8. Regressie

Een introductie

Al vaak is genoemd dat statistische modellen allemaal neerkomen op uitkomst = model + error. Dit model kun je ook gebruiken om de uitkomst te voorspellen, met een correlatie als model. Dit model ziet er dan uit als:

Uitkomst = (b_0+b_1X) + error.

De b is hier de correlatiecoëfficiënt. Als je de errorterm even wegdenkt, kun je deze vergelijking ook zo schrijven:

Uitkomst = ax + b

Dit is de algemene formule voor een lineaire functie. Een rechte lijn wordt bepaald door twee dingen, de richtingscoëfficiënt, b_1 (slope in het Engels) en het snijpunt met de y-as, b_0 (intercept). Deze b_0 en b_1 worden ook wel regressiecoëfficiënten genoemd. Een model met een positieve slope heeft een positieve relatie, een lijn met een negatieve b_1 beschrijft een negatieve relatie. De slope vertelt je hoe het model eruit ziet, de intercept vertelt waar het model zich bevindt in een grafiek.

Als je deze regressiecoëfficiënten weet, kun je de uitkomst voorspellen. Er is echter altijd een error, de voorspelling is niet 100% accuraat, het is een *voorspelde waarde*.

Je kunt ook een lineair model hebben met meer dan twee variabelen. Meerdere predictoren voorspellen dan de uitkomst. Dit heet *multipele regressie*. Deze predictoren hebben dan elk een eigen regressiecoëfficiënt.

Het model geeft een voorspelde waarde, die iets afwijkt van de werkelijke data. Het verschil tussen de door het model voorspelde waardes en de werkelijke waardes heet de error, of de *residuen* (*residuals*).

Goodness of fit
Om voorspellingen te maken hebben we het model, dus de coëfficiënten nodig dat het beste bij de gegevens past. Om de parameters te vinden die het beste bij de gegevens past wordt de least squares methode toegepast. Je kunt dan kijken hoe goed de fit is van het model. Dit doe je door te kijken naar de residuen. Dit kan met de volgende formule:

Totale error = Σ (geobserveerd – model)^2

Omdat je in regressie spreekt van residuen in plaats van error, noem je deze sum of squares van de error de *residual sum of squares* of *residuensom* (SS_R). Als de SS_R klein is, past de lijn goed bij de data. Als de SS_R groot is, is de lijn helemaal niet representatief voor de data.

Het model dat het best past bij je data, is het model dat de kleinste SS_R oplevert. Dit doen we met de least squares method, waarbij de computer de parameters (de b's in je model) schat die de kleinste residuensom oplevert.

Als je het beste model hebt, moet je kijken hoe goed het model is. De beste lijn kan immers nog steeds een hele slechte fit hebben met de data.

De SS_R geeft aan hoe groot je error is, maar je hebt een vergelijkpunt nodig om te kijken of je model beter is dan niets. Als je geen model hebt, is je gemiddelde de enige nuttige schatting die je kunt maken. Het gemiddelde gebruik je dus als model waarbij er geen relatie is tussen de variabelen.

Als je het gemiddelde als model neemt, kun je de verschillen uitrekenen tussen de data en het gemiddelde, en daarmee een sum of squares uitrekenen. Dit heet de *totale kwadratensom (SS_T)*, de totale hoeveelheid verschillen wanneer het basismodel, het gemiddelde, wordt toegepast op de gegevens.

De *residuensom (SS_R)* is het verschil tussen het regressiemodel en de data, dus de error wanneer het best mogelijke model op de data wordt toegepast. Het verschil tussen de SS_T en de SS_R is de verbetering in voorspelling die het model biedt boven het gemiddelde. Deze verbetering is de sum of squares van het model, de SS_M.

