Eksamen i STK1110 høsten 2020 - løsningsforslag

Oppgave 1

a

Vi vet at $\bar{X} = 1/n \sum_{i=1}^{28} X_i \sim N(\mu, \sigma^2/n)$, slik at

$$\frac{\bar{X} - \mu}{S/\sqrt{28}} \sim t_{28-1},$$

der $S = \sqrt{1/(28-1)\sum_{i=1}^{28}(X_i - \bar{X})^2}$ er det empiriske standardavviket. Vi får

$$\begin{aligned} &\mathrm{P}\left(-t_{0.005,27} \leq \frac{\bar{X} - \mu}{S/\sqrt{28}} \leq t_{0.005,27}\right) = 0.99 \\ \rightarrow &\mathrm{P}\left(\bar{X} - t_{0.005,27}\frac{S}{\sqrt{28}} \leq \mu \leq \bar{X} + t_{0.005,27}\frac{S}{\sqrt{28}}\right) = 0.99. \end{aligned}$$

Et 95% konfidensintervall for μ er dermed gitt ved

$$\bar{x} \pm t_{0.005,27}\frac{s}{\sqrt{28}}.$$

Vi setter inn og får $(87.6, 107.4)$.

b

Vi bruker

$$T = \frac{\bar{X} - 115}{S/\sqrt{28}},$$

og da den alternative hypotesen er at $\mu < 115$, er det naturlig å forkaste H_0 dersom $T \leq c$, der c er slik at signifikansnivået på testen er 1%. Med $c = -t_{0.01,27}$ får vi

$$\begin{aligned} \mathrm{P}(\text{Forkaste } H_0 \mid H_0 \text{ er sann}) =& \mathrm{P}\left(\frac{\bar{X} - 115}{S/\sqrt{28}} \leq -t_{0.01,27} \;\middle|\; \mu \geq 115\right) \\ \leq& \mathrm{P}\left(\frac{\bar{X} - 115}{S/\sqrt{28}} \leq -t_{0.01,27} \;\middle|\; \mu = 115\right) \\ =& 0.01, \end{aligned}$$

som ønsket. Vi setter inn og får $t_{obs} = \frac{97.52-115}{18.95/\sqrt{28}} = -4.88$. Da $t_{obs} \leq -t_{0.01,27} = -2.473$, forkaster vi H_0 ved 1% signifikansnivå og konkluderer med at eplene

er mindre enn vanlig størrelse. Det er ikke så overraskende da vi fra a) vet at μ med stor sikkerhet ligger mellom 87.6 og 107.4 gram, som ligger et godt stykke nedenfor 115.

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 $\mu = 115$, for at vi vil få en verdi av testobsetvatorene T som er lik $t_{obs} = -4.88$ eller mindre, dvs

$$\mathrm{P}(T \leq t_{obs} \mid \mu = 115).$$

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

$$-t_{0.00001,27} \leq t_{obs} \leq -t_{0.0001,27},$$

hvilket betyr at

$$0.00001 \leq \mathrm{P}(T \leq t_{obs} \mid \mu = 115) \leq 0.0001,$$

altså at P-verdien ligger mellom 0.00001 og 0.0001, og er betydelig mindre enn signifikansnivået 1% fra b).

Oppgave 2

a

Vi finner momentestimatoren $\tilde{\theta}$ ved å løse ligningen vi får når vi setter første teoretiske moment $\mathrm{E}(X_i)$ lik første empiriske moment $\bar{X} = 1/n \sum_{i=1}^{n} X_i$. Vi får:

$$\frac{1}{1-\tilde{\theta}} = \bar{X}.$$

Det gir

$$\tilde{\theta} = \frac{\bar{X} - 1}{\bar{X}}.$$

b

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

$$L(\theta) \stackrel{uif}{=} \prod_{i=1}^{n} f(x_i; \theta) = \prod_{i=1}^{n} \frac{e^{-\theta x_i}(\theta x_i)^{x_i - 1}}{x_i!} = e^{-\theta \sum_{i=1}^{n} x_i} \theta^{\sum_{i=1}^{n} x_i - n} \prod_{i=1}^{n} \frac{x_i^{x_i - 1}}{x_i!}.$$

