Norges teknisk–
naturvitenskapelige universitet
Institutt for matematiske fag

LØSNINGSFORSLAG TIL EKSAMEN I FAG TMA4240 STATISTIKK

Mandag 12. desember 2011

Oppgave 1 Oljeleting

a) Siden $P(A \cap B) = P(A) + P(B) - P(A \cup B) = 0$, så er hendelsene disjunkte.
Siden $P(A \cap B) = 0 \neq 0{,}3 \cdot 0{,}3 = P(A)P(B)$, så er hendelsene ikke uavhengige.

b) Hendelsene med sannsynligheter er tegnet inn i Venndiagrammet i Figur 1.

Sannsynligheten for olje på felt 1, dvs. hendelsen A, kan beregnes fra lov om total sannsynlighet,

$$P(A) = P(A \cap B) + P(A \cap B^{\mathcal{C}}) = 0{,}05 + 0{,}15 = 0{,}20.$$

Hendelsen olje på felt 1 gitt olje er funnet på felt 2 kan skrives $A|B$ og sannsynligheten beregnes fra Bayes lov,

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} = \frac{P(A \cap B)}{P(B \cap A) + P(B \cap A^{\mathcal{C}})} \\ &= \frac{0{,}05}{0{,}05 + 0{,}1} = \frac{1}{3}. \end{aligned}$$

Her er lov om total sannsynlighet brukt for å beregne $P(B)$.

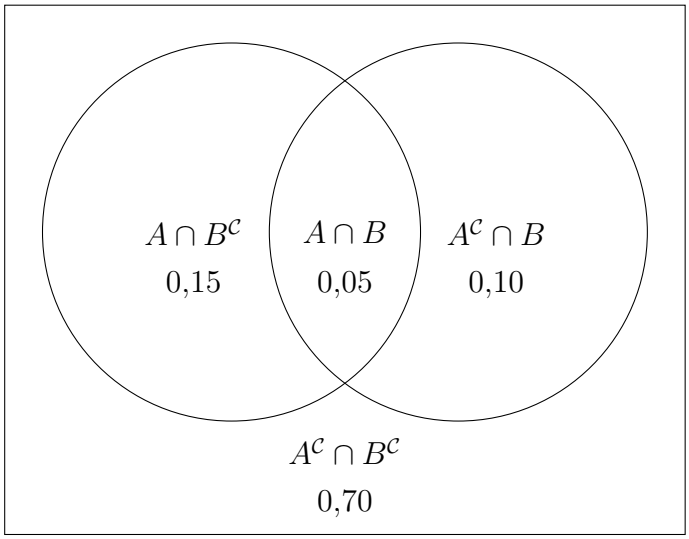
Sannsynligheten for hendelsen ikke olje på felt 2 kan sees fra ligningen over

$$P(B^{\mathcal{C}}) = 1 - P(B) = 1 - 0{,}15 = 0{,}85.$$

Dette kan settes inn i Bayes lov for å få sannsynlighet for olje på felt 1 gitt ingen olje funnet på felt 2,

$$\begin{aligned} P(A|B^{\mathcal{C}}) &= \frac{P(A \cap B^{\mathcal{C}})}{P(B^{\mathcal{C}})} \\ &= \frac{0{,}15}{0{,}85} \approx 0{,}176. \end{aligned}$$

Fra de tidligere beregningene ser vi at $P(A|B) \neq P(A|B^{\mathcal{C}})$, dermed er A og B ikke uavhengige.



