Løsningsforslag til eksamen i STK1110 10. desember 2010

Løsningsforslaget har med flere detaljer enn det vil bli krevd til eksamen.

Oppgave 1

a) Det er tilpasset en multippel lineær regresjonsmodell av formen

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$$

Her er responsen Y_i bensinforbruket for bil nummer i, mens forklaringsvariablene x_{i1}, x_{i2} og x_{i3} er henholdsvis antall hestekrefter for motoren, vekten til bilen og slagvolumet for motoren. Videre angir ϵ_i-ene de tilfeldige avvikene, som antas å være uavhengige og $N(0, \sigma^2)$-fordelte.

I `R`-utskriften står det følgende i kolonnene i tabellen:

`Estimate:`	Minste kvadraters estimater for β_0, β_1, β_2 og β_3.
`Std. Error:`	Estimerte standardfeil til estimatene nevnt i forrige punkt.
`t value:`	Testobservatorer for testing av $H_0 : \beta_j = 0$ mot $H_a : \beta_j \neq 0$; $j = 0, 1, 2, 3$. Testobservatorene er forholdet mellom estimatene og de estimerte standardfeilene.
`Pr(>|t|):`	P-verdier for testene nevnt i forrige punkt.

Under tabellen er `Residual standard error` oppgitt. Det er estimatet for σ. Videre er det gitt `Multipple R-squared`. Denne står for R^2, som er et mål for hvor stor andel av variasjonen i responsen (dvs. bensinforbruket) som er forklart av regresjonsmodellen (dvs. av antall hestekrefter, vekt og slagvolum).

b) Vi lar som i forrige punkt β_3 være regresjonskoeffisienten for slagvolum. Vi vil teste nullhypotesen $H_0 : \beta_3 = 0$ mot den alternative hypotesen $H_0 : \beta_3 \neq 0$. Vi bruker da testobservatoren $T = \hat{\beta}_3 / S_{\hat{\beta}_3}$, der $S_{\hat{\beta}_3}$ er en estimator for standardfeilen til $\hat{\beta}_3$. Hvis nullhypotesen er sann, er testobservatoren t-fordelt med $19 - 4 = 15$ frihetsgrader.

En test med signifikansnivå α forkaster H_0 hvis

$$T \leq -t_{\alpha/2,\,15} \qquad \text{eller} \qquad T \geq t_{\alpha/2,\,15}$$

Av `R`-utskriften i punkt a er minste kvadraters estimat for regresjonskoeffisienten $\hat{\beta}_3 = -0.23979$, mens estimatet for standardfeilen er $s_{\hat{\beta}_3} = 0.07335$. Testobservatoren T får dermed verdien $t = -0.23979/0.07335 = -3.27$. Fra tabellen over t-fordelingene, finner vi at $t_{0.005,\,15} = 2.947$ og $t_{0.001,\,15} = 3.733$. Vi forkaster altså nullhypotesen hvis signifikansnivået er $2 \cdot 0.005 = 0.01 = 1\%$, men ikke hvis signifikansnivået er $2 \cdot 0.001 = 0.002 = 0.2\%$. P-verdien for testen (jf. oppgave 2b) er dermed mellom 0.2% og 1%. Vi forkaster nullhypotesen og konkludere med at slagvolumet har en signifikant effekt for bensinforbruket.

Estimatet $\hat{\beta}_3 = -0.24$ forteller oss at hvis slagvolumet økes med én liter og antall hestekrefter og vekten av bilen ikke endres, så vil forventet bensinforbruk bli redusert med 0.24 liter per mil. Hvis vi sammenligner biler med samme motorstyrke og samme vekt, vil vi atså forvente at den bilen som har størst slagvolum bruker minst bensin. Av plottet først i oppgaven ser det ut til at bensinforbruket øker med slagvolumet. Men det kommer av at en bil som har stort slagvolum typisk også vil veie mer og ha en sterkere motor enn en bil med mindre slagvolum.

c) Som i punkt a lar vi β_1 være regresjonskoeffisienten for hestekrefter. Det betyr at β_1 er forventet endring i bensinforbruket når styrken til motoren økes med én hestekraft og vekt og slagvolum ikke endres. Vi vil bestemme et 95% konfidensintervall for $\theta = 10\beta_1$.