Een grote SS_M betekent dat het regressiemodel erg verschillend is van alleen het gemiddelde gebruiken om de uitkomstvariabele te voorspellen. Met deze kwadratensommen kun je de proportie van verbetering uitrekenen:

$$R^2 = \frac{SS_M}{SS_T}$$

R^2 is hier de proportie van verbetering en als je het vermenigvuldigt met honderd krijg je er een percentage uit. R^2 is de proportie verklaarde variantie dat verklaard wordt door het model tegenover hoeveel variantie er in totaal is. Als je hieruit de wortel trekt krijg je de Pearson correlatie.

Een tweede manier om de kwadratensom te gebruiken is voor het berekenen van de F-toets. De F-toets is gebaseerd op de ratio van de verbetering door het model (SS_M) en het verschil tussen het model en de geobserveerde gegevens (SS_R). Omdat de sum of sqaures afhangen van de steekproefgrootte, gebruik je de *gemiddelde kwadratensom (mean square, MS)*.

Om de mean square te krijgen deel je de SS door het aantal vrijheidsgraden. Voor SS_M is het aantal variabelen het aantalvrijheidsgraden (dus bij een simpele regressie heb je df_M=1) en voor SS_R is het aantal vrijheidsgraden het aantal observaties min het aantal parameters (dit komt terug in hoofdstuk 11). De F-ratio is een meting van hoeveel het model de voorspelling van de uitkomst verbetert in vergelijking met de onnauwkeurigheid in het model.

$$F = \frac{MS_M}{MS_R}$$

Een goed model heeft een grote F-waarde, want dat betekent dat er een grote verbetering is in de voorspelling (een grote MS_M) en dat het verschil tussen de voorspelling en de data klein is (een kleine MS_R).

Individuele voorspellers

Elke predictor in het regressiemodel heeft een coëfficiënt (b). In simpele regressie, waarbij er maar één predictor is, is deze b de richtingscoëfficiënt van de regressielijn.

De b geeft dan de verandering in uitkomst aan die komt door de verandering in de predictor. Als je het gemiddelde gebruikt als model, is er geen verandering in de uitkomst bij een verandering in de predictor. Wat de waarde van de predictor ook is, de geschatte uitkomstwaarde is altijd het gemiddelde. Hierbij heb je een richtingscoëfficiënt (b) van 0. De regressielijn is horizontaal.

Als een variabele significant een uitkomst wil voorspellen, dan moet het dus een b-waarde hebben die significant verschilt van 0. Dit kan getoetst worden met de *t-toets*. Hierbij toets je de nulhypothese dat de waarde van b gelijk is aan 0. De t-toets is net als de F-toets een ratio van de verklaarde variantie tegen de onverklaarde variantie van meetfouten. Om error van de b te schatten, gebruik je de standard error. De formule is:

$t = b_{geobserveerd} - b_{verwacht} / SE_b$

Vanuit de nulhypothese ga je ervan uit dat $b_{verwacht} = 0$, dus dan wordt de formule:

$t = b_{geobserveerd} / SE_b$

Bij regressie is het aantal vrijheidsgraden N-p-1. N is hierbij de totale steekproefgrootte en p is het aantal predictors. Bij simpele regressie heb je dus df = N-2. Als je een grote t-waarde hebt, die groter is dan de kritieke waarde (zie de tabel in de appendix), verwerp je de nulhypothese, b wijkt dan significant af van 0. De predictor heeft dan een significante bijdrage in het voorspellen van de uitkomst.

Bias in regressiemodellen

Nadat het model gemaakt is, zijn er twee belangrijke vragen die gesteld moeten worden:

1. Wordt het model beïnvloed door een klein aantal gevallen?
2. Kan het model gegeneraliseerd worden naar andere steekproeven?

Voor het beantwoorden van de eerste vraag kun je kijken naar uitschieters en invloedrijke gevallen. Uitschieters kunnen de schattingen van de parameters in het model sterk beïnvloeden. Uitschieters zijn te herkennen aan een groot residu, een grote afwijking van de trend.