Figur 1: Venndiagram med hendelsene $A \cap B$, $A \cap B^{\mathcal{C}}$, $A^{\mathcal{C}} \cap B$ og $A^{\mathcal{C}} \cap B^{\mathcal{C}}$ og tilhørende sannsynligheter påtegnet.

c) La F_A være den stokastiske variabelen «Fortjeneste fra felt 1». Da er

$$\mathrm{E}[F_A] = 400P(A) - 100P(A^{\mathcal{C}}) = 400 \cdot 0{,}20 - 100 \cdot 0{,}80 = 0.$$

La $F_A|B$ være den stokastiske variabelen «Fortjeneste fra felt 1 gitt funn på felt 2». Da er

$$\mathrm{E}[F_A|B] = 400P(A|B) - 100P(A^{\mathcal{C}}|B) = 400 \cdot \frac{1}{3} - 100\frac{2}{3} = \frac{200}{3} \approx 66{,}67.$$

Vi begynner med de følgende to observasjonene. Forventningsverdien for hver strategi er bestemt av fortjenesten ved hver av hendelsene $A \cap B$, $A \cap B^{\mathcal{C}}$, $A^{\mathcal{C}} \cap B$ og $A^{\mathcal{C}} \cap B^{\mathcal{C}}$, og den beste strategien er å ikke lete noe sted, å begynne å lete på felt 1 eller å begynne å lete på felt 2. Det er dermed tre muligheter vi må vurdere.

Første mulighet. Vi leter ingen steder, dette gir umiddelbart forventningsverdi 0.

Andre mulighet. Vi begynner med å lete på felt 1. Dette gir oss informasjon om felt 1 og vi kan velge mellom følgende strategier for felt 2. Ikke lete uansett om det er olje på felt 1 eller ikke, lete uansett om det er olje på felt 1 eller ikke, lete hvis og bare hvis vi finner olje på felt 1 eller lete hvis og bare hvis vi ikke finner olje på felt 1. Kall de stokastiske variablene som beskriver fortjenesten ved hver av disse strategiene for henholdsvis F_1,

F_2, F_3 og F_4. Da har vi

$$\begin{aligned}
\mathrm{E}[F_1] &= \mathrm{E}[F_A] = 0 \\
\mathrm{E}[F_2] &= 300P(A \cap B^{\mathcal{C}}) + 1400P(A \cap B) + 900P(A^{\mathcal{C}} \cap B) - 200P(A^{\mathcal{C}} \cap B^{\mathcal{C}}) = 65, \\
\mathrm{E}[F_3] &= 300P(A \cap B^{\mathcal{C}}) + 1400P(A \cap B) - 100P(A^{\mathcal{C}} \cap B) - 100P(A^{\mathcal{C}} \cap B^{\mathcal{C}}) = 35, \\
\mathrm{E}[F_4] &= 400P(A \cap B^{\mathcal{C}}) + 400P(A \cap B) + 900P(A^{\mathcal{C}} \cap B) - 200P(A^{\mathcal{C}} \cap B^{\mathcal{C}}) = 30.
\end{aligned}$$

Tredje mulighet. Vi begynner med å lete på felt 2. Dette gir oss informasjon om felt 2 og vi kan velge mellom følgende strategier for felt 1. Ikke lete uansett om det er olje på felt 2 eller ikke, lete uansett om det er olje på felt 2 eller ikke, lete hvis og bare hvis vi finner olje på felt 2 eller lete hvis og bare hvis vi ikke finner olje på felt 2. Kall de stokastiske variablene som beskriver fortjenesten ved hver av disse strategiene for henholdsvis F_5, F_6, F_7 og F_8. Da har vi

$$\begin{aligned}
\mathrm{E}[F_5] &= 1000P(B) - 100P(B^{\mathcal{C}}) = 65 \\
\mathrm{E}[F_6] &= 300P(A \cap B^{\mathcal{C}}) + 1400P(A \cap B) + 900P(A^{\mathcal{C}} \cap B) - 200P(A^{\mathcal{C}} \cap B^{\mathcal{C}}) = 65, \\
\mathrm{E}[F_7] &= -100P(A \cap B^{\mathcal{C}}) + 1400P(A \cap B) + 900P(A^{\mathcal{C}} \cap B) - 100P(A^{\mathcal{C}} \cap B^{\mathcal{C}}) = 75, \\
\mathrm{E}[F_8] &= 300P(A \cap B^{\mathcal{C}}) + 1000P(A \cap B) + 1000P(A^{\mathcal{C}} \cap B) - 200P(A^{\mathcal{C}} \cap B^{\mathcal{C}}) = 55.
\end{aligned}$$

Utregningene for hver av mulighetene viser at beste strategi er å lete først i felt 2 og så lete videre i felt 1 hvis og bare hvis man finner olje på felt 2. Dette gir en forventet fortjeneste på 75 millioner.

