Hoofdstuk 7

7.1

De meeste simpele lineaire regressie is een rechte lijn, Y = bX + a. In deze formule is X de voorspellende variabele (predictor variable). Deze wordt gebruikt om de criterium variabele (criterion variable) Y te bepalen. De helling van de lijn is b, en dit laat de zien hoeveel -waarden de lijn verandert wanneer er 1 eenheid van X veranderd. De Y-intercept wordt a genoemd, dit is het punt waar de lijn kruist met de Y-as. Deze term noemen we intercept. De helling kan als volgt worden berekend:

$$b = \frac{\Delta Y}{\Delta X} = \frac{Y_2 - Y_1}{X_2 - X_1} \qquad (51)$$

Wanneer je twee punten weet van de lijn, kan je de helling berekenen met de voorgaande formule. Wanneer je de b van de formule weet, kan je het tweede punt invullen in de formule Y = bX + a. Op deze manier kan je a vinden. We zullen deze theorie nu combineren met correlatie. Er is sprake van een positieve correlatie als wanneer als X groter wordt, Y ook groter wordt. De helling van een lijn zal dan ook positief zijn. Bij een correlatie van 0, dus als Y gelijk blijft wanneer X groter wordt, is de helling van de lijn ook nul. Bij een negatieve correlatie zal de helling van de grafiek ook negatief zijn. Ditlaat dus zien dat het teken van de helling gelijk is aan het teken van de correlatie.

7.2

We zullen nu deze concepten gebruiken in de enkelvoudige lineaire regressie (simple linear regression). We definiëren het lineaire regressiemodel als een formule voor een rechte lijn. Het populatie regressiemodel voor Y (afhankelijke variabele) dat wordt voorspeld door X (onafhankelijke variabele) is als volgt:

$$Y_i = \beta_{yx} X_i + \alpha_{yx} + \varepsilon_i \qquad (52)$$

In deze formule:
Y is de criterium variabele
X is de voorspellende variabele
β_{yx} is de populatie helling voor Y die voorspeld is door X
α_{yx} is de populatie intercept voor Y die voorspeld is door X
ε_i zijn de populatie residuen, of fouten van het voorspellen (errors of prediction), dus de delen van Y die niet zijn voorspeld door X.
i staat voor de index van een bepaald geval.

De index i kan waarden aannemen van 1 tot N, waarbij N de totale grootte van de populatie is. Dus i = 1,..,N.

Het population prediction model is:

$$Y'_i = \beta_{yx} X_i + \alpha_{yx} \quad (53)$$

Waar Y'_i de voorspelde waarde van Y is voor een specifieke waarde van X. Dus Y_i is echt waarde of de geobserveerde waarde, terwijl Y'_i is de voorspelde waarde. Dus de populatie predictor error is:

$$\varepsilon_i = Y_i - Y'_i \quad (54)$$

Het verschil tussen de regressie en prediction modellen is dat het regressie model uitdrukkelijke de prediction error (voorspellingsfout) meeneemt als ε_i,. Terwijl het prediction model deze error als vanzelfsprekend meeneemt als een deel van de waarde Y'_i.

Een makkelijke methode om de populatie helling (β_{yx}) en intercept (α_{yx}) te bepalen is als volgt:

$$\beta_{yx} = \rho_{xy} \frac{\sigma_Y}{\sigma_x}$$

$$\alpha_{yx} = \mu_y - \beta_{yx}\mu_x \quad (54)$$

Waarin
σ_Y en σ_x zijn de populatie standaard deviaties voor Y en X.
ρ_{xy} is de populatie correlatie tussen X en Y.
μ_y en μ_x zijn de populatie gemiddelden voor Y en X.

Deze methode kan je niet gebruiken om de helling en intercept van een rechte lijn te bepalen in een regressie analyse met echte gegevens!