Det gir log-likelihood-funksjonen

$$\log L(\theta) = -n\theta\bar{x} + n(\bar{x} - 1)\log(\theta) + \sum_{i=1}^{n}(x_i - 1)\log(x_i) - \sum_{i=1}^{n}\log(x_i!).$$

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

$$-n\bar{x} + n\frac{\bar{x} - 1}{\hat{\theta}} = 0,$$

som gir maksimum likelihood-estimatet

$$\hat{\theta} = \frac{\bar{x} - 1}{\bar{x}},$$

og tilsvarende maksimum likelihood-estimator

$$\hat{\theta} = \frac{\bar{X} - 1}{\bar{X}}.$$

Vi ser at maksimum likelihood-estimatoren er den samme som momentestimatoren.

c

Fisher informasjonen i én observasjon er gitt ved

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

Vi har

$$\log(f(X_i;\theta)) = -\theta X_i + (X_i - 1)\log(\theta) + (X_i - 1)\log(X_i) - \log(X_i!),$$

slik at

$$\frac{\partial \log f(X_i;\theta)}{\partial \theta} = -X_i + \frac{X_i - 1}{\theta}$$

og

$$\frac{\partial^2 \log f(X_i;\theta)}{\partial \theta^2} = -\frac{X_i - 1}{\theta^2}.$$

Vi får dermed

$$\begin{aligned} I(\theta) =& -\operatorname{E}\left(-\frac{X_i - 1}{\theta^2}\right) \\ =& \frac{\operatorname{E}(X_i) - 1}{\theta^2} \\ =& \frac{1}{\theta^2}\left(\frac{1}{1-\theta} - 1\right) = \frac{1}{\theta(1-\theta)}. \end{aligned}$$

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

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

Oppgave 3

a
Konstantleddet β_0 i den enkle lineære regresjonsmodellen er forventet billettsalg når $x_{i1} - \bar{x}_1 = 0$, dvs. når produksjonskostnadene er lik gjennomsnittet (som er 8.74 millioner USD). Fra R-utskriften ser vi at $\hat{\beta}_0 = 85.2$, så når produksjonskostnadene er gjennomsnittlige, vil en forvente at billettsalget er på 85.2 millioner USD. Stigningstallet β_1 er forventet endring i billettsalget når produksjonskostnadene øker med 1 million USD. Her er $\hat{\beta}_1 = 7.98$, så hvis produksjonskostnadene øker med 1 million USD, vil en forvente at billettsalget øker med 7.98 millioner USD.
Videre vet vi at $(\hat{\beta}_1 - \beta_1)/S_{\hat{\beta}_1} \sim t_{10-2}$, der $S^2_{\hat{\beta}_1}$ er den forventningsrette estimatoren for $\operatorname{Var}(\hat{\beta}_1)$. Et 95% kondidensintervall for β_1 er dermed gitt ved

$$\hat{\beta}_1 \pm t_{0.025,8} s_{\hat{\beta}_1}.$$

Når vi setter tall fra R-utskriften, får vi

$$7.98 \pm 2.306 \cdot 1.223 = (5.15, 10.80).$$

Vi ser at verdien 10 er innenfor konfidensintervallet, dog i nærheten av øvre grense. Det betyr at om vi hadde testet $H_0 : \beta_1 = 10$ mot $H_a : \beta_1 \neq 10$, altså

om forventet økning i bilettsalget per ekstra million USD investert i filmen er 10 millioner USD, ville vi ikke kunne forkaste H_0 ved 5% signifikansnivå.

b

Det første plottet viser residualene $e_i = Y_i - \hat{Y}_i$ mot $\hat{Y}_i$. Det er ikke noe åpenbart mønster i plottet. Antakelsen om at sammenhengen mellom billettsalg og produksjonskostnader er lineær ser altså ut til å være grei.

