TMA4240 Statistikk
Høst 2015

Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

Øving nummer 12, blokk II
Løsningsskisse

Oppgave 1

a) - β angir bilens bensinforbruk (i liter/mil)

- Rimelig med $\alpha = 0$ fordi med $x = 0$ (ingen kjøring) brukes ingen bensin

- en tur av lengde $x_1 = x$ kan tenkes sammensatt av to turer på $x_2 = x/2$ og $x_3 = x/2$. La Y_1, Y_2, Y_3 være tilhørende bensinforbruk. Det er da rimelig å kreve at

$\text{Var}(Y_1) = \text{Var}(Y_2) + \text{Var}(Y_3)$.

Dette oppnås ved å velge

$\text{Var}(Y) = x\sigma^2$

b) $\beta = 0.75 \quad , x = 5.0 \quad , \sigma^2 = 0.1^2$

Dette betyr at

$Y \sim n(y; \beta x, \sqrt{x\sigma^2}) \sim n(y; 3.75, \sqrt{0.05})$

$$\begin{aligned} P(Y > 4) &= 1 - P(Y \leq 4) = 1 - P\left(\frac{Y - 3.75}{\sqrt{0.05}} \leq \frac{4 - 3.75}{\sqrt{0.05}}\right) \\ &= 1 - \Phi(1.12) = 1 - 0.869 = \underline{\underline{0.131}} \end{aligned}$$

Ser så på to kjøreturer

$Y_1 \sim n(y; 3.75, \sqrt{0.05})$ og

$Y_2 \sim n(y; 7.5, \sqrt{0.1})$

P.g.a. uavhengighet har vi at $Z = Y_1 + Y_2 \sim n(z; 3.75 + 7.5, \sqrt{0.05 + 0.10})$.

$$\begin{aligned} P(Z < 12) &= P\left(\frac{z - 11.25}{\sqrt{0.15}} \leq \frac{12 - 11.25}{\sqrt{0.15}}\right) = \Phi(1.94) \\ &= \underline{\underline{0.974}} \end{aligned}$$

$$\begin{aligned} U = Y_2 - 2Y_1 &\sim n(z; 0, \sqrt{0.1 + 4 \cdot 0.05}) \\ P(Y_2 - 2Y_1 > 0) &= P(U > 0) = \underline{\underline{0.5}} \end{aligned}$$

Siden fordelingen til U er symmetrisk om $u = 0$.

c) Studerer to estimatorer $\hat{\beta}$ og $\tilde{\beta}$

$$\begin{aligned}
\mathrm{E}(\hat{\beta}) &= \mathrm{E}\left(\frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n x_i}\right) = \frac{\mathrm{E}(\sum_{i=1}^n Y_i)}{\sum_{i=1}^n x_i} = \frac{\sum_{i=1}^n \mathrm{E}(Y_i)}{\sum_{i=1}^n x_i} \\
&= \frac{\sum_{i=1}^n \beta x_i}{\sum_{i=1}^n x_i} = \beta \frac{\sum_{i=1}^n x_i}{\sum_{i=1}^n x_i} = \underline{\underline{\beta}} \\
\mathrm{Var}(\hat{\beta}) &= \mathrm{Var}\left(\frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n x_i}\right) = \frac{\mathrm{Var}(\sum_{i=1}^n Y_i)}{(\sum_{i=1}^n x_i)^2} = \frac{\sum_{i=1}^n \mathrm{Var}(Y_i)}{(\sum_{i=1}^n x_i)^2} = \frac{\sum_{i=1}^n x_i\sigma^2}{\sum_{i=1}^n x_i} \\
&= \sigma^2 \frac{\sum_{i=1}^n x_i}{(\sum_{i=1}^n x_i)^2} = \underline{\underline{\frac{\sigma^2}{\sum_{i=1}^n x_i}}}
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}(\tilde{\beta}) &= \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^n \frac{Y_i}{x_i}\right) = \frac{1}{n}\sum_{i=1}^n \mathrm{E}\left(\frac{Y_i}{x_i}\right) = \frac{1}{n}\sum_{i=1}^n \frac{\mathrm{E}(Y_i)}{x_i} \\
&= \frac{1}{n}\sum_{i=1}^n \frac{\beta x_i}{x_i} = \frac{\beta}{n}\sum_{i=1}^n \frac{x_i}{x_i} = \frac{\beta}{n} n = \underline{\underline{\beta}} \\
\mathrm{Var}(\tilde{\beta}) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n \frac{Y_i}{x_i}\right) = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}\left(\frac{Y_i}{x_i}\right) = \frac{1}{n^2}\sum_{i=1}^n \frac{\mathrm{Var}(Y_i)}{x_i^2} \\
&= \frac{1}{n^2}\sum_{i=1}^n \frac{x_i\sigma^2}{x_i^2} = \underline{\underline{\frac{\sigma^2}{n^2}\sum_{i=1}^n \frac{1}{x_i}}}
\end{aligned}$$

