

## Onderzoekspracticum 2, aantekeningen college 10

### Enkelvoudige lineaire regressie

Het doel van regressie is om een verband te vinden tussen een onafhankelijke variabele  $x$  en een afhankelijke variabele  $y$ . Enkelvoudige lineaire regressie heeft de volgende kenmerken:

- De onafhankelijke variabele  $x$  en de afhankelijke variabele  $y$  zijn beide continu. Dit houdt in dat  $x$  en  $y$  alle mogelijke waarden van een continuüm kunnen aannemen.
- Het verband tussen  $x$  en  $\mu_y$  wordt beschreven door een rechte lijn, de populatie-regressielijn:  $\mu_y = \beta_0 + \beta_1 x$ . Hierin is  $\beta_0$  de intercept, oftewel de waarde waar de regressielijn de  $y$ -as snijdt.  $\beta_1$  is de richtingscoëfficiënt en de waarde hiervan geeft aan met hoeveel eenheden  $y$  toeneemt als  $x$  toeneemt met één. We kijken dus naar hoezeer de vele gemiddelden van  $y$  ( $\mu_y$ ) veranderen als  $x$  verandert.
- De geobserveerde waarden van  $y$  variëren rondom  $\mu_y$ . Deze variatie wordt weergegeven met de standaarddeviatie van de populatie ( $\sigma$ ). De aanname bij regressieanalyse is dat deze standaarddeviatie voor alle waarden van  $x$  hetzelfde is.
- In de praktijk moeten de waarden van  $\beta_0$ ,  $\beta_1$  en  $\sigma$  geschat worden uit de data, omdat de waarden ervan onbekend zijn.
- Het enkelvoudige lineaire regressiemodel beschrijft het verband tussen de observaties  $y_i$  en  $x_i$  van een persoon  $i$  als volgt:  $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ .

De parameters van bovenstaand model zijn  $\beta_0$ ,  $\beta_1$ , en  $\sigma$ . Als  $x$  gelijk is aan  $x_i$ , dan is  $\beta_0 + \beta_1 x_i$  het gemiddelde van  $y$ . Met  $\varepsilon_i$  wordt de error, oftewel de afwijking, van persoon  $i$  bedoeld. Er wordt verondersteld dat de afwijkingen normaal verdeeld en onafhankelijk zijn met een gemiddelde nul en standaarddeviatie  $\sigma$ . Het enkelvoudige lineaire regressiemodel komt overeen met het enkelvoudig ANOVA-model (DATA = FIT + RESIDU).

### Schatten van regressieparameters

Het schatten van de regressielijn doe je volgens het kleinste-kwadratenprincipe. Hierbij zoek je naar de lijn waarvan de som van de gekwadrateerde afwijkingen tot die lijn minimaal is. De geschatte regressielijn heeft als formule:  $\hat{y} = b_0 + b_1 x$ . Hierin staat  $\hat{y}$  dus voor de geschatte waarde van  $y$ . Voor de waarden  $b_0$  en  $b_1$  geldt:

$$b_0 = \bar{y} - b_1 \bar{x} \quad \text{en} \quad b_1 = r \frac{s_y}{s_x}$$

Hierin is  $r$  de correlatie tussen  $x$  en  $y$ . Als  $x$  toeneemt met waarde 1, dan neemt  $y$  gemiddeld toe met  $b_1$  eenheden. Het gaat hierbij om de *gemiddelde* toename, over individuen kun je niets zeggen omdat deze allemaal iets zullen afwijken van de regressielijn.

Het schatten van de residuen  $\varepsilon_i$  wordt gedaan met behulp van  $e_i$ . De waarde van  $e_i$  is gelijk aan de waarde van de geobserveerde respons min de waarde van de voorspelde respons, oftewel  $y_i - \hat{y}_i$ . Dit kan weer verder worden uitgewerkt tot  $y_i - b_0 - b_1 x_i$ .

De schatter voor  $\sigma^2$  is  $s^2$  en wordt ook wel de gemiddelde kwadratensom van de error (MSE) genoemd. De schatter voor  $\sigma$  is de wortel uit  $s^2$ . Er geldt:

$$s^2 = \frac{\sum_{i=1}^n e_i^2}{n-2} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2}$$

### Betrouwbaarheidsintervallen en significantietoets

Regressiecoëfficiënten hebben een betrouwbaarheidsinterval. Dit betrouwbaarheidsinterval is gebaseerd op

de normale steekproevenverdeling van de schattingen van  $b_0$  en  $b_1$ . Omdat  $\sigma$  onbekend is, wordt  $s$  gebruikt en gaan we uit van een t-verdeling met  $n - 2$  vrijheidsgraden.