7.3

We zullen nu terug gaan naar de echt wereld van de steekproef statistieken, en we zullen gaan kijken naar de steekproef enkelvoudige lineaire regressie (sample simple linear regression). We gebruiken altijd voor de populatie parameter Griekse letter, en voor de steekproef statistieken de normale Engelse (Nederlandse) letters. De steekproef enkelvoudige regressiemodel is als volgt:

$$Y_i = b_{yx} X_i + a_{yx} + e_i \quad (55)$$

Waarin:
Y en X zoals hierboven uitgelegd.
b_{yx} is de steekproef helling voor Y voorspeld door X
a_{yx} is de steekproef intercept voor Y voorspeld door X
e_i zijn steekproef residuen of fouten van voorspelling.
i staat voor de index van een bepaald geval. Waarden tussen i = 1,…., n.

Het steekproef prediction model is als volgt:

$$Y'_i = b_{yx} X_i + a_{yx} \quad (56)$$

Met opnieuw de populatie predictor fout als:

$$e_i = Y_i - Y'_i \qquad (57)$$

Dit is precies hetzelfde als bij het populatie regressiemodel, en populatie predictor model. Alleen nu gebruiken we steekproeven in plaats van populaties.

Dus de helling (b_{yx}) en intercept (a_{yx}) kunnen als volgt worden berekend:

$$b_{yx} = r_{xy} \frac{s_y}{s_x}$$

$$a_{yx} = \bar{Y} - b_{yx}\bar{X} \qquad (58)$$

waarbij
s_y en s_x zijn de steekproef standaard deviaties voor Y en X.
r_{yx} is de steekproef correlatie tussen X en Y
$\bar{Y}$ en $\bar{X}$ zijn de steekproef gemiddelden voor Y en X.

Naar de steekproefhelling, beta, wordt ook wel gerefereerd als (a) de verwachte of voorspelde verandering in Y voor een 1 eenheid verandering in X en (b) de niet gestandaardiseerde of onbewerkte regressie coëfficiënt.

Naar de de steekproef intercept, alpha, wordt gerefereerd als (a) het punt waar de regressielijn kruist met de Y-as en (b) de waarde van Y wanneer X 0 is.

Tot nu toe hebben we gekeken naar de berekeningen in de enkelvoudige lineaire regressie wanneer we gebruik maken van onbewerkte (niet gestandaardiseerde) scores. Dus dit is een unstandardized regression model. De voorspelde waarde van de helling is een onbewerkte regressie helling omdat het de voorspelde verandering is in onbewerkte Y-waarden voor een 1 eenheid verandering in onbewerkte X-waarden.

We kunnen het regressiemodel ook in gestandaardiseerde waarden uitdrukken, in z-scores:

$$z(X_i) = \frac{X_i - \bar{X}}{s_x}$$

$$z(Y_i) = \frac{Y_i - \bar{Y}}{s_Y} \qquad (59)$$

De steekproef gestandaardiseerde lineaire prediction model word als volgt, waar $z(Y'_i)$ de gestandaardiseerde voorspelde waarde van Y is:

$$z(Y'_i) = b *_{yx} z(X_i) = r_{xy}\, z(X_i) \qquad (60)$$

De gestandaardiseerde regressie helling, $b *_{yx}$, soms ook *beta weight* genoemd, is gelijk aan r_{xy}. En de gestandaardiseerde intercept is gelijk aan 0.

Het is heel onwaarschijnlijk dat de voorspelling van Y door X precies goed is. Alleen wanneer er perfecte correlatie (correlatiecoëfficiënt =1.0) is, zal dit gebeuren. De residuen ei, worden ook wel fouten van de voorspelling (errors of estimate) genoemd. Dit is het deel van Y dat niet wordt voorspeld door X. De residuenwaarden zijn random waarden die uniek zijn voor elk individu of voorwerp.

In figuur 7.2 op pagina 330, staat een scatterplot van een regressie-analyse afgebeeld. Je ziet een rechte diagonale lijn. Individuen die boven de regressielijn vallen hebben positieve residuen. Dit betekent dat het verschil tussen de geobserveerde score groter is dan de voorspelde waarde, wat getoond wordt door de regressielijn. Wanneer individuen onder de regressielijn vallen hebben ze negatieve residuen, dit betekent dat het verschil tussen de geobserveerde score kleiner is dan de voorspelde waarde.

