Hoofdstuk 15. Logistic Regression: Modeling Categorical Responses

In dit hoofdstuk worden lineaire modellen voor categorische responsvariabelen besproken. Deze paragrafen bespreken het logistische regressiemodel voor binaire responsvariabelen.

Logistische regressie
De twee categorieën van een binaire responsvariabele y worden aangegeven met 1 en 0 (meestal met succes en falen aangeduid). Regressiemodellen voor binaire resoponsen beschrijven de populatieproporties. De populatie proportie van succes representeert de kans P(y=1) van een random geselecteerde proefpersoon. Modellen voor binaire data nemen vaak een binomiale verdeling voor de responsvariabele aan.

Lineair Probability Model
Voor 1 verklarende variabele, is het simpele model:

$$P(y=1) = \alpha + \beta x$$

Dit duidt op een linaire functie van x. Dit wordt het lineaire probability model genoemd. Dit model is simpel maar vaak ongepast. Het impliceert dat de probabilities onder de 0 liggen of boven de 1 wanneer de x-waarde heel klein of groot is, terwijl ze tussen 0 en 1 moeten vallen.

Logistisc Regression Model for Binary Responses
Er bestaat ook een meer realistische responscurve: S-vorm. Bij deze curves, valt de kans op succes tussen 0 en 1 bij alle mogelijke x-waardes. Deze curvilineaire relaties worden beschreven door de formule:

De ratio P(y=1) / (1 – P(y=1)) staat gelijk aan de odds. Wanneer bijvoorbeeld P(y=1) = 0.75, dan is de odds 0.75/0.25 = 3.0, wat betekent dat de akns op succes 3 keer zo groot is als de kans op falen. De log van de odds wordt gebruikt, dit wordt ook *wel logistische transformatie* genoemd, of *logit*. Het model wordt ook wel geschreven als:

logit (P (y=1)) =

Dit is het logistische regressie model. De bepaald of de curve naar boven of beneden gaat wanneer *x* groter wordt. Als > 0, dan zal *P(y=1)* groter worden wanneer *x* groter wordt. Als < 0, dan zal *P(y=1)* kleiner worden als *x* groter wordt. Voor beide geldt, dat hoe groter , hoe steiler de curve. Wanneer = 0, dan zal *P(y=1)* niet veranderen als *x* verandert. De curve wordt platter tot een horizontale lijn.
Wanneer P(y=1) = 0.50, dan is de odds P(y=1) / (1-P(y=1)) =1, en log (P(y=1)) / (1-P (y=1)) = 0. Dus om de waarde van *x* te vinden waarbij P (y=1) = 0.50, stellen we de log odds waarde van 0 gelijk aan en lossen we dit op om *x* te vinden. Hier vind je P (y=1) = 0.50, met *x* = - / .

Logistische Regressievergelijking voor Probabilities
Een alternatieve vergelijking voor logistische regressie drukt de kans op succes direct uit:

$$P(y=1) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

Hier worden natuurlijke logaritmes gebruikt. We gebruiken deze formule om de waardes van P(y=1) te voorspellen, bij bepaalde waardes van de predictor. Het getal dat uit deze formule komt is de kans dat iemand wél succes heeft.

Interpreteren van het Logistisch Regressie Model
β is hier niet de helling voor de verandering in P(y=1) als *x* verandert. Omdat de curve een S-vorm heeft, hangt het van de waarde van *x* af in hoeverre de helling stijgt of daalt.
De makkelijkste manier om β te gebruiken bij het interpreteren van de steilheid van de curve, is door een rechte lijn schatting te gebruiken voor de logistische regressie curve. Een rechte raaklijn vaan de curve heeft een sloop van βP (y=1) (1 – P (y=1)), waarbij P(y=1) de kans is op dat punt. De helling is het grootste wanneer P(y=1) = 1/2, waar β (1/2)(1/2) = β/4. Wanneer P(y=1) vlakbij 1/2 ligt, is 1/4 van het β-effect de snelheid waarmee P(y=1) verandert per 1 unit groter worden van *x*.

Een andere manier om het effect van *x* te beschrijven is om P(y=1) op twee verschillende waardes van *x* te vergelijken. We hebben gezien dat wanneer *x* groter wordt, van de kleinste tot de grootste waarde in de steekproef, P(y=1) ook steeds groter wordt. Dit is een sterk effect wanneer er grote verandering is in P(y=1).

Interpretatie die gebruik maakt van de Odds en Odds Ratio
Een andere interpretatie van de logistische regressie parameter β gebruikt de odds ratio measure of association. De volgende vergelijking wordt gebruikt:

$$e^{\alpha+\beta x} = e^{\alpha}(e^{\beta})^{x}$$

De rechter exponentiële vergelijking impliceert dat elke elke vergroting in x een multiplicatief effect van e^{β} op de odds.