Vi har at $(\hat{\beta}_1 - \beta_1)/S_{\hat{\beta}_1}$ er t-fordelt med 15 frihetsgrader. Det gir

$$P\left(-t_{0.025,\,15} < \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} < t_{0.025,\,15}\right) = 0.95$$

Av dette følger på vanlig måte at

$$P\left(\hat{\beta}_1 - t_{0.025,\,15}\, S_{\hat{\beta}_1} < \beta_1 < \hat{\beta}_1 + t_{0.025,\,15}\, S_{\hat{\beta}_1}\right) = 0.95$$

Vi ganger med 10 på begge sider av ulikhetene og får (siden $\theta = 10\beta_1$)

$$P\left(10\hat{\beta}_1 - 10\,t_{0.025,\,15}\,S_{\hat{\beta}_1} < \theta < 10\hat{\beta}_1 + 10\,t_{0.025,\,15}\,10S_{\hat{\beta}_1}\right) = 0.95$$

Et 95% konfidensintervall for θ er dermed gitt som

$$\left(10\hat{\beta}_1 - 10\,t_{0.025,\,15}\,S_{\hat{\beta}_1}\ ,\ 10\hat{\beta}_1 + 10\,t_{0.025,\,15}\,S_{\hat{\beta}_1}\ \right)$$

Fra R-utskriften finner vi estimatene $\hat{\beta}_1 = 0.00604$ og $s_{\hat{\beta}_1} = 0.00148$, og av tabellen for t-fordelingene finner vi $t_{0.025,\,15} = 2.131$. Det gir følgende 95% konfidensintervall for θ

$$(10 \cdot 0.00604 - 10 \cdot 2.131 \cdot 0.00148\ ,\ 10 \cdot 0.00604 - 10 \cdot 2.131 \cdot 0.00148\ )$$

dvs. $(0.029\ ,\ 0.092\ )$. Det betyr at vi med stor grad av sikkerhet kan si at når mororstyrken økes med 10 hestekrefter (og vekt og slagvolum ikke endres), så vil forventet bensinforbruk øke med mellom 0.029 og 0.092 liter per mil.

d) I punkt a skrev vi opp modellen for den multiple lineære regresjonen. Modellen bygger på følgende forutsetninger:

(i) Forventet bensinforbruk avhenger lineært av de tre forklaringsvariablene (x_1 = hestekrefter, x_2 = vekt og x_3 = slagvolum)

(ii) De tilfeldige avvikene (dvs. ϵ_i-ene) har alle samme varians.

(iii) De tilfeldige avvikene er normalfordelte.

(iv) De tilfeldige avvikene er uavhengige.

For å kontrollere disse forutsetningene kan vi lage ulike plott av residualene. Residualene er gitt ved

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3})$$

der y_i-ene er de observerte forbrukene av bensin. Residualene er prediksjoner av de uobserverte ϵ_i-ene, og hvis forutsetningene for regresjonsanalysen er rimelig godt oppfylt, skal residualene oppføre seg (omtrent) som uavhengige og $N(0, \sigma^2)$-fordelte variabler.

Plottet øverst til venstre er et normalfordelingsplott av residualene. Hvis de tilfeldige avvikene er normalfordelte vil punktene i dette plottet ligge noenlunde på en rett linje. Det er tilfellet her, så forutsetning (iii) er rimelig godt oppfylt.

Plottet øverst til høyre er et plott av residualene mot de tilpassede verdiene ($\hat{y}_i$-ene). Hvis de tilfeldige avvikene alle har samme varians, vil residualene fordele seg tilfeldig (i y-retningen), mens en "viftefasong" av residualene vil være et tegn på at de tilfeldige avvikene ikke har samme varians. Her fordeler residualene seg noenlunde tilfeldig, så forutsetning (ii) er rimelig godt oppfylt.

De tre nederste plottene er plott av residualene mot hver av forklaringsvariablene. Hvis forventet bensinforbruk avhenger lineært av en forklaringsvariabel, vi residualene fordele seg tilfeldig (i y-retningen), mens et systematisk mønster for residualene (f. eks. en krumning) vil tyde på at sammenhengen ikke er lineær. Her fordeler residualene seg noenlunde tilfeldig, så forutsetning (i) er rimelig godt oppfylt for alle forklaringsvariablene (muligens med unntak av slagvolum der det er antydning til en viss krumming).

Ingen av residualplottene som er gitt i oppgaven egner seg til å kontrollere forutsetning (iv). Men det er antageligvis en uproblematisk forutsetning her.

