Kapittel 11: Enkel lineær regresjon og korrelasjon

TMA4245 Statistikk

11.4-11.5: Eigenskapar til MKE for regresjonskoeffisientane.

Enkel lineær regresjon - modell

- Ein stokastisk variabel Y med $\mathrm{E}(Y) = \mu$ og $\mathrm{Var}(Y) = \sigma^2$ kan skrivast

$$Y = \mu + \epsilon, \quad \text{der } E(\epsilon) = 0, \text{ og } \mathrm{Var}(\epsilon) = \sigma^2.$$

- **Enkel lineær regresjon:** Relaterer avhengig variabel Y (stokastisk) til uavhengig variabel (forklaringsvariabel) x (ikkje-stokastisk) ved

$$Y = \alpha + \beta x + \epsilon, \quad \text{der } E(\epsilon) = 0, \text{ og } \mathrm{Var}(\epsilon) = \sigma^2,$$

$$\begin{aligned} \mathrm{E}(Y|x) &= \mu_{Y|x} = \alpha + \beta x, \\ \mathrm{Var}(Y|x) &= \sigma^2_{Y|x} = \sigma^2. \end{aligned}$$

Enkel lineær regresjon - tilpasning

- **Tilpassa regresjonslinje:** $\hat{y} = a + bx$, der a og b er estimat for regresjonskoeffisientane α og β.
- **Minste kvadratsums estimat** for α og β, basert på data $(x_i, y_i),\ i = 1, \ldots, n$: Verdiar a og b som minimerer

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2, \qquad e_i = \text{residual}$$

- **Minste kvadratsums estimatorar** A og B (stokastiske variablar):

$$B = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$A = \bar{Y} - B\bar{x}$$

Estimator for σ^2

- Ein forventningsrett estimator for σ^2 er

$$S^2 = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n}(Y_i - A - Bx_i)^2}{n-2}$$

- Dette er *ikkje* sannsynsmaksimeringsestimatoren, SME for σ^2 er $\frac{n-2}{n}S^2$.
- Har at $\frac{(n-2)S^2}{\sigma^2}$ er kjikvadrat-fordelt med $n-2$ fridomsgrader.

Eigenskapar til MKE A og B

Situasjon:

- Utvalg $(x_i, Y_i),\ i = 1, \ldots n.$
- For oss er x_i ein konstant, observert utan støy, dvs. $E(x_i) = x_i$ og $\mathrm{Var}(x_i) = 0$.
- $Y_i = \alpha + \beta x_i + \epsilon_i$
- Antar ϵ_i normalfordelt, dvs. $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.
- Sidan Y_i er ein lineær funksjon av ϵ_i, er Y_i også normalfordelt, med

$$\begin{aligned} E(Y_i) &= E(\alpha + \beta x_i + \epsilon_i) = \alpha + \beta x_i + 0 = \alpha + \beta x_i \\ \mathrm{Var}(Y_i) &= \mathrm{Var}(\alpha + \beta x_i + \epsilon_i) = 0 + 0 + \sigma^2 = \sigma^2 \end{aligned}$$

- ϵ_i'ane uavhengige $\Rightarrow$ Y_i og Y_j er uavhengige for $i \neq j$.
- Får at

$$\begin{aligned} \mathrm{E}(\bar{Y}) &= \mathrm{E}(\frac{1}{n}\sum_{i=1}^{n} Y_i) = \alpha + \beta\bar{x} \\ \mathrm{Var}(\bar{Y}) &= \mathrm{Var}(\frac{1}{n}\sum_{i=1}^{n} Y_i) = \frac{\sigma^2}{n} \end{aligned}$$

Estimatoren B for stigningstalet β

- Estimator: $B = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$
- B er normalfordelt med $E(B) = \beta$ og $\mathrm{Var}(B) = \sigma^2 / \sum_{i=1}^{n}(x_i - \bar{x})^2$ (avhengig av design, dvs. valg av x_i-ane).
- Bereknar konfidensintervall og baserer hypotesetesting på at

$$T = \frac{B - \beta}{S/\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

er t-fordelt med $n - 2$ fridomsgrader.