Det andre plottet er et normalfordelingsplott av de standardiserte residualene e_i^*. Når disse punktene ikke ligger nær en rett linje, har vi en indikasjon på at feilleddene ikke er normalfordelt. Punktene i plottet ligge omtrent langs en rett linje, med unntak av de største verdiene. Det kan tyde på at antakelsen om at feilleddene er normalfordelt ikke stemmer helt, og at de har en fordeling som er mer tunghalet.

Det tredje plottet viser de standardiserte residualene $e_i^* = e_i/s_{e_i}$ mot $\hat{Y}_i$, og er konstruert for undersøke om variansen til ϵ_i i avhenger av $\mathrm{E}(Y_i)$, og altså varierer. Det er ikke noe åpenbart mønster i dette plottet heller, så det ser ikke ut til å være noe arvorlig avvik fra antakelsen om konstant varians.

c

Da $\epsilon_1, \ldots, \epsilon_{10} \overset{uif}{\sim} N(0, \sigma^2)$, er også $Y_i$ene uavhengige og normalfordelt med varians σ^2. Da $\hat{Y}$ er en lineærkombinasjon av Y_i-ene, må denne også være normafordelt. Vi får

$$\begin{aligned}
\sigma^2_{\hat{Y}} = \mathrm{Var}(\hat{Y}) &= \mathrm{Var}\left(\sum_{i=1}^{10}\left(\frac{1}{10} + \frac{(x_1^* - \bar{x}_1)(x_{i1} - \bar{x}_1)}{S_{xx}}\right) Y_i\right) \\
&\overset{uavh.}{=} \left(\sum_{i=1}^{10}\frac{1}{n^2} + \frac{2}{10}\frac{(x_1^* - \bar{x}_1)}{S_{xx}}\sum_{i=1}^{10}(x_{i1} - \bar{x}_1) + \frac{(x_1^* - \bar{x}_1)^2}{S_{xx}^2}\sum_{i=1}^{10}(x_{i1} - \bar{x}_1)^2\right)\mathrm{Var}(Y_i) \\
&= \sigma^2\left(\frac{1}{10} + \frac{(x_1^* - \bar{x}_1)^2}{S_{xx}}\right).
\end{aligned}$$

Til slutt vet vi at $\hat{\beta}_0$ og $\hat{\beta}_1$ er forventningsrette for β_0 og β_1, slik at

$$\mathrm{E}(\hat{Y}) = \mathrm{E}(\hat{\beta}_0 + \hat{\beta}_1(x_1^* - \bar{x})) = \mathrm{E}(\hat{\beta}_0) + \mathrm{E}(\hat{\beta}_1)(x_1^* - \bar{x}) = \beta_0 + \beta_1(x_1^* - \bar{x}) = \mu_{Y|x^*}.$$

Altså er $\hat{Y} \sim N(\mu_{Y|x^*}, \sigma^2_{\hat{Y}})$.

d

La $S^2_{\hat{Y}} = S^2\left(\frac{1}{10} + \frac{(x_1^* - \bar{x}_1)^2}{S_{xx}}\right)$. Da er $S_{\hat{Y}}/\sigma^2_{\hat{Y}} = S^2/\sigma^2$. Vi får

$$T = \frac{\hat{Y} - \mu_{Y|x^*}}{S_{\hat{Y}}} = \frac{(\hat{Y} - \mu_{Y|x^*})/\sigma_{\hat{Y}}}{S_{\hat{Y}}/\sigma_{\hat{Y}}} = \frac{(\hat{Y} - \mu_{Y|x^*})/\sigma_{\hat{Y}}}{\sqrt{((10-2)S^2/\sigma^2)/(10-2)}} = \frac{Z}{\sqrt{U/8}}$$

der $Z \sim N(0,1)$ og $U = (10-2)S^2/\sigma^2 \sim \chi^2_8$. Dermed er $T \sim t_8$. Vi får

$$\begin{aligned} &\mathrm{P}\left(-t_{\alpha/2,8} \leq \frac{\hat{Y} - \mu_{Y|x^*}}{S_{\hat{Y}}} \leq t_{\alpha/2,8}\right) = 1-\alpha \\ \rightarrow &\mathrm{P}\left(\hat{Y} - t_{\alpha/2,8}S_{\hat{Y}} \leq \mu_{Y|x^*} \leq \hat{Y} + t_{\alpha/2,8}S_{\hat{Y}}\right) = 1-\alpha. \end{aligned}$$

