse, kunnen we de totale steekproefgrootte bepalen die we nodig hebben voor de meervoudige lineaire regressie wanneer we de geschatte grootte, f2, alpha level, gewilde power, en het aantal predictoren weten. Een klein effect: r2=0.02, gemiddeld effect: r2=0.15 en groot effect: r2=0.35.