Estimatoren B (forts.)

- Aktuell hypotesetest: Test

$$H_0 : \beta = 0 \quad \text{mot} \quad H_1 : \beta \neq 0.$$

  Nullhypotesen betyr "ingen lineær samanheng" mellom $E(Y|x)$ og x.

- Dersom vi antar at både X og Y er stokastiske, og at simultanfordelinga er binormal med korrelasjonskoeffisient ρ, så er

$$\beta = \rho \frac{\sigma_Y}{\sigma_X} \quad \text{(Kap. 11.12)}.$$

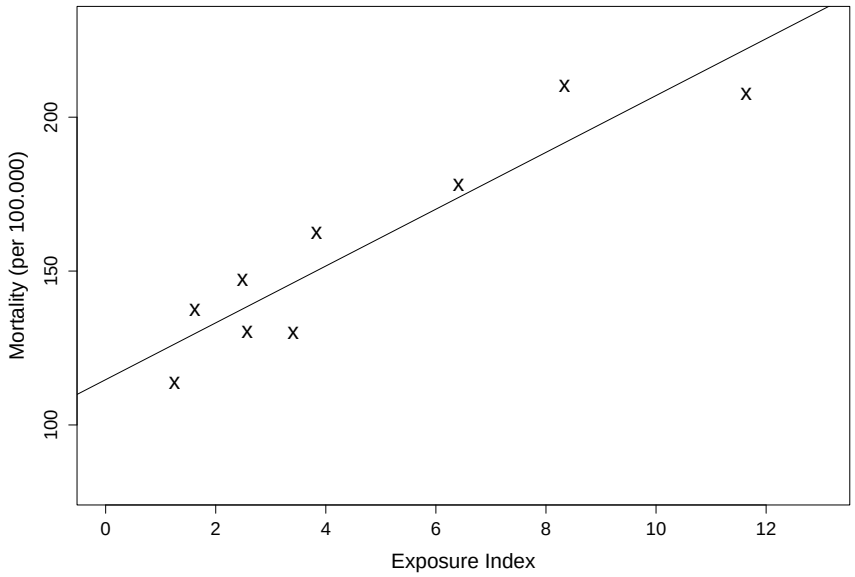
Estimatoren A for skjæringspunktet α

- Estimator $A = \bar{Y} - B\bar{X}$
- A er normalfordelt med $E(A) = \alpha$ og $\text{Var}(A) = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}$.
- Bereknar konfidensintervall og baserer hypotesetesting på at

$$T = \frac{A - \alpha}{S\sqrt{\frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}}}$$

  er t-fordelt med $n - 2$ fridomsgrader.

Radioaktivitet og kreftdødsfall



- $a = 114.70$ og $b = 9.23$
- $\sum_{i=1}^{9}(x_i - \bar{x})y_i = 900.13, \quad \sum_{i=1}^{9}(x_i - \bar{x})^2 = 97.51$
- $\sum_{i=1}^{9} x_i^2 = 289.42$, $\sum_{i=1}^{9}(y_i - a - bx_i)^2 = 1373.95$
- $\bar{x} = 4.62$ og $\bar{y} = 157.34$

"Bestemmelseskoeffisienten" R^2

- "Coeffi cient of determination", R^2
- Blir brukt som eit mål på kor godt den lineære regresjonsmodellen passar til dataene.
- Uttrykker kor stor andel av den totale variasjonen i dataene som den lineære regresjonsmodellen forklarer.
- R^2 er gitt ved

$$R^2 = \frac{\text{SST} - \text{SSE}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}, \quad \text{der}$$

  - $\text{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2$, "total sum of squares"
  - $\text{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, "error sum of squares".
- (Dersom X og Y er simultant binormalfordelte vil $R^2 = \hat{\rho^2}$, der ρ er korrelasjonskoeffi sienten mellom X og Y).