

Hoofdstuk 7

BS HC 7B – 28 oktober

Dichotome variabele: een variabele kan maar twee waarden aannemen, bijvoorbeeld geslacht (man of vrouw). Je kunt van een variabele een dichotome variabele maken door de nadruk te leggen op het wel of niet aanwezig zijn van een kenmerk. Hierbij krijgt de ene waarde een 0 en de andere waarde een 1. Wel een vrouw zijn = 1, niet een vrouw zijn is 0. Hierdoor kun je met de variabele rekenen op rationiveau. We noemen hem echter nog steeds een nominale variabele!

Tau, lambda, Eta en R^2 zijn alle vier gebaseerd op de proportie voorspellingsverbetering. Tau en lambda gebruik je bij een categorisch meetniveau en R^2 bij een numeriek meetniveau. Eta gebruik je als je een categorische onafhankelijke variabele hebt en een numerieke afhankelijke variabele.

Pearson productmomentcorrelatiecoëfficiënt (r) is een associatiemaat op interval- en rationiveau (numeriek). Deze associatiemaat wordt gebruikt voor het berekenen van symmetrische relaties. De maat varieert tussen de -1 (perfecte negatieve samenhang) en de 1 (perfecte positieve samenhang). Hij is gebaseerd op het verschil tussen de werkelijke waarden en het rekenkundig gemiddelde van twee variabelen. De spreiding van de gegevens kun je zien in een spreidingsdiagram. In de spreidingsdiagram kun je al zien hoe groot de samenhang ongeveer is en in welke richting die gaat. Hoe kleiner de spreiding is, hoe hoger de samenhang.

De **covariantie** is de spreiding tussen twee variabelen. Bij een positieve covariantie scoort de ene variabele hoger als de andere variabele ook hoger scoort. Bij een negatieve covariantie scoort de ene variabele lager als de andere variabele hoger scoort.

De formule voor de covariantie:

$$Cov(x, y) = \frac{\sum (x - \bar{x})(y - \bar{y})}{n - 1}$$

De formule voor de standaarddeviatie was:

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

Met de covariantie en de standaarddeviatie kun je r uitrekenen:

$$r = \frac{Cov(x, y)}{s_x s_y}$$

Een **regressieanalyse** is de enige statistische analyse die een voorspelling kan meten. De regressielijn is de best passende lijn door een puntenwolk heen. Het doel van een regressieanalyse is een voorspelling doen over de afhankelijke variabele (y), wanneer je de waarde van de onafhankelijke variabele (x) invult. De regressieanalyse gaat dus over het effect van x op y .

De formule voor de lijn is: $\hat{y} = a + b(x)$. De a is het snijpunt met de y -as en dus de waarde van y als x 0 is. b is de **regressiecoëfficiënt**: de toename van de waarde van y als x met 1 toeneemt (de helling). Om dit goed te kunnen begrijpen, kun je het best een voorbeeld van een regressielijn in een spreidingsdiagram bekijken.

Een regressieanalyse bestaat altijd uit drie tabellen. In de eerste tabel staat de R^2 : de proportie verklaarde variantie. Deze is afgeleid van de Eta, maar nu gaat het om twee variabelen. In de eerste tabel staat ook de covariantie. In de tweede tabel staat de intercept (a) (als x 0 is, is y a) en de ongestandaardiseerde regressiecoëfficiënt (b) (als x toeneemt met 1, y toeneemt/daalt met b).

R^2 (de proportie verklaarde variantie) is de mate waarin de varia(n)tie in de afhankelijke variabele verklaard wordt door de varia(n)tie in de onafhankelijke variabele. Varia(n)tie betekent hier hetzelfde als het verschil. Bijvoorbeeld: de mate waarin de verschillen in televisiekijktijd (= de variantie in televisiekijktijd) verklaard worden door de verschillen in sekse (= de variantie in sekse). De formule voor R^2 is hetzelfde als die van Eta: $(E1 - E2) / E1$. E1 is weer het aantal voorspellingsfouten als je alleen uitgaat van y: *totale variatie*. E1 is dus de afstand van een waarneming t.o.v. het gemiddelde. E2 is weer het aantal voorspellingsfouten als je ook rekening houdt met x: *onverklaarde variatie*.

R^2 in formule:

$$R^2 = \frac{\sum (y - \bar{y})^2 - \sum (y - \hat{y})^2}{\sum (y - \bar{y})^2}$$

Het roodgedrukte is de E2. Om deze te berekenen moet je de $\hat{y}$ berekenen. De formule voor $\hat{y}$ is:

$$\hat{y} = a + bx$$

Om deze formule in te kunnen vullen moet je a en b berekenen. a is de intercept, het snijpunt met de y-as. a geeft dus de voorspelde waarde aan van $\hat{y}$ als de waarde van x nul is. De formule voor a:

$$a = \bar{y} - b\bar{x}$$

b is het effect van de onafhankelijke variabele op de afhankelijke variabele. Het is de voorspelde waarde die $\hat{y}$ toe- of afneemt wanneer x met één eenheid stijgt. De formule voor b:

$$b = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

De regressievergelijking geeft een voorspelling van $\hat{y}$ wanneer je de x invult. R^2 geeft de proportie verklaarde variantie aan: de mate waarin de variantie (verschillen) in de onafhankelijke variabelen de variantie (verschillen) in de afhankelijke variabele verklaart.