Eksamen i STK1110 høsten 2022 - løsningsforslag

Oppgave 1

a

Vi vet at $\bar{X} = 1/15 \sum_{i=1}^{15} X_i \sim N(\mu, \sigma^2/15)$, slik at

$$\frac{\bar{X} - \mu}{S/\sqrt{15}} \sim t_{15-1},$$

der $S = \sqrt{1/(15-1)\sum_{i=1}^{15}(X_i - \bar{X})^2}$ er det empiriske standardavviket. Vi får

$$\begin{aligned} &\mathrm{P}\left(-t_{0.025,14} \leq \frac{\bar{X} - \mu}{S/\sqrt{14}} \leq t_{0.025,14}\right) = 0.95 \\ \rightarrow &\mathrm{P}\left(\bar{X} - t_{0.025,14}\frac{S}{\sqrt{15}} \leq \mu \leq \bar{X} + t_{0.025,14}\frac{S}{\sqrt{15}}\right) = 0.95. \end{aligned}$$

Et 95% konfidensintervall for μ er dermed gitt ved

$$\bar{x} \pm t_{0.025,14}\frac{s}{\sqrt{15}}.$$

Vi setter inn og får $(6.1, 7.2)$.

b

Vi bruker

$$T = \frac{\bar{X} - 6.2}{S/\sqrt{15}}.$$

Testobservatoren T er da t_{15-1}-fordelt dersom $\mu = 6.2$. Da den alternative hypotesen er at $\mu \neq 6.2$, er det naturlig å forkaste H_0 dersom $T \geq c$ eller $T \leq -c$, der c er slik at signifikansnivået på testen er 5%. Med $c = t_{0.025,14}$ får vi signifikansnivået

$$\begin{aligned} &\mathrm{P}(\text{Forkaste } H_0 \mid H_0 \text{ er sann}) \\ =&\mathrm{P}\left(\frac{\bar{X} - 6.2}{S/\sqrt{15}} \leq -t_{0.025,14} \cup \frac{\bar{X} - 6.2}{S/\sqrt{15}} \geq t_{0.025,14} \;\middle|\; \mu = 6.2\right) \\ =&1 - \mathrm{P}\left(-t_{0.025,14} \leq \frac{\bar{X} - 6.2}{S/\sqrt{15}} \leq t_{0.025,14} \;\middle|\; \mu = 6.2\right) \\ =&1 - \left(\mathrm{P}\left(\frac{\bar{X} - 6.2}{S/\sqrt{15}} \leq t_{0.025,14} \;\middle|\; \mu = 6.2\right) - \mathrm{P}\left(\frac{\bar{X} - 6.2}{S/\sqrt{15}} \leq -t_{0.025,14} \;\middle|\; \mu = 6.2\right)\right) \\ =&1 - (0.975 - 0.025) = 0.05, \end{aligned}$$

som ønsket. Vi setter inn og får $t_{obs} = \frac{6.63-6.2}{0.967/\sqrt{15}} = 1.72$. Da
$-2.145 = -t_{0.025,14} \leq t_{obs} \leq t_{0.025,14} = 2.145$, kan vi ikke forkaste H_0 ved 5% signifikansnivå. I Oppgave a) fikk vi et 95% konfidensintervall som inneholder 6.2, noe som er i overensstemmelse med resultatene vi får her.

c

Generelt er P-verdien lik sannsynligheten, beregnet under forutsetning av 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 $\mu = 6.2$, for at vi vil få en verdi av testobservatorene T som er større enn eller lik $|t_{obs}| = 1.72$ eller mindre enn eller lik $-|t_{obs}| = -1.72$, dvs

$$\begin{aligned}\mathrm{P}(-|t_{obs}| \leq T \ \cup \ T \geq |t_{obs}| \mid \mu = 6.2) =& 1 - \mathrm{P}(-|t_{obs}| \leq T \leq |t_{obs}| \mid \mu = 6.2) \\ =& 2 \cdot (1 - \mathrm{P}(T \leq |t_{obs}| \mid \mu = 6.2)) \,.\end{aligned}$$