Multipele Logistische Regressie
Multipele logistische regressiemodel heeft de vorm:

$$\text{Logit } (P(y=1)) = \alpha + \beta_1 x_1 + \ldots + \beta_k x_k$$

De formule voor de kans zelf is:

$$P(y=1) = \frac{e^{\alpha+\beta 1x1+\cdots+\beta kxk}}{1+ e^{\alpha+\beta 1x1+\cdots+\beta kxk}}$$

Een bèta parameter exponentiëren zorgt voor controleren van andere variabele. Hoe meer een βi van de 0 afwijkt, hoe sterker het effect van de predictor *xi*, omdat de odds ratio verder van de 1 valt.

Om te zorgen voor categorische verklarende variabelen, moet je dummy variabelen opzetten. Zie voorbeeld 15.2 op bladzijde 489.

Effecten op Odds
De parameter schattingen voor het logistische regressiemodel zijn lineaire effecten, maar op de schaal van de log van de odds. Effecten op de odds schaal zijn makkelijker te begrijpen dan effecten op de log odds schaal.
Neem bijvoorbeeld de volgende vergelijking over doodstraf:

$$\text{Log}(P(y=1) / (1-P (y=1)) = -3.596 – 0.868d + 2.404v$$

Deze vergelijking refeeert naar de log odds. De corresponderende vergelijking voor de odds is :

$$\text{Odds} = e^{-3.596 – 0.868d + 2.40v} = e^{-3.596} e^{- 0.868d} e^{2.404v}$$

Voor witte verdachten, d=1 en is de geschatte odds gelijk aan $e^{-3.596} e^{- 0.868} e^{5.404v}$. Voor zwarte verdachten, d=0, is de odds $e^{-3.596}e^{2.404v}$. De geschatte odds voor witte verdachten gedeeld door de geschatte odds voor zwarte verdachten is gelijk aan $e^{- 0.868} = 0.42$. Dit laat zien waarom antilog van het coëfficiënt voor d in de vergelijking, de geschatte odds ratio tussen het ras van de verdachten

en de uitspraak voor de doodstraf is, voor elk ras van het slachtoffer. Het effect van het witte ras van de verdachte krijg je door de geschatte odds van een ‘ja’ doodstrafbepaling te vermenigvuldigen met $e^{-0.868} = 0.42$, in vergelijking met de waarde voor zwarte verdachten. De echte waardes van de odds hangen af van het ras van de slachtoffers, maar de ratio van de odds zijn voor beide hetzelfde.

De antilogs van de parameters worden vermenigvuldigd om odds te krijgen. Dit kan je gebruiken om odds uit te rekenen voor elke combinatie van ras van de verdachte en slachtoffers, bijvoorbeeld wanneer de verdachte zwart is (d=0), en slachtoffer wit (v=1), de geschatte odds van de doodstraf is:

$$\text{Odds} = e^{-3.596} e^{-0.868d} e^{2.404v} = e^{-3.596} e^{-0.868(0)} e^{2.404(1)} = e^{-1.192} = 0.304$$

Omdat de formule voor de geschatte kans op doodstraf is:

Deze kans is om te zetten in Odds:
Dit is een formule die al in 8.4 is gebruikt. Wanneer de geschatte odds bijvoorbeeld 0.304 is zoals in dit voorbeeld, dan: d = 0, v = 1, dan P(y=1)= 0.304 / (1+0.304) = 0.233. dit is de geschatte kans op doodstraf.

Effecten op Probabilities
We kunnen dus de effecten van predictoren samenvatten door odds ratio’s. veel vinden het makkelijker om de effecten te bekijken door naar probability scales te kijken. Deze rapporteren geschatte probabilities op bepaalde waardes van een predictor. Dit wordt gedaan op bepaalde vastgezette waardes, bijvoorbeeld het gemiddelde of waarden die hier van belang zijn.
Je kunt de verandering in de geschatte probability P(y=1), wanneer een predictor een bepaald aantal groter wordt, bijvoorbeeld door een 1) een fixed value (bv. 1), 2) een standaarddeviatie, 3) van de laagste waarde in de range naar de grootste, of 4) inter-kwartiel range van lagere kwartiel naar hoge kwartiel. 4 is niet zoals 1 beïnvloed door keuze van de schaal, en niet zoals 2 en 3 beïnvloed door de outliers.

