

TMA4245 – Parallell 1

Geir-Arne Fuglstad

24. mars 2025

Enkel lineær regresjon – Problemformulering

Utgangspunkt:

Det finnes to størrelser y og x som vi kan måle. Vi er mest interessert i y , men tror at x kan forklare en del av variasjonen i y .

Eksempeler:

- Fødselsvekt (y) og lengde på svangerskap (x)
- Kvalitet på produsert vare (y) og en produksjonsparameter (x)
- Fortjeneste (y) og penger brukt på reklame (x)

Mål:

1. Forstå sammenhengen mellom y og x **med usikkerhet**.
2. Bruke sammenhengen til å si noe om en ny måling y^* for $x = x^*$ **med usikkerhet**.

Enkel lineær regresjon – Modell

Utgangspunkt:

Vi ser på Y som stokastisk og antar en lineær avhengighet på x i forventningsverdi

$$E[Y] = \alpha + x\beta, \quad \text{for } x \text{ i et rimelig område.}$$

Observasjoner:

Vi samler inn n målinger $(x_1, y_1), \dots, (x_n, y_n)$ ved **kjente** $x_1, \dots, x_n$ og **stokastiske** $Y_1, \dots, Y_n$, og antar

1. homogen varians

$$\text{Var}[Y_i] = \sigma^2, \quad i = 1, \dots, n.$$

2. ukorrelerte målinger

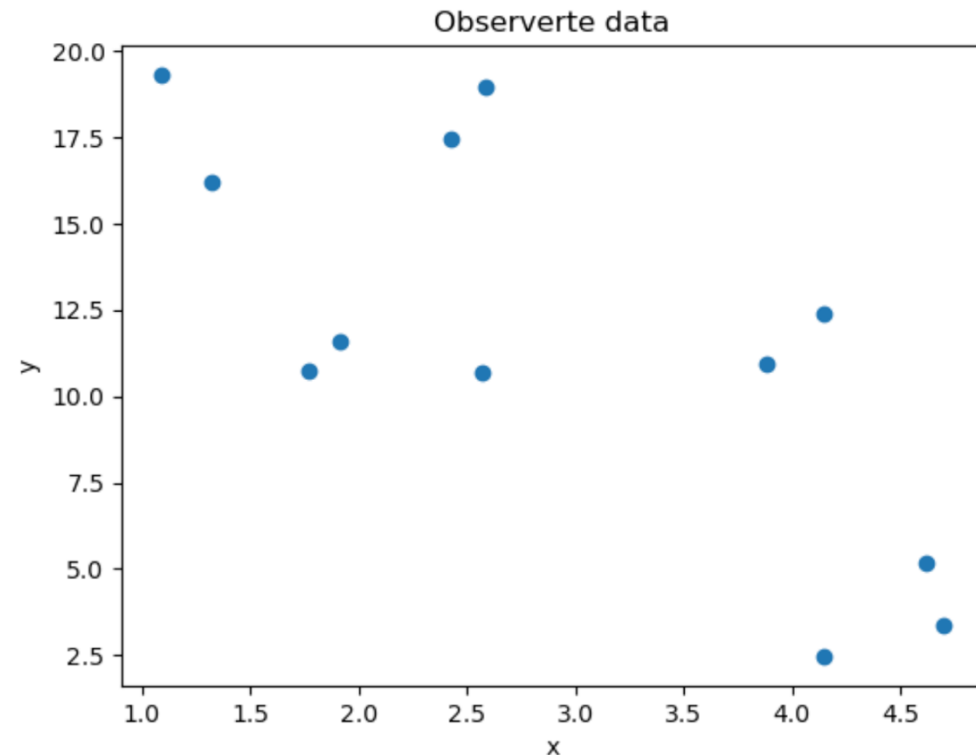
$$\text{Cov}[Y_i, Y_j] = 0, \quad i, j = 1, \dots, n.$$

Oppsummert:

$$Y_i = \alpha + x_i\beta + \epsilon_i, \quad i = 1, \dots, n,$$

der

1. $\epsilon_1, \dots, \epsilon_n$ er ukorrelerte, og alle har forventning 0 og varians σ^2 .
2. α, β , og $\sigma^2 > 0$ er **ukjente tall**.



Enkel lineær regresjon – Interessante størrelser

Modell:

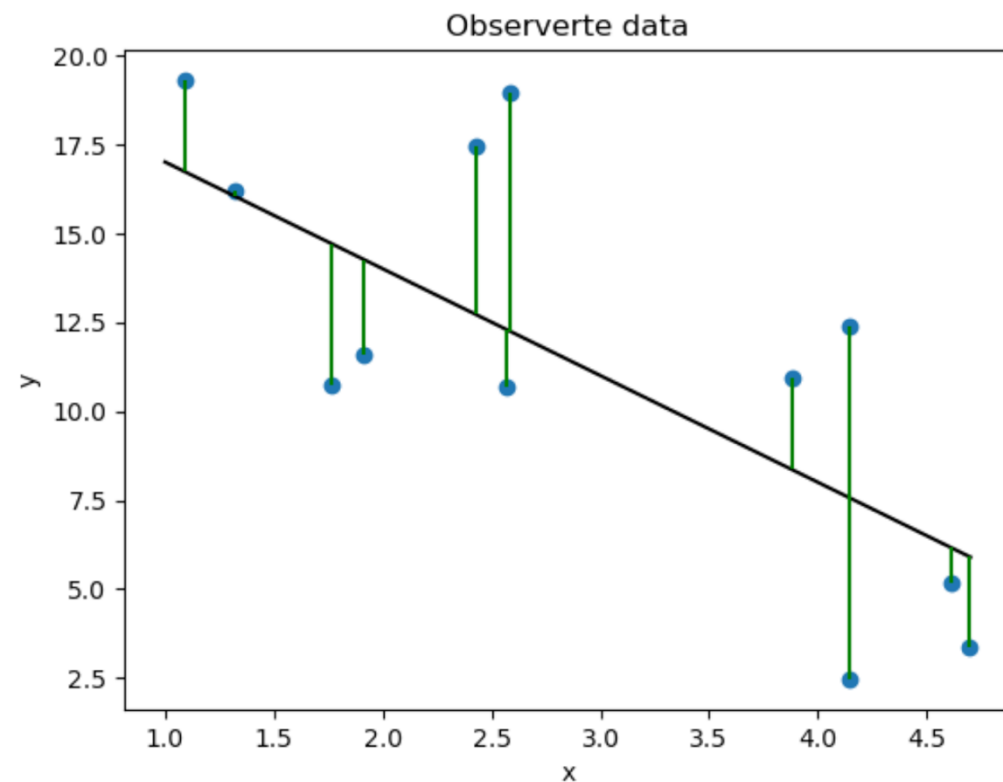
$$Y_i = \alpha + x_i\beta + \epsilon_i, \quad i = 1, \dots, n,$$

der

1. $\epsilon_1, \dots, \epsilon_n$ er ukorrelerte, og alle har forventning 0 og varians σ^2 .
2. α, β , og $\sigma^2 > 0$ er **ukjente tall**.

Interessante størrelser:

1. Sant **skjæringspunkt** α
2. Sant **stigningstall** β
3. Sann **(måle)variens** σ^2
4. Sann **regresjonslinje** $\mu(x) = \alpha + x\beta$
5. Sanne **residualer** $\epsilon_i = Y_i - \mu(x_i)$
6. Sann **prediktiv fordeling** til Y^* for x^* , $N(\alpha + x^*\beta, \sigma^2)$



Sannhet: $\mu(x) = 20 - 3x$ og $\sigma = 3$

Enkel lineær regresjon – Estimering

Modell:

$$Y_i = \alpha + x_i\beta + \epsilon_i, \quad i = 1, \dots, n,$$

der

1. $\epsilon_1, \dots, \epsilon_n$ er ukorrelerte, og alle har forventning 0 og varians σ^2 .
2. α, β , og $\sigma^2 > 0$ er **ukjente tall**.

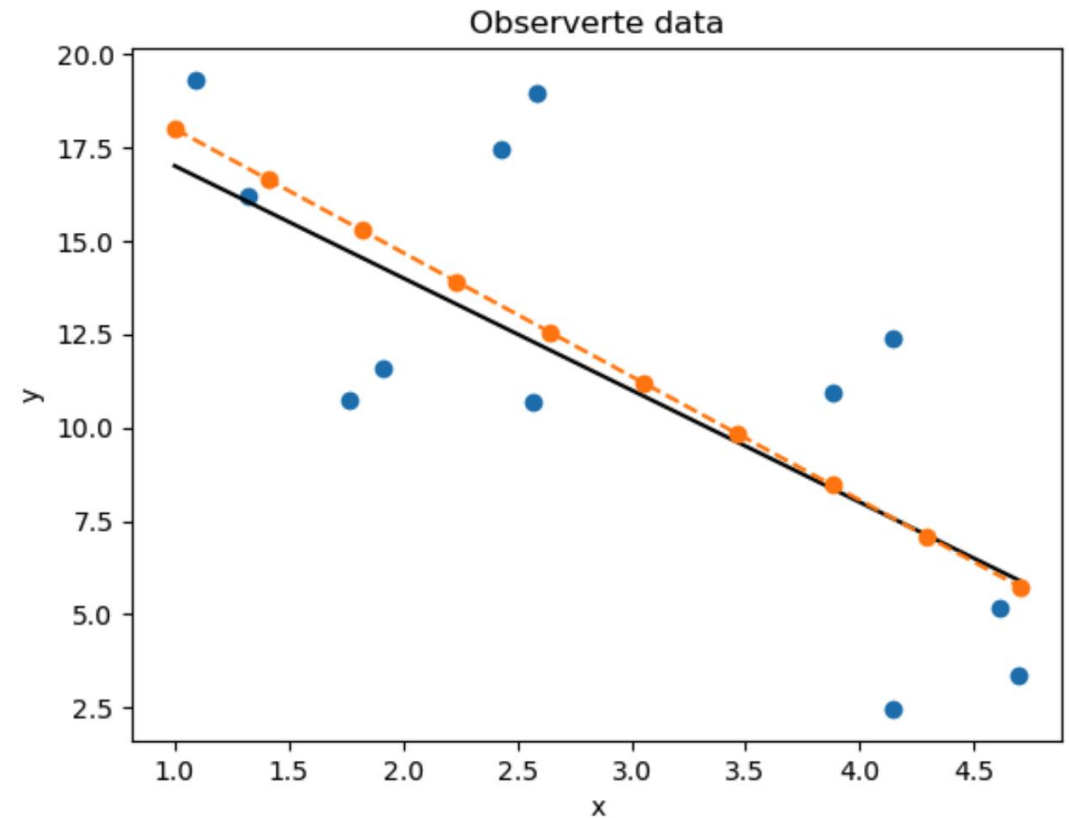
Denne uken:

- Estimator for **skjæringspunkt** $\hat{A}$
- Estimator for **stigningstall** $\hat{B}$
- Estimator for **(måle)variens** S^2
- Estimator for forventningsverdi av Y når $x = x^*$,
$$\hat{y}(x^*) = \hat{\mu}(x^*) = \hat{A} + \hat{B}x^*$$

Metoder:

Vi skal diskutere to metoder: minste-kvadraters-metode og sannsynlighetsmaksimering.

NB: Sannsynlighetsmaksimering, krever **flere antagelser**.



Sannhet:

$$\mu(x) = 20 - 3x \text{ og } \sigma = 3$$

Estimat:

$$\hat{y}(x) = \hat{\mu}(x) = 21.30 - 3.31x \text{ og } s = 3.98$$

Minste-kvadraters-metode

Prinsipp:

Minimer $SSE = \sum_{i=1}^n (y_i - a - x_i b)^2$ med hensyn på a og b .

Resultat:

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

Vanlig estimator for varians er:

$$S^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{A} - \hat{B} x_i)^2$$

NB: S^2 kan ikke utledes fra minste-kvadraters-metode, men er forventningsrett.

Sannsynlighetssmaksimering

Modell:

$$Y_i = \alpha + x_i\beta + \epsilon_i, \quad i = 1, \dots, n,$$

der

1. $\epsilon_1, \dots, \epsilon_n$ er ukorrelererte, og alle har forventning 0 og varians σ^2 .
2. α, β , og $\sigma^2 > 0$ er **ukjente tall**.
3. $\epsilon_1, \dots, \epsilon_n$ er normalfordelte.

Likelihood:

Vi ser at $Y_i \sim N(\alpha + x_i\beta, \sigma^2)$, $i = 1, \dots, n$, er uavhengige. Dette gir en likelihood $l(\alpha, \beta, \sigma^2; \mathbf{y}) = \prod_{i=1}^n f(y_i; \alpha, \beta, \sigma^2)$.

SME:

Maksimering av $l(\alpha, \beta, \sigma^2; \mathbf{y})$ med hensyn på α, β, σ^2 gir

$$\hat{B} = \frac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \hat{A} = \bar{Y} - \hat{B}\bar{x}, \quad \text{og} \quad \widehat{\sigma^2} = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{A} - \hat{B}x_i)^2$$

NB: $\hat{A}$ og $\hat{B}$ er like som minste-kvadraters-metode. Man bruker vanligvis S^2 og ikke $\widehat{\sigma^2}$.

Egenskaper til estimatorer

Modell:

$$Y_i = \alpha + x_i \beta + \epsilon_i, \quad i = 1, \dots, n,$$

der

1. $\epsilon_1, \dots, \epsilon_n$ er ukorrelerte, og alle har forventning 0 og varians σ^2 .
2. α, β , og $\sigma^2 > 0$ er **ukjente tall**.
3. $\epsilon_1, \dots, \epsilon_n$ er normalfordelte.

Fordelinger:

Under disse tre antagelsene har vi:

$$\begin{aligned}\hat{A} &\sim N\left(\alpha, \frac{\sum_{i=1}^n x_i^2}{n \cdot \sum_{i=1}^n (x_i - \bar{x})^2} \sigma^2\right) \\ \hat{B} &\sim N\left(\beta, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) \\ \frac{n-2}{\sigma^2} S^2 &\sim \chi_{n-2}^2\end{aligned}$$

I tillegg er 1) S^2 uavhengig av $\hat{A}$ og $\hat{B}$, og 2) $\bar{Y}$ er uavhengig av $\hat{B}$.

Videre

Fredag:

- Oppgaver med regning på sannsynlighetssmaksimeringsestimatoren
- Andre eksamensrelevante oppgaver

Neste uke:

- Vurdere antagelsene basert på plott av residualene
- Lage konfidensintervall og hypotesetester basert på fordelingene
- Bruke fordelingene til å lage prediksjonsintervaller for nye observasjoner