Oppgave 2 Lønninger

a) Vi har kumulativ fordeling

$$F(y) = P(Y \leq y) = P(Y < y) = 1 - P(Y \geq y) = 1 - \left(\frac{k}{y}\right)^{\theta}, \qquad y \geq k.$$

Dermed blir sannsynlighetstettheten

$$f(y) = \frac{\mathrm{d}}{\mathrm{d}y}F(y) = -k^{\theta}\frac{\mathrm{d}}{\mathrm{d}y}\left(\frac{1}{y}\right)^{\theta} = \theta k^{\theta}\frac{1}{y^{\theta+1}}, \qquad y \geq k.$$

Forventningsverdi blir

$$\begin{aligned}
\mathrm{E}[Y] &= \int_k^{\infty} y f(y)\,\mathrm{d}y = \int_k^{\infty} \theta k^{\theta}\frac{1}{y^{\theta}}\,\mathrm{d}y \\
&= \theta k^{\theta}\left[\frac{y^{-\theta+1}}{-\theta+1}\right]_{y=k}^{\infty} \\
&= \frac{\theta}{\theta-1}k.
\end{aligned}$$

Vi starter med å beregne det andre momentet til Y

$$\begin{aligned}\mathrm{E}[Y^2] &= \int_k^\infty y^2 f(y)\,\mathrm{d}y = \int_k^\infty \theta k^\theta \frac{1}{y^{\theta-1}}\,\mathrm{d}y \\ &= \theta k^\theta \left[\frac{y^{-\theta+2}}{-\theta+2}\right]_{y=k}^\infty \\ &= \frac{\theta}{\theta-2}k^2.\end{aligned}$$

Fra dette og middelverdien finner vi

$$\begin{aligned}\mathrm{Var}[Y] &= \mathrm{E}[Y^2] - \mathrm{E}[Y]^2 = \frac{\theta}{\theta-2}k^2 - \frac{\theta^2}{(\theta-1)^2}k^2 \\ &= k^2\theta\left(\frac{(\theta-1)^2 - \theta(\theta-2)}{(\theta-2)(\theta-1)^2}\right) \\ &= k^2\frac{\theta}{(\theta-1)^2(\theta-2)}.\end{aligned}$$

b) Vi finner rimelighetsfunksjonen

$$L(\theta; y_1, y_2, \ldots, y_n) = \prod_{i=1}^n \frac{\theta k^\theta}{y_i^{\theta+1}} = \theta^n k^{n\theta} \prod_{i=1}^n y_i^{-\theta-1}.$$

Dette gir

$$\begin{aligned}l(\theta; y_1, y_2, \ldots, y_n) &= \ln\left(L(\theta; y_1, y_2, \ldots, y_n)\right) \\ &= n\ln(\theta) + n\theta\ln(k) + (-\theta-1)\sum_{i=1}^n \ln(y_i).\end{aligned}$$

For å finne sannsynlighetsmaksimeringsestimatoren (SME) må denne funksjonen maksimeres med hensyn på θ. Vi begynner med å finne ekstremalpunktene fra

$$\begin{aligned}0 &= \frac{\mathrm{d}}{\mathrm{d}\theta} l(\hat{\theta}; y_1, y_2, \ldots, y_n) \\ 0 &= \frac{n}{\hat{\theta}} + n\ln(k) - \sum_{i=1}^n \ln(y_i) \\ \frac{n}{\hat{\theta}} &= \sum_{i=1}^n \ln(y_i) - n\ln(k) \\ \hat{\theta} &= \frac{n}{\sum_{i=1}^n \ln(y_i) - n\ln(k)}.\end{aligned}$$

