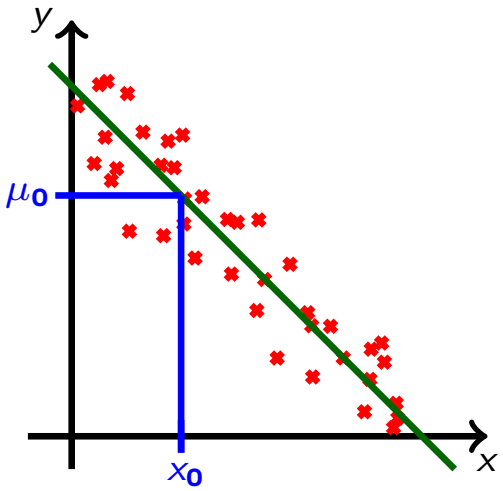
Prediksjon og prediksjonsintervall i enkel lineær regresjon

TMA4240/TMA4245 Statistikk

Håkon Tjelmeland
Institutt for matematiske fag
Norges teknisk-naturvitenskapelige universitet

Prediksjon og prediksjonsintervall i enkel lineær regresjon

- Situasjon: Har observasjonspar $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- Tilpasser en rett linje til de observerte parene
- Sannsynlighetsmodell: $Y_i \sim N(\alpha + \beta x_i, \sigma^2)$,
  $Y_1, Y_2, \ldots, Y_n$ uavhengige
- Husk:
  - α og β er ukjente parametre (tall)
  - betrakter Y_i-ene som stokastiske
  - betrakter x_i som tall (altså ikke stokastiske variabler)

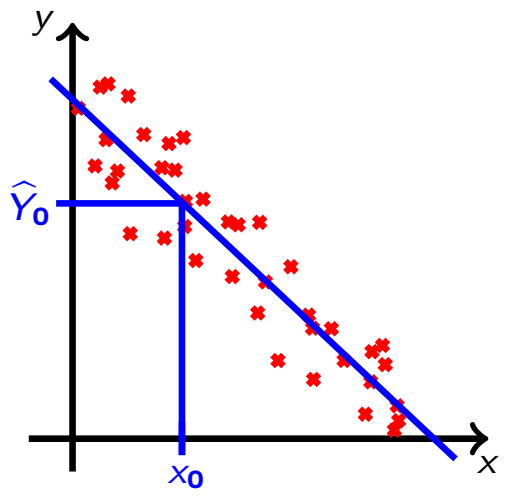


- Spørsmål: Hvis vi skal gjøre en ny observasjon Y_0 for $x = x_0$, hva kan vi si om Y_0?
  - samme modell: $Y_0 \sim N(\alpha + \beta x_0, \sigma^2)$
  - forventningsverdi: $\mu_0 = E[Y_0|x_0] = \alpha + \beta x_0$
  - hva er vårt gjett (prediksjon) på hva vi vil observere for Y_0?
  - kan vi finne et prediksjonsintervall for Y_0?
- Husk: Et prediksjonsintervall $\left[\widehat{Y}^\star_{0,L}(y_1, y_2, \ldots, y_n), \widehat{Y}^\star_{0,U}(y_1, y_2, \ldots, y_n)\right]$ skal være slik at

$$P\left(\widehat{Y}^\star_{0,L}(Y_1, Y_2, \ldots, Y_n) \leq Y_0 \leq \widehat{Y}^\star_{0,U}(Y_1, Y_2, \ldots, Y_n)\right) = 1 - \alpha$$

  - utleder prediksjonsintervall på lignende måte som et konfidensintervall

Prediksjon av Y_0



- Husk: $Y_0 \sim N(\alpha + \beta x_0, \sigma^2)$
  - alternativt: $Y_0 = \alpha + \beta x_0 + \varepsilon_0$ der $\varepsilon_0 \sim N(0, \sigma^2)$
  - vi estimerer verdien til α med $\widehat{\alpha}$
  - vi estimerer verdien til β med $\widehat{\beta}$
  - vi predikerer ε_0 med 0
- Predikterer Y_0 med

$$\widehat{Y}_0 = \widehat{\alpha} + \widehat{\beta} x_0$$

- Hvilke egenskaper har $\widehat{Y}_0 = \widehat{\alpha} + \widehat{\beta} x_0$?
  - husk: har tidligere funnet estimator for $\mu_0 = E[Y_0|x_0] = \alpha + \beta x_0$

$$\widehat{\mu}_0 = \widehat{\alpha} + \widehat{\beta} x_0 \sim N\left(\mu_0, \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)\right)$$

  - vi har altså at

$$\widehat{Y}_0 \sim N\left(\mu_0, \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)\right)$$

- For å utlede prediksjonsintervall for Y_0 skal vi ta utgangspunkt i

$$Y_0 - \widehat{Y}_0 \sim N\left(0, \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)\right)$$

Prediksjonsintervall for Y_0

- Har tidligere funnet at

$$Y_0 - \widehat{Y}_0 \sim N\left(0, \sigma^2\left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)\right), \qquad \frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}$$

