Løsningsforslag til eksamen i STK1110 16. desember 2019

Oppgave 1

a) Hvis vi tester

$$H_0 : \mu \leq 5.0 \qquad \text{mot} \qquad H_{\text{a}} : \mu > 5.0,$$

er nullhypotesen at mengden organisk karbon i drikkevannet ikke overstiger grenseverdien 5.0 mg/l, mens den alternative hypotesen er at mengden organisk karbon oversiger grenseverdien. Hvis vi forkaster nullhypotesen, vil vi være rimelig sikre på at drikkevannet inneholder for mye organisk karbon.

Hvis vi tester

$$H_0 : \mu \geq 5.0 \qquad \text{mot} \qquad H_{\text{a}} : \mu < 5.0.$$

er nullhypotesen at mengden organisk karbon i drikkevannet overstiger grenseverdien 5.0 mg/l, mens den alternative hypotesen er at mengden organisk karbon ikke oversiger grenseverdien. Hvis vi forkaster nullhypotesen, vil vi være rimelig sikre på at drikkevannet ikke inneholder for mye organisk karbon.

b) Vi vil teste

$$H_0 : \mu \geq 5.0 \qquad \text{mot} \qquad H_{\text{a}} : \mu < 5.0.$$

Vi har at $\overline{X} \sim N(\mu, \sigma^2/5)$ og at $(\overline{X} - \mu)/(\sigma/\sqrt{5}) \sim N(0,1)$. Siden alternativet er $\mu < 5$, er det rimelig å forkaste nullhypotesen hvis

$$Z = \frac{\overline{X} - 5}{\sigma/\sqrt{5}} \leq c$$

for en passende valgt c. For at testen skal få signifikansnivå 5% må vi velge $c = -z_{0.05} = -1.645$. For da har vi at [der Φ er den kumulative standardnormalfordelingen]

$$\begin{aligned} P(\text{forkast } H_0 \,|\, H_0 \text{ er sann}) &= P\left(\frac{\overline{X} - 5}{\sigma/\sqrt{5}} \leq -z_{0.05} \,|\, \mu \geq 5\right) \\ &\leq P\left(\frac{\overline{X} - 5}{\sigma/\sqrt{5}} \leq -z_{0.05} \,|\, \mu = 5\right) \\ &= \Phi(-z_{0.05}) \\ &= 0.05 \end{aligned}$$

som viser at testen har signifikansnivå 5%.

Med tallene i oppgaven finner vi at $\overline{x} = 4.66$. Med $\sigma = 0.5$, får testobservatoren verdien

$$z = \frac{4.66 - 5}{0.5/\sqrt{5}} - 1.52$$

Altså er $z > -z_{0.05} = -1.645$, så vi forkaster ikke nulhypotesen på nivå 5%.

c) Generelt er P-verdien lik sannsynligheten, beregnet under forutsetning at nullhypotesen er sann, for at vi vil få en verdi av testobservatoren som er minst like mye i motsetning til nullhypotesen som den verdien vi faktisk fikk. Så her er P-verdien lik sannsynligheten, når $\mu = 5$, for at vi vil få en verdi av testobsetvatorene Z som er lik -1.52 eller mindre.

Ved å bruke tabellen over den kumulative standardnormalfordelingen,har vi at $\Phi(-1.52) = 0.064$, så P-verdien er 6.4%.

d) Hvis vi tar n vannprøver, får vi en test med signifikansnivå 5% hvis vi forkaster H_0 så sant

$$Z = \frac{\overline{X} - 5}{\sigma/\sqrt{n}} \leq -z_{0.05}$$

Feil av type II betyr at vi ikke forkaster nullhypotesen når den er gal. Vi har her at

$$\begin{aligned}
\beta(\mu') &= P(\text{feil av type II} \,|\, \mu = \mu') \\
&= P(\text{forkaster ikke } H_0 \,|\, \mu = \mu') \\
&= P\left(\frac{\overline{X} - 5}{\sigma/\sqrt{n}} > -z_{0.05} \,|\, \mu = \mu'\right) \\
&= P\left(\frac{\overline{X} - \mu'}{\sigma/\sqrt{n}} > -z_{0.05} + \frac{5 - \mu'}{\sigma/\sqrt{n}} \,|\, \mu = \mu'\right) \\
&= 1 - \Phi\left(-z_{0.05} + \frac{5 - \mu'}{\sigma/\sqrt{n}}\right)
\end{aligned}$$