Je kunt bijvoorbeeld kijken naar het inkomen van een echtgenoot. Bedenk dat de kans P(y=1) van een eigen huis wanneer de vrouw 50.000 dollar verdient, en ze 3 jaar getrouwd zijn. de vrouw werkt al 2 jaar en ze hebben 0 kinderen, ook twee jaar later hebben ze nog geen kinderen (0). Ze heeft 16 jaar educatie gehad en de ouders hebben een eigen huis. Het inkomen van de echtgenoot is 20.000 dollar. Dan geldt:

= 0.41

De waardes hierboven zijn te halen uit de gegeven tabel. P(y=1) neemt toe met 0.55 als de inkomsten van de echtgenoot toenemen tot 30.000 dollar. Tot 0.79 als dit inkomen tot 50.000 verhoogt en tot 0.98 als dit inkomen tot 100.000 toeneemt. Het effect is vrij sterk.

Inference for logistic regression models
Inferentie neemt meestal randomisatie aan van de data en een binomiale verdeling voor de responsvariabele. Y moet een binomiale verdeling hebben en de logit link functie wordt voor P(y=1) gebruikt, welke het gemiddelde is van y.

Wald en Likelihood-ratio Tests van Onafhankelijkheid
Voor het bivariate logistische regressie model geldt:

$$\text{Logit}(P(y=1)) = \alpha + \beta x$$

Wanneer H_0: $\beta = 0$, dan geldt dat x geen effect heeft op P(y=1). Dit is de onafhankelijkheidshypothese. Behalve voor kleine steekproefen, kunnen we H0 testen met een z-statistic, waarbij de maximum likelihood schatter β gedeeld wordt door zijn standaard error. Het kwadraat van deze statistic wordt soms een Wald statistic genoemd. Dit heeft een chi-kwadraat verdeling met df=1. Het heeft dezelfde P-waarde als de z-statistic voor de tweezijdige Ha: $\beta \neq 0$.

Vaak wordt ook een andere hypothese voor deze test gerapporteerd: de likelihood-ratio test. Dit is een manier om twee modellen te vergelijken, een complex model en een simpeler model. Het test of de extra parameters in het complexe gelijk zijn aan 0. Bijvoorbeeld, voor bivariate logistische regressie test het H0: $\beta = 0$, door de model Logit (P(y=1)) = $\alpha + \beta x$ te vergelijken met Logit (P(y=1)) = α. Het gebruikt de likelihood function (ℓ). Dit geeft de kans aan op de geobserveerde data als functie van de parameter waardes. De maximum likelihood maximaliseert de functie.

ℓ_0 is het maximum van de likelihood functie als H0 waar is en ℓ_1 is het maximum zonder aanname. De formule voor de likelihood ratio test statistic is:

$$-2\log () = (-2\log \ell_0) - (-2\log \ell_1)$$

Het vergelijkt de gemaximaliseerde waardes van ($-2\log \ell_0$) wanneer H0 waar is en wanneer het niet waar moet zijn.

Voor H_0: $\beta = 0$ met grote steekproefen, produceren de Wald test en de likelihood ratio test vaak gelijke resultaten. Bij kleine tot gemiddelde steekproeven, is de likelihood ratio statistic vaak groter en heeft meer power dan de Wald statistic. Gebruik deze dus sneller dan de Wald statistic.

Inference in multiple logistic regression
Om het partial effect van een predictor in multiple logistic regression model te testen, moet je de de parameter schatter delen door de standaard error. Dit is een z-test statistic. Wanneer je hiervan het kwadraat neemt, de Wald statistic, krijg je een chi-kwadraat statistic met df=1. De meeste software rapporteert ook de likelihood-ratio tests, waarbij de ($-2\log \ell$) waardes worden vergelijken met en zonder predictor in het model. Dit is handig wanener een categorische predictor meerdere levels heeft, waarbij er dus meerdere dummy variabelen. Zie voor een rekenvoorbeeld bladzijde 495.

Likelihood-ratio test comparing Logistic Regression Models
Om een model te vergelijken met een set predictoren, met een simpeler model met minder predictoren, gebruikt de likelihood-ratio test de verschillen in de waardes van ($-2\log \ell$) voor de twee modellen. Dit is een chi-kwadraat statistic met df gegeven door het aantal extra parameters in het complexe model. Deze test is analoog aan de F test voor complete en reduced regression modellen te vergelijken. Bijvoorbeeld het model in voorbeeld 15.3 (blz 491) over een eigen huis hebben, heeft ($-2\log \ell$) = 2931.2. Nadat er 5 variabelen worden toegevoegd aan het model die gerelateerd zijn aan de huizenmarkt, bijvoorbeeld median verkoopprijs van bestaande huizen in de buurt, verlaagt ($-2\log \ell$) tot 2841.1. het verschil (2931.2 – 2841.1) = 90.1. dit is een chi-kwadraat statistic met df=5, omdat het model 5 toegevoegde parameters heeft. Dit laat een sterk bewijs zien voor betere fit van het complexere model ($P<0.0001$). Ten minste 1 van de variabelen zorgen voor verbetering van de predictieve power.