Introduction to the practice of statistics, Moore, McGabe & Craig, 2012, 7e druk

Hoofdstuk 14 Logistische regressie

In dit hoofdstuk worden enkelvoudige en multipele regressiemethoden besproken die gebruikt worden wanneer de responsvariabele maar twee mogelijke waarden (1, bijvoorbeeld succes en 0, mislukking) kan aannemen. Het gemiddelde is de proportie van enen (p = P(succes)). Wat er nieuw is, is dat we nu data hebben voor een onafhankelijke variabele x. Er wordt bestudeerd hoe p van x afhangt.

Het Logistische Regressiemodel

Logistische regressie werkt meer met kansverhoudingen (odds) dan met proporties. Een kansverhouding is de verhouding van de proporties van de twee mogelijke uitkomsten $\hat{p}$ en $1-\hat{p}$. $\hat{p}$ staat voor populatiekansverhoudingen.

$$odds = \frac{\hat{p}}{1-\hat{p}}$$

Bij enkelvoudige lineaire regressie wordt het gemiddelde m van de responsvariabele y beschreven als een lineaire functie van de onafhankelijke variabele: $m = b_0 + b_1 X$.
Bij logistische regressie zijn we geïnteresseerd in het gemiddelde van de responsvariabele $p = b_0 + b_1 X$.

Dit is echter geen goed model. Zolang $b_1 \neq 0$, zouden extreme waarden van x waarden opleveren die niet tussen 0 en 1 zijn.
De oplossing hiervoor is het transformeren van p naar een kansverhouding. Vervolgens wordt het logaritme genomen van de kansverhouding. De term logaritmische kansverhouding (log odds) wordt hiervoor gebruikt.

$$\log\left(\frac{p}{1-p}\right) = b_0 + b_1 X$$

Dit is het Logistische Regressiemodel.

Logistische regressie met een indicator is een speciaal geval. Een indicator is een geslachtsvariabele; 0 = vrouw, 1 = man. Voor mannen is het model:

$$\log\left(\frac{p_{mannen}}{1-p_{mannen}}\right) = b_0 + b_1$$

En voor vrouwen:

$$\log\left(\frac{p_{vrouwen}}{1-p_{vrouwen}}\right) = b_0$$

b_1 ontbreekt in het model voor vrouwen, want x = 0 hier.

De helling in het logistische regressiemodel is het verschil tussen de log (odds) voor mannen en de log (odds) voor vrouwen. Het is lastig om te denken in de log (odds) schaal, daarom wordt er een transformatie gebruikt die het interpreteren van de resultaten eenvoudiger maakt:

$$X = \frac{odds_{mannen}}{odds_{vrouwen}}.$$

De logistische regressie wordt hierin getransformeerd tot een odds-verhouding en maakt het logaritme ongedaan.
Dit is ook uit te drukken als:

$$odds_{mannen} = X \cdot odds_{vrouwen}$$

Methoden voor Logistische Regressie

De methoden voor logistische regressie lijken sterk op de methoden voor enkelvoudige lineaire regressie. Er worden schattingen gemaakt van de modelparameters en van standaardfouten. Ook betrouwbaarheidsintervallen worden op dezelfde manier gevormd. Alleen worden standaardnormale z-waarden meer gebruikt dan kritische waarden van de t-verdelingen. De verhouding van de geschatte standaardfouten is de basis voor hypothesetoetsen.

Betrouwbaarheidsintervallen en significante toetsen voor Logistische Regressieparamaters

Het betrouwbaarheidsinterval voor de helling b_1 is:

$$b_1 \pm z^* SE_{b_1}$$

Het betrouwbaarheidsinterval voor de odds-verhouding e^{b_1} is:

$$e^{b_1 - z^* SE_{b_1}}, \ldots e^{b_1 + z^* SE_{b_1}}$$

z^* is de waarde voor de standaardnormale dichtheidscurve met een gebied tussen $-z^*$ en $+z^*$.

Om de nulhypothese $H_0 : b_1 = 0$ moet men de toetsstatistic uitrekenen.

$$X^2 = \left(\frac{b_1}{SE_{b_1}} \right)^2$$

De P-waarde voor een toets van de nulhypothese tegen de alternatieve hypothese is:

$$P\left(x^2 \geq X^2\right)$$

Vaak wordt een 95%-betrouwbaarheidsinterval gehanteerd en een significantieniveau van 0.05. Het betrouwbaarheidsinterval geeft het resultaat van het toetsen van de nulhypothese, die stelt dat de odds-verhouding 1 is. Wanneer 1 niet in het betrouwbaarheidsinterval voorkomt, wordt H_0 verworpen. De odds voor de twee groepen zijn dan verschillend.

5. Multipele logistische regressie

Multipele logistische regressie wordt toegepast wanneer er sprake is van meer dan één onafhankelijke variabelen. Andere onafhankelijke variabelen kunnen aanvullende informatie bevatten, waardoor een betere voorspelling gedaan kan worden.
De statistische concepten zijn hetzelfde als bij enkelvoudige lineaire regressie, maar de berekeningen zijn complexer.

De nulhypothese is hier: $H_0 : b_1 = b_2 = b_3 = ... = b_i$