Dette stemmer med estimatoren oppgitt i oppgaven. [Kommentar: Dette er alltid maksimum av rimelighetsfunksjonen hvis θ kan være i $(0, \infty)$, men basert på opplysningene i oppgave a) så må $\theta > 2$. Dette medfører at hvis en mindre verdi enn dette oppnås fra $\hat{\theta}$ så finnes ikke SME.]

Vi beregner estimatet

$$\hat{\theta}_e = \frac{30}{174{,}7 - 30\ln(214{,}9)} \approx 2{,}21.$$

Dette gir sannsynligheten

$$P(Y < 472{,}2) = 1 - P(Y \geq 472{,}2) = 1 - \left(\frac{214{,}9}{472{,}2}\right)^{2{,}21} \approx 0{,}824.$$

c) La μ betegne forventningsverdien av Y. Da ønsker vi å teste om μ har økt til $435 \cdot 1{,}036 = 450{,}7$ eller om μ er mindre enn 450,7. Vi formulerer dette som en hypotesetest der det vi ønsker å teste om har skjedd velges som alternativ hypotese. Dette gir

$$H_0 : \mu \leq 450{,}7$$
$$H_1 : \mu > 450{,}7.$$

Ettersom fordelingen til Y er vanskelig å jobbe med ønsker vi å gjøre denne testen basert på $\bar{Y}$ på vanlig måte. For at dette skal være meningsfullt må vi anta $\bar{Y}$ har en fordeling som er nærme en normalfordeling. Basert på sentralgrenseteoremet er det rimelig at dette vil være en tilstrekkelig tilnærming hvis vi har nok målinger. Vi må dermed anta 30 målinger er nok til at dette skal være en god tilnærming.

Under denne antagelsen kan man under nullhypotesen (H_0) bruke

$$\mathrm{E}[\bar{Y}] = 450{,}7$$

og

$$\mathrm{Var}[\bar{Y}] = \frac{\sigma^2}{30} = \frac{516^2}{30} \approx 8875$$

til å teste på

$$Z = \frac{\bar{Y} - 450{,}7}{\sqrt{8875}} \sim N(0, 1).$$

d) Vi har en ensidig test på signifikansnivå 0,05 og forkaster H_0 hvis vi observerer

$$z > z_{0{,}05} = 1{,}645.$$

Fra tallene i oppgavene finner vi $\bar{y} = 13611/30 = 453{,}7$ som gir observert verdi

$$z = \frac{453{,}7 - 450{,}7}{\sqrt{8875}} = 0{,}03 < 1{,}645 = z_{0{,}05}.$$

Ergo forkaster vi ikke nullhypotesen. Det er ikke grunnlag for å si at lønnen har steget med 3,6 % på signifikansnivå 0,05.

Fra figuren virker det som det er betydelig skjevhet (skewness) i fordelingen. På ønsket signifikansnivå må vår ensidige test ha 0,05 sannsynlighet i øvre hale av fordelingen. På grunn av skjevheten i fordelingen virker det klart at den virkelige grensen for testen må være mye lengre mot høyre enn indikert av normalfordelingstilnærmingen. Det virker ikke som utvalgsstørrelsen er stor nok til å bruke sentralgrenseteoremet.

Oppgave 3 Gruvedrift

a) For å forenkle uttrykkene innfører vi $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$.

