

Institutt for matematiske fag

Eksamensoppgave i TMA4240 Statistikk

Faglig kontakt under eksamen: Håkon Tjelmeland^a, Sara Martino^b

Tlf: ^a48 22 18 96, ^b99 40 33 30

Faglig kontakt møter i eksamenslokalet:

Eksamensdato: Fredag 22. desember 2023

Eksamenstid (fra–til): 09.00–13.00

Hjelpemiddelkode/Tillatte hjelpemidler: C

Tabeller og formles i statistikk, Akademika,
Bestemt, enkel kalkulator,
Gult stemplet A5-ark med egne håndskrevne notater.

Annen informasjon:

Alle svar skal begrunnes og besvarelsen skal inneholde naturlig mellomregning slik at det er helt klart hva som er gjort.

Målform/språk: bokmål

Antall sider (uten forside): 12

Antall sider vedlegg: 0

Kontrollert av:

Informasjon om trykking av eksamensoppgave

Originalen er:

1-sidig ☐ 2-sidig ☒

sort/hvit ☒ farger ☐

skal ha flervalgskjema ☐

Dato

Sign

Oppgave 1

Ved å standardisere får vi at

$$P(X \geq 2) = 1 - P(X < 2) = 1 - P(X \leq 2) = 1 - \Phi(2.00) = 1 - 0.9772 = \underline{\underline{0.0228}}.$$

Siden $13X - 2Y$ er en lineær funksjon av uavhengige og normalfordelte variabler er denne også normalfordelt. Forventningsverdi og varians for $13X - 2Y$ blir

$$\begin{aligned} E[13X - 2Y] &= E[13X] - E[2Y] = 13E[X] - 2E[Y] = 13 \cdot 0 - 2 \cdot 1 = -2, \\ \text{Var}[13X - 2Y] &= \text{Var}[13X] + \text{Var}[-2Y] = 13^2 \text{Var}[X] + (-2)^2 \text{Var}[Y] = 169 \cdot 1^2 + 4 \cdot 3^2 = 205. \end{aligned}$$

Dermed får vi at

$$P(13X - 2Y \leq 12) = P\left(\frac{13X - 2Y - (-2)}{\sqrt{205}} \leq \frac{12 - (-2)}{\sqrt{205}}\right) = \Phi\left(\frac{14}{\sqrt{205}}\right) = \Phi(0.98) = \underline{\underline{0.8365}}.$$

Siden X og Y er uavhengige får vi at

$$\begin{aligned} P(Y \geq 4 | X \geq 2) &= P(Y \geq 4) = 1 - P(Y < 4) = 1 - P(Y \leq 4) \\ &= 1 - P\left(\frac{Y - 1}{3} \leq \frac{4 - 1}{3}\right) = 1 - \Phi(1.00) = 1 - 0.8413 = \underline{\underline{0.1587}}. \end{aligned}$$

Oppgave 2

Fra python-koden ser vi at man genererer en realisasjon y ved formelen

$$y = \varphi\sqrt{u}. \tag{1}$$

Siden vi vet at $u \in [0, 1]$ får vi at de mulige verdiene for y er $y \in [0, \varphi]$.

Dessuten vet vi at $y = F_Y^{-1}(u) \Leftrightarrow u = F_Y(y)$ slik at vi kan finne $F_Y(y)$ ved å løse (1) med hensyn på u . Fra (1) får vi at

$$y = \varphi\sqrt{u} \Leftrightarrow \left(\frac{y}{\varphi}\right)^2 = u.$$

Følgelig har vi at

$$F_Y(y) = \left(\frac{y}{\varphi}\right)^2 = \frac{y^2}{\varphi^2} \quad \text{for } y \in [0, \varphi],$$

```

import numpy as np

def simX(n, theta):
    u = np.random.uniform(size=n)
    x = np.sqrt(-theta * np.log(1-u))

    return x

```

Figur 1: Python-funksjon som genererer n uavhengige realisasjoner av den kontinuerlige stokastiske variabelen X .