Er zijn statistische criteria die helpen bepalen welke methode we moeten gebruiken om de helling en intercept te bepalen. Het criterium dat het meest wordt gebruikt in lineaire regressie analyse is het least squares criterion. Dit criterium zegt dat de som van de gekwadrateerde voorspelde fouten (errors) of residuen het kleinst moet zijn. Dus we willen een lijn vinden met een bepaalde helling en intercept, waarbij de som van de gekwadrateerde residuen het kleinst is. Deze methode wordt ook wel de least squares estimation genoemd, omdat b en a de steekproefwaarden laten zien van de populatie parameters .

We kunnen de utiliteit (utillity) van een predictor variabele bepalen door de partiële kwadratensom van Y, die wordt SStotal genoemd. Het proces lijkt erg op de partiële kwadratensom in ANOVA. In de enkelvoudige lineaire regressie kunnen we SStotal verdelen in:

SStotal = SSreg + SSres

$$\sum_{i=1}^{n}(Y - \bar{Y})^2 = \sum_{i=1}^{n}(Y' - \bar{Y})^2 + \sum_{i=1}^{n}(Y - Y')^2 \qquad (61)$$

Waarin:

- SStotal is de totale kwadratensom van Y
- SSreg is de kwadratensom van de regressie van Y voorspeld door X
- SSres is de kwadratensom van de residuen

In makkelijke bewoordingen: SStotal staat voor de totale variatie in de geobserveerde Y-scores, SSreg staat voor de variatie in Y voorspeld door X, en SSres is de variatie in Y die niet is voorspeld door X. SSreg kijkt hoe goed de best gekozen lijn past in vergelijking met het gemiddelde van Y. SSres laat zien hoe onnauwkeurig het model is. Hoe meer het nadert naar 0, hoe beter het model past.

$r_{xy}^2 = SS_{reg}/SS_{total}$, we can write SS_{total}, SS_{reg}, and SS_{res} as follows:

$$SS_{total} = n\sum_{i=1}^{n} Y^2 - (\sum_{i=1}^{n} Y^2)/n$$
$$SS_{reg} = r_{xy}^2 \; SS_{total}$$
$$SS_{res} = (1 - r_{xy}^2)\; SS_{total} \qquad (62)$$

Waarin de gekwadrateerde correlatie is tussen X en Y. Dit wordt meestal de coefficient of determination genoemd. Dit vertelt ook welke proportie van de totale variatie van de afhankelijke variabele die uitgelegd kan worden door het regressiemodel. De coefficient of determination kan worden gebruikt worden om de effectgrootte te berekenen of als significantie toets . Cohen heeft laten zien wanneer het een klein, gemiddeld of groot effect is:

- Klein effect: r = 0.10, or r2 = 0.01.
- Gemiddeld effect: r = 0.30, or r2 = 0.09
- Groot effect: r = 0.50 or r2 =0.25.

We zullen nu gaan kijken naar vier procedures die worden gebruikt in de enkelvoudige lineaire regressiemodel. De eerste twee zijn significantie toetsen die testen of X wel of niet een significante voorspeller is van Y. Daarna zullen we twee betrouwbaarheidsinterval technieken gebruiken:

Test of significance

Het is belangrijk dat het niet gelijk is aan 0, want dan kunnen we geen goede voorspelling doen. De nul en alternatieve hypothese zijn als volgt:

$H_0: \rho^2_{xy} = 0$
$H_1: \rho^2_{xy} \neq 0$

Deze toets is gebaseerd op de volgende toetsingsgrootheid

$$F = \frac{r^2/m}{(1-r^2)/(n-m-1)} \qquad (63)$$

Waarin:

- F laat zien dat het de F-toets is
- r2 is coefficient of determination
- r2 is het deel van de variatie in Y die niet is voorspeld door X.
- m is het aantal voorspellers (predictors) (in enkelvoudige lineaire regressie altijd 1).
- n is de steekproefgrootte.

De F-toets wordt vergeleken met de kritische waarde van F. Dit is altijd een eenzijdige toets (one-tailed test), en met significantie level alpha. Met vrijheidsgraden gelijk aan m en (n-m-1). Kritische waarde kan je halen uit tabel A.4, Fm, (n – m – 1).