Vi skal bestemme n slik at $\beta(4.5) \leq 0.10$. Nå har vi at $z_{0.05} = 1.645$ og $\sigma = 0.5$, så for $\mu' = 4.5$ skal vi bestemme n slik at

$$1 - \Phi\left(-1.645 + \frac{5 - 4.5}{0.5/\sqrt{n}}\right) \leq 0.10$$

dvs. slik at

$$\Phi(-1.645 + \sqrt{n}) \geq 0.90$$

Nå har vi at $\Phi(1.28) = 0.90$, så vi må velge n slik at

$$-1.645 + \sqrt{n} \geq 1.28$$

eller

$$\sqrt{n} \geq 1.28 + 1.645 = 2.925$$

som gir

$$n \geq 2.925^2 = 8.56$$

Det betyr at ingeniørene må ta minst 9 vannprøver.

e) Når σ er ukjent må vi estimere variansen σ^2 med

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Vi bruker da testobservatoren

$$T = \frac{\overline{X} - 5}{S/\sqrt{5}}$$

Når $\mu = 5$ er testobservatoren t-fordelt med $5 - 1 = 4$ frihetsgrader, så vi får en test med signifikansnivå 5% hvis vi forkaster nullhypotesen så sant $T \leq -t_{0.05,4}$.

Oppgave 2

a) Vi finner momentestimatoren ved å løse ligningen vi får når vi setter første teoretiske moment, dvs. $E(X_i)$, lik første empiriske moment, dvs. $\overline{X}$. Det er gitt i oppgaven at $E(X_i) = \theta\sqrt{\frac{\pi}{2}}$, så her er momentestimatoren $\widetilde{\theta}$ gitt ved ligningen

$$\widetilde{\theta}\sqrt{\frac{\pi}{2}} = \overline{X}$$

Det gir

$$\widetilde{\theta} = \overline{X}\sqrt{\frac{2}{\pi}}$$

Det er videre gitt i oppgaven at $E(X_i^2) = 2\theta^2$. Av det får vi at

$$\mathrm{V}(X_i) = E(X_i^2) - [E(X_i)]^2 = 2\theta^2 - \left(\theta\sqrt{\frac{\pi}{2}}\right)^2 = 2\theta^2 - \frac{\pi}{2}\theta^2 = \frac{4-\pi}{2}\theta^2$$

Av sentralgrensesetningen følger det nå at $\overline{X}$ er tilnærmet $N(\mu_X, \sigma_X^2/n)$-fordelt, der

$$\mu_X = E(X_i) = \theta\sqrt{\frac{\pi}{2}} \quad \text{og} \quad \sigma_X^2 = \mathrm{V}(X_i) = \frac{4-\pi}{2}\theta^2$$

Dermed er $\widetilde{\theta} = \sqrt{\frac{2}{\pi}}\overline{X}$ tilnærmet normalfordelt med forventningsverdi lik

$$\mu_{\widetilde{\theta}} = \sqrt{\frac{2}{\pi}}\mu_X = \sqrt{\frac{2}{\pi}}\theta\sqrt{\frac{\pi}{2}} = \theta$$

og varians lik

$$\sigma_{\widetilde{\theta}}^2 = \left(\sqrt{\frac{2}{\pi}}\right)^2 \frac{\sigma_X^2}{n} = \frac{2}{\pi} \cdot \frac{4-\pi}{2} \cdot \frac{\theta^2}{n} = \frac{4-\pi}{n\pi}\theta^2$$

b) La $x_1, \ldots, x_n$ være de observerte verdiene av $X_1, \ldots, X_n$. Da er likelihood funksjonen gitt ved

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta) = \prod_{i=1}^{n} \frac{1}{\theta^2} x_i e^{-x_i^2/(2\theta^2)} = \frac{1}{\theta^{2n}}\left(\prod_{i=1}^{n} x_i\right)\exp\left(-\frac{1}{2\theta^2}\sum_{i=1}^{n} x_i^2\right)$$