mens for $y < 0$ må vi at ha $F_Y(y) = 0$, og for $y > \varphi$ må vi ha $F_Y(y) = 1$. Vi kan da finne sannsynlighetstettheten ved å derivere $F_Y(y)$,

$$f_Y(y) = \begin{cases} \frac{2y}{\varphi^2} & \text{for } y \in [0, \varphi], \\ 0 & \text{ellers.} \end{cases}$$

Merk at en alternativ fremgangsmåte for å bestemme $f_Y(x)$ er å benytte transformasjonsformelen.

For å bestemme hvordan vi kan simulere en realisasjon x trenger vi å finne $F_X^{-1}(u)$. Dette gjør vi ved å løse $u = F_X(x)$ med hensyn på x . Fra formelen for $F_X(x)$ ser vi at vi må ha $x \geq 0$, og vi løser derfor ligningen under forutsetning av at $x \geq 0$. Vi får da

$$\begin{aligned}
 u &= 1 - \exp\left\{-\frac{x^2}{\theta}\right\} \\
 1 - u &= \exp\left\{-\frac{x^2}{\theta}\right\} \\
 \ln(1 - u) &= -\frac{x^2}{\theta} \\
 -\theta \ln(1 - u) &= x^2 \\
 x &= \sqrt{-\theta \ln(1 - u)}.
 \end{aligned}$$

Vi kan dermed generere en realisasjon x ved først å generere en realisasjon u fra uniformfordelingen på intervallet $[0, 1]$ og deretter beregne $x = \sqrt{-\theta \ln(1 - u)}$. En python-funksjon som genererer n uavhengige realisasjoner av X er gitt i figur 1.

Oppgave 3

a) Med $\theta = 100$ får vi at

$$\begin{aligned} P(X \geq 12) &= 1 - P(X < 12) = 1 - P(X \leq 12) = 1 - F_X(12) \\ &= 1 - \left(1 - \exp\left\{-\frac{12^2}{100}\right\}\right) = \exp\left\{-\frac{144}{100}\right\} = \exp\{-1.44\} = \underline{0.2369}, \\ P(X \geq 15|X \geq 12) &= \frac{P(X \geq 15 \cap X \geq 12)}{P(X \geq 12)} = \frac{P(X \geq 15)}{P(X \geq 12)}. \end{aligned}$$

Siden vi allerede har funnet $P(X \geq 12)$ trenger vi å bestemme $P(X \geq 15)$. Ved tilsvarende regning som for $P(X \geq 12)$ får vi at

$$\begin{aligned} P(X \geq 15) &= 1 - P(X < 15) = 1 - P(X \leq 15) = 1 - F_X(15) \\ &= 1 - \left(1 - \exp\left\{-\frac{15^2}{100}\right\}\right) = \exp\left\{-\frac{225}{100}\right\} = \exp\{-2.25\}. \end{aligned}$$

Dermed får vi at

$$\begin{aligned} P(X \geq 15|X \geq 12) &= \frac{\exp\{-2.25\}}{\exp\{-1.44\}} \\ &= \exp\{-2.25 + 1.44\} = \exp\{-0.81\} = \underline{0.4449}. \end{aligned}$$

Vi finner sannsynlighetstettheten til X ved å derivere. For $x < 0$ har vi åpenbart $f_X(x) = 0$. For $x \geq 0$ får vi ved å bruke kjerneregelen at

$$\begin{aligned} f_X(x) &= -\exp\left\{-\frac{x^2}{\theta}\right\} \cdot \left(-\frac{2x}{\theta}\right) \\ &= \underline{\underline{\frac{2x}{\theta} \exp\left\{-\frac{x^2}{\theta}\right\}}}. \end{aligned}$$