Test

Dit is de toets van de helling. In andere woorden, is de niet gestandaardiseerde regressie coëfficiënt significant anders dan 0? Dezelfde tijd wordt uitgevoerd voor b*, de gestandaardiseerde regressie coëfficiënt. De nul en alternatieve hypothese zijn als volgt:

$H_0: \beta_{yx} = 0$
$H_1: \beta_{yx} \neq 0$

Om te toetsen of de regressie coëfficiënt gelijk is aan 0, hebben de standaardfout voor de helling b nodig. Daarom moeten we een paar nieuwe concepten bespreken. De eerste is de variatie fout van de schatting (variance error of estimate, variance of the residuals), is gedefinieerd als:

$$s^2_{res} = \sum e_i^2 / df_{res} = SS_{res}/df_{res} = MS_{res} \qquad (64)$$

Waarin dfres = (n – m- 1). Deze waarde laat de variatie in de residuen zien. Een relatieve grote variance of error, betekent dat er bepaalde hele grote residuen zijn, dat betekent een slechte voorspelling. Een relatieve kleine variance of error, laat zien dat de voorspelling goed is.

Het volgende nieuwe concept is de standaardfout van de voorspelling (standard error of estimate, root mean square error). Dit is het kwadraat van de variatie fout van de voorspelling. Dus dit is de standaarddeviatie van de residuen. De standaarderror wordt genoteerd als sres.

Het laatste nieuwe concept is de standaardfout van b. Deze wordt genoteerd als sb en gedefinieerd als volgt:

$$s_b = s_{res} / \sqrt{[n \sum X^2 - (\sum X)^2]/n} = SS_{res} / \sqrt{SS_x} \qquad (65)$$

We willen dat sb klein is wanneer we Ho willen verwerpen. Dus we hebben een kleine sres en een grote SSx nodig. In andere bewoordingen, we willen dat er een grote spreiding is in de scores van X.

Wanneer we deze concepten samenvoegen in een toetsingsgrootheid voor de significantie van de helling b, is dat als volgt:

$$t = \frac{b}{s_b} \qquad (66)$$

We vergelijken deze waarde met de kritische waarde uit tabel A.2. Een tweezijdige test voor een non-directional H1. Voor het significantie level en met de vrijheidsgraden (n-m-1).

We kunnen ook het betrouwbaarheidsinterval om de helling van b maken:

$$CI\ (b) = b \pm (\alpha/2)\ t_{(n-m-1)sb} \qquad (67)$$

Confidence interval for the predicted mean value of Y (betrouwbaarheidsinterval voor de voorspelde gemiddelde waarde van Y)

Third procedure is to develop a CI for the predicted mean value of Y, denoted by $\bar{Y}'_0$.
The standard error of $\bar{Y}'_0$ is:

$$s(\bar{Y}'_0) = s_{res} \sqrt{(1/n) + [(X_0 - \bar{X})^2 / SS_X}} \qquad (68)$$

Uit deze formule kunnen we opmaken dat we de beste voorspellingen doen in het midden van de verdeling van de X-scores. En we maken de slechtste voorspellingen voor de extreme waarden van X. Het betrouwbaarheidsinterval wordt als volgt berekend:

values of X. A CI around $\bar{Y}'_0$ is formed as follows:

$$CI(\bar{Y}'_0) = \bar{Y}'_0 \pm (\alpha/2)\ t_{(n-2)} s(\bar{Y}'_0) \qquad (69)$$

Prediction interval for individual values of Y

De laatste methode is het maken van een prediction interval (PI). Dus de voorspelde score weten we voor dat bepaalde individu, maar de criterium score hebben we nog niet geobserveerd. Dit is anders dan bij het betrouwbaarheidsinterval, waar de scores van het individu al wel zijn geobserveerd. Dus het betrouwbaarheidsinterval maakt gebruikt van voorspelde waarden, terwijl de PI gebruikt maakt van de voorspelde waarde van een individu die nog niet is geobserveerd.