Oppgave 2

a) La X være antall seksere vi får når marmorterningen blir kastet 204 ganger. Da er X binomisk(n, p)-fordelt, der $n = 204$ og p er sannsynligheten for å få sekser når vi kaster marmorterningen.

Vi vil teste nullhypotesen $H_0 : p = p_0$ mot den alternative hypotesen $H_\text{a} : p > p_0$, der $p_0 = \frac{1}{6}$. Vi innfører

$\hat{p} = X/n$ og merker oss at $np_0 = 204 \cdot \frac{1}{6} = 34$ og $n(1-p_0) = 204 \cdot \frac{5}{6} = 170$ begge er større enn 10. Det følger da at

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{\hat{p} - \frac{1}{6}}{\sqrt{\frac{\frac{1}{6}\cdot\frac{5}{6}}{204}}} = \frac{\hat{p} - 1/6}{\sqrt{5/7344}}$$

er tilnærmet standardnormalfordelt hvis H_0 er sann.

La z_α være som gitt i læreboka. Da er $P(Z \geq z_\alpha) = \alpha$ hvis H_0 er sann. Vi får derfor en test med signifikansnivå 5% hvis vi forkaster H_0 såsant $Z \geq z_{0.05} = 1.645$.

Da marmorterningen ble kastet 204 ganger, fikk en sekser i 54 av kastene. Testobservatoren Z får da verdien

$$z = \frac{54/204 - 1/6}{\sqrt{5/7344}} = 3.76$$

Siden $z = 3.76 \geq 1.645$ forkaster vi nullhypotesen. Konklusjonen er da at sannsynligheten for å få sekser med marmorterningen er større enn en seksdel.

b) P-verdien til en test er det minste signifikansnivået α som gir forkastning av nullhypotesen. Litt mer uformelt, kan vi si at P-verdien er sannsynligheten for å få et resultat som det vi har fått, eller et resultat som er enda mer i favør av den alternative hypotesen, bare på grunn av tilfeldigheter (dvs. hvis H_0 er sann).

Tabellen over den kumulative standardnormalfordelingen $\Phi(z)$ som er gitt i læreboka dekker verdier av z fra -3.49 til 3.49. Siden testobservatoren har verdien 3.76, klarer vi ikke å bestemme P-verdien helt nøyaktig av tabellen. Men av tabellen ser vi at vi får en test med signifikansnivå 0.02% hvis vi forkaster H_0 såsant $Z \geq 3.49$. Vi vil altså forkaste nullhypotesen på dette nivået, så P-verdien er mindre enn 0.02%. Det betyr at det er svært usannsynlig å få 54 eller flere seksere i 204 kast hvis sannsynligheten for sekser er 1/6. Vi er derfor veldig sikre på at sannsynligheten for å få sekser med marmorterningen er større enn en seksdel.

c) Da terningen av jern ble kastet 204 ganger, fikk en sekser i 42 av kastene. Testobservatoren Z i punkt a får da verdien

$$z = \frac{42/204 - 1/6}{\sqrt{5/7344}} = 1.50$$

Siden $z = 1.50 < 1.645$ forkaster vi ikke nullhypotesen på signifikansnivå 5%. Vi kan derfor ikke konkludere med at sannsynligheten for å få sekser med terningen av jern er større enn en seksdel. (Men siden P-verdien er såpass liten som 6.7% har vi en indikasjon på at sannsynligheten for sekser kan være større enn en seksdel, men at testen ikke har nok styrke til å oppdage det.)

d) Vi gjør en feil av type II hvis vi ikke forkaster nullhypotesen når den er gal. Vi forkaster ikke H_0 på signifikansnivå 5% hvis $Z < 1.645$. Hvis sannsynligheten for å få sekser med terningen av jern er $p = 0.20$, blir sannsyligheten for feil av type II

$$\begin{aligned}
\beta(0.20) &= P(Z < 1.645 \,|\, p = 0.20) \\
&= P\left(\frac{\hat{p} - 1/6}{\sqrt{5/7344}} < 1.645 \,\middle|\, p = 0.20 \right) \\
&= P\left(\hat{p} < 1/6 + 1.645\sqrt{5/7344} \,\middle|\, p = 0.20 \right) \\
&= P\left(\frac{\hat{p} - 0.20}{\sqrt{0.20 \cdot 0.80/204}} < \frac{1/6 - 0.20 + 1.645\sqrt{5/7344}}{\sqrt{0.20 \cdot 0.80/204}} \,\middle|\, p = 0.20 \right) \\
&= P\left(\frac{\hat{p} - 0.20}{\sqrt{0.20 \cdot 0.80/204}} < 0.34 \,\middle|\, p = 0.20 \right) \\
&= 0.633
\end{aligned}$$