Ved bruk av delvis integrasjon kan vi da finne $E[X^2]$,

$$\begin{aligned} E[X^2] &= \int_{-\infty}^{\infty} x^2 f_X(x) dx = \int_0^{\infty} x^2 \cdot \frac{2x}{\theta} \exp\left\{-\frac{x^2}{\theta}\right\} dx \\ &= \left[x^2 \cdot \left(-\exp\left\{-\frac{x^2}{\theta}\right\}\right) \right]_0^{\infty} - \int_0^{\infty} 2x \cdot \left(-\exp\left\{-\frac{x^2}{\theta}\right\}\right) dx \\ &= 0 - 0 + \theta \int_0^{\infty} \frac{2x}{\theta} \exp\left\{-\frac{x^2}{\theta}\right\} dx \\ &= \theta \int_0^{\infty} f_X(x) dx = \underline{\underline{\theta}}. \end{aligned}$$

- b) Sentralgrenseteoremet antar at vi har uavhengige og identisk fordelte stokastiske variabler $V_1, V_2, \dots, V_n$ med $\mu = E[V_i]$ og $\sigma^2 = \text{Var}[V_i]$. Teoremet definerer så

$$\bar{V} = \frac{1}{n} \sum_{i=1}^n V_i$$

og

$$Z = \frac{\bar{V} - \mu}{\sqrt{\frac{\sigma^2}{n}}}.$$

Når $n \rightarrow \infty$ vil da sannsynlighetsfordelingen til Z konvergere mot en standard normalfordeling når $n \rightarrow \infty$.

Den praktiske konsekvensen av teoremet er at når n er stor vil Z være tilnærmet normalfordelt. Siden $\bar{V} = \mu + \sqrt{\frac{\sigma^2}{n}} Z$ er en lineær funksjon av Z vil også $\bar{V}$ være tilnærmet normalfordelt når n er stor. Dessuten får man at

$$\begin{aligned} E[\bar{V}] &= E\left[\mu + \sqrt{\frac{\sigma^2}{n}} Z\right] = \mu + \sqrt{\frac{\sigma^2}{n}} E[Z] = \mu, \\ \text{Var}[\bar{V}] &= \text{Var}\left[\mu + \sqrt{\frac{\sigma^2}{n}} Z\right] = \left(\sqrt{\frac{\sigma^2}{n}}\right)^2 \text{Var}[Z] = \frac{\sigma^2}{n}, \end{aligned}$$

som betyr at

$$\bar{V} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

når n er stor.

Ved å sette $V_i = X_i^2$ kan vi anvende sentralgrenseteoremet på situasjonen i oppgaven. Da blir $\mu = E[V_i] = E[X_i^2] = \theta$ og $\sigma^2 = \text{Var}[V_i] = \text{Var}[X_i^2] = 2\theta^2$. Dermed har vi at

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^n X_i^2 = \frac{1}{n} \sum_{i=1}^n V_i \approx N\left(\theta, \frac{\theta^2}{n}\right).$$

- c) Siden vi vet at $E[X^2] = \theta$, vet vi at levetiden X tenderer til å være stor dersom θ er stor. Alternativ hypotese blir derfor $H_1 : \theta > \theta_0$, og dermed blir nullhypotesen $H_0 : \theta = \theta_0$. Vi skal altså teste

$$\underline{\underline{H_0 : \theta = \theta_0 \quad \text{mot} \quad H_1 : \theta > \theta_0.}}$$

Fra deloppgave c) har vi at

$$\hat{\theta} \approx N\left(\theta, \frac{\theta^2}{n}\right).$$

Ved å standardisere får vi at

$$\frac{\hat{\theta} - \theta}{\sqrt{\frac{\theta^2}{n}}} \approx N(0, 1).$$

Siden $\theta = \theta_0$ når H_0 er sann får vi en naturlig testobservator ved å erstatte den ukjente verdien θ i uttryket over med den kjente verdi θ_0 ,

$$Z = \frac{\hat{\theta} - \theta_0}{\sqrt{\frac{\theta_0^2}{n}}}$$

og vi har at $Z \approx N(0, 1)$ når H_0 er sann.