Residuen laten de meetfout in het model zien. De *ongestandaardiseerde residuen* (de normale residuen) zijn in dezelfde schaal als de uitkomstvariabele gemeten en zijn dus moeilijk te gebruiken in andere modellen. *Gestandaardiseerde residuen* zijn residuen die tot z-scores zijn getransformeerd en kunnen bij meerdere modellen gebruikt worden als standaard. Het voordeel is dat er voor deze gestandaardiseerde residuen richtlijnen zijn over welke residuen acceptabel zijn en welke niet.

Er is ook nog de *studentized residu* die varieert van punt tot punt. Ze hebben dezelfde proporties als de gestandaardiseerde residuen alleen geven ze een iets preciezere schatting van de meetfout in een specifiek geval.

Invloedrijke gevallen

Soms heb je een of enkele scores die de schattingen van de parameters van het model enorm beïnvloeden. Met verschillende statistieken kunnen deze scores gevonden worden. De *aangepaste voorspelde waarde* is zo'n statistiek. Deze techniek berekent het model zonder een bepaalde score. Wanneer de score weinig invloed heeft zal de aangepaste voorspelde waarde ongeveer overeenkomen met de voorspelde waarde.

Het *verwijderde residu* is het verschil tussen de aangepaste voorspelde waarde en de geobserveerde waarde. Als je dit residu deelt door de standaard deviatie krijg je de *Studentized verwijderde residu*.

Deze techniek van het verwijderde residu geeft alleen aan hoeveel invloed de score heeft op hoe goed het model die ene score kan voorspellen. Het geeft geen informatie over hoe die score het hele model beïnvloed. De *Cook's distance* kijkt wel naar het effect van één score op het gehele model.

Een tweede manier om de invloed te meten is *leverage* (hoofd waardes) die de geobserveerde waardes boven de voorspelde waardes stelt. Het kan uitgerekend worden met (k+1)/n, waarbij k staat voor het aantal voorspellers. Er zijn ook nog de *Mahalanobis distances* die de afstand meet van de scores tot het gemiddelde van de predictorvariabele.

Je kunt een regressieanalyse uitvoeren met een score erbij en daarna zonder die score, en dan kijken hoe groot het verschil in regressiecoëfficiënten is. Het verschil tussen een schatting van de parameters met alle scores en de schattingen waarbij een score is verwijderd, heet de *DFBeta.*

Om geen invloed van de schalen te hebben wordt de *gestandaardiseerde DFBeta* gebruikt. Waardes boven de 1 hebben veel invloed op de modelparameters.

Een gerelateerde statistiek is de *DFFit,* het verschil tussen de voorspelde waarde voor een score wanneer het model is berekend met die score erbij en wanneer het model is berekend zonder die score. Wanneer een score geen invloed heeft, is de DFFit 0. Ook hier is een *gestandaardiseerde DFFit* mogelijk.

Als laatste is er ook nog de *covariantie ratio* (CVR). Dit meet of een score invloed heeft op de variantie van de regressieparameters. Een score heeft weinig invloed wanneer de waarde van de CVR dicht bij 1 ligt.

De hierboven genoemde statistieken zijn een manier om te kijken hoe goed het model bij de gegevens past. Het is niet een manier om te kijken welke punten handig zijn om erin te houden of erbuiten te laten om een b-waarde significant te laten zijn.

Generalisatie

In de sociale wetenschappen wil men graag de bevindingen kunnen generaliseren naar een hele populatie. Hiervoor moet aan alle assumpties voor regressieanalyse zijn voldaan.

Lineariteit en optelbaarheid: De uitkomstvariabele moet in werkelijkheid lineair samenhangen met alle predictors, en als je meerdere predictorvariabelen hebt, moet hun gecombineerde effect het best beschreven worden door het optellen van hun effecten.

