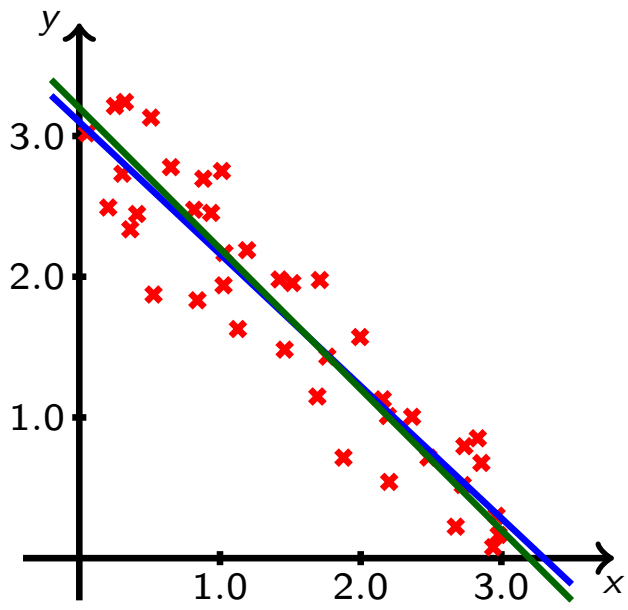
# Enkel lineær regresjon og minste kvadraters metode

TMA4240/TMA4245 Statistikk

Håkon Tjelmeland
Institutt for matematiske fag
Norges teknisk-naturvitenskapelige universitet

# Enkel lineær regresjon

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

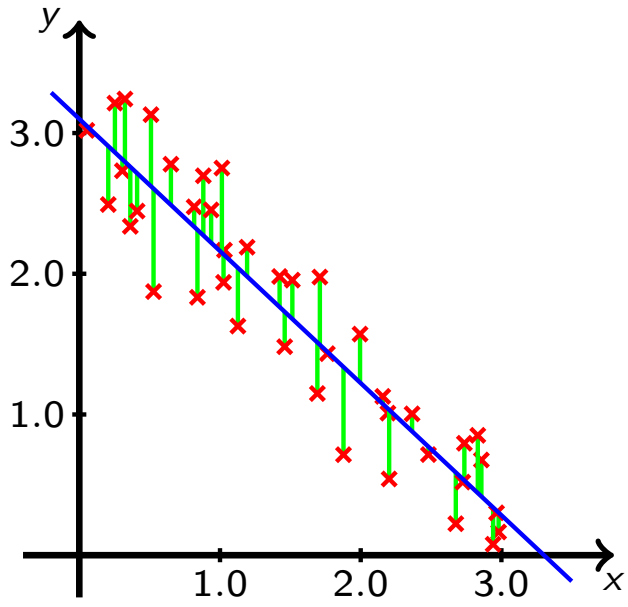



- Ønsker å tilpasse en rett linje til de observerte parene
- Sannsynlighetsmodell: $Y_i = \alpha + \beta x_i + \varepsilon_i$ der $E[\varepsilon_i] = 0$ og $\text{Var}[\varepsilon_i] = \sigma^2$
  - $\alpha$ og $\beta$ er ukjente parametre, vi ønsker å estimere disse
  - vi betrakter $y_i$'er som realisasjoner av av stokastiske variabler $Y_1, Y_2, \ldots, Y_n$
  - vi betrakter $x_i$ som tall (altså ikke stokastiske variabler)
- Hvordan finne gode estimatorer for $\alpha$ og $\beta$?
  - minste kvadraters metode
- Merk: Vi har to linjer
  - den ukjente sanne linja $y = \alpha + \beta x$
  - den estimerte linja $y = \widehat{\alpha} + \widehat{\beta} x$

# Minste kvadraters metode

- Idé: Måler «avstanden» mellom observasjonene og estimert linje ved

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




  - velger verdier for $\widehat{\alpha}$ og $\widehat{\beta}$ slik at SSE minimeres

- Matematisk: Finner $\widehat{\alpha}$ og $\widehat{\beta}$ ved å minimere SSE med hensyn på $\widehat{\alpha}$ og $\widehat{\beta}$.
  - må løse ligningssystemet

$$\frac{\partial \text{SSE}}{\partial \widehat{\alpha}} = 0, \quad \frac{\partial \text{SSE}}{\partial \widehat{\beta}} = 0$$

    med hensyn på $\widehat{\alpha}$ og $\widehat{\beta}$

# Utregning av estimatorer

- Husk: Vi minimere

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

- Partiellderiverte

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

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

- Ligningssystem (normalligningene)

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

- Løsning

$$\widehat{\beta} = \frac{\sum_{i=1}^{n} x_i Y_i - \frac{1}{n} \left( \sum_{i=1}^{n} x_i \right) \left( \sum_{i=1}^{n} Y_i \right)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left( \sum_{i=1}^{n} x_i \right)^2} \qquad \text{og} \qquad \widehat{\alpha} = \frac{1}{n} \sum_{i=1}^{n} Y_i - \widehat{\beta} \cdot \frac{1}{n} \sum_{i=1}^{n} x_i$$

- Omskriving (bruker $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ og $\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$)

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

# 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 $E[\varepsilon_i] = 0$ og $\text{Var}[\varepsilon_i] = \sigma^2$

- Minste kvadraters metode
  - bestemmer estimatorer $\widehat{\alpha}$ og $\widehat{\beta}$ ved å minimere

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

  - estimatorer

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

- Merk: Vi har enda ikke diskutert om det er fornuftig å tilpasse ei rett linje

- Hvorfor kalles det **enkel lineær regresjon**?
  - regresjon: ønsker å finne hvordan $y$ varierer som funksjon av $x$
  - lineær: $Y_i$ er en lineær funksjon av parametrene $\alpha$ og $\beta$
  - enkel: modellen har kun en forklaringsvariabel, $x$