Når H_1 er sann er θ stor, og når θ er stor vil $\hat{\theta}$ tendere til å være stor, og når $\hat{\theta}$ tenderer til å være stor vil Z tendere til å være stor. Dermed er det rimelig å forkaste H_0 når $Z > k$, der k er en kritisk verdi.

For å finne p-verdien må vi først bestemme observert verdi for testobservatoren. Med de observerte verdiene får vi

$$\hat{\theta} = \frac{1}{50} \cdot 6084.853 = 121.6971$$

som gir

$$z_{\text{obs}} = \frac{121.6971 - 100}{\sqrt{\frac{100^2}{50}}} = 1.53422.$$

Dermed blir p-verdien gitt som

$$\begin{aligned} p &= P(Z > z_{\text{obs}} | H_0 \text{ sann}) = 1 - P(Z \leq z_{\text{obs}} | H_0 \text{ sann}) \\ &= 1 - \Phi(z_{\text{obs}}) = 1 - \Phi(1.53) = 1 - 0.9370 = \underline{\underline{0.0630}}. \end{aligned}$$

Vi ser at p-verdien er relativt liten. Sannsynligheten for å observere det vi har observert eller noe mer ekstremt når H_0 er sann er altså 0.0630. Om man skal forkaste H_0 eller ikke kommer an på hvilket signifikansnivå man velger. Man skal forkaste H_0 dersom p-verdien er mindre enn signifikansnivået α . Så dersom man for eksempel velger $\alpha = 0.10$ blir konklusjonen at man skal forkaste H_0 , mens dersom man velger $\alpha = 0.05$ blir konklusjonen at man ikke skal forkaste H_0 .

- d) Som diskutert over har vi at p-verdien er gitt som

$$p = P(Z > z_{\text{obs}} | H_0 \text{ sann}).$$

Siden vi nå er interessert i den eksakte p-verdien kan vi ikke regne som om Z er standard normalfordelt. Vi må i stedet simulere mange realisasjoner av Z (under forutsetning av at H_0 er sann) og så bruke disse realisasjonene til å estimere sannsynligheten $P(Z > z_{\text{obs}} | H_0 \text{ sann})$. For å simulere uavhengige realisasjoner $z_1, z_2, \dots, z_m$ av Z må vi benytte python-funksjonen `simX` som vi definerte i løsningen av oppgave 2. Funksjonen `simX` må kalles med $n = 50$ siden det var 50 observasjoner vi faktisk gjorde, og med $\theta = 100$ siden under H_0 er θ lik 100. For hver $i = 1, 2, \dots, m$ må vi først kalle `simX` og få returnert verdier $x_1, x_2, \dots, x_n$. Basert på disse simulerte verdiene $x_1, x_2, \dots, x_n$ beregner vi så $\hat{\theta}$ og observert verdi for testobservatoren Z , og denne utregnede verdien kaller vi z_i . Vi har da verdier $z_1, z_2, \dots, z_m$ som er uavhengige realisasjoner fra (den eksakte) sannsynlighetsfordelingen til Z . Et estimat for p-verdien blir da

$$\hat{p} = \frac{1}{m} \sum_{i=1}^m \mathbb{I}(z_i > z_{\text{obs}}),$$

der $\mathbb{I}(A)$ er indikatorfunksjonen som er lik 1 dersom A er sann og lik 0 dersom A ikke er sann.

I figur 2 er det en python-funksjon som returnerer estimert p-verdi. Funksjonen må kalles med `n=50`, `theta0=100` og `zObs=1.53422`. Input-parameteren `m` må være et stort positivt heltall.