Onafhankelijke meetfouten: Voor elke twee observaties moet het residu ongecorreleerd zijn. Dit wordt ook beschreven als gebrek aan *autocorrelatie*. Het kan getest worden met de *Durbin-Watson toets*. Deze toets test seriële correlaties tussen meetfouten. De waardes liggen tussen de 0 en de 4 en de waarde 2 betekent dat de residuen ongecorreleerd zijn. Een waarde groter dan 2 betekent een negatieve correlatie en lager dan 2 een positieve correlatie.

Homoscedasticiteit: De residuen moeten op elk niveau dezelfde variantie hebben. Wanneer de varianties niet gelijk zijn wordt dit *heteroscedasticiteit* genoemd.

Normaal verdeelde meetfouten: De residuen zijn random, normaal verdeelde variabelen en hebben een gemiddelde van 0.

Voorspellers moeten ongecorreleerd zijn met externe variabelen: Externe variabelen zijn variabelen die wel invloed hebben op de uitkomstvariabele, maar niet in de regressieanalyse opgenomen zijn. Dit lijkt op het 'derde variabele probleem' bij de correlatie. Als er wel een correlatie is worden de conclusies minder betrouwbaar.

Soort variabelen: Alle voorspellervariabelen moeten kwantitatief (op intervalniveau) of categorisch (met twee categorieën) zijn. De uitkomstvariabele moet kwantitatief, continu en onbegrensd zijn.

Geen perfecte *multicollineariteit*: Er mag geen perfect lineaire relatie tussen twee of meer predictors zijn. De predictors mogen dus niet sterk correleren.

Geen variantie van 0: De predictorvariabelen moeten variantie hebben.

Als het model aan de assumpties voldoet, dan is het regressiemodel van de steekproef gemiddeld hetzelfde als het model van de populatie. Het kan nog steeds dat het model van de steekproef afwijkt van die van de populatie, maar de kans dat de modellen overeenkomen, is een stuk groter wanneer aan de assumpties is voldaan.

Cross-validatie van het model

Cross-validatie is de nauwkeurigheid van een model bepalen bij verschillende steekproeven. Wanneer een model gegeneraliseerd kan worden, zou het ook op andere steekproeven dezelfde uitslagen moeten geven. Nadat we het regressiemodel hebben bepaald, zijn er twee methoden voor de cross-validatie:

Aangepaste R^2: Deze statistiek wordt in SPSS ook weergegeven. Het schat het verlies van de voorspelde power in (krimp). De aangepaste waarde geeft aan hoeveel variantie in Y wordt verklaard als het model voor de hele populatie zou gelden. Met de formule van Stein kan de aangepaste R^2 ook berekend worden (te vinden op pagina 312).

Gegevens splitsen: Dit betekent dat je de data random in tweeën splitst en dus eigenlijk twee kleine regressievergelijkingen maakt. De vergelijkingen kunnen dan weer met elkaar vergeleken worden.

Steekproefgrootte

Er zijn een aantal algemene regels die gelden voor de steekproefgrootte. Zo moet je minimaal 10 of 15 deelnemers per voorspellervariabele hebben. Verder hangt de steekproefgrootte af van de effectgrootte van het effect dat je wil meten en van de power. Het belangrijkste motto: Hoe groter de steekproefgrootte, hoe beter!

Het beste is om de grafiek van Miles en Shevlin (2000) te gebruiken om de steekproefgrootte te bepalen, waarbij rekening wordt gehouden met de power en de effectgrootte. Zie hiervoor bladzijde 314.

Simpele regressie in SPSS

Voordat je gaat analyseren, moet je een scatterplot maken van je data om de assumptie van lineariteit te controleren en om te kijken of je uitschieters hebt. Ook de andere assumpties moet je controleren als je het model wil kunnen generaliseren.