Her får vi den siste likheten fra tabellen over den kumulative standardnormalfordelingen siden

$$\frac{\hat{p} - 0.20}{\sqrt{0.20 \cdot 0.80/204}}$$

er (tilnærmet) standardnormalfordelt når $p = 0.20$. Sannsynligheten for feil av type II er altså 63.3% hvis sannsynligheten for å få sekser med terningen av jern er 20%.

e) Hvis vi kaster terningen av jern n ganger, vil en test med signifikansnivå 5% forkaste H_0 såsant

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{\hat{p} - \frac{1}{6}}{\sqrt{\frac{\frac{1}{6}\cdot\frac{5}{6}}{n}}} = \frac{\hat{p} - 1/6}{\sqrt{5/(36n)}} \geq 1.645$$

jf. punkt a. Hvis sannsynligheten for å få sekser med terningen av jern er $p = 0.20$, blir sannsynligheten for feil av type II (jf. punkt d):

$$\begin{aligned}
\beta(0.20) &= P(Z < 1.645 \,|\, p = 0.20) \\
&= P\left(\frac{\hat{p} - 1/6}{\sqrt{5/(36n)}} < 1.645 \,\middle|\, p = 0.20 \right) \\
&= P\left(\hat{p} < 1/6 + 1.645\sqrt{5/(36n)} \,\middle|\, p = 0.20 \right) \\
&= P\left(\frac{\hat{p} - 0.20}{\sqrt{0.20 \cdot 0.80/n}} < \frac{1/6 - 0.20 + 1.645\sqrt{5/(36n)}}{\sqrt{0.20 \cdot 0.80/n}} \,\middle|\, p = 0.20 \right) \\
&= P\left(\frac{\hat{p} - 0.20}{\sqrt{0.20 \cdot 0.80/n}} < -0.0833\sqrt{n} + 1.533 \,\middle|\, p = 0.20 \right)
\end{aligned}$$

Hvis $p = 0.20$ er $(\hat{p} - 0.20)/\sqrt{0.20 \cdot 0.80/n}$ tilnærmet standardnormalfordelt. Av tabellen over den kumulative standardnormalfordelingen har vi dermed at

$$P\left(\frac{\hat{p} - 0.20}{\sqrt{0.20 \cdot 0.80/n}} < -0.84 \,\middle|\, p = 0.20 \right) = 0.20$$

For at sannsynligheten for feil av type II skal være 20% hvis $p = 0.20$ må derfor n bestemmes slik at

$$-0.0833\sqrt{n} + 1.533 = -0.84$$

Det gir

$$n = \left(\frac{1.533 + 0.84}{0.0833} \right)^2 = 811.5$$

Hvis sannsynligheten for feil av type II skal være 20% når $p = 0.20$ må vi kaste terningen omtrent 810 ganger.

Oppgave 3

a) Minste kvadraters estimatorer for β_0 og β_1 er de verdiene av b_0 og b_1 som minimerer kvadratsummen

$$\mathrm{SS}(b_0, b_1) = \sum_{i=1}^{n} (Y_i - b_0 - b_1(x_i - \overline{x}))^2$$

Vi deriverer med hensyn på b_0 og finner

$$\frac{\partial}{\partial b_0} \mathrm{SS}(b_0, b_1) = -2 \sum_{i=1}^{n} (Y_i - b_0 - b_1(x_i - \overline{x}))$$

$$
\begin{aligned}
&= -2\left(\sum_{i=1}^{n} Y_i - nb_0 - b_1\sum_{i=1}^{n}(x_i-\overline{x})\right)\\
&= -2\left(\sum_{i=1}^{n} Y_i - nb_0\right)
\end{aligned}
$$

der den siste likheten følger siden $\sum_{i=1}^{n}(x_i-\overline{x})=0$. Vi deriverer også med hensyn på b_1 og får

