Hoofdstuk 9

9.1

Als je een binaire (maar twee categorieën hebbende) responsvariabele hebt zijn meeste regressie technieken waardeloos. In het geval van een binaire verklarende variabele kan er gebruik gemaakt worden van dummy coding, maar dit is niet een optie als het de responsvariabele is die binair is. Het is in deze situatie ook niet mogelijk om gebruik te maken van OLS regressie, omdat deze vorm van regressie gebaseerd is op de assumptie van een lineaire relatie.

Maar, met het gebruik van logistische regressie ("logistic regression") kunnen er nog steeds voorspellingen over een binaire responsvariabele worden gemaakt. Dit is echter alleen het geval als de verklarende variabele 2 categorieën heeft. Als er drie of meer categorieen zijn kan er gebruik gemaakt worden van "multinomial logistic regression", maar dat wordt hier niet besproken.

9.2

Zoals we weten wordt in regressie informatie over de verklarende variabele gebruikt om de waardes van de responsvariabele te voorspellen. Dit concept is altijd terug te vinden in the regressie formule. Helaas is de regressie formule van logistische regressie iets meer ingewikkeld. Om te begrijpen hoe deze regressie formule werkt is het eerst belangrijk om de volgende drie concepten te begrijpen: kans ('probability'), 'odds', en 'logit'.

Laten we beginnen met kans. Dit zijn gegenereerde waarden tussen 0 en 1, die, zoals de naam impliceert, de kans voorspellen dat die data in de ene of the andere categorie van de responsvariabele valt. Dit laat dus voorspelling d.m.v. regressie toe, maar waarden hebben alleen betekenis binnen het gebied tussen 0 en 1. Om dit op te lossen gaan we door met odds.

Odds kunnen gedefinieerd worden als the ratio van de kans van de twee categorieën van de responsvariabele. Het wordt als volgt berekend:

$$Odds(Y=1) = \frac{P(Y=1)}{1-P(Y=1)}$$

Met het gebruik van odds wordt het bereik van betekenisvolle waardes uitgebreid tot alle positieve nummers. Maar, alle waardes onder 0 houden dus nog steeds geen betekenis. Hiervoor gaan we dus verder met logit.

Logit (ook wel 'log odds') wordt verkregen door het natuurlijke logaritme van de odds te nemen, en creëert zo waardes van de responsvariabele die zowel positief als negatief kunnen zijn. De logit wordt dus als volgt berekend:

$$Logit(Y) = ln\frac{P(Y=1)}{1-P(Y=1)}$$

Interpretatie van de logit vergelijking is lijkt eigenlijk erg op de interpretatie van OLS: Voor elke verandering van 1 eenheid in de verklarende variabele, laat de logistische regressie coëfficiënt de gerelateerde verandering in de responsvariabele zien. Hier neemt die verandering de vorm van de voorspelde log odds van het horen in een van beide categorieën van de responsvariabele.

Maar hoewel deze veranderingen in een normaal regressie model constant zijn, is dit bij logistische regressie niet het geval. Dit komt om dat de originele s-curve van het model lineair gemaakt wordt door de natuurlijke log. Een verandering van 1 eenheid heeft dan een groter effect als het in het centrum van de reeks gebeurt.

In tegenstelling tot OLS regressie is het standaardiseren van coëfficiënten niet gebruikelijk in logtische regressie modellen. Dit is zo omdat een interpretatie van een verandering in standaarddeviatie misschien handig is als de responsvariabele continu is, maar dit is niet het geval met een binaire responsvariabele. Hiernaast is dit gebruik niet erg compatibel met log odds. Zoals eerder uitgelegd is interpretatie van de logistische vergelijking niet te ingewikkeld omdat het erg lijkt op OLS, maar log odds zelf zijn wel iets meer ingewikkeld om als waarde mee te werken. Hierom converteren we logg odds weer naar odds als we een specifieke relatie tussen een verklarende- en responsvariabele proberen te begrijpen. Dit wordt gedaan door machtsverheffing, zoals in de volgende formule te zien is.

$$Odds(Y=1)=e^{logit(Y)}=e^{ln[Odds(Y=1)]}=e^{\alpha+\beta 1X1+\beta 2X2+\ldots+\beta mXm}=\left(e^{\alpha}\right)\left(e^{\beta 1X1}\right)\left(e^{\beta 2X2}\right)\ldots.\left(e^{\beta mXm}\right)$$

Deze machtsverheffing maakt de vergelijking vermenigvuldigend (i.p.v. toevoegend), wat de interpretatie van de coëfficiënten verandert. De waarden van de odds (en dus de uitkomst) zullen niet veranderen als de coëfficiënt 1 is. Een coëfficiënt groter dan 1 zal een toename in de odds veroorzaken, en een coëfficiënt kleiner dan 1 zal een afname veroorzaken. Hoe verder de waarde van 1 af is, hoe groter de verandering in de odds zal zijn.