Vi ser at begge estimatorene er forventingsrette. Vi foretrekker den med minst varians. Med oppitte tall for x_i'ene får vi

$\mathrm{Var}(\hat{\beta}) = \sigma^2 \cdot 0.00299$ og $\mathrm{Var}(\tilde{\beta}) = \sigma^2 \cdot 0.0107$

Det vil si at vi $\underline{\underline{\text{foretrekker } \hat{\beta}}}$

d)

$$H_0 : \beta = 0.56 \quad \text{mot } H_1 : \beta > 0.56$$

$\hat{\beta}$ blir normalfordelt siden den er en lineærkombinasjon av uavhengige, normalfordelte variabler.

Under H_0 vil en ha at $\mathrm{E}(\hat{\beta}) = 0.56$ og $\mathrm{Var}(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^n x_i}$

Vi benytter testobservatoren

$$U = \frac{\hat{\beta} - 0.56}{\sqrt{\frac{\sigma^2}{\sum_{i=1}^n x_i}}} \sim n(u; 0, 1) \quad \text{under } H_0$$

Vi forkaster H_0 dersom $U > k$, der k bestemmes fra kravet

$$P(\text{Forkast } H_0 \text{ når } H_0 \text{ er riktig }) = 0.05$$

det vil si at $k = u_{0.05} = 1.645$

Innsatt observasjonene:

$$\hat{\beta} = 0.584 \quad \sigma^2 = 0.1^2 \quad \sum_{i=1}^{n} x_i = 335 \quad \Rightarrow U = \frac{0.584 - 0.56}{\sqrt{\frac{0.1^2}{335}}} = 4.38 > k$$

Det vil si $\underline{\underline{\text{Forkast } H_0}}$. Vi vil da påstå at bilen bruker mer bensin enn forhandleren sier.

e) Vet at

$$V = \frac{\hat{\beta} - \beta}{\sqrt{\frac{\sigma^2}{\sum_{i=1}^{n} x_i}}} \sim n(v; 0, 1)$$

$$\begin{aligned} P\left(-u_{0.025} \leq V \leq u_{0.025}\right) &= 0.95 \\ P\left(-u_{0.025} \leq \frac{\hat{\beta} - \beta}{\sqrt{\frac{\sigma^2}{\sum_{i=1}^{n} x_i}}} \leq u_{0.025}\right) &= 0.95 \\ P\left(\hat{\beta} - u_{0.025}\sigma\sqrt{\frac{1}{\sum_{i=1}^{n} x_1}} \leq \beta \leq \hat{\beta} + u_{0.025}\sigma\sqrt{\frac{1}{\sum_{i=1}^{n} x_1}}\right) &= 0.95 \end{aligned}$$

Vi finner da et 95% konfidensintervall for β

$$\left[\hat{\beta} - u_{0.025}\sigma\sqrt{\frac{1}{\sum_{i=1}^{n} x_1}} \; , \; \hat{\beta} + u_{0.025}\sigma\sqrt{\frac{1}{\sum_{i=1}^{n} x_1}}\right]$$

Innsatt for tallverdiene $\hat{\beta} = 0.584$, $\sigma = 0.1$, $\sum_{i=1}^{n} x_i = 335$ og $u_{0.025} = 1.96$ får vi da

$$\underline{\underline{[0.573, 0.595]}}$$

Oppgave 2

a) Minste kvadraters metode minimerer $\text{SSE}(\beta) = \sum_{i=1}^{11}(y_i - \beta x_i)^2$.

$$\frac{d\text{SSE}}{d\beta} = 0$$

$$\sum_{i=1}^{11} y_i x_i - \beta \sum_{i=1}^{11} x_i^2 = 0$$

Dette tilsvarer: $\sum_{i=1}^{11} y_i x_i = \beta \sum_{i=1}^{11} x_i^2$ som gir svaret.
Innsetting gir $\hat{\beta} = 0.0567$.