De standaardfout is:

$$s(Y'_0) = s_{res}\sqrt{1 + \left(\frac{1}{n}\right) + [\{X_0 - \bar{X})^2/SS_X]} \qquad (70)$$

The PI around Y'_0 is formed as follows:

$$PI\ (Y'_0) = Y'_0 \pm (\alpha/2)\ t_{(n-2)}s(Y'_0) \qquad (71)$$

We zullen nu gaan kijken naar de aannames die worden gemaakt bij de enkelvoudige lineaire regressie: (a) onafhankelijkheid, (b) homogeniteit, (c) normaliteit, (d) lineair en (e) een vastgestelde X.

Onafhankelijkheid

Deze aanname hebben we ook al gehad met het ANOVA model. Een andere manier om deze aanname uit te leggen in het regressiemodel is dat the fouten (errors) in de voorspelling of de residuen random en onafhankelijk moeten zijn. Dus er moet geen patroon zijn in de fouten (errors). Er zijn verschillende typen residuen. De eerste is e, onbewerkt (raw) residu. Dit is een onbewerkt residu omdat X en Y ook onbewerkte scores zijn, dus gebruiken de originele schaal. Sommige onderzoekers vinden het niet fijn om onbewerkte scores te gebruiken. Daarom zijn er verschillenden gestandaardiseerde residuen ontworpen. Deze waarden zijn gemeten op een z-score schaal, met een gemiddelde van 0 en een variatie van 1. En ongeveer 95% van de waarde ligt tussen 2 waarden van 0 af. We zullen in de uitleg van SPSS studentized residuen gebruikt. Dit is een vorm van gestandaardiseerde residuen die gevoeliger is voor het ontdekken van uitbijters.

De makkelijkste manier om te kijken of er onafhankelijkheid is, is door het scatterplot (Y vs. X) of een plot van de residuen te bekijken. Wanneer er onafhankelijkheid is, is er een random weergave van de punten. Wanneer er geen onafhankelijkheid is, zullen de punten een patroon weergeven. Het niet waarmaken van deze aannamen kan komen door 3 situaties: (1) wanneer de observaties zijn verzameld na verloop van tijd, (2) wanneer de observaties zijn gemaakt in blokken, zodat de observaties in een bepaald blok meer op elkaar lijken dan de observaties van verschillende blokken, (3) wanneer de observatie is herhaald. Wanneer er geen onafhankelijkheid is, heeft dit invloed of the voorspelde standaardfouten, die kunnen onder- of overschat zijn. Wanneer deze aanname helemaal niet wordt waargemaakt kan je generalized of weighted least squares gebruiken.

Homogeniteit

Tweede aanname is dat er homogeniteit van de variatie is. Deze aanname moet een beetje anders worden genoemd wanneer we het gebruiken in het regressiemodel. We gebruiken het concept conditional distribution. In de regressie analyse, wordt de conditional distribution gedefinieerd als de distributie van Y voor een bepaalde waarde van X. Dus de aanname van homogeniteit is dat de conditional distribution een constante variatie moet hebben voor alle waarden van X. In de plot van de Y-waarden of de residuen tegen de X-scores, kan dit worden bekeken. Wanneer er geen homogeniteit is zal de variatie van de residuen groter worden wanneer X groter wordt.

Wanneer er geen homogeniteit is, zijn de schattingen van de standaardfout groter, en is de validiteit van de toets aangetast. Ook zorgen grotere standaardfouten ervoor dat het moeilijker is om de nulhypothese te verwerpen, dus een grotere Type II fout.

Wanneer de aanname heel erg is geschonden, kan je een soort van transformatie gebruiken, ook wel variatie stabilisatie transformatie genoemd. Meestal wordt ofwel een log ofwel een kwadraat van Y gebruikt. Dit kaner ook voor zorgen dat de normaliteit wordt verbeterd. Een tweede oplossing is om generalized of weighted least squares te gebruiken. Een derde oplossing is om een vorm van robuuste voorspelling te gebruiken.