Oppgave 4

- a) For å bestemme hvilken estimator som er best starter vi med å undersøke hvilke estimatorer som er forventningsrette ved å regne ut forventningsver-

```

import numpy as np

def estimerPverdi(n, theta0, zObs, m):
    antallStorreEnnZobs = 0
    for i in np.arange(m):
        x = simX(n, theta0)
        thetaHat = sum(x**2)/n
        zSim = (thetaHat - theta0)/np.sqrt(theta0**2/n)
        if zSim > zObs:
            antallStorreEnnZobs = antallStorreEnnZobs + 1

    pHat = antallStorreEnnZobs / m

    return pHat

```

Figur 2: Python-funksjon som estimerer eksakt p-verdi.

diene til hver estimator,

$$\begin{aligned}
 E[\beta^*] &= E\left[\frac{1}{n} \sum_{i=1}^n Y_i\right] = \frac{1}{n} E\left[\sum_{i=1}^n Y_i\right] = \frac{1}{n} \sum_{i=1}^n E[Y_i] = \frac{1}{n} \sum_{i=1}^n \beta x_i = \beta \bar{x}, \\
 E[\tilde{\beta}] &= E\left[\frac{1}{n} \sum_{i=1}^n \frac{Y_i}{x_i}\right] = \frac{1}{n} E\left[\sum_{i=1}^n \frac{Y_i}{x_i}\right] = \frac{1}{n} \sum_{i=1}^n E\left[\frac{Y_i}{x_i}\right] = \frac{1}{n} \sum_{i=1}^n \frac{1}{x_i} \cdot E[Y_i] \\
 &= \frac{1}{n} \sum_{i=1}^n \frac{1}{x_i} \cdot \beta x_i = \frac{1}{n} \sum_{i=1}^n \beta = \frac{1}{n} \cdot n\beta = \beta, \\
 E[\tilde{\tilde{\beta}}] &= E\left[\frac{\sum_{i=1}^n Y_i x_i}{\sum_{i=1}^n x_i^2}\right] = \frac{1}{\sum_{i=1}^n x_i^2} \cdot E\left[\sum_{i=1}^n Y_i x_i\right] = \frac{\sum_{i=1}^n E[Y_i x_i]}{\sum_{i=1}^n x_i^2} \\
 &= \frac{\sum_{i=1}^n x_i E[Y_i]}{\sum_{i=1}^n x_i^2} = \frac{\sum_{i=1}^n x_i \cdot \beta x_i}{\sum_{i=1}^n x_i^2} = \frac{\beta \sum_{i=1}^n x_i^2}{\sum_{i=1}^n x_i^2} = \beta.
 \end{aligned}$$

Vi ser at $\tilde{\beta}$ og $\tilde{\tilde{\beta}}$ er forventningsrette, mens β^* er forventningsskjev. Vi foretrekker en forventningsrett estimator. For å bestemme hvilken av $\tilde{\beta}$ og $\tilde{\tilde{\beta}}$ som er best må vi bestemme variansen til disse. I disse utregningen benytter vi

at $Y_1, Y_2, \dots, Y_n$ er antatt å være uavhengige stokastiske variabler,

$$\begin{aligned}\text{Var}[\tilde{\beta}] &= \text{Var}\left[\frac{1}{n} \sum_{i=1}^n \frac{Y_i}{x_i}\right] = \left(\frac{1}{n}\right)^2 \text{Var}\left[\sum_{i=1}^n \frac{Y_i}{x_i}\right] = \frac{1}{n^2} \sum_{i=1}^n \text{Var}\left[\frac{Y_i}{x_i}\right] \\ &= \frac{1}{n^2} \sum_{i=1}^n \left(\frac{1}{x_i}\right)^2 \text{Var}[Y_i] = \frac{1}{n^2} \sum_{i=1}^n \frac{\sigma^2}{x_i^2} = \frac{\sigma^2}{n^2} \sum_{i=1}^n \frac{1}{x_i^2}, \\ \text{Var}[\tilde{\beta}] &= \text{Var}\left[\frac{\sum_{i=1}^n Y_i x_i}{\sum_{i=1}^n x_i^2}\right] = \left(\frac{1}{\sum_{i=1}^n x_i^2}\right)^2 \text{Var}\left[\sum_{i=1}^n Y_i x_i\right] \\ &= \frac{\sum_{i=1}^n \text{Var}[Y_i x_i]}{(\sum_{i=1}^n x_i^2)^2} = \frac{\sum_{i=1}^n x_i^2 \text{Var}[Y_i]}{(\sum_{i=1}^n x_i^2)^2} = \frac{\sum_{i=1}^n x_i^2 \sigma^2}{(\sum_{i=1}^n x_i^2)^2} \\ &= \frac{\sigma^2 \sum_{i=1}^n x_i^2}{(\sum_{i=1}^n x_i^2)^2} = \frac{\sigma^2}{\sum_{i=1}^n x_i^2}.\end{aligned}$$