Odds kunnen ook weer terug-geconverteerd worden naar kans, met de volgende formule:

$$P(Y=1)=\frac{Odds(Y=1)}{1+Odds(Y=1)}=\frac{e^{\alpha+\beta 1X1+\beta 2X2+\ldots+\beta mXm}}{1+e^{\alpha+\beta 1X1+\beta 2X2+\ldots+\beta mXm}}$$

Met kans geldt dat hoe dichter de waarde bij 1 is, hoe groter de kans is.

9.3

De volgende stap is het beoordelen van de vergelijking en vaststellen of het model past. Met logistische regressie maken we gebruik ban de “Maximum likelikood (ML) estimation” om dit te doen. Om de waardes van de parameters (of: logistische coëfficiënten) vast te stellen wordt ML op het model toegepast, en deze schat dan de odds nadat deze tot logit zijn getransformeerd. Of beter gezegd: ML schat de meest waarschijnlijke parameters op de basis van de patronen van de data van de steekproef.

Deze ML schatting resulteert in de “log of the likelihood (LL) function”. De LL reflecteert de waarschijnlijkheid dat deze steekproef statistieken geobserveerd zouden worden als de populatie parameters waar zijn. Op deze manier laat de LL zien hoeveel het model nog niet verklaard met de huidige schatting van de parameters. Hierom wordt LL gebruikt als een indicatie voor hoe goed het model past. Waardes van de LL vallen in het gebied van 0 en onder, waarbij waardes dichterbij 0 een beter passend model aangeven.

ML schatting begint meestal met een conservatieve schatting, en creëert dan doorlopend betere paramaters om betere waarschijnlijkheden aan te geven. Dit proces stopt zodra de toename in waarschijnlijkheid te klein is om significant toe te voegen.

9.4

In logistische regressie zijn er twee significantie testen om te doen. De eerste is het toetsen van de significantie van het globale regressie model, waarbij geëvalueerd wordt hoe goed het model past en tot hoeverre de voorspelde waardes de geobserveerde waardes goed representeren. Dit kan op meerdere manieren gedaan worden:

De “likelihood ratio” test: Een test gebaseerd op de verandering in de LL functie bij de verandering van een klein model naar een groter model (met meer verklarende variabelen). Deze test kan dus gebruikt worden om de veranderingen in model gepastheid tussen verschillende modellen te analyseren. Deze test lijkt erg op de algemene F-test in OLS regressie, waarbij de nul hypothese dat alle regressie coëfficiënten gelijk zijn aan 0 wordt getest. Berekeningen gaan als volgt:

$$\chi^2 = -2(LL_{model} - LL_{baseline})$$

Het kleinere model wordt altijd gebruikt als het “baseline”-model. Hoe groter het verschill in LL waarden tussen de “baseline” en het geteste model, hoe beter het model past. Deze test maakt helaas wel de assumptie van genestelde modellen, wat betekent dat alle waardes die in het “baseline”-model aanwezig zijn, ook in het geteste model aanwezig moeten zijn.

De Hosmer-Lemesho “Goodness-of-fit” test: Voor deze test moet een Hosmer-Lemeshow (HL) statistiek berekend worden, wat gedaan wordt door gevallen in 10 groepen (“deciles”) te verdelen gebaseerd op voorspelde kansen. Een chi-kwadraat waarde wordt dan berekend gebaseerd op de geobserveerde en verwachte frequenties. Een niet-statistische significante uitkomst van deze test geeft aan dat het model goed past (omdat het model dan niet statistisch significant verschilt van de geobserveerde waardes).