Fra tabellen over kritiske verdier for t-fordelingen ser vi at

$$t_{0.10,14} \leq t_{obs} \leq t_{0.05,14},$$

hvilket betyr at

$$0.20 \leq \mathrm{P}(-|t_{obs}| \leq T \ \cup \ T \geq |t_{obs}| \mid \mu = 6.2) \leq 0.10,$$

altså at P-verdien ligger mellom 0.20 og 0.10, og er større enn signifikansnivået 5% fra b).

Oppgave 2

a

Likelihood-funksjonen gitt ved

$$L(\beta) \overset{uif}{=} \prod_{i=1}^{n} f(x_i;\beta) = \prod_{i=1}^{n} f(x_i;\beta) = \prod_{i=1}^{n} \frac{1}{6\beta^4} x_i^3 e^{-\frac{x_i}{\beta}} = \frac{1}{6^n \beta^{4n}} e^{-\frac{1}{\beta}\sum_{i=1}^{n} x_i} \prod_{i=1}^{n} x_i^3.$$

Det gir log-likelihood-funksjonen

$$\log L(\beta) = -n\log(6) - 4n\log(\beta) + 3\sum_{i=1}^{n} \log(x_i) - \frac{1}{\beta}\sum_{i=1}^{n} x_i.$$

Vi får maksimum likelihood-estimatet $\hat{\beta}$ ved å løse ligningen $\frac{\partial \log L(\beta)}{\partial \beta} = 0$. Altså er $\hat{\beta}$ løsningen av ligningen

$$-\frac{4n}{\hat{\beta}} + \frac{1}{\hat{\beta}^2}\sum_{i=1}^{n} x_i = 0,$$

som gir maksimum likelihood-estimatet

$$\hat{\beta}_{ML} = \frac{1}{4n}\sum_{i=1}^{n} x_i = \frac{\bar{x}}{4},$$

med $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, og tilsvarende maksimum likelihood-estimator

$$\hat{\beta}_{ML} = \frac{\bar{X}}{4}.$$

b

Fisher informasjonen i én observasjon er gitt ved

$$I(\beta) = -\mathrm{E}\left(\frac{\partial^2 \log f(X_i;\beta)}{\partial \beta^2}\right).$$

Vi har

$$\log(f(X_i;\beta)) = -\log(6) - 4\log(\beta) + 3\log(X_i) - \frac{1}{\beta}X_i,$$

slik at

$$\frac{\partial \log f(X_i;\beta)}{\partial \beta} = -\frac{4}{\beta} + \frac{1}{\beta^2}X_i$$

og

$$\frac{\partial^2 \log f(X_i;\beta)}{\partial \beta^2} = \frac{4}{\beta^2} - \frac{2}{\beta^3}X_i.$$

Vi får dermed

$$\begin{aligned} I(\beta) &= -\mathrm{E}\left(\frac{4}{\beta^2} - \frac{2}{\beta^3}X_i\right) \\ &= -\frac{4}{\beta^2} + \frac{2}{\beta^3}\mathrm{E}(X_i) \\ &= -\frac{4}{\beta^2} + \frac{2}{\beta^3}\cdot 4\beta = \frac{4}{\beta^2}. \end{aligned}$$

c

Det er kjent at maksimum likelihood-estimatoren $\hat{\beta}_{ML}$ (under visse regularitetsbetingelser som er oppfylt her) er tilnærmet normalfordelt $N(\beta, \sigma^2_{\hat{\beta}})$ med $\sigma^2_{\hat{\beta}} = 1/(nI(\beta))$. Her får vi