Bij regressie staan de uitkomstvariabele en de voorspellers in verschillende kolommen en elke rij geeft de onafhankelijke waardes weer. Voor een regressieanalyse ga je naar analyze - regression - linear. Bij dependent plaats je de uitkomstvariabele en bij independent(s) plaats je de voorspeller (zie bladzijde 317). Bij bootstrap kun je een bootstrap betrouwbaarheidsinterval voor de regressiecoëfficiënten krijgen. Selecteer hiervoor Bias corrected accelerated (BCa).

Interpretatie

Bij de output vind je als eerste een tabel met de samenvatting die R en R^2 weergeeft. De R is bij simpele regressie de correlatie tussen de predictor en de uitkomst. Ook vind je een ANOVA tabel. Daarin vind je de MS en de F-ratio en de p-waarde van die F-ratio. Bij significantie kunnen we zeggen dat het regressiemodel significant beter voorspelt dan alleen het gemiddelde. Voor de output zie bladzijde 318.

Model parameters

De ANOVA tabel zegt of het model in het algemeen een goede voorspelling geeft van de uitkomstvariabele, maar vertelt niets over de individuele bijdrage van de variabelen in het model. Bij simpele regressie heb je maar één variabele, dus als het een goed model is, was het ook een goede predictorvariabele.

De schattingen van de parameters in het model vindt je in de tabel coefficients. b_0 is terug te vinden in de SPSS tabel bij B (constant). b_1 staat ook onder B bij de predictorvariabele (zie blz 319). Deze waarde geeft dus aan met hoeveel de uitkomstvariabele verandert als de predictor verandert met een bepaalde eenheid. Er staat een t-toets achter, die test of de b-waarde significant afwijkt van 0.

In de tabel bootstrap staat het bootstrap betrouwbaarheidsinterval. Bij BCa 95% Confidence Interval staan de grenzen waartussen de b-waardes waarschijnlijk liggen. Als 0 niet in dit interval zit, is er een significante relatie tussen de variabelen.

Met dit model kun je voorspellingen doen. Dit doe je door de b-waardes in te vullen in de standaard regressieformule. Nu kun je een X-waarde invullen om een bepaalde uitkomst te voorspellen.

Multiple regressie

Multiple regressie is hetzelfde als de simpele regressie alleen worden hier meerdere voorspellers gebruikt. Wanneer je meerdere voorspellers aan de vergelijking toevoegt kan je daarmee de proportie verklaarde variantie vergroten, maar dit hoeft niet altijd het geval te zijn. De predictors die je meeneemt in een model, en de manier waarop je ze meeneemt, hebben veel invloed. Dus je moet nooit oneindig veel predictors random in een regressieanalyse gooien en hopen dat er iets goeds uit komt, de keuze voor de voorspeller en methode moet een theoretische basis hebben.

Regressiemethoden

Wanneer de predictorvariabelen helemaal niet gecorreleerd zijn met elkaar, maakt de methode die je kiest niet uit. Dit komt echter bijna nooit voor, dus dan zijn er cruciale verschillen. Er zijn de volgende methodes voor het selecteren van voorspellers:

Hiërarchische methode (blockwise entry): Hier worden voorspellers toegevoegd gebaseerd op de ervaring van de onderzoeker. De onderzoeker bepaalt de rangorde van de voorspellers, waarbij als eerste predictors worden toegevoegd die gebleken zijn uit eerder onderzoek, en daarna pas de nieuwe voorspellers. De nieuwe voorspellers kunnen allemaal tegelijk ingevoerd worden, stapsgewijs of hiërarchisch (waarbij de waarschijnlijk beste voorspeller als eerste wordt toegevoegd).

Gedwongen methode (forced entry of enter): Hier wordt ook rekening gehouden met de theorie waarop de voorspellers gebaseerd zijn, alleen worden hier alle voorspellers tegelijk toegevoegd en wordt er geen onderscheid gemaakt in rangorde.