Vi må så avgjøre hvilken varians som er minst, $\text{Var}[\tilde{\beta}]$ eller $\text{Var}[\tilde{\beta}]$. Den enkleste måten å gjøre dette på er å sette inn de observerte verdiene for x_i^2 i uttrykkene for de to variansene,

$$\begin{aligned}\text{Var}[\tilde{\beta}] &= \frac{\sigma^2}{31^2} \cdot 13.5248 = 0.01407\sigma^2, \\ \text{Var}[\tilde{\beta}] &= \frac{\sigma^2}{193.3769} = 0.00517\sigma^2.\end{aligned}$$

Vi ser at variansen til $\tilde{\beta}$ er minst, og siden vi foretrekker den (forventningsrette) estimatoren som har minst varians betyr dette at $\tilde{\beta}$ er den beste estimatoren.

Man kan også vise at $\tilde{\beta}$ er det beste estimatoren ikke bare for de x_i -verdiene gitt over, men for alle verdier av $x_1, x_2, \dots, x_n$. For å vise dette kan man regne på forholdet mellom de to variansene,

$$\begin{aligned}\frac{\text{Var}[\tilde{\beta}]}{\text{Var}[\tilde{\beta}]} &= \frac{\frac{\sigma^2}{n^2} \sum_{i=1}^n \frac{1}{x_i^2}}{\frac{\sigma^2}{\sum_{i=1}^n x_i^2}} = \frac{1}{n^2} \left(\sum_{i=1}^n \frac{1}{x_i^2}\right) \cdot \left(\sum_{j=1}^n x_j^2\right) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \frac{x_j^2}{x_i^2} \\ &= \frac{1}{n^2} \left[\sum_{i=1}^n \frac{x_i^2}{x_i^2} + \sum_{i=2}^n \sum_{j=1}^{i-1} \left(\frac{x_i^2}{x_j^2} + \frac{x_j^2}{x_i^2}\right) \right] = \frac{1}{n^2} \left[n + \sum_{i=2}^n \sum_{j=1}^{i-1} \left(z_{ij} + \frac{1}{z_{ij}}\right) \right], \quad (2)\end{aligned}$$

der vi har innført

$$z_{ij} = \frac{x_i^2}{x_j^2}$$

og hvor vi skal legge merke til at $z_{ij} \geq 0$. Vi kan så analysere funksjonen

$$h(z_{ij}) = z_{ij} + \frac{1}{z_{ij}}; z_{ij} \geq 0.$$

Vi finner funksjonens ekstremalpunkter ved å sette den deriverte lik null,

$$h'(z_{ij}) = 1 - \frac{1}{z_{ij}^2} = 0 \Leftrightarrow z_{ij} = 1.$$

Vi regner ut den andrederiverte for å sjekke om dette er et maksimums- eller minimumspunkt,

$$h''(z_{ij}) = -(-2)z_{ij}^{-3}.$$

Vi ser at $h''(z_{ij}) > 0$ for alle $z_{ij} > 0$, som betyr at funksjonen $h(z_{ij})$ har sitt minimumspunkt for $z_{ij} = 1$. Dermed har vi at

$$h(z_{ij}) \geq h(1) = 1 + \frac{1}{1} = 2.$$

Ved å sette dette resultatet inn i (2) får man

$$\frac{\text{Var}[\tilde{\beta}]}{\text{Var}[\tilde{\tilde{\beta}}]} \geq \frac{1}{n^2} \left[n + \sum_{i=2}^n \sum_{j=1}^{i-1} 2 \right].$$