Forventning og varians blir

$$E[\hat{\beta}] = \frac{\sum_{i=1}^{11} x_i E[Y_i]}{\sum_{i=1}^{11} x_i^2} = \frac{\sum_{i=1}^{11} x_i^2 \beta}{\sum_{i=1}^{11} x_i^2} = \beta$$

$$Var[\hat{\beta}] = \frac{\sum_{i=1}^{11} x_i^2 Var[Y_i]}{(\sum_{i=1}^{11} x_i^2)^2} = \frac{\sum_{i=1}^{11} x_i^2 \sigma^2}{(\sum_{i=1}^{11} x_i^2)^2} = \frac{\sigma^2}{\sum_{i=1}^{11} x_i^2}$$

b) Predikert verdi er $\hat{y}_0 = x_0\hat{\beta} = 900 \cdot 0.0567 = 51.03$.

Merk at $\hat{y}_0$ er predikert verdi som vi finner ved å følge regresjonslinja medd $x = x_0$ innsatt *estimatet* for β basert på observasjonene i oppgaven. $\hat{y}_0$ er en realisasjon av den stokastiske variablen $\hat{Y}_0 = \hat{\beta}x_0$, der $\hat{\beta}$ er den stokastiske *estimatoren* for β. Videre er y_0 den sanne hastigheten til galaksen, som er en realisasjon av den stokastiske variabelen Y_0, og vi har antatt at Y_0 er gitt ved $Y_0 = \beta x_0 + \varepsilon$. Vi har at forskjellen $\hat{y}_0 - y_0$ er en realisering av forskjellen mellom de to stokastiske variablene; $\hat{Y}_0 - Y_0$. Videre er

$\mathrm{E}(\hat{Y}_0 - Y_0) = \mathrm{E}(\hat{\beta}x_0 - (\beta x_0 + \varepsilon)) = \beta x_0 - \beta x_0 - 0 = 0.$
$\mathrm{Var}(\hat{Y}_0 - Y_0) = \mathrm{Var}(\hat{\beta}x_0 - (\beta x_0 + \varepsilon)) = \mathrm{Var}(\hat{\beta}x_0 - \varepsilon) = \frac{x_0^2\sigma^2}{\sum_{i=1}^{11} x_i^2} + \sigma^2,$

Et estimat for σ er $s = \sqrt{\frac{1}{10}9.87} = 0.993$. Vi har at $T = \frac{\hat{Y}_0 - Y_0}{s\sqrt{1+\frac{900^2}{\sum_{i=1}^{11} x_i^2}}} \sim t_{10}$.
Da blir et 95% prediksjonsintervall for denne sanne hastigheten y_0 gitt ved

$$\left(\hat{y}_0 - t_{10,0.025}s\sqrt{1+\frac{900^2}{\sum_{i=1}^{11} x_i^2}}, \hat{y}_0 + t_{10,0.025}s\sqrt{1+\frac{900^2}{\sum_{i=1}^{11} x_i^2}}\right),$$

der $t_{0.025,10} = 2.23$.
Innsetting gir $(48.5, 53.5)$.

Oppgave 3

a) (TIPS: se notat til øving 12 på hjemmesiden for utledning av andre likhetstegn)

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})y_i}{\sum_{i=1}^n (x_i - \bar{x})x_i} = \frac{\sum_{i=1}^n x_i y_i - \bar{x}\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i^2 - \bar{x}\sum_{i=1}^n x_i} \\ &= \frac{\sum_{i=1}^n x_i y_i - \frac{1}{n}\sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sum_{i=1}^n x_i^2 - \frac{1}{n}\left(\sum_{i=1}^n x_i\right)^2} = \frac{2\,752\,667 - \frac{1}{24}\cdot 925 \cdot 71\,194}{35\,727 - \frac{1}{24}\cdot 925^2} = \underline{\underline{115}}.\end{aligned}$$

$$\hat{\beta}_0 = \frac{1}{n}\sum_{i=1}^n y_i - \hat{\beta}_1 \cdot \frac{1}{n}\sum_{i=1}^n x_i = \frac{71\,194}{24} - \frac{115 \cdot 925}{24} = \underline{\underline{-1465}}.$$

We have $\mathrm{Var}(\hat{\beta}_1) = \sigma^2/(\sum_{i=1}^n (x_i - \bar{x})^2)$. Then $T = \frac{\hat{\beta}_1 - \beta_1}{s_b}$, where

$$s_b = \frac{s}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} = \frac{s}{\sqrt{\sum_{i=1}^n x_i^2 - (1/n)(\sum_{i=1}^n x_i)^2}}.$$