Log-likelihooden blir

$$\ln L(\theta) = -2n \ln \theta + \sum_{i=1}^{n} x_i - \frac{1}{2\theta^2} \sum_{i=1}^{n} x_i^2$$

Vi deriver med hensyn på θ og får at

$$\frac{\partial}{\partial \theta} \ln L(\theta) = -\frac{2n}{\theta} + \frac{1}{\theta^3} \sum_{i=1}^{n} x_i^2$$

Vi får maksimum likelihood estimatet $\widehat{\theta}$ ved å løse ligningen $\frac{\partial}{\partial \theta} \ln L(\theta) = 0$. Altså er $\widehat{\theta}$ løsningen av ligningen

$$-\frac{2n}{\widehat{\theta}} + \frac{1}{(\widehat{\theta})^3} \sum_{i=1}^{n} x_i^2 = 0$$

som gir

$$(\widehat{\theta})^2 = \frac{1}{2n} \sum_{i=1}^{n} x_i^2$$

Maksimum likelihood estimatet blir dermed

$$\widehat{\theta} = \sqrt{\frac{1}{2n} \sum_{i=1}^{n} x_i^2}$$

Vi får maksimum likelihood estimatoren ved å erstatte de obseverte x_i-ene med de stokastiske X_i-ene, så maksimum likelihood estimatoren er

$$\widehat{\theta} = \sqrt{\frac{1}{2n} \sum_{i=1}^{n} X_i^2}$$

c) Fisher informasjonen i én observasjon er gitt ved

$$I(\theta) = -E\left(\frac{\partial^2}{\partial \theta^2} \ln f(X_i; \theta) \right)$$

For $X_i > 0$ har vi nå at

$$\ln f(X_i; \theta) = -2 \ln \theta + \ln X_i - \frac{1}{2\theta^2} X_i^2$$

Dermed er

$$\frac{\partial}{\partial \theta} \ln f(X_i; \theta) = -\frac{2}{\theta} + \frac{1}{\theta^3} X_i^2$$

og

$$\frac{\partial^2}{\partial \theta^2} \ln f(X_i; \theta) = \frac{2}{\theta^2} - \frac{3}{\theta^4} X_i^2$$

Det er gitt i oppgaven at $E(X_i^2) = 2\theta^2$. Dermed er Fisher informasjonen i én observasjon gitt ved

$$I(\theta) = -E\left(\frac{2}{\theta^2} - \frac{3}{\theta^4}X_i^2\right) = -\frac{2}{\theta^2} + \frac{3}{\theta^4}E(X_i^2) = -\frac{2}{\theta^2} + \frac{3}{\theta^4}2\theta^2 = -\frac{2}{\theta^2} + \frac{6}{\theta^2} = \frac{4}{\theta^2}$$

d) Det er kjent at maksimum likelihood estimatoren $\widehat{\theta}$ (under visse regularitetsbetingelser som er oppfylt her) er tilnærmet normalfordelt med forventningsverdi θ og varians $1/[nI(\theta)]$. Her får vi variansen

$$\sigma^2_{\widehat{\theta}} = \frac{1}{nI(\theta)} = \frac{1}{n(4/\theta^2)} = \frac{\theta^2}{4n}$$

Både momentestimatoren $\widetilde{\theta}$ og maksimum likelihood estimatoren $\widehat{\theta}$ er tilnærmet normalfordelt med forventning θ. For å sammenligne estimatorene, kan vi dermed sammenligne variansene. Vi finner at

$$\frac{\sigma^2_{\widetilde{\theta}}}{\sigma^2_{\widehat{\theta}}} = \frac{\frac{4-\pi}{n\pi}\theta^2}{\frac{\theta^2}{4n}} = \frac{4(4-\pi)}{\pi} = 1.093$$

Det betyr at variansen til momentestimatoren $\widetilde{\theta}$ er 9.3% større enn variansen til maksimum likelihood estimatoren $\widehat{\theta}$, så maksimum likelihood estimatoren er å foretrekke.