Ved å innse at antall ledd i dobbeltsummen i dette uttrykket er lik antall elementer under diagonalen i en $n \times n$ -matrise får vi at antall ledd er $n(n-1)/2$. Dermed får man at

$$\frac{\text{Var}[\tilde{\beta}]}{\text{Var}[\tilde{\tilde{\beta}}]} \geq \frac{1}{n^2} \left[n + 2 \cdot \frac{n(n-1)}{2} \right] = 1,$$

som betyr at

$$\text{Var}[\tilde{\beta}] \geq \text{Var}[\tilde{\tilde{\beta}}].$$

Man kan også legge merke til at likhet inntreffer kun dersom alle verdier x_i^2 er like. Som over kan vi konkludere at siden vi foretrekker den (forventningsrette) estimatoren som har minst varians får vi at $\tilde{\tilde{\beta}}$ er den beste estimatoren.

- b)** For å finne sannsynlighetsmaksimeringsestimatoren starter vi med rimelighetsfunksjonen,

$$L(\beta, \sigma^2) = f(y_1, y_2, \dots, y_n; \beta, \sigma^2) = \prod_{i=1}^n \left[\frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sqrt{\sigma^2}} \exp \left\{ -\frac{1}{2\sigma^2} (y_i - \beta x_i)^2 \right\} \right].$$

Finner så log-rimelighetsfunksjonen ved å ta $\ln$,

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

For å finne maksimumspunktet for $\ell(\beta, \sigma^2)$ finner vi først de partiellderiverte,

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

Vi finner så maksimumspunktet ved å sette de partiellderiverte lik null og får da to ligninger med to ukjente, nemlig β og σ^2 ,

$$\frac{\partial \ell}{\partial \beta} = 0 \Leftrightarrow \sum_{i=1}^n y_i x_i = \beta \sum_{i=1}^n x_i^2 \quad (3)$$

$$\frac{\partial \ell}{\partial (\sigma^2)} = 0 \Leftrightarrow n = \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \beta x_i)^2. \quad (4)$$

For å løse ligningssystemet bestående av (3) og (4) med hensyn på β og σ^2 kan vi legge merke til at (3) kun inneholder β . Vi kan dermed først løse (3) med hensyn på β og deretter bruke denne løsningen for β når vi løser (4) med hensyn på σ^2 . Dette gir

$$\begin{aligned}\beta &= \frac{\sum_{i=1}^n y_i x_i}{\sum_{i=1}^n x_i^2} \\ \sigma^2 &= \frac{1}{n} \sum_{i=1}^n (y_i - \beta x_i)^2.\end{aligned}$$

Sannsynlighetsmaksimeringsestimatorene blir dermed

$$\begin{aligned}\hat{\beta} &= \frac{\sum_{i=1}^n Y_i x_i}{\sum_{i=1}^n x_i^2} \\ \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{\beta} x_i)^2.\end{aligned}$$

Vi ser her at $\hat{\beta}$ er identisk med $\tilde{\beta}$ og at $\hat{\sigma}^2$ er på formen det blir spurt om i oppgaveteksten.

c) Fra modellantagelsene har vi at

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

Ved å standardisere får vi da at

$$\frac{Y_i - \beta x_i}{\sigma} \sim N(0, 1).$$

Så vet vi at dersom $Z \sim N(0, 1)$ vil vi ha at $z^2 \sim \chi_1^2$. Dette gir at

$$\left(\frac{Y_i - \beta x_i}{\sigma} \right)^2 \sim \chi_1^2.$$

Vi vet også at dersom $V_1, V_2, \dots, V_n$ er uavhengige og $V_i \sim \chi_{\nu_i}^2$ for $i = 1, 2, \dots, n$ har vi at $\sum_{i=1}^n V_i \sim \chi_{\sum_{i=1}^n \nu_i}^2$. Ved å benytte dette resultatet med $\nu_i = 1$ og at vi har at Y_i 'ene er uavhengige får vi

$$\sum_{i=1}^n \left(\frac{Y_i - \beta x_i}{\sigma} \right)^2 \sim \chi_n^2, \quad (5)$$

som gitt i oppgaveteksten.