$$\sigma^2_{\hat{\beta}} = 1/(nI(\beta)) = \frac{\beta^2}{4n}.$$

Her kan en alternativt bruke sentralgrenseteoremet, som sier at for store n er $\bar{X} = 4\hat{\beta}_{ML}$ tilnærmet normalfordelt med forventning $\mathrm{E}(\bar{X}) = \mathrm{E}(X_i) = 4\beta$ og varians $\mathrm{V}(\bar{X}) = \frac{1}{n}\mathrm{V}(X_i) = \frac{4\beta^2}{n}$, slik at $\hat{\beta}_{ML}$ er tilnærmet normalfordelt med forventning $\frac{1}{4}\mathrm{E}(\bar{X}) = \beta$ og varians $\sigma^2_{\hat{\beta}} = \frac{1}{4^2}\mathrm{V}(\bar{X}) = \frac{\beta^2}{4n}$.

d

Grunnet invariansprinsippet for maksimum likelihood-estimatoren er maksimum likelihood-estimatoren for $\psi = g(\beta) = \frac{1}{\beta}$

$$\hat{\psi}_{ML} = g(\hat{\beta}_{ML}) = 4/\bar{X}.$$

e

En apriorifordeling er konjugert med fordelingen til dataene dersom aposteriorifordelingen tilhører samme fordelingsfamilie som apriorifordelingen. Vi har:

$$\begin{aligned}
\pi(\psi) =& \frac{1}{\beta_0^{\alpha_0}\Gamma(\alpha_0)}\psi^{\alpha_0-1}e^{-\frac{1}{\beta_0}\psi} \\
f(x_1,\ldots,x_n|\psi) \overset{uif}{=} & \prod_{i=1}^{n} f(x_i;\psi) = \frac{\psi^{4n}}{6^n}e^{-\psi\sum_{i=1}^{n}x_i}\prod_{i=1}^{n}x_i^3.
\end{aligned}$$

Vi vet at tettheten til aposteriorifordelingen til ψ er gitt ved

$$\begin{aligned}
\pi(\psi|x_1,\ldots,x_n) =& k\cdot\pi(\psi)\cdot f(x_1,\ldots,x_n|\psi) \\
=& k\cdot\frac{1}{\beta_0^{\alpha_0}\Gamma(\alpha_0)}\psi^{\alpha_0-1}e^{-\frac{1}{\beta_0}\psi}\cdot\frac{\psi^{4n}}{6^n}e^{-\psi\sum_{i=1}^{n}x_i}\prod_{i=1}^{n}x_i^3 \\
\propto& \psi^{\alpha_0+4n-1}e^{-\left(\frac{1}{\beta_0}+\sum_{i=1}^{n}x_i\right)\psi},
\end{aligned}$$

som er proporsjonal med tettheten til $Gamma\left(\alpha_0+4n, \frac{1}{\frac{1}{\beta_0}+\sum_{i=1}^{n}x_i}\right)$. Altså må $[\psi|X_1=x_1,\ldots,X_n=x_n] \sim Gamma\left(\alpha_0+4n, \frac{1}{\frac{1}{\beta_0}+\sum_{i=1}^{n}x_i}\right)$, som i likhet

med aprioriforde lingen er en gammafordelingen. Det betyr at aprioriforde lingen er konjugert med fordelingen til dataene.

Bayes-estimatoren for ψ er gitt ved

$$\begin{aligned}\hat{\psi}_{Bayes} = \mathrm{E}(\psi|X_1,\ldots,X_n) =&(\alpha_0+4n)\cdot\frac{1}{\frac{1}{\beta_0}+\sum_{i=1}^n X_i}\\ =&\frac{\alpha_0+4n}{\frac{1}{\beta_0}+\sum_{i=1}^n X_i} = \frac{\alpha_0/n+4}{\frac{1}{n\beta_0}+\bar{X}}.\end{aligned}$$

Vi ser at når n blir stor, blir Bayes-estimatoren $\hat{\psi}_{Bayes}$ essensielt den samme som maksimum likelihood-estimatoren $\hat{\psi}_{ML}$.