De “pseudo-variance explained” index: Dit is een index erg gelijkend op de meervoudige R2 in OLS regressie. Deze waardes worden beschouwd als “pseudovariance explained” omdat de variantie in een binaire uitkomst natuurlijk anders is dan in een continue uitkomst (zoals in OLS het geval is). Er zijn meerdere manieren om dit te berekenen, een paar hiervan zijn:

- Cox & Snell (automatisch berekend door SPSS): Berekend als de ratio van de waarschijnlijkheid waardes, tot de macht van 2/n.
- Nagelkerke (automatisch berekend door SPSS): Hetzelfde als Cox & Snell, maar zo bijgesteld dat de maximum waarde 1 is.
- Hosmer & Lemeshow: Berekend met de ratio van het model tot de “baseline” van -22L. De waarde valt binnen het bereik van 0 tot 1, en geeft een indicatie van hoe de gepastheid van het model verbeterd wordt met het toevoegen van meer verklarende variabelen.

$$R_L^2 = \frac{-2LL_{model}}{-2LL_{baseline}}$$

Harell: Hetzelfde als Hosmer & Lemeshow, maar bijgesteld voor de hoeveelheid parameters in het model.

$$R_{LA}^2 = \frac{(-2LL_{model}) - 2m}{-2LL_{baseline}}$$

Aldrich & Nelson: Geeft een waarde die equivalent is aan de “squared contingency coeffient”.

$$pseudo\ R^2 = \frac{-2LL_{model}}{-2LL_{model} + n}$$

Traditionele R2: Berekend door de geobserveerde waardes van de binaire responsvariabele te correleren met de voorspelde waardes van het logistische regressie model. Deze gecorreleerde waarde wordt dan gekwadrateerd.

Voorspellen van groep lidmaatschap: Hier wordt er getoetst door het verschil tussen de voorspelde en geobserveerde groep lidmaatschap te evalueren. Met een afbreekpunt van 0,5 wordt een voorspelde kans van 0,5 of meer aangegeven als een waarde van 1, en een voorspelde kans van minder dan 0,5 als een waarde van 0 (zodat er weer een binaire verdeling gecreëerd wordt). Dan wordt de voorspelde groep lidmaatschap en geobserveerde groep lidmaatschap vergeleken om de zien of de voorspelde waardes correct geclassificeerd zijn. Uit deze vergelijking komt dan een frequentie en percentage van correct geclassificeerde gevallen. Een perfect model krijgt een score van 100%, terwijl een model met een 50% score waardeloos is, omdat deze voorspellingen niet beter zijn dan kans. Press’ Q (en chi-kwadraat statistiek) kan gebruikt worden als een formele test van classificatie nauwkeurigheid:

$$Q = \frac{[N - (nK)]^2}{N(K-1)}$$

Hier is N de totale steekproef grootte; n is de hoeveelheid correct geclassificeerde gevallen; en K is de hoeveelheid groepen.

Zwaktepunten van deze test zijn dat het (1) gevoelig is voor steekproef grootte, en (2) onaanvaardbare classificaties van 1 of meer groepen misschien over het hoofd ziet. Hierom is het belangrijk om ook de classificaties van elke individuele groep de evalueren. Hierbij gelden de volgende termen:

- Gevoeligheid (“Sensitivity”): De kans dat een geval dat gecodeerd is als 1 m.b.t. de responsvariabele correct gecodeerd is (oftewel het percentage van correcte voorspellingen van 1).
- Specificiteit (“Specificity”): De kans dan een geval dat gecodeerd is als 0 m.b.t. de responsvariabele correct gecodeerd is (oftewel het percentage van correcte voorspellingen van 0).
- Vals positief ratio (“False positive rate”): De kans dat een geval dat gecodeerd is als 0 m.b.t. de responsvariabele incorrect gecodeerd is (oftewel het percentage van gevallen met incorrecte voorspellingen van 1, terwijl het 0 hoort te zijn).
- Vals negatief ratio (“False negative rate”): De kans dat een geval dat gecodeerd is als 1 m.b.t. de responsvariabele incorrect gecodeerd is (oftewel het percentage van gevallen met incorrecte voorspellingen van 0, terwijl het 1 hoort te zijn).

- Kruisvalidatie: Een aanbevolen techniek in logistische regressie, hoewel het alleen gebruikt kan worden als de steekproef groot genoeg is. Kruisvalidatie is het testen van het model op twee steekproeven, een primaire steekproef (wat bestaat uit 75-80% van de originele steekproef) en een "holdout" steekproef (die bestaat uit 20-25% van de originele steekproef). Als het verschil in classificatie nauwkeurigheid 10% of minder zijn, dan bewijst dit het nut van het logistisch regressie model.