Variansen til $\hat{\beta}_1$ er

$$\begin{aligned}
\mathrm{Var}[\hat{\beta}_1] &= \mathrm{Var}\left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{S_{xx}}\right] \\
&= \frac{1}{S_{xx}^2}\mathrm{Var}\left[\sum_{i=1}^{n}(x_i - \bar{x})Y_i\right] \\
&= \frac{1}{S_{xx}^2}\sum_{i=1}^{n}\mathrm{Var}\left[(x_i - \bar{x})Y_i\right] \\
&= \frac{1}{S_{xx}^2}\sum_{i=1}^{n}(x_i - \bar{x})^2\sigma^2 \\
&= \frac{\sigma^2}{S_{xx}}.
\end{aligned}$$

Ved andre likhet flyttes konstanten ut, ved tredje likhet brukes uavhengigheten til de ulike Y_i, ved fjerde likhet flyttes konstanten ut og $\mathrm{Var}[Y_i] = \sigma^2$ benyttes.

Forventningsverdien til $\hat{\beta}_1$ er

$$\begin{aligned}
\mathrm{E}[\hat{\beta}_1] &= \frac{\sum_{i=1}^n (x_i - \bar{x})\mathrm{E}[Y_i]}{S_{xx}} \\
&= \frac{\sum_{i=1}^n (x_i - \bar{x})(\beta_0 + \beta_1 x_i)}{S_{xx}} \\
&= \frac{\beta_0 \sum_{i=1}^n (x_i - \bar{x}) + \beta_1 \sum_{i=1}^n (x_i - \bar{x})x_i}{S_{xx}} \\
&= \frac{\beta_0 \cdot 0 + \beta_1 \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2\right)}{S_{xx}} \\
&= \frac{\beta_1 S_{xx}}{S_{xx}} = \beta_1.
\end{aligned}$$

Dermed er $\hat{\beta}_1$ forventningsrett.

b) Fra uttrykket for $\hat{\beta}_1$ ser vi at $\hat{\beta}_1$ er en lineærkombinasjon av uavhengige Y_i som er normalfordelte. Dette betyr at $\hat{\beta}_1$ også er normalfordelt, og vi har $\hat{\beta}_1 \sim N(\beta_1, \sigma^2/S_{xx})$. Vi kan derfor lage et konfidensintervall basert på den stokastiske variabelen

$$Z = \frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}} \sim N(0,1).$$

For et 90 % konfidensintervall trenger vi z som er slik at

$$P(-z < Z < z) = 0.90.$$

Denne verdien er som kjent $z_{0,05} = 1{,}645$. Deretter jobber vi oss fra ulikhetene

$$-z_{0,05} < Z < z_{0,05}$$

til et intervall for β_1. Vi får intervallet.

$$\begin{aligned}
-z_{0,05} < \frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}} &< z_{0,05}, \\
-z_{0,05}\frac{\sigma}{\sqrt{S_{xx}}} < \hat{\beta}_1 - \beta_1 &< z_{0,05}\frac{\sigma^2}{S_{xx}}, \\
\hat{\beta}_1 - z_{0,05}\frac{\sigma}{\sqrt{S_{xx}}} < \beta_1 &< \hat{\beta}_1 + z_{0,05}\frac{\sigma}{\sqrt{S_{xx}}}.
\end{aligned}$$

Vi finner $\bar{x} = 2$, $S_{xx} = 4$ og

$$\hat{\beta}_1 = \frac{-y_1 - y_2 + y_6 + y_7}{4} = \frac{-2{,}7 - 3{,}1 + 4{,}1 + 5{,}5}{4} = 0{,}95.$$

Dermed blir intervallet

$$(0{,}95 - 1{,}645 \cdot 0{,}5/2, 0{,}95 + 1{,}645 \cdot 0{,}5/2) \Rightarrow (0{,}54, 1{,}36).$$

Basert på uttrykket for konfidensintervallet som vi har funnet så er lengden av intervallet bestemt av $z_{0,05}\sigma/S_{xx}$. Av faktorene i dette uttrykket er det bare S_{xx} vi kan påvirke. Denne faktoren inngår i nevneren slik at jo større S_{xx} er jo mindre er lengden av konfidensintervallet. Med andre ord vil det beste være å gjøre S_{xx} størst mulig under begrensningen at vi skal måle 7 punkter fordelt på bergart 1, 2 og 3. Siden S_{xx} betegner summen av kvadratavvikene virker det rimelig at det beste er å fordele målingene omtrent jevnt i bergart 1 og bergart 3.