$$
\begin{aligned}
\frac{\partial}{\partial b_1}\mathrm{SS}(b_0,b_1) &= -2\sum_{i=1}^{n}\left(Y_i - b_0 - b_1(x_i-\overline{x})\right)(x_i-\overline{x})\\
&= -2\left(\sum_{i=1}^{n}(x_i-\overline{x})\,Y_i - b_0\sum_{i=1}^{n}(x_i-\overline{x}) - b_1\sum_{i=1}^{n}(x_i-\overline{x})^2\right)\\
&= -2\left(\sum_{i=1}^{n}(x_i-\overline{x})\,Y_i - b_1\sum_{i=1}^{n}(x_i-\overline{x})^2\right)\\
&= -2\left(\sum_{i=1}^{n}(x_i-\overline{x})\,Y_i - S_{xx}\,b_1\right)
\end{aligned}
$$

der $S_{xx}=\sum_{i=1}^{n}(x_i-\overline{x})^2$.

Minste kvadraters estimatorer for $\hat{\beta}_0$ og $\hat{\beta}_1$ er løsningene av de ligningene vi får når vi setter de partiell deriverte lik null. Det gir

$$
\sum_{i=1}^{n} Y_i - n\hat{\beta}_0 = 0
$$

$$
\sum_{i=1}^{n}(x_i-\overline{x})\,Y_i - S_{xx}\,\hat{\beta}_1 = 0
$$

Minste kvadraters estimatorer er dermed

$$
\begin{aligned}
\hat{\beta}_0 &= \frac{1}{n}\sum_{i=1}^{n} Y_i = \overline{Y}\\
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i-\overline{x})\,Y_i}{S_{xx}} = \sum_{i=1}^{n}\frac{x_i-\overline{x}}{S_{xx}}\,Y_i
\end{aligned}
$$

b) Vi bruker at $E(Y_i)=\beta_0+\beta_1(x_i-\overline{x})$ og får

$$
\begin{aligned}
E(\hat{\beta}_0) &= E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i)\\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\beta_0+\beta_1(x_i-\overline{x})\right)\\
&= \frac{1}{n}\,n\beta_0 + \frac{1}{n}\beta_1\sum_{i=1}^{n}(x_i-\overline{x})\\
&= \beta_0 + 0 = \beta_0
\end{aligned}
$$

Det viser at $\hat{\beta}_0$ er forventningsrett.

Vi finner videre at

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

Det viser at også $\hat{\beta}_1$ er forventningsrett.

c) Vi bruker at $V(Y_i) = \sigma^2$ og får siden Y_i-ene er uavhengige

$$V(\hat{\beta}_0) = V\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(Y_i) = \frac{1}{n^2}\, n\sigma^2 = \frac{\sigma^2}{n}$$

og

$$\begin{aligned}
V(\hat{\beta}_1) &= V\left(\sum_{i=1}^{n} \frac{x_i - \overline{x}}{S_{xx}}\, Y_i\right) = \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{S_{xx}}\right)^2 V(Y_i) \\
&= \frac{\sigma^2}{S_{xx}^2} \sum_{i=1}^{n} (x_i - \overline{x})^2 = \frac{\sigma^2}{S_{xx}^2}\, S_{xx} = \frac{\sigma^2}{S_{xx}}
\end{aligned}$$

For kovariansen får vi når vi bruker hintet

$$\begin{aligned}
\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) &= \text{Cov}\left(\sum_{i=1}^{n} \frac{1}{n} Y_i\,,\, \sum_{i=1}^{n} \frac{x_i - \overline{x}}{S_{xx}}\, Y_i\right) \\
&= \sum_{i=1}^{n} \frac{1}{n}\, \frac{x_i - \overline{x}}{S_{xx}}\, V(Y_i) \\
&= \frac{\sigma^2}{n S_{xx}} \sum_{i=1}^{n} (x_i - \overline{x}) = 0
\end{aligned}$$

d) Vi bruker resultatet i punkt b og finner at

$$\begin{aligned}
E\left(\hat{\mu}_{Y\cdot x^*}\right) &= E\left(\hat{\beta}_0 + \hat{\beta}_1(x^* - \overline{x})\right) \\
&= E(\hat{\beta}_0) + E(\hat{\beta}_1)(x^* - \overline{x}) \\
&= \beta_0 + \beta_1(x^* - \overline{x}) = \mu_{Y\cdot x^*}
\end{aligned}$$