Oppgave 3

a

Konstantleddet β_0 i den enkle lineære regresjonsmodellen er forventet import når $x_{i1}-\bar{x}_1 = 0$, dvs. når forbruket er lik gjennomsnittet (som er 167 millioner FRF). Fra R-utskriften ser vi at $\hat{\beta}_0 = 30.1$, så når forbruket er gjennomsnittlig, vil en forvente at importen er på 30.1 millioner FRF. Stigningstallet β_1 er forventet endring i importen når forbruket øker med 1 million FRF. Her er $\hat{\beta}_1 = 0.296$, så hvis forbruket øker med 1 million FRF, vil en forvente at billettsalget øker med 0.296 millioner FRF.
Videre vet vi at $(\hat{\beta}_1 - \beta_1)/S_{\hat{\beta}_1} \sim t_{18-2}$, der $S^2_{\hat{\beta}_1}$ er den forventningsrette estimatoren for $\mathrm{Var}(\hat{\beta}_1)$. Et 95% konfidensintervall for β_1 er dermed gitt ved

$$\hat{\beta}_1 \pm t_{0.025,16} s_{\hat{\beta}_1}.$$

Når vi setter inn tall fra R-utskriften, får vi

$$0.296 \pm 2.120 \cdot 0.0131 = (0.268, 0.323).$$

Vi ser at verdien 0.25 er utenfor konfidensintervallet, dog i nærheten av nedre grense. Det betyr at om vi hadde testet $H_0 : \beta_1 = 0.25$ mot $H_a : \beta_1 \neq 0.25$, altså om forventet økning i importen øker med omtrent 0.25 millioner FRF når forbruket øker med 1 million FRF, ville vi forkaste H_0 ved 5% signifikansnivå.

b

Antakelsene vi gjør i modell (1) er

1. Sammenhengen er lineær: $\mathrm{E}(Y|x) = \beta_0 + \beta_1(x_1 - \bar{x}_1)$.
2. Feilleddene ϵ_i er normalfordelt.
3. Variansen er lik for alle verdier av x: $\mathrm{V}(\epsilon_i) = \sigma^2$.
4. Feilleddene ϵ_i er uavhengige.

Den første antakelsen kan vi sjekke ved å se på plottet av residualene $e_i = Y_i - \hat{Y}_i$ mot de tilpassede verdiene $\hat{Y}_i$. Dersom sammenhengen er lineær, skal det ikke være noe mønster i dette plottet. Den andre antakelsen kan vi sjekke ved å se på normalfordelingsplottet for de standardiserte residualene. Dersom punktene i plottet ligger omtrent langs diagonalen, kan en gå ut fra at antakelsen om normalfordeling er rimelig. Den tredje antakelsen sjekker en ved å se på plottet av de standardiserte residualene $e_i^* = e_i/s_{e_i}$ mot $\hat{Y}_i$. Dersom antakelsen om lik varians holder, skal det ikke være noe mønster i dette plottet.

I residualplottene for modell (1), ser vi fra normalfordelingsplottet av de standardiserte residualene at antakelsen om normalfordeling ser grei ut. Imidlertid har de to andre plottene et tydelig parabelmønster, som kan tyde på at det er en ikke-lineær, og mer spesifikt en kvadratisk effekt av forbruket på importen. Det medfører også at antakelsen om konstant varians blir brutt.

c
Anta at forbruket er x_1^*, slik at $x_2 = (x_1^* - \bar{x}_1)^2$. Forventet import er nå gitt ved:

$$\begin{aligned}\mathrm{E}(Y|x_1^*) =& \mathrm{E}(\gamma_0 + \gamma_1(x_1^* - \bar{x}_1) + \gamma_2(x_1^* - \bar{x}_1)^2 + \varepsilon) \\ =& \gamma_0 + \gamma_1(x_1^* - \bar{x}_1) + \gamma_2(x_1^* - \bar{x}_1)^2 + \underbrace{\mathrm{E}(\varepsilon)}_{=0} \\ =& \gamma_0 + \gamma_1(x_1^* - \bar{x}_1) + \gamma_2(x_1^* - \bar{x}_1)^2,\end{aligned}$$