Stapsgewijze methodes (stepwise): Hierbij wordt de volgorde van de predictors in het model bepaald met wiskundige criteria. Bij de voorwaartse methode wordt de variabele dat als best de uitkomst variabele voorspelt (de hoogste correlatie) als eerste toegevoegd. Als dit een verbetering is van het model, gaat de computer op zoek naar de volgende predictor om toe te voegen, de predictor die het grootste deel van de overgebleven variantie verklaard, die niet verklaard wordt door de eerste predictor. Tegelijk wordt steeds bekeken of een predictor niet beter weggehaald kan worden.
De achterwaartse methode is het tegenovergestelde van de voorwaartse methode. Hierin worden eerst alle voorspellers in de vergelijking toegevoegd en worden ze later één voor één verwijderd.

Het kiezen van een methode
De stapsgewijze methode kun je beter niet gebruiken, omdat het werkt op basis van wiskundige criteria in plaats van theorie. Het bekijkt of een predictor in een model past, op basis van de predictors die al in het model zitten. De achterwaartse methode werkt iets beter dan de voorwaartse methode omdat hier beter rekening wordt gehouden met *onderdrukkende effecten*. Bij voorwaartse methode is de kans groter dat er voorspellers niet bijzitten die wel effect hebben ook is een groter risico voor type II fouten.

Modellen vergelijken

Bij de hiërarchische (en stapsgewijze, maar die gebruiken we niet) methode voeg je predictors in fases toe aan het model. Je wil daarom weten of deze extra toevoegingen het model ook werkelijk verbeteren. Als de extra voorspellers zorgen voor een beter model, verklaart het meer variantie, dus is de R^2 groter.
Omdat de R^2 altijd groter is bij meer voorspellers, kun je het *Aikake informatie criterium (AIC)* gebruiken, wat de fit van het model beschrijft, terwijl het compenseert voor het feit dat je meerdere predictors gebruikt. Als de AIC kleiner wordt, heeft je model een betere fit.

Multicollineariteit

Er is sprake van multicollineariteit als twee of meer voorspelles in een regressiemodel sterk met elkaar correleren. *Perfecte collineariteit* is wanneer twee voorspellers een perfect lineaire relatie met elkaar hebben. Dit maakt het onmogelijk om unieke schattingen van de regressiecoëfficiënten te maken.
Wanneer de collineariteit groot is kunnen de volgende problemen ontstaan:

Onbetrouwbare b's: De b wordt minder betrouwbaar, omdat de standaardfout toeneemt. Dit betekent dus dat de b-waardes meer variëren tussen steekproeven, en dus kun je minder betrouwbare schattingen voor de populatie doen.

Er komen grenzen aan de grootte van R: Wanneer twee voorspellers aan elkaar correleren kunnen ze haast geen uniek deel van de variantie verklaren. Het percentage variantie dat de ene predictor verklaart, komt voor een heel groot deel overeen met de variantie die de andere predictor verklaart.

Het belang van voorspellers: Wanneer voorspellers sterk met elkaar correleren weten we niet wat het belang is van elke voorspeller. Het is om het even of je de ene of de andere predictor in het model opneemt.

In SPSS kan je met de *variantie inflatie factor (VIF)* de multicollineariteit berekenen. Een waarde van tien geeft aan dat er veel multicollineariteit is. Tolerantie is een andere waarde voor de multicollineariteit (1/VIF).

Multiple regressie in SPSS

Ook bij multipele regressie moet je voor de analyse de voorwaarden controleren. De regressieanalyse kun je uitvoeren via analyze – regression – linear.

Opties
Voor de hiërarchische methode heeft SPSS blokken voor elke stap. Bij next kom je in een volgend blok. Bij verschillende blokken is het mogelijk verschillende methoden te kiezen, in het afrolmenu bij Method (blz. 327). Je kunt twee of meer voorspellers tegelijk invoeren door de control toets ingedrukt te houden terwijl je de variabelen selecteert en naar Independents sleept.