så estimatoren er forventningsrett. Videre får vi av resultatet i punkt c

$$\begin{aligned}
V\left(\hat{\mu}_{Y\cdot x^*}\right) &= V\left(\hat{\beta}_0 + \hat{\beta}_1(x^* - \overline{x})\right) \\
&= V(\hat{\beta}_0) + V\left(\hat{\beta}_1(x^* - \overline{x})\right) + 2\,\text{Cov}\left(\hat{\beta}_0\,,\, \hat{\beta}_1(x^* - \overline{x})\right) \\
&= V(\hat{\beta}_0) + (x^* - \overline{x})^2\, V(\hat{\beta}_1) + 2\,(x^* - \overline{x})\,\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) \\
&= \frac{\sigma^2}{n} + (x^* - \overline{x})^2\, \frac{\sigma^2}{S_{xx}} + 2\,(x^* - \overline{x}) \cdot 0
\end{aligned}$$

$$= \quad \sigma^2 \left(\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}} \right)$$

Alternativt kan vi finne variansen ved å skrive $\hat{\mu}_{Y \cdot x^*}$ som en lineærkombinasjon av Y_i-ene (jf. punkt e) og bruke formelen for variansen til en lineærkombinasjon av uavhengige variabler.

e) Av resultatet i punkt a har vi at

$$\begin{aligned}
\hat{\mu}_{Y \cdot x^*} &= \hat{\beta}_0 + (x^* - \overline{x})\hat{\beta}_1 \\
&= \frac{1}{n} \sum_{i=1}^{n} Y_i + (x^* - \overline{x}) \sum_{i=1}^{n} \frac{x_i - \overline{x}}{S_{xx}} Y_i \\
&= \sum_{i=1}^{n} \left(\frac{1}{n} + \frac{(x^* - \overline{x})(x_i - \overline{x})}{S_{xx}} \right) Y_i
\end{aligned}$$

Altså er $\hat{\mu}_{Y \cdot x^*}$ en lineærkombinasjon av normalfordelte stokastiske variabler, så $\hat{\mu}_{Y \cdot x^*}$ er normalfordelt.

Det følger av dette at

$$Z = \frac{\hat{\mu}_{Y \cdot x^*} - E(\hat{\mu}_{Y \cdot x^*})}{\sqrt{V(\hat{\mu}_{Y \cdot x^*})}} = \frac{\hat{\mu}_{Y \cdot x^*} - \mu_{Y \cdot x^*}}{\sigma \sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}}}$$

er standardnormalfordelt.

Siden $\hat{\mu}_{Y \cdot x^*}$ er en funksjon av $\hat{\beta}_0$ og $\hat{\beta}_1$, som er uavhengige av S^2, er Z og $V = (n-2)S^2/\sigma^2$ uavhengige. Videre er V kji-kvadratfordelt med $n-2$ frihetsgrader. Dermed er

$$\frac{Z}{\sqrt{V/(n-2)}} = \frac{\frac{\hat{\mu}_{Y \cdot x^*} - \mu_{Y \cdot x^*}}{\sigma \sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}}}}{\sqrt{\frac{(n-2)S^2}{\sigma^2} \Big/ (n-2)}} = \frac{\hat{\mu}_{Y \cdot x^*} - \mu_{Y \cdot x^*}}{S \sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}}}$$

t-fordelt med $n-2$ frihetsgrader.

La nå $t_{\alpha/2,\, n-2}$ være som gitt i læreboka. Da har vi at

$$P\left(-t_{\alpha/2,\, n-2} < \frac{\hat{\mu}_{Y \cdot x^*} - \mu_{Y \cdot x^*}}{S \sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}}} < t_{\alpha/2,\, n-2} \right) = 1 - \alpha$$

Vi omformer ulikhetene og får at dette er det samme som

$$P\left(-t_{\alpha/2,\, n-2}\, S \sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}} < \hat{\mu}_{Y \cdot x^*} - \mu_{Y \cdot x^*} < t_{\alpha/2,\, n-2}\, S \sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}} \right) = 1 - \alpha$$

som igjen er det samme som

$$P\left(\hat{\mu}_{Y \cdot x^*} - t_{\alpha/2,\, n-2}\, S \sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}} < \mu_{Y \cdot x^*} < \hat{\mu}_{Y \cdot x^*} + t_{\alpha/2,\, n-2}\, S \sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}} \right) = 1 - \alpha$$

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

$$\left(\hat{\mu}_{Y \cdot x^*} - t_{\alpha/2,\, n-2}\, S \sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}}\ ,\ \hat{\mu}_{Y \cdot x^*} + t_{\alpha/2,\, n-2}\, S \sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}} \right)$$