Karin Dahmström

Regresssion och korrelation

Vi har n st talpar (x_i, y_i), i =1,2,…,n. X och Y är på minst intervallskalenivå.
Anpassa linjen $\hat{y} = a + bx$

Bilda residualer $e_i = y_i - \hat{y}_i = y_i - [a + bx_i]$ i = 1,2,…,n

Minstakvadratmetoden (MK-metoden) innebär att a och b bestäms så att

$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ minimeras. Detta medför att

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

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

a och b kallas för regressionskoefficienterna.

Mått på styrkan av det linjära sambandet mellan X och Y på minst intervallskala:
Korrelationskoefficienten r:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{\sqrt{\left[n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2\right] \cdot \left[n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2\right]}}$$

Observera skillnaden mellan $(\sum_{i=1}^{n} x_i)^2$ och $\sum_{i=1}^{n} x_i^2$ samt mellan $\sum_{i=1}^{n} x_i y_i$ och $\sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i$!

Det gäller alltid att $-1 \le r \le 1$.

Residualvariansen

$$s_e^2 = \frac{\sum_{i=1}^{n}(e_i - \bar{e})^2}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n-2} = \frac{SSE}{n-2}$$

Total variation i Y

$$s_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = \frac{SST}{n-1}$$

Andel förklarad variation (determinationskoefficienten)

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$ = 1- andelen oförklarad variation

Vid enkel regression gäller att $r^2 = R^2$.

Använd regression även vid anpassning av linjära och exponentiella trender.

Ex: $\hat{y} = a + bt$
Om t = tiden kodas så att $\sum t = 0$, fås

$$b = \frac{\sum yt}{\sum t^2} \quad \text{och} \quad a = \bar{y}$$

Ex: $\hat{y} = a \cdot b^t$ Logaritmera!
log y = log a + t· log b

$y' = a' + b' \cdot t \quad \Rightarrow \quad b' = \frac{\sum y' t}{\sum t^2}$ och $a' = \bar{y}'$

Antilogaritmering ger $a = 10^a$ och $b = 10^b$

Säsongrensning

i = år, k = säsong ε = slump

Additiv modell: $y_{ik} = T_{ik} + S_{ik} + \varepsilon_{ik}$
Multiplikativ modell: $y_{ik} = T_{ik} \cdot S_{ik} \cdot \varepsilon_{ik}$

I båda modellerna antas att $S_{ik} = S_k$ för alla i.

Trenden skattas med hjälp av centrerade, glidande medeltal, $\hat{T}_{ik}$
Vikter:

Kvartalsdata: $\frac{1}{8}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{8}$ (5-leds)

Halvårsdata: $\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$ (3-leds)

Tertialdata: $\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$ (3-leds)

Veckodata: $\frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}$ (7-leds)

Krav och korrigeringar för S_k:

Additiv modell: $\sum S_k = 0 \Rightarrow \hat{S}_k = \bar{S}_k - \frac{\sum \bar{S}_k}{\textit{antal säsonger}}$

Multiplikativ modell: $\sum S_k = \textit{antal}$ säsonger $\Rightarrow \hat{S}_k = \bar{S}_k \cdot \frac{\textit{antal}\ \text{säsonger}}{\sum \bar{S}_k}$