Normaliteit

In de regressie analyse is de aanname van normaliteit dat de conditional distribution van of de Y-scores of de verwachte fouten (residuen) een normale verdeling moeten hebben. Uitbijters zorgen er meestal voor dat er geen normaliteit is. De voorspelde waarden van het regressiemodel zijn gevoelig voor uitbijters. Meestal zal de regressielijn naar de uitbijter toe uitweiden, omdat het least squares principe altijd probeert de lijn te vinden die het beste bij de punten past. Uitbijters kunnen een resultaat zijn van (a) een eenvoudige opname of een fout met het invoeren van de gegevens, (b) een fout in de observatie, (c) een instrument dat niet goed werkt, (d) verkeerd gebruik van de administratie instructies, of (e) een echte uitbijter.

Een makkelijke manier om te gebruiken bij uitbijters is om twee regressie analyses te doen, eentje met de uitbijter en eentje zonder de uitbijter. Daarna kan je deze twee analyses vergelijken.

Er zijn twee manieren om te kijken of er normaliteit is. De makkelijkste is om de symmetrie te bekijken in een histrogram of boxplot. Je kan ook kijken naar de platheid en scheefheid, maar een nonzero kurtosis (niet nul platheid) heeft weinig effect op de verwachte waarden. Nonzero skewness (niet nul scheefheid) heeft hier meer effect op. Een regel is dat wanneer de waarde van de scheefheid groter is dan 1.5 of 2.0, moet je verder kijken of het effect heeft.

Je kan ook een Q-Q plot gebruiken. Wanneer er normaliteit is, zullen de punten in een rechte, diagonale lijn vallen. Er zijn ook verschillende testen die kunnen worden gebruikt. Ook transformaties kunnen worden gebruikt. De meest gebruikte transformaties zijn (a) om de afhankelijke variabele te transformeren door log (te gebruiken bij positieve scheefheid), (b) of kwadraat (te gebruiken bij positieve of negatieve scheefheid).

Lineariteit

Dit betekent dat er een lineaire relatie is tussen X en Y. Wanneer deze relatie lineair is, zullen de helling en intercept onbevooroordeeld zijn.

Om de aanname van lineariteit te checken kan een scatterplot van Y en X worden geanalyseerd. Wanneer er een lineaire relatie is zal er een patroon in de punten zijn.

Er zijn twee oplossingen wanneer de relatie niet lineair is. De eerste is om de een van de variabelen te transformeren. Daarna kan je methode van de least squares worden toegepast. Maar wanneer je variabelen gaat transformeren, heb je een andere schaal, wat het moeilijke maakt, want ze moeten eigenlijk op een originele schaal zijn. Een tweede oplossing is om een nonlinear model te gebruiken.

Vastgestelde X

Dit betekent dat de X is vastgesteld en niet random is. Dit zorgt ervoor dat het regressiemodel alleen valide is voor die bepaalde waarden van X die echt zijn geobserveerd in de analyse. Twee situaties komen dan op, namelijk extrapolation en interpolation van de waarden van X. Over het algemeen mogen we geen voorspellingen doen over individuen die een X-score hebben buiten de waarden die we hebben gebruikt om het model te voorspellen (extrapolating).

Het kan ook zijn dat we geen belang hebben bij het maken van voorspellingen over individuen die een X-score hebben binnen de waarden van het model (interpolation) In deze situatie verwachten we de prediction fouten kleiner te zijn dan in de extrapolating situatie.

Deze tabel geeft een korte samenvatting van de aannames en de effecten wanneer deze niet worden nageleefd.

Aanname	Effect wanneer niet nageleefd
Onafhankelijkheid	Heeft effect op de standaardfouten van het model
Homogeniteit	Vooroordeel in variaties van de residuen Kan de standaarfouten opblazen en dus de kans of een Type II fout vergroten Kan zorgen voor niet normale verdelingen
Normaliteit	Minder precieze helling, intercept en R2
Linear	Vooroordeel in de helling en intercept Verachte verandering in Y is niet constant, en is afhankelijk van de waarde van X Verminderde grootte van de coefficient of determination
Waarden van X zijn vastgelegd	Extrapolating: prediction fouten zijn groter, en kan ook een vooroordeel zijn in de helling en intercept Interpolating: kleiner effect dan extrapolating. Wanneer alle andere aannames worden nagekomen is dit effect te verwaarlozen.

