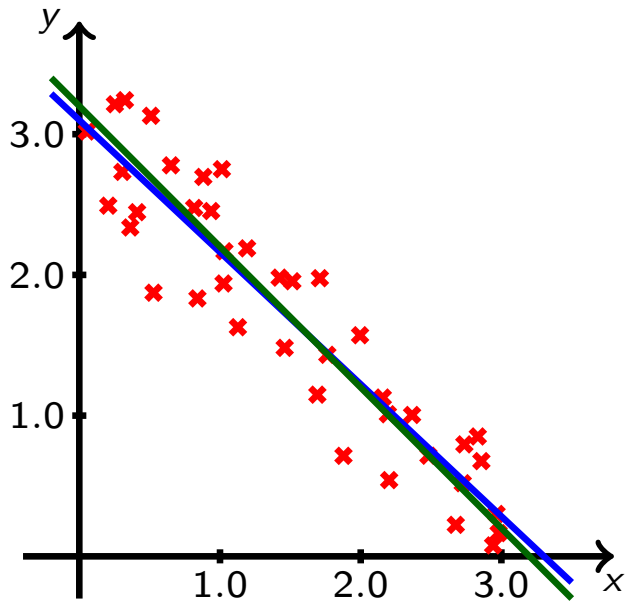
Enkel lineær regresjon og SME

TMA4240/TMA4245 Statistikk

Håkon Tjelmeland
Institutt for matematiske fag
Norges teknisk-naturvitenskapelige universitet

Enkel lineær regresjon

- Situasjon: Har observasjonspar $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$



- Ønsker å tilpasse en rett linje til de observerte parene
- Sannsynlighetsmodell: $Y_i = \alpha + \beta x_i + \varepsilon_i$ der $E[\varepsilon_i] = 0$ og $\text{Var}[\varepsilon_i] = \sigma^2$
  - α og β er ukjente parametre, vi ønsker å estimere disse
  - vi betrakter y_i'er som realisasjoner av stokastiske variabler $Y_1, Y_2, \ldots, Y_n$
  - vi betrakter x_i som tall (altså ikke stokastiske variabler)

- For å bruke sannsynlighetsmaksimeringsprinsippet må vi også anta hvilken sannsynlighetsfordeling ε_i'ene har
  - anta $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ uavhengige og $\varepsilon_i \sim N(0, \sigma^2)$
- Dermed: $Y_1, Y_2, \ldots, Y_n$ er uavhengige og

$$Y_i \sim N(\alpha + \beta x_i, \sigma^2)$$

- Merk: Vi har tre parametre vi trenger å estimere, α, β og σ^2

SME i enkel lineær regresjon

Normalfordeling: $f(x) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} \exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\}$

- Husk: $Y_1, Y_2, \ldots, Y_n$ er uavhengige og $Y_i \sim N(\alpha + \beta x_i, \sigma^2)$
- Rimelighetsfunksjon

$$L(\alpha, \beta, \sigma^2) = \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\left(y_i - (\alpha + \beta x_i)\right)^2\right\}\right]$$

- Log-rimelighetsfunksjonen

$$\begin{aligned}\ell(\alpha, \beta, \sigma^2) &= \sum_{i=1}^{n} \left[-\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\left(y_i - (\alpha + \beta x_i)\right)^2\right] \\ &= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - (\alpha + \beta x_i)\right)^2\end{aligned}$$

- Partiellderiverte

$$\frac{\partial \ell}{\partial \alpha} = -\frac{1}{2\sigma^2}\sum_{i=1}^{n} 2\left(Y_i - (\alpha + \beta x_i)\right) \cdot (-1) = \ldots = \frac{1}{\sigma^2}\left[\sum_{i=1} nY_i - n\alpha - \beta\sum_{i=1}^{n} x_i\right]$$

$$\frac{\partial \ell}{\partial \beta} = \frac{1}{2\sigma^2}\sum_{i=1}^{n} 2\left(Y_i - (\alpha + \beta x_i)\right) \cdot (-x_i) = \ldots = \frac{1}{\sigma^2}\left[\sum_{i=1}^{n} x_i Y_i - \alpha\sum_{i=1}^{n} x_i - \beta\sum_{i=1}^{n} x_i^2\right]$$

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}\left(y_i - (\alpha + \beta x_i)\right)^2 = \frac{1}{2\sigma^2}\left[-n + \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(y_i - (\alpha + \beta x_i)\right)^2\right]$$

SME i enkel lineær regresjon

- Ligningssystemet vi må løse blir

$$n\alpha + \beta \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} Y_i$$

$$\alpha \sum_{i=1}^{n} x_i + \beta \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i Y_i$$

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2 = n$$

  - merk: De to første er normalligningene fra minste kvadraters metode

- Log-rimelighetsfunksjonen har sitt maksimum for

$$\beta = \frac{\sum_{i=1}^{n}(x_i - \overline{x})y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$\alpha = \overline{y} - \beta\overline{x}$$

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$$

- SME er

$$\widehat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})Y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$\widehat{\alpha} = \overline{Y} - \widehat{\beta}\overline{x}$$

$$\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - (\widehat{\alpha} + \widehat{\beta} x_i) \right)^2$$

Oppsummering

- Enkel lineær regresjon
  - har observasjonspar $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
  - ønsker å finne en rett linje som passer med dataene
  - betrakter $y_1, y_2, \ldots, y_n$ som realisasjoner av stokastiske variabler $Y_1, Y_2, \ldots, Y_n$
  - betrakter $x_1, x_2, \ldots, x_n$ som tall
  - sannsynlighetsmodell: $Y_i = \alpha + \beta x_i + \varepsilon_i$ der $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ uavhengige og $\varepsilon_i \sim N(0, \sigma^2)$
- SME
  - bestemmer estimatorer $\widehat{\alpha}$, $\widehat{\beta}$ og $\widehat{\sigma}^2$ ved å maksimere log-rimelighetsfunksjonen
  - estimatorer

$$\widehat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})Y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$\widehat{\alpha} = \overline{Y} - \widehat{\beta}\overline{x}$$

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\Big(Y_i - (\widehat{\alpha} + \widehat{\beta}x_i)\Big)^2$$

- Merk:
  - vi har enda ikke diskutert om det er fornuftig å tilpasse ei rett linje
  - vi skal senere se at $\widehat{\sigma}^2$ er forventningsskjev