Statistieken

Bij statistieken kan je een aantal opties kiezen (blz.328):

- Estimates: Geeft de geschatte b-waardes samen met de t-toets weer.
- Confidence Intervals: Geeft de betrouwbaarheidsintervallen van de ongestandaardiseerde regressiecoëfficiënten weer.
- Covariance matrix: produceert een matrix van de covarianties, de correlatiecoëfficiënten en de varianties van de regressiecoëfficiënten van de variabelen in het model.
- Model fit: Deze heb je altijd nodig, en is ook al standaard aangevinkt. Het geeft de F-waarde en R weer.
- R squared change: Geeft de verandering van R^2 weer wanneer er een nieuwe voorspeller bijkomt.
- Descriptives: Geeft de standaardafwijking, het gemiddelde en de n voor alle variabelen. Geeft bovendien een correlatiematrix, die nuttig is om multicollineairiteit te checken.
- Part en partial correlations: Geeft de zero order correlaties, de gewone bivariate Pearson correlatie, tussen de voorspellers en de uitkomstvariabele. Daarnaast geeft het de part en partial correlaties van de voorspellers.
- Collinearity diagnostics: Geeft VIF of tolerantie weer.
- Durbin-Watson toets: Toetst de assumptie van onafhankelijke meetfouten.
- Casewise diagnostics: Geeft de geobserveerde uitkomstwaarde, de voorspelde uitkomstwaarde, het verschil hiertussen (het residu) en het gestandaardiseerde residu.

Bij plots kan je verschillende grafieken maken. Zie blz. 330. Bij save kan je aangeven welke diagnostieken je allemaal erbij wilt hebben, bijvoorbeeld om invloedrijke scores te ontdekken.

Bij options kan je de stapmethode voor stapsgewijze regressie kiezen en aangeven hoe je met missende waarden omgaat. Exclude cases listwise betekent dat als een persoon op één variabele een missende waarde heeft, deze persoon uit de analyse wordt gehaald.

Exclude cases pairwise betekent dat als iemand een missende waarde heeft op één variabele, diegene alleen uitgesloten wordt in de analyses over die variabele. Dit kan leiden tot vreemde resultaten, bijvoorbeeld een negatieve R^2 of een R^2 groter dan 1. Dus deze optie kan beter niet gebruikt worden.

Replace with mean betekent dat voor de missende waarde het gemiddelde wordt ingevuld. Dit zorgt voor een kleinere standaardafwijking, wat kan ertoe kan leiden dat resultaten sneller significant zijn. Dit is zeker in kleine steekproeven dus niet verstandig.

De tabel van het bootstrap betrouwbaarheidsinterval wordt niet weergegeven als je de save optie gebruikt hebt voor de diagnostieken.

Interpretatie van multiple regressie

Wanneer je bij statistics de optie descriptives hebt aangevinkt krijg je een tabel zoals op blz. 334. In deze tabel staat een samenvatting van de gegevens. Ook kan men ruwweg zien hoe de variabelen met elkaar correleren en of er multicollineariteit is. dit zie je door in de tabel te kijken naar de correlaties van de onafhankelijke variabelen met elkaar.

Het volgende deel in de output (blz.335) laat de samenvatting van het hele model zien. Dit is een van de belangrijkste onderdelen, daarom is de optie Model fit, waarmee je deze tabel produceert, standaard aangevinkt. Hier zie je de R van de predictoren, de R^2 en de aangepaste R^2. Als je dit hebt aangevinkt, kunnen ook de R^2 change en of deze R^2 change significant is worden weergegeven. Ook de Durbin-Watson test staat in deze tabel. Het laat zien of de aanname van onafhankelijke meetfouten verdedigbaar is. Hoe dichter de waarde bij 2 ligt, hoe beter.

De ANOVA tabel in de output (blz. 337) die test of het model (of de modellen, bij een hiërarchische regressie) significant beter is dan het gemiddelde als beste gok.