For 3 prøver i bergart 1, 1 prøve i bergart 2 og 3 prøver i bergart 3 får vi $S_{xx} = 6$ som er bedre enn punktene brukt i oppgaven. Dette er rimelig siden målingene i bergart 2 ikke hadde noen innvirkning på S_{xx} siden $\bar{x} = 2$. For 1 prøve i bergart 1, 5 i bergart 2 og 1 i bergart 3 for vi $S_{xx} = 2$ som er dårligere enn punktene brukt i oppgaven. Dette er rimelig siden de fleste x-verdiene er lik 2 som ikke har noen påvirkning på S_{xx} når $\bar{x} = 2$. Ettersom det er ønskelig med mest mulig spredning lønner det seg å flytte også den siste måling i bergart 2 til noe annet. Med 4 målinger i bergart 1 og 3 målinger i bergart 3 får vi $S_{xx} = 6.86$.

Disse resultatene er rimelige fordi $\hat{\beta}_1$ er estimatoren til et stigningstall. Når vi skal estimere stigningstallet er vi mindre påvirket av usikkerheten i y-verdiene jo lenger avstand vi har mellom x-punktene. Dette er fordi σ blir mindre relativt til økningen i y. Med liten avstand mellom x-punktene vil det være nærmest umulig å skille en faktisk økning på grunn av $\beta_1 \Delta x$ og en forandring på grunn av usikkerheten i de to punktene vi har målt. Derimot hvis $\beta_1 \Delta_x$ er mye større enn σ er målingen av økningen nærmest upåvirket av usikkerheten i y-punktene. I virkelige situasjoner er det ikke nødvendigvis mulig å gjøre avstanden veldig stor i forhold til σ. Et annet viktig punkt er at denne tankegangen krever at man faktisk stoler på at modellen er korrekt. Hvis dette ikke er tilfellet er det bedre å fordele målingene utover de mulige verdiene enn å bare bruke endepunktene.

c) La $\hat{Y}_0$ være prediktoren basert på $(x_1, Y_1), (x_2, Y_2), \ldots, (x_7, Y_7)$ ved x_0, gitt ved

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_i, \tag{1}$$

og la

$$Y_0 = \beta_0 + \beta_1 x_0 + \epsilon_0,$$

hvor $\epsilon_0 \sim N(0, \sigma^2)$, være en ny måling ved x_0 uavhengig av $Y_1, Y_2, \ldots, Y_7$.

Legg merke til at vi umiddelbart har $Y_0 \sim N(\beta_0 + \beta_1 x_0, \sigma^2)$. For $\hat{Y}_0$ setter vi inn $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$ i Ligning (1),

$$\hat{Y}_0 = \bar{Y} + \hat{\beta}_1 (x_0 - \bar{x}).$$

For å finne prediksjonintervallet har vi lyst til å basere oss på $\hat{Y}_0 - Y_0$, men for å gjøre dette trenger vi forventningsverdi og varians til $\hat{Y}_0$.