Når vi i uttrykket (5) erstatter den (ukjente) sanne verdien β med vår estimator $\hat{\beta}$ «mister vi en frihetsgrad» slik at

$$\sum_{i=1}^n \left(\frac{Y_i - \hat{\beta} x_i}{\sigma} \right)^2 \sim \chi_{n-1}^2. \quad (6)$$

Vi vet at forventningsverdien til en χ^2 -fordelt variabel er lik antall frihetsgrader. Fra (6) følger det dermed at

$$E \left[\sum_{i=1}^n \left(\frac{Y_i - \hat{\beta} x_i}{\sigma} \right)^2 \right] = n - 1.$$

Ved å bruke at

$$\frac{n\hat{\sigma}^2}{\sigma^2} = \frac{n}{\sigma^2} \cdot \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{\beta} x_i)^2 = \sum_{i=1}^n \left(\frac{Y_i - \hat{\beta} x_i}{\sigma} \right)^2,$$

gir dette at

$$E\left[\frac{n\hat{\sigma}^2}{\sigma^2}\right] = n - 1 \Leftrightarrow \frac{n}{\sigma^2}E[\hat{\sigma}^2] = n - 1 \Leftrightarrow E[\hat{\sigma}^2] = \frac{n-1}{n}\sigma^2.$$

Siden $E[\hat{\sigma}^2] \neq \sigma^2$ er altså $\hat{\sigma}^2$ en forventningsskjev estimator for σ^2 . En forventningsrett estimator for σ^2 blir

$$S^2 = \frac{n}{n-1}\hat{\sigma}^2 = \frac{n}{n-1} \cdot \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{\beta}x_i)^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \hat{\beta}x_i)^2.$$

Viser at S^2 faktisk er en forventningsrett estimator for σ^2 ved å regne ut $E[S^2]$,

$$E[S^2] = E\left[\frac{n}{n-1}\hat{\sigma}^2\right] = \frac{n}{n-1}E[\hat{\sigma}^2] = \frac{n}{n-1} \cdot \frac{n-1}{n}\sigma^2 = \sigma^2.$$

d) Regresjonsmodellen i denne oppgaven kan formuleres ved

$$Y_i = \beta x_i + \varepsilon_i,$$

der $\varepsilon_i \sim N(0, \sigma^2)$ kalles residualet. Fra dette får vi at

$$\varepsilon_i = Y_i - \beta x_i.$$

Da verdien til β er ukjent kan vi ikke beregne ε_i , men vi kan estimere den realiserte verdien ved å erstatte β med estimatet $\hat{\beta}$ og den stokastiske verdien Y_i med den observerte verdien y_i ,

$$\hat{\varepsilon}_i = y_i - \hat{\beta}x_i.$$

I et residualplott plotter man x_i langs x-aksen og (estimert) residual $\hat{\varepsilon}_i$ langs y-aksen.

I residualplottet gitt i oppgaven ser vi at det ser ut som om variabiliteten til $\hat{\varepsilon}_i$ øker med økende verdi for x_i . Dersom den antatte modellen var korrekt skulle variabiliteten vært konstant som funksjon av x_i .

Det er dermed grunn til å tro at en regresjonsmodell hvor variansen øker med økende verdi av x_i vil passe bedre for dataene som analyseres i denne oppgaven. Man kan for eksempel anta at standardavviket til Y_i er proporsjonal med x_i , dvs at $Y_1, Y_2, \dots, Y_n$ er uavhengige og

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