Model parameters

Bij hiërarchische regressie komen de modelparameters ook in de output. Bij coëfficiënts kan de individuele contributie van de variabelen bekeken worden. Voor elke voorspeller kan je de bijdrage zien terwijl de andere voorspellers dan constant gehouden worden. Door naar de b-waardes te kijken kan je zien hoe belangrijk de bijdrage van de voorspeller is. Er staat een t-toets bij die toetst of de predictor een significante bijdrage levert aan het model.

De gestandaardiseerde b-waardes (Beta in de SPSS output) zijn makkelijker te interpreteren, omdat ze in standaardafwijkingen gemeten zijn, en dus kun je de predictors makkelijker met elkaar vergelijken. De Beta waarde geeft aan hoeveel standaardafwijkingen de uitkomst verandert voor een standaardafwijking verandering in de predictor.

SPSS geeft ook een tabel met de variabelen die nog niet in de regressie-analyse zijn opgenomen. Bij de hiërarchische methode staan hier de variabelen die er in een latere stap nog bij komen. Bij de stapsgewijze methode staan hier de variabelen die SPSS overweegt in het model op te nemen. In deze tabel kan je zien wat de bijdrage van de variabele is als die wel in de analyse zou komen.

De assumptie van multicollineariteit

Deze statistieken zijn te vinden in de tabel coëfficiënten van de output. Voor het gemiddelde van VIF, tel je de VIF statistieken bij elkaar op en deel je door het aantal statistieken. Wanneer het gemiddelde dicht bij 1 ligt zijn er geen problemen in de multicollineariteit.

SPSS produceert een tabel genaamd Collinearity Diagnostics. In deze tabel kijk je naar hoe de variantieproporties zijn verdeeld over de dimensies van eigenwaardes. Als er geen multicollineariteit is, hebben de verschillende voorspellers hoge variantieproporties (richting de 1) op verschillende dimensies. In de output op bladzijde 343 is te zien dat het promotiebudget een variantieproportie van 0.96 heeft op dimensie 2, het aantal radio uitzendingen heeft 0.93 op dimensie 3 en de aantrekkelijkheid van de band heeft 0.92 op dimensie 4. Dat is een goed voorbeeld waarbij er geen sprake is van multicollineariteit.

Casewise diagnostics

Deze tabel laat de extreme scores zien (zie blz. 345). Naar scores die het regressiemodel kunnen beïnvloeden kan extra gekeken worden.

Bekijk de gestandaardiseerde residuen. 95% van de sample hoort een standardized residual te hebben tussen de -2 en +2. Als meer dan 5% hier buiten valt, moet hier kritisch naar gekeken worden.

Bekijk de Cook’s distance. Waardes boven de 1 betekent dat die score beïnvloedend kan zijn.

Bereken het gemiddelde leverage. Waardes die 2 tot 3 keer zo groot zijn als de gemiddelde leverage (k+1)/n kunnen problematisch zijn.

Bekijk de Mahalanobis distance en waardes boven de 25 in grootte steekproeven (500) en waardes boven de 15 in kleine steekproeven (100) zijn reden voor zorgen.

Kijk voor absolute waardes van DFBeta groter dan 1.

Bereken de grenzen van CVR (zie de formules op bladzijde 347). Scores die ver buiten dit interval vallen, zijn problematisch.

Op blz. 349 staat een aantal voorbeeldgrafieken waarnaar gekeken kan worden om de assumpties te checken.

Een assumptie schenden

Het is interessant wanneer je het regressiemodel kan generaliseren. Dit kan pas als aan alle assumpties is voldaan. Als een assumptie geschonden wordt bij regressie, kun je een robuuste regressie uitvoeren door middel van het bootstrap betrouwbaarheidsinterval. Let erop dat dit alleen werkt als de opties bij save allemaal uit staan.