Forventningsverdi finner vi ved

$$\mathrm{E}[\hat{Y}_0] = \mathrm{E}\left[\frac{1}{7}\sum_{i=1}^{7} Y_i + \hat{\beta}_1(x_0 - \bar{x})\right].$$

Her er forventningsverdi til $\hat{\beta}_1$ kjent og $(x_0 - \bar{x})$ er en konstant slik at vi får

$$\begin{aligned}\mathrm{E}[\hat{Y}_0] &= \frac{1}{7}\sum_{i=1}^{7}\mathrm{E}[Y_i] + \beta_1(x_0 - \bar{x}) = \frac{1}{7}\sum_{i=1}^{7}(\beta_0 + \beta_1 x_i) + \beta_1(x_0 - \bar{x}) \\ &= \beta_0 + \beta_1\bar{x} + \beta_1(x_0 - \bar{x}) = \beta_0 + \beta_1 x_0.\end{aligned}$$

For å beregne variansen trenger vi å vite at $\bar{Y}$ og $\hat{\beta}_1$ er uavhengige, eller vise det ved

$$\begin{aligned}\mathrm{Cov}(\bar{Y}, \hat{\beta}_1) &= \mathrm{Cov}\left(\frac{1}{7}\sum_{i=1}^{7} Y_i, \frac{\sum_{j=1}^{7}(x_j - \bar{x})Y_j}{S_{xx}}\right) \\ &= \frac{1}{7S_{xx}}\sum_{i=1}^{7}\sum_{j=1}^{7}(x_j - \bar{x})\mathrm{Cov}(Y_i, Y_j).\end{aligned}$$

Her er $\mathrm{Cov}(Y_i, Y_j) = 0$ når $i \neq j$ og $\mathrm{Cov}(Y_i, Y_i) = \mathrm{Var}(Y_i) = \sigma^2$. Dette gir

$$\begin{aligned}\mathrm{Cov}(\bar{Y}, \hat{\beta}_1) &= \frac{1}{7S_{xx}}\sum_{j=1}^{7}(x_j - \bar{x})\mathrm{Cov}(Y_j, Y_j) \\ &= \frac{1}{7S_{xx}}\sigma^2\sum_{j=1}^{7}(x_j - \bar{x}) \\ &= 0.\end{aligned}$$

Med denne informasjonen kan vi regne ut

$$\begin{aligned}\mathrm{Var}[\hat{Y}_0] &= \mathrm{Var}[\bar{Y} + \beta_1(x_0 - \bar{x})] = \mathrm{Var}[\bar{Y}] + (x_0 - \bar{x})^2\mathrm{Var}[\hat{\beta}_1] \\ &= \frac{\sigma^2}{7} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\sigma^2,\end{aligned}$$

hvor vi har brukt at $\bar{Y}$ og $\hat{\beta}_1$ er ukorrelerte ved likhet nummer 2.

Vi har totalt at

$$\mathrm{E}[\hat{Y}_0 - Y_0] = 0$$

og at

$$\text{Var}[\hat{Y}_0 - Y_0] = \text{Var}[\hat{Y}_0] + \text{Var}[Y_0] = \frac{\sigma^2}{7} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\sigma^2 + \sigma^2.$$

Dermed er

$$Z = \frac{\hat{Y}_0 - Y_0}{\sigma\sqrt{1 + \frac{1}{7} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim N(0,1),$$

og vi kan gjenta utregningene for konfidensintervallet i b) for å få prediksjonintervallet

$$\hat{Y}_0 - z_{0,05}\sigma\sqrt{1 + \frac{1}{7} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} < Y_0 < \hat{Y}_0 + z_{0,05}\sigma\sqrt{1 + \frac{1}{7} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}.$$

Ved å sette inn tall får vi

$$\hat{Y}_0 = \bar{y} - \hat{\beta}_1(x_0 - \bar{x}) = 3{,}814 + 0{,}95(3 - 2) \approx 4{,}76.$$

I tillegg har vi

$$z_{0,05}\sigma\sqrt{1 + \frac{1}{7} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} = 1{,}645 \cdot 0{,}5\sqrt{1 + \frac{1}{7} + \frac{1^2}{4}} \approx 0{,}97.$$

Dette gir intervallet

$$(4{,}76 - 0{,}97, 4{,}76 + 0{,}97) \Rightarrow (3{,}79, 5{,}73).$$