Et $100 \cdot (1-\alpha)\%$ konfidensintervall for $\mu_{Y|x^*}$ er dermed gitt ved

$$\hat{y} \pm t_{\alpha/2,8}s_{\hat{Y}},$$

der $\hat{y}$ og $s_{\hat{Y}}$ er observerte verdier av $\hat{Y}$ og $S_{\hat{Y}}$.
For $x_1^* = 301$ får vi

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1(x_1^* - \bar{x}) = 85.24 + 7.98 \cdot (301 - 8.74) = 2417.$$

Altså er forventet billettsalg utrolige 2.49 milliarder USD i henhold til modellen. Tilhørende 95% konfidensintervall er gitt ved, der vi får $s = 14.26$ fra R-utskriften,

$$\hat{y} \pm t_{0.025,8}s_{\hat{Y}} = 2417 \pm 2.306 \cdot 14.26\sqrt{\frac{1}{10} + \frac{(301-8.74)^2}{135.864}} = 2417 \pm 824,$$

altså $(1592, 3241)$. Dette er et usedvanlig bredt konfidensintervall, hvilket betyr at det er knyttet svært stor usikkerhet til anslaget $\hat{y}$ for forventet billettsalg.

Verdien $x^* = 301$ ligger langt bortenfor alle observerte verdier x_i. En bør være varsom med å ekstrapolere modellen så langt forbi området med data, og en bør dermed ta estimatet med en klype salt.

e

Konstantleddet β_0 i den multiple lineære regresjonsmodellen er forventet billettsalg når både $x_{i1} - \bar{x}_1 = 0$ og $x_{i2} = 0$, dvs. når produksjonskostnadene er lik gjennomsnittet (som er 8.74 millioner USD) og en ikke har brukt penger på markedsføring. Fra R-utskriften ser vi at $\hat{\beta}_0 = 48.8$, så når produksjonskostnadene er gjennomsnittlige og en bruker penger på markedsføring, vil en forvente at billettsalget er på 48.8 millioner USD. Nå er β_1 er forventet endring i billettsalget når produksjonskostnadene øker med 1 million USD **og** markedsføringskostnadene holdes konstant. Vi ser at $\hat{\beta}_1 = 4.23$, så hvis produksjonskostnadene øker med 1 million USD mens markedsføringskostnadene er de samme, vil en forvente at billettsalget øker med 4.23 millioner

USD. Altså har $\hat{\beta}_1$ endret seg nokså mye fra den enkle lineære regresjonsmodellen, og det skyldes konfundering på grunn av nokså sterk (positiv) korrelasjon mellom produksjons- og markedsføringskostnadene, som vi ser av matrisespredningsplottet. Videre er β_2 forventet endring i billettsalget når markedsføringskostnadene øker med 1 million USD **og** produksjonskostnadene holdes konstant. Vi ser at $\hat{\beta}_2 = 7.44$, så hvis markedsføringskostnadene øker med 1 million USD mens produksjonskostnadene er de samme, vil en forvente at billettsalget øker med 7.44 millioner USD.
Hypotesetesten dreier seg om modellen som inkluderer markedsføringskostnader er bedre enn den enkle, første modellen uten, dvs.

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

Vi ser at den tilsvarende P-verdien er på bare 0.0045, hvilket betyr at vi forkaster H_0 med god margin ved 5% signifikansnivå. Det betyr at markedsføringskostnader bidrar signifikant til å forklare billetsalget, hvilket vi også ser av den ekvivalente t-testen for om $\beta_2 \neq 0$ i den øverste utskriften fra R. Dermed velger vi den nye modellen, som inneholder både produksjons- og markedsføringskostnader.