Een C%-betrouwbaarheidsinterval voor de intercept  $\theta_0$  wordt berekend door:

$$SE_{b_0} = s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

$b_0 \pm t^* SE_{b_0}$ . Hierin geldt:

Een C%-betrouwbaarheidsinterval voor de coëfficiënt  $\theta_1$  wordt berekend door:

$$SE_{b_1} = \frac{s}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

$b_1 \pm t^* SE_{b_1}$ . Hierin geldt:

De nulhypothese dat  $\theta_1 = 0$  houdt in dat  $y$  op geen enkele manier verband houdt met  $x$ . Voor het toetsen van deze nulhypothese wordt een t-verdeling gebruikt met  $n - 2$  vrijheidsgraden. De alternatieve hypothese kan eenzijdig zijn (gevonden p-waarde niet vermenigvuldigen met 2), of tweezijdig (gevonden p-waarde wel vermenigvuldigen met 2). De t-waarde wordt berekend door:

$$t = \frac{b_1}{SE_{b_1}}$$

### Betrouwbaarheidsintervallen voor de gemiddelde respons

Voor de gemiddelde respons  $\mu_y$  kan er een C%-betrouwbaarheidsinterval worden berekend wanneer  $x$  een bepaalde waarde  $x^*$  aanneemt. Je bepaalt dan tussen welke grenzen deze bepaalde  $x^*$ -waarde varieert:

$$SE_{\hat{\mu}_y} = s \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

$\hat{\mu}_y \pm t^* SE_{\hat{\mu}_y}$  en

Het dakje op  $\mu_y$  betekent dat het om een geschatte waarde gaat. Verder wordt ook bij dit betrouwbaarheidsinterval gebruik gemaakt van de t-verdeling met  $n - 2$  vrijheidsgraden.

### Voorspellingsintervallen voor toekomstige observaties

De voorspelde waarde van  $y$  voor een individu dat een bepaalde score  $x^*$  heeft behaald, wordt als volgt in een formule weergegeven:  $\hat{y} = b_0 + b_1 x^*$ . Een bruikbare voorspelling bevat ook altijd een foutenmarge. Deze foutenmarge wordt ook wel het voorspellingsinterval genoemd en houdt het volgende in:

- Je trekt een steekproef van  $n$  observaties  $(x_i, y_i)$  en één extra observatie  $(x^*, y)$ .
- Dit herhaal je vele keren en voor elke keer bereken je bijvoorbeeld het 95%-voorspellingsinterval.
- De extra observatie valt in 95% van de gevallen binnen het voorspellingsinterval.

Het C%-voorspellingsinterval heeft weer een t-verdeling met  $n - 2$  vrijheidsgraden en wordt berekend door:

$$SE_{\hat{y}} = s \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

$\hat{y} \pm t^* SE_{\hat{y}}$  en

## Variantieanalyse bij enkelvoudige lineaire regressie

Variantieanalyse voor regressie is gebaseerd op het model van DATA = FIT + RESIDU. Er zijn twee bronnen van spreiding in y:

1. Spreiding als gevolg van variatie in x.
2. Individuele spreiding rondom de geschatte y-waarde ( $\hat{y}_i$ ), voor een vaste waarde van  $x_i$ . Hieruit volgt

dat  $(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$ . Kwadrateren van deze factoren en sommeren over alle observaties leidt tot de volgende formule:

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad . \text{ Dit is gelijk aan } SST = SSM + SSE.$$

Bij elke kwadratensom hoort een bepaald aantal vrijheidsgraden:

- DFM is het aantal onafhankelijke variabelen. Bij enkelvoudige regressie: DFM = 1.
- DFE is gelijk aan  $N - 2$
- DFT is gelijk aan  $N - 1$  (DFM + DFE).

Net als bij variantieanalyse kunnen nu gemiddelde kwadratensommen worden berekend:

- MSM is de gemiddelde kwadratensom van het model (= SSM/DFM)
- MSE is de gemiddelde kwadratensom van de error (= SSE/ DFE)

De formules die horen bij MSM en MSE worden dan:

$$MSM = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 / 1 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \quad \text{en} \quad MSE = s^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - 2}$$

De nulhypothese dat  $\beta_1 = 0$  kan getoetst worden met de F-toets:  $F = MSM/MSE$ . F heeft een F-verdeling met in de teller 1 vrijheidsgraad en in de noemer  $n - 2$  vrijheidsgraden. De alternatieve hypothese is tweezijdig:  $\beta_1 \neq 0$ . Bij enkelvoudige regressie geldt dat  $F = t^2$ . De t-toets heeft dan de voorkeur omdat je daarmee ook eenzijdig kunt toetsen.

Het percentage verklaarde variantie geeft weer welk percentage van de variantie in y kan worden verklaard door het effect van x. Dit wordt weergegeven met  $r^2$ . Bij enkelvoudige regressie is  $r^2$  gelijk aan de gekwadrateerde correlatie tussen x en y. Verder geldt:

$$r^2 = \frac{SSM}{SST} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

De correlatie van de populatie wordt weergegeven met  $\rho$ . Als  $\rho = 0$ , dan is er in de populatie geen verband tussen x en y. Om de nulhypothese  $\rho = 0$  te toetsen (eenzijdig of tweezijdig), wordt een t-toets met  $n - 2$  vrijheidsgraden gebruikt. De t-waarde wordt berekend door:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad \text{en verder geldt bij enkelvoudige regressie:} \quad t = \frac{b_1}{SE_{b_1}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$