De tweede significantie test gaat om het toetsen van de significantie van de logistische regressie coëfficiënten. SPSS gebruikt hier het Wald statistiek (die gebruik maakt van chi-kwadraat distributie) als de toets statistiek voor regressie coëfficiënten. Continue verklarende variabelen worden zo als volgt berekend (door het kwadrateren van de ratio van de regressie coëfficiënt, en deze dan te delen door de standaard fout):

$$W = \frac{\beta_k^2}{SE_{\hat{\imath}}} 2$$

Een minpunt van deze toets is dat, vanwege afrondingsfouten, grote regressie coëfficiënten onnauwkeurigheid in de schatting van de standaard fout creëren. Dit leidt tot onjuistheden bij het testen van de nul hypothese, en een toename in Type 2 fouten (falen in het verwerpen van de nul hypothese terwijl hij wel fout is).

Een alternatief voor de Wald test is de "log likelihood (LL)" test die al eerder besproken werd.

Nog een ander alternatief is de "Bayesian information criterion (BIC)" die door Raferty voorgesteld werd. De BIS representeert het verschil tussen de chi-kwadraat waarde en de natuurlijke log van de steekproef grootte, maar kan ook gebruikt worden om logistische regressie coëfficiënten te testen. De formule is als volgt:

$$BIC = \chi^2 - \ln n$$

De BIC moet positief zijn om de nul hypothese te verwerpen.

Na het vaststellen van de statistische significantie van de individuele verklarende variabelen, kan het ook een goed idee zijn om te beoordelen welke verklarende variabelen het meest toevoegen aan het model. Helaas heeft SPSS geen gestandaardiseerde regressie coëfficiënten voor logistische regressie, maar gelukkig zijn deze wel makkelijk te berekenen. Je moet simpelweg de verklarende variabelen standaardiseren voordat je het logistische regressie model genereert, en dan run je het model. De logistische regressie coëfficiënten kunnen dan geïnterpreteerd worden als gestandaardiseerde regressie coëfficiënten (zoals in OLS).

Een andere optie is om een betrouwbaarheidsinterval ("Confidence interval (CI)") om de logistische regressie coëfficiënt (bk) te vormen. Deze CI formule is hetzelfde als in OLS regressie:

$$CI\left(b_k\right) = b \pm t_{(n-m-1)} s_b$$

9.5

De assumpties van logistische regressie zijn iets losser dan die van OLS. Maar er zijn nog steeds wel vier primaire assumpties om rekening mee te houden:

Geen co-lineairiteit ("Non-colinearity"): Deze assumptie is toepasselijk bij elk regressie model met meerdere verklarende variabelen. Multi-co-lineairiteit kan gedetecteerd worden door een OLS regressie model te creëren in SPSS met dezelfde variabelen als het logistische regressie model, en dan de co-lineairiteit statistieken aan te vragen. Tolerantie statistieken van minder dan 0,2 suggereren dat er multi-co-lineairiteit is, en waardes lager dan 0,1 suggereren serieuze multi-co-lineairiteit. Elke waarde boven de 10 geeft aan dat er een schending is van deze assumptie. (Zie hoofdstuk 8 voor meer details over deze assumptie).

Lineairiteit ("Linearity"): In logistische regressie refereert deze assumptie naar de lineairiteit tussen de logit van de responsvariabele en de continue verklarende variabelen. Deze assumptie kan op meerdere manieren getoetst worden, maar de makkelijkste is de Box-Tidwell transformatie. Dit wordt gedaan door een logistisch regressie model te creëren met alle verklarende variabelen die van interesse zijn, elke gekoppeld aan een interactie term. Deze interactie term wordt gecreëerd door de continue verklarende variabelen met zijn natuurlijke log te vermenigvuldigen. Non-lineairiteit wordt gesuggereerd door statistisch significante interactie termen. Schending van deze assumptie kan leiden tot biased parameter schattingen, en dat de verwachte veranderingen van Y niet constant zijn over X.
NB: Deze assumptie geldt alleen voor continue verklarende variabelen.

Onafhankelijkheid van fouten ("Independence of errors"): Deze assumptie is op een logistisch regressie model op dezelfde manier toepasselijk als bij OLS regressie. Schending van deze assumptie kan leiden tot onderschatte standaard fouten, met verscheidende consequenties van den. (Zie hoofdstuk 7 en 8 voor meer details).