  - S^2 er uavhengig av $\widehat{\mu}_0 = \widehat{\alpha} + \widehat{\beta} x_0 = \widehat{Y}_0$
  - S^2 og Y_0 er opplagt uavhengige
  - dermed er S^2 uavhengig av $Y_0 - \widehat{Y}_0$
- Husk definisjonen av t-fordeling: La $Z \sim N(0,1)$ og $V \sim \chi^2_\nu$ uavhengige. Da er

$$T = \frac{Z}{\sqrt{\frac{V}{\nu}}} \sim t_\nu$$

- Fra dette følger det at

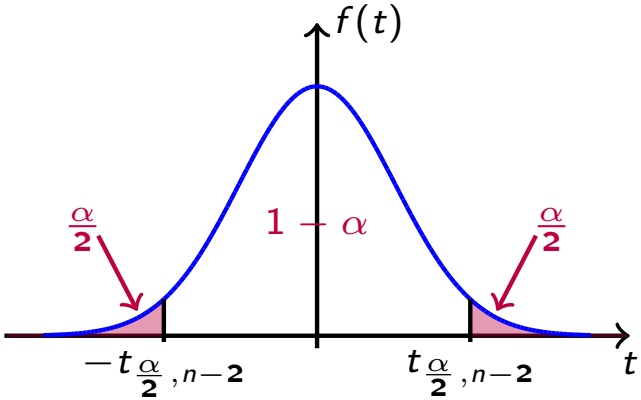
$$\frac{Y_0 - \widehat{Y}_0}{\sqrt{\sigma^2\left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)}} \sim N(0,1)$$

$$T = \frac{Y_0 - \widehat{Y}_0}{\sqrt{S^2\left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)}} = \frac{\frac{Y_0 - \widehat{Y}_0}{\sqrt{\sigma^2\left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)}}}{\sqrt{\frac{\frac{(n-2)S^2}{\sigma^2}}{n-2}}} \sim t_{n-2}$$

Prediksjonsintervall for Y_0

- Husk:

$$T = \frac{Y_0 - \widehat{Y}_0}{\sqrt{S^2\left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)}} \sim t_{n-2}$$



- Dermed får vi

$$P\left(-t_{\frac{\alpha}{2},n-2} \leq \frac{Y_0 - \widehat{Y}_0}{\sqrt{S^2\left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)}} \leq t_{\frac{\alpha}{2},n-2}\right) = 1 - \alpha$$

- Løser hver ulikhet med hensyn på Y_0

$$\widehat{Y}_0 - t_{\frac{\alpha}{2},n-2}\sqrt{S^2\left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)} \leq Y_0$$

$$Y_0 \leq \widehat{Y}_0 + t_{\frac{\alpha}{2},n-2}\sqrt{S^2\left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)}$$

- Dermed har vi at

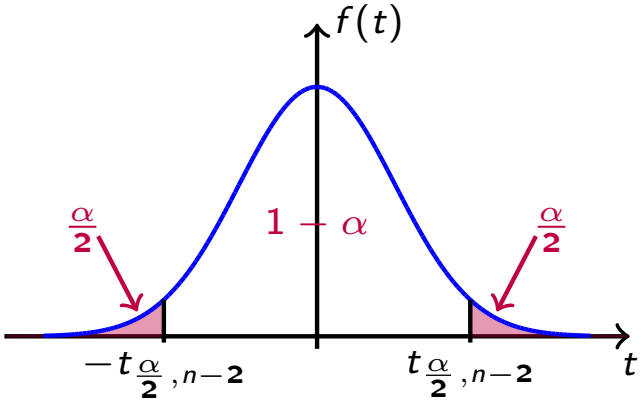
$$P\left(\widehat{Y}_0 - t_{\frac{\alpha}{2},n-2}\sqrt{S^2\left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)} \leq Y_0\right.$$

$$\left. \leq \widehat{Y}_0 + t_{\frac{\alpha}{2},n-2}\sqrt{S^2\left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)}\right) = 1 - \alpha$$

Prediksjonsintervall for Y_0

- Husk:

$$T = \frac{Y_0 - \widehat{Y}_0}{\sqrt{S^2\left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)}} \sim t_{n-2}$$



- Dermed får vi

$$P\left(-t_{\frac{\alpha}{2},n-2} \leq \frac{Y_0 - \widehat{Y}_0}{\sqrt{S^2\left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)}} \leq t_{\frac{\alpha}{2},n-2}\right) = 1 - \alpha$$

- Løser hver ulikhet med hensyn på Y_0

$$\widehat{Y}_0 - t_{\frac{\alpha}{2},n-2}\sqrt{S^2\left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)} \leq Y_0$$

$$Y_0 \leq \widehat{Y}_0 + t_{\frac{\alpha}{2},n-2}\sqrt{S^2\left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)}$$

- Prediksjonsintervallet blir

$$\left[\widehat{y}_0 - t_{\frac{\alpha}{2},n-2}\sqrt{s^2\left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)}, \widehat{y}_0 + t_{\frac{\alpha}{2},n-2}\sqrt{s^2\left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)}\right]$$

Oppsummering

- For en enkel lineær regresjonsmodell
  - har funnet hvordan vi skal predikere verdien til en fremtidig observasjon Y_0 for $x = x_0$
  - har utledet prediksjonsintervall for en fremtidig observasjon Y_0 for $x = x_0$
  - tok utgangspunkt i $Y_0 - \widehat{Y}_0$
  - prediksjonsintervallet ble

$$\left[\widehat{y}_0 - t_{\frac{\alpha}{2},n-2}\sqrt{s^2\left(1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right)}, \widehat{y}_0 + t_{\frac{\alpha}{2},n-2}\sqrt{s^2\left(1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right)}\right]$$

- Husk tolkning av prediksjonsintervall!