som er en kvadratisk funksjon av forbruket x_1^*. Vi får

$$\begin{aligned}&\mathrm{E}(Y|x_1^* + 1) - \mathrm{E}(Y|x_1^*) \\ &= \gamma_0 + \gamma_1(x_1^* + 1 - \bar{x}_1) + \gamma_2(x_1^* + 1 - \bar{x}_1)^2 - (\gamma_0 + \gamma_1(x_1^* - \bar{x}_1) + \gamma_2(x_1^* - \bar{x}_1)^2) \\ &= \gamma_1 + \gamma_2(x_1^* - \bar{x}_1)^2 + 2\gamma_2(x_1^* - \bar{x}_1) + \gamma_2 - \gamma_2(x_1^* - \bar{x}_1)^2 = \gamma_1 + \gamma_2(1 + 2(x_1^* - \bar{x}_1)).\end{aligned}$$

Altså øker forventet import med $\gamma_1 + \gamma_2(1 + 2(x_1^* - \bar{x}_1))$ når forbruket øker fra x_1^* til $x_1^* + 1$.

Vi har fortsatt $\mathrm{E}(Y|\bar{x}_1) = \gamma_0$, slik at konstantleddet γ_0 fortolkes slik som β_0 i modell (1), altså som forventet import når forbruket er gjennomsnittlig. Fortolkningen av γ_1 er imidlertid forskjellig fra fortolkningen av β_1, da forventet import øker med $\gamma_1 + \gamma_2(1 + 2(x_1^* - \bar{x}_1))$ når forbruket øker med én enhet fra nivået x_1^*. Koeffisienten γ_1 angir altså bare en del av effekten av forbruket på importen, og denne effekten avhenger forøvrig av nivået på forbruket.

d

Hypotesetesten dreier seg om modell (2), som modellerer effekten av forbruk på importen som kvadratisk, er bedre enn den enkle, første modellen med en lineær effekt, dvs.

$$H_0 : \ \beta_2 = 0 \text{ mot } H_a : \ \beta_2 \neq 0.$$

Vi ser at den tilsvarende P-verdien er på bare $1.53 \cdot 10^{-6}$, hvilket betyr at vi forkaster H_0 med god margin ved 5% signifikansnivå. Det betyr at det ekstra kvadratiske leddet i modell (2) bidrar signifikant til å forklare importen, hvilket vi også ser av den ekvivalente t-testen for om $\beta_2 \neq 0$ i R-utskriften fra tilpasningen av modell (2). Dermed velger vi modell (2) med kvadratisk effekt av forbruket på importen. Vi legger også merke til at justert R^2 har økt fra 0.9679 i tilpasningen til modell (1) til 0.9930 i tilpasningen til modell (2).

Vi bruker så residualplottene til å sjekke modellantakelsen, som her er

1. Sammenhengen er kvadratisk: $\mathrm{E}(Y|x) = \gamma_0 + \gamma_1(x_1 - \bar{x}_1) + \gamma_2(x_1 - \bar{x}_1)^2$.
2. Feilleddene ε_i er normalfordelt.
3. Variansen er lik for alle verdier av x_1: $\mathrm{V}(\varepsilon_i) = \tau^2$.
4. Feilleddene ε_i er uavhengige.

Vi ser av normalfordelingsplottet av de standardiserte residualene at antakelsen om normalfordeling fortsatt ser grei ut. Nå er det imidlertid ikke noe åpenbart mønster i plottene av residualene og av de standardiserte residualene mot de tilpassede verdiene. Dette tyder på at hele effekten av forbruket på importen er fanget opp med en kvadratisk effekt, slik at antakelsen om

konstant varians nå ser rimelig ut. Dette gir en bekreftelse på at en bør velge modell (2) framfor modell (1).