

# TMA4245 – Parallel 1

Geir-Arne Fuglstad, IMF

31. mars 2025

# Informasjon om kode på eksamen

- Dere skal kunne tolke Python kode
- Dere skal kunne skrive enkel Python kode
  - Den trenger ikke være syntaktisk perfekt men den må bruke funksjoner og funksjonalitet som faktisk finnes
  - Dere trenger ikke å huske navnet perfekt på alle funksjoner
- Dere skal kunne tolke output fra kode

Se eksamenener høst 2023, vår 2024, sommer 2024 og høst 2024 for eksempler.

# Viktige ting dere må vite/kunne utlede

**Modellen (står indirekte i formelsamlingen):**

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

- Residualene ( $\epsilon_1, \dots, \epsilon_n$ ) 1) har forventning 0, 2) har varians  $\sigma^2$ , 3) er ukorrelerte, og 4) er normalfordelte.
- $\alpha, \beta, \sigma^2$  er ukjente konstanter

**NB:** Antagelse 4) kan unnlates, men da får vi ikke kjente fordelinger på estimatorene.

**Estimatorene (står i formelsamlingen):**

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \hat{\alpha} = \bar{Y} - \hat{\beta} \bar{x} \quad \text{og} \quad S^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta} \bar{x})^2$$

**Fordelingene ( $\hat{\alpha}$  og  $\hat{\beta}$  står ikke i formelsamlingen):**

$$\hat{\alpha} \sim N\left(\alpha, \frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} \sigma^2\right), \quad \hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) \quad \text{og} \quad \frac{(n-2)S^2}{\sigma^2} \sim \chi_{n-2}^2$$

**NB:** Dere trenger ikke utlede fordelingen til  $S^2$ .

**Egenskaper (disse skal typisk ikke utledes):**

- $S^2$  er uavhengig av  $\hat{\alpha}$  og  $\hat{\beta}$  (**NB:** Dette står i formelsamlingen)
- $\bar{Y}$  og  $\hat{\beta}$  er ukorrelerte/uavhengige (**NB:** Står ikke i formelsamlingen og er veldig tungvint å vise med teknikkene dere kjenner)

# Vurdering av antagelsene

Etter modellen er estimert kan man regne ut empiriske/observerte residualer

$$e_i = \hat{\epsilon}_i = y_i - (\hat{a} + \hat{b}x_i), \quad i = 1, \dots, n.$$

Disse kan plottes for å vurdere antagelsene på residualene:

- Kryssplott av  $e_i$  mot  $i$  for å vurdere 1) forventning lik 0, 2) homogenitet, og 3) ukorrelert/uavhengig.
- Normalfordelingsplott av  $e_1, \dots, e_n$  for å vurdere normalfordelingsantagelsen.

Hvis man er usikker kan man simulere nye observasjoner fra den estimert modellen, og gjenta estimering og utregning av observerte residualer for å se hvor mye variasjon man kan forvente i disse plottene hvis den estimerte modellen var sann.

Målet er ikke å bevise at antagelsene er sanne. Målet er å oppdage om noen av antagelsene åpenbart ikke er oppfylt.

# Konfidensintervaller og hypotesetester

## Utgangspunkt:

$$\hat{A} \sim N\left(\alpha, \frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} \sigma^2\right), \quad \hat{B} \sim N\left(\beta, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) \quad \text{og} \quad \frac{(n-2)S^2}{\sigma^2} \sim \chi_{n-2}^2$$

Der vi antar kjent

- $S^2$  er uavhengig av  $\hat{A}$  og  $\hat{B}$
- $\bar{Y}$  og  $\hat{B}$  er ukorrelerte/uavhengige

## Oppsummering:

- For  $\hat{A}$ ,  $\hat{B}$  og  $\hat{A} + \hat{B}x$  baserer vi oss på pivotal med  $N(0,1)$ -fordeling når  $\sigma^2$  er kjent, men må bytte til  $t_{n-2}$ -fordeling når  $\sigma^2$  er ukjent (og byttes med estimator  $S^2$ ).
  - Vi vet at  $\hat{A}$ ,  $\hat{B}$  og  $\hat{A} + \hat{B}x$  er normalfordelte
  - Utfordringen er å kjenne/utlede variansen
- For  $S^2$  baserer vi oss på  $\chi_{n-2}^2$ -fordeling

# Prediksjonsintervall – Del 1

Hvis vi kjente alle parametere kunne vi direkte skrive opp prediktiv fordeling for en ny observasjon ved  $x^*$

$$Y^* \sim N(\alpha + \beta x^*, \sigma^2).$$

Under kjente  $\alpha$ ,  $\beta$ , og  $\sigma^2$  ville man basert prediksjonsintervallet på

$$Z = \frac{Y^* - (\alpha + \beta x^*)}{\sigma} \sim N(0,1). \quad (1)$$

Problemet er at  $\alpha$ ,  $\beta$ , og  $\sigma^2$  er ukjente og vi må bytte dem med estimatorer

$$\tilde{T} = \frac{Y^* - (\hat{A} + \hat{B}x^*)}{S}$$

men hva blir da fordelingen???

# Prediksjonsintervall – Del 2

Vi må ha inn en skalering for at

$$T = \frac{\tilde{T}}{\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} = \frac{Y^* - (\hat{A} + \hat{B}x^*)}{S \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \sim t_{n-2}. \quad (2)$$

**Ergo:** Det er to kostnader ved å gå fra ligning (1) til ligning (2) :

1. Bredden på prediksjonsintervaller øker med en faktor

$$\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}.$$

Dette skyldes at  $\alpha$  og  $\beta$  byttes med  $\hat{A}$  og  $\hat{B}$ .

2. Bredden på prediksjonsintervallet øker ved å skifte fra  $N(0,1)$  til  $t_{n-2}$ .  
Dette skyldes at  $\sigma^2$  byttes med  $S^2$ .