Oppgave 3

a) Konstantleddet β_0 i regresjonsmodellen (1) er forventet utslipp når $x_{i1} = 0$, dvs. når lufttrykket er 750 mmHg. Av R-utskriften har vi estimatet $\widehat{\beta}_0 = 1.03$, så når lufttrykket er 750 mmHg vil vi forvente at utslippet er 1.03 ppm (parts per milion). Signingskoeffisienten β_1 er forventet endring i utslippet når lufttrykket øker med 1 mmHg. Her har vi estimatet $\widehat{\beta}_1 = 0.0182$, så hvis lufttrykket øker med 1 mmHg vil vi forvente at utslippet øker med 0.0182 ppm.

Vi har at $(\widehat{\beta}_1 - \beta_1)/S_{\widehat{\beta}_1}$ er t-fordelt med $n - 2 = 20 - 2 = 18$ frihetsgrader. [Her er $S^2_{\widehat{\beta}_1}$ den vanlige forventningsrette estimatoren for $\mathrm{V}(\widehat{\beta}_1)$.] Av dette følger det ved et standard argument at et 99% konfidensintervall for β_1 er gitt ved

$$\widehat{\beta}_1 \pm t_{0.005,18} \cdot s_{\widehat{\beta}_1}$$

Av tabellen først i oppgavesettet har vi at $t_{0.005,18} = 2.878$ og av R-utskriften har vi at $s_{\widehat{\beta}_1} = 0.00338$. Vi får dermed konfidensintervallet

$$0.0182 \pm 2.878 \cdot 0.00338$$

dvs. fra 0.0085 ppm til 0.0279 ppm.

b) Vi ser at i modell (1) er $\widehat{\beta}_1 = 0.0182$, mens i modell (2) er $\widehat{\beta}_1 = 0.0065$. Estimatet for modell (1) er altså nesten tre ganger så stort som estimatet for modell (2). Forklaringen på dette er at regresjonskoeffisienten β_1 ikke betyr det samme i modell (1) og modell (2). I modell (1) er β_1 forventet endring i utslippet når lufttrykket øker med 1 mmHg. Men i modell (2) er β_1 forventet endring i utslippet når lufttrykket øker med 1 mmHg *og luftfuktigheten forblir uendret*. Av matrise-spredningsplottet gitt i oppgaven ser vi at det er negativ korrelasjon mellom

lufttrykk og luftfuktighet. Så når lufttrykket øker, vil typisk luftfuktigheten bli mindre. Og lavere luftfuktighet vil føre til større utslipp. Estimatet for β_1 i modell (1) vil i tillegg til betydningen av lufttrykk også fange opp noe av effekten av luftfuktighet.

c) Hvis lufttrykket er 740 mmHg og luftfuktigheten er 80%, blir forklaringsvariablene $x_1^* = 740-750 = -10$ og $x_2^* = 80-50 = 30$. Forventet utslipp for disse verdiene av forklaringsvariablene er

$$\mu_{Y\cdot x_1^*, x_2^*} = \beta_0 + \beta_1 x_1^* + \beta_2 x_2^* = \beta_0 - 10\beta_1 + 30\beta_2$$

og en estimator for denne forventningsverdien er

$$\widehat{\mu}_{Y\cdot x_1^*, x_2^*} = \widehat{\beta}_0 - 10\widehat{\beta}_1 + 30\widehat{\beta}_2 = 0.9738 - 10 \cdot 0.00645 + 30 \cdot (-0.002655) = 0.830$$

Standardfeilen til estimatoren er

$$\sigma_{\widehat{\mu}_{Y\cdot x_1^*, x_2^*}} = \sqrt{V(\widehat{\mu}_{Y\cdot x_1^*, x_2^*})}$$

der variansen gitt ved

$$\begin{aligned} \mathrm{V}(\widehat{\mu}_{Y\cdot x_1^*, x_2^*}) &= \mathrm{V}(\widehat{\beta}_0 - 10\widehat{\beta}_1 + 30\widehat{\beta}_2) \\ &= \mathrm{V}(\widehat{\beta}_0) + (-10)^2 \cdot \mathrm{V}(\widehat{\beta}_1) + 30^2 \cdot \mathrm{V}(\widehat{\beta}_2) \\ &\quad + 2 \cdot (-10) \cdot \mathrm{Cov}(\widehat{\beta}_0, \widehat{\beta}_1) + 2 \cdot 30 \cdot \mathrm{Cov}(\widehat{\beta}_0, \widehat{\beta}_2) + 2 \cdot (-10) \cdot 30 \cdot \mathrm{Cov}(\widehat{\beta}_1, \widehat{\beta}_2) \end{aligned} \tag{A}$$