Waarden van X zijn vast ("Values of X are fixed"): Deze assumptie is bij logistische regressie niet anders dan bij OLS. (Zie hoofdstuk 7 en 8 voor meer details).

Logistische regressie moet zich ook aan de volgende condities houden:

- Non-nul cel waardes ("Nonzero cell counts";): Een nul cel waarde komt voor wanneer de uitkomst constant is voor 1 of meer categorieën van het nominale variabele. Omdat dit betekent dat een gehele groep individuen odds van 0 of 1 heeft, leidt dit tot hoge standaard fouten. Verschillende manieren om nul cel waarden te verwijderen zijn: (1) her-coderen van de categorieën, of (2) het toevoegen van een constante. Nul cel waarden kunnen behouden blijven als het over het algemeen geen effect heeft op de relatie tussen de verklarende- en responsvariabele, maar de aanwezigheid hiervan moet wel herkend worden.
- Non-scheiding van data ("Non-seperation of data"): Wanneer de responsvariabele perfect voorspeld is, kan er complete scheiding voorkomen, wat resulteert in een onvermogen om de modellen te schatten. Als de scheiding niet geheel compleet is wordt dit quasi-complete scheiding genoemd, wat resulteert in erg grote coëfficiënten en standaard fouten. Deze condities kunnen voorkomen als de hoeveelheid variabelen gelijk (of bijna gelijk) is aan de hoeveelheid gevallen in de dataset.

- Gebrek aan influentiele punten (“Lack of influential points”): Net zoals in OLS zijn uitbijters (“outliers”) en influentiele punten problematisch in logistische regressie. Dezelfde middelen die in OLS gebruik worden, zoals “residual analysis” en andere diagnostische tests, kunnen hier gebruikt worden. (Zie hoofdstuk 7 en 8 voor meer details).
- Voldoende steekproef grootte (“Sufficient sample size”): Logistische regressie kan het best gebruikt worden met grote steekproeven. Om de tests van significantie goed uit te voeren zijn hier steekproeven van minstens 100 nodig.

9.6

Het is belangrijk om ook even stil te staan bij het statistiek “odds ratio (OR)”, wat gebruikt kan worden als een index van de effect grootte (vergelijkbaar aan R2). De OR wordt berekend door de logistische regressie coëfficiënt, ebk, te kwadrateren. Een OR van 1 zou aangeven dat er geen relatie is tussen de verklarende variabele en de responsvariabele.

Dus, wanneer we testen voor effect grootte willen we weten of de OR significant verschilt van 1. Wanneer de OR groter is dan 1, dan vergroot de verklarende variabele de kans dat de responsvariabele voorkomt. Wanneer de OR kleiner is dan 1, verkleint de verklarende variabele de kans dat de responsvariabele voorkomt.

SPSS output labelt OR als “Exp(B)” onder “Variables in the Equation”.

OR waardes kunnen ook geconverteerd worden naar Cohen’s d op de volgende manier:

$$d = \frac{ln(OR)}{1.81}$$

9.7

Met logistische regressie kunnen er drie soorten model-bouwtechnieken gebruikt worden.

De eerste is gelijktijdige logistische regressie (“simultaneous logistic regression”), waarbij alle verklarende variabelen gelijktijdig aan het model worden toegevoegd. Dit model wordt meestal gebruikt wanneer er niet gedacht wordt dat sommige verklarende variabelen belangrijker dan anderen zouden zijn. Op deze manier kunnen alle verklarende variabelen geëvalueerd worden alsof ze als laatste aan het model zijn toegevoegd. Dit kan echter wel leiden tot te hoog berekende correlaties tussen de verklarende- en responsvariabele (omdat de verklarende variabelen elkaar versterken). In SPSS wordt deze methode “Enter” genoemd.

De tweede is stapsgewijze logistische regressie (“stepwise logistic regression”), waar de computer variabelen toevoegt of verwijdert op een stapsgewijze manier, zoals eerder besproken.

De derde is hiërarchische regressie (“hierarchial regression”), waar de verklarende variabelen toegevoegd/verwijderd worden aan de hand van een volgorde die van te voren door de researcher wordt vast gesteld. Deze volgorde kan voorwaarts, achterwaarts, of stapsgewijs zijn. (Zie hoofdstuk 8 voor meer details).