7.4

Om een enkelvoudige lineaire regressie uit te voeren, moet je gegevens hebben die uit twee variabelen bestaan, namelijk een afhankelijke variabel en een onafhankelijke variabele. We zullen nu kijken naar het stappenplan om een enkelvoudige linearie regressie uit te voeren in SPSS:

- Ga naar “Analyze” en selecteer “regression” en selecteer “Linear”
- Klik de afhankelijke variabele in de “dependent” box, en de onafhankelijke variabele in de “Independent(s)” box.
- Vanuit de “Linear regressie” box, door de kliken op “statistics” geeft het opties om te selecteren. Selecteer (1) estimates, (2) confidence intervals, (3) model fit, (4) descriptive, (5) Durbin-Watson, (6) case wise diagnostics. Klik op “continue”.
- Klik op “plots” om de volgende dingen te selecteren: (1) histogram, en (2) normal probability plot”. Klik op “continue”
- Klik op “save” en selecteer de volgende dingen: (1) unstandardized en (2) studentized. Onder het kopje heading of distances, selecteerd het volgende (1) mahalanobis en (2) Cook’s. Onder het kopje influence statistics, selecteer het volgende (1) DFBETA(s), en (2) standardizes DFBETA(s). Klik op “continue” en dan op “OK”.

We zullen nu verder kijken naar de waarden die we hebben opgeslagen in ons bestand (zie pagina 351):

- 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 voor een voorspelde waarde van de standaarddeviatie. 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 mahalnobis afstanden 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 regen, wanneer de Cook's waarde groter is dan 1.0 geeft dit aan dat het problematisch kan zijn.
- DFB0_1 en DFB1_1 zijn niet gestandardiseerde DFBETA waarden voor de intercept en helling. Deze waarden laten een voorspelling zien van de intercept en helling wanneer dat geval is verwijderd.
- SDB0_1 en SDB1_1 zijn gestandaardiseerde DFBETA waarden. Deze kan je makkelijke 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.

We kunnen een grafiek maken met de residuen tegen de waarden van X om te kijken of er onafhankelijkheid is. Deze aanname is waargemaakt wanneer de punten in een random patroon vallen in een gebied van -2.0 tot +2.0. We kunnen dezelfde grafiek gebruiken om te kijken naar homogeniteit. Wanneer deze aanname is waargemaakt, zijn de residuen constant verdeeld over het gebied van de X-waarden. Wanneer de verdeling van de residuen vergroot of verkleind in de grafiek kan dit laten zien dat er geen homogeniteit is. Wanneer we maar 1 onafhankelijke variabelen hebben is een enkelvoudige bivariate scatterplot handig om te kijken voor lineariteit. De afhankelijke variabele op de Y-as en de onafhankelijke variabele op de X-as. Je kunt ook kijken naar een grafiek van de studentized residuen tegen de X-waarden. Wanneer hier een random weergaven van de punten is met een absolute waarde van 2 of 3 is de aanname waargemaakt.

We kunnen de residuen bekijken voor normaliteit. We kunnen ook de scheefheid en platheid gebruiken. Wanneer beide waarden in een gebied met een absolute waarde van 2.0 zitten is er normaliteit. Je kan ook de Shapiro-Wilk (S-W) toets gebruiken. Ook Q-Q plots kunnen worden gebruikt.

7.5

Ook hier de priori en post hoc power analyse worden gedaan in G*power. Hierin moet je goede test familie selecteren. We voeren een enkelvoudige lineaire regressie uit. Om dit te vinden klik je op "tests" daarna "correlation and regression", daarna "Linear bivariate regression: size of slope". Wanneer je dit hebt geselecteerd zal de test familie automatisch veranderen in een t-toets. De input parameters voor de post-hoc toets zijn: (1) number of tails, (2) effect size, slope H1, (3) alpha level, (4) total sample size, (5) slope H0, (6) standard deviation of X, (7) standard deviation of Y.