For å bestemme et estimat for variansen (og dermed standardfeilen), trenger vi estimater for variansene og kovariansene i formel (A). Vi kan få estimater for variansene fra `R`-utskriften, men ikke estimater for kovariansene.

d) Vi setter $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}$. Det er gitt i oppgaven at

$$\mathbf{C} = \begin{pmatrix} c_{00} & c_{01} & c_{02} \\ c_{10} & c_{11} & c_{12} \\ c_{20} & c_{21} & c_{22} \end{pmatrix} = \begin{pmatrix} 0.1460 & 0.0212 & 0.00313 \\ 0.0212 & 0.00483 & 0.000612 \\ 0.00313 & 0.000612 & 0.000138 \end{pmatrix}$$

Vi har at $\mathrm{V}(\widehat{\beta}_j) = \sigma^2 c_{jj}$ for $j = 0, 1, 2$ og $\mathrm{Cov}(\widehat{\beta}_j, \widehat{\beta}_\ell) = \sigma^2 c_{j\ell}$ for $j \neq \ell$.

Av `R`-utskriften har vi at et estimat for σ^2 er $s^2 = 0.0521^2$.

Ved formel (A) har vi da at et estimat for variansen til $\widehat{\mu}_{Y\cdot x_1^*, x_2^*}$ er

$$\begin{aligned} \widehat{\mathrm{V}}(\widehat{\mu}_{Y\cdot x_1^*, x_2^*}) &= \widehat{\mathrm{V}}(\widehat{\beta}_0) + (-10)^2 \cdot \widehat{\mathrm{V}}(\widehat{\beta}_1) + 30^2 \cdot \widehat{\mathrm{V}}(\widehat{\beta}_2) \\ &\quad + 2 \cdot (-10) \widehat{\mathrm{Cov}}(\widehat{\beta}_0, \widehat{\beta}_1) + 2 \cdot 30 \cdot \widehat{\mathrm{Cov}}(\widehat{\beta}_0, \widehat{\beta}_2) + 2 \cdot (-10) \cdot 30 \cdot \widehat{\mathrm{Cov}}(\widehat{\beta}_1, \widehat{\beta}_2) \\ &= s^2 \left\{ c_{00} + 100 \cdot c_{11} + 900 \cdot c_{22} - 20 \cdot c_{01} + 60 \cdot c_{02} - 600 \cdot c_{12} \right\} \\ &= 0.0521^2 \left\{ 0.1460 + 100 \cdot 0.00483 + 900 \cdot 0.000138 - 20 \cdot 0.0212 + 60 \cdot 0.00313 - 600 \cdot 0.000612 \right\} \\ &= 0.0004066 \end{aligned}$$

Et estimat for standardfeilen er dermed

$$s_{\widehat{\mu}_{Y\cdot x_1^*, x_2^*}} = \sqrt{\widehat{V}(\widehat{\mu}_{Y\cdot x_1^*, x_2^*})} = \sqrt{0.0004066} = 0.0202$$

Et et 95% konfidensintervall for $\mu_{Y \cdot x_1^*, x_2^*}$ er gitt ved

$$\widehat{\mu}_{Y \cdot x_1^*, x_2^*} \pm t_{0.025,17} \cdot s_{\widehat{\mu}_{Y \cdot x_1^*, x_2^*}}$$

Av tabellen først i oppgavesettet har vi at $t_{0.025,17} = 2.110$. Vi får dermed konfidensintervallet

$$0.830 \pm 2.110 \cdot 0.0202$$

dvs. fra 0.787 ppm til 0.873 ppm.