Based on the t-distribution $P(-t_{n-2,\alpha/2} < T < t_{n-2,\alpha/2}) = 1-\alpha$. This gives:

$$P(\hat{\beta}_1 - t_{n-2,\alpha/2}s_b < \beta_1 < \hat{\beta}_1 + t_{n-2,\alpha/2}s_b) = 1-\alpha.$$

Using $t_{22,0.025} = 2.07$ we get: $(115 \pm 2.07 \cdot 194/\sqrt{35727-(1/24)(925)^2})$, $= (115 \pm 46) = \underline{\underline{(69, 161)}}$.

b) Prediction at week 40:

$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 40 = -1465 + 115 \cdot 40 = \underline{\underline{3135}}$.

$\hat{Y}_0 - Y_0$ is Gaussian distributed. Because of unbiased estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, we have mean $E(\hat{Y}_0) = \beta_0 + \beta_1 40$. We also have $E(Y_0) = \beta_0 + \beta_1 40$. This means $E(\hat{Y}_0 - Y_0) = 0$.

We next use that $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = \bar{Y} + \hat{\beta}_1(x_0 - \bar{x})$. Recall that $\bar{Y}$ and $\hat{\beta}$ have zero covariance, i.e. $\text{Cov}(\bar{Y}, \hat{\beta}) = 0$ (you do not have to show this). The mean of the data is $\bar{x} = (1/24)(925) = 38.5$.

The variance is

$$\text{Var}(\hat{Y}_0 - Y_0) = \text{Var}(\hat{Y}_0) + \text{Var}(Y_0)$$

$$= \text{Var}(\bar{Y}) + (40-\bar{x})^2\text{Var}(\hat{\beta}_1) + \text{Var}(Y_0) = \sigma^2/n + \sigma^2 \frac{(40-\bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} + \sigma^2$$

We estimate σ^2 with s^2 and we get

$$s_0 = s\sqrt{1 + \frac{1}{n} + \frac{(40-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}} = 194\sqrt{1 + \frac{1}{24} + \frac{(40-38.5)^2}{35727-(1/24)(925)^2}} = 201$$

and get a statistic with a T-distribution: $\frac{\hat{Y}_0 - Y_0}{s_0} \sim t(n-2)$. Then,

$$P(-t_{n-2,\alpha/2} < \frac{\hat{Y}_0 - Y_0}{s_0} < t_{n-2,\alpha/2}) = 1-\alpha,$$

and this gives:

$$P(\hat{Y}_o - t_{n-2,\alpha/2} \cdot s_0 < Y_0 < \hat{Y}_0 + t_{n-2,\alpha/2} \cdot s_0) = 1-\alpha.$$

The significance level $\alpha = 0.1$ means we set $t_{22,0.95} = 1.72$. This gives prediction interval $3135 \pm 1.72 \cdot 201 \approx \underline{\underline{(2\,789, 3\,481)}}$.

The interval at week 42 has the same form, but now we have a term $(42-38.5)^2$ instead of $(40-38.5)^2$ in the variance.

This gives only a small increase in the width since $\frac{(42-38.5)^2}{35727-(1/24)(925)^2} = 0.16$ and $\frac{(40-38.5)^2}{35727-(1/24)(925)^2} = 0.0296$ are both relatively small compared with 1. We get width at week 40:

$$2 \cdot 1.72 \cdot 194 \cdot \sqrt{1 + (1/24) + 0.03} = \underline{\underline{690}},$$

and width at week 42:

$$2 \cdot 1.72 \cdot 194 \cdot \sqrt{1 + (1/24) + 0.16} = \underline{\underline{732}}.$$

The prediction interval is narrowest at the mean of the week data, i.e., at week 39.

c) From the plot we clearly see that boys are heavier than girls on average. The increase with gestational weeks seem similar, and for this reason we fit the same slope. The error level seems to be about the same for boys and girls, and there is no reason to use a different variance.

The model would predict a non-zero intercept term and different weight for boys and girls at 0 weeks, which is unntatural. Perhaps a model with different slopes and intercept 0 would be more appropriate. Then again the linear regression model is valid only within the region where data is available.

$$SSE = \sum_{i=1}^{n_b}(y_i - \beta_b - \beta_1 x_i)^2 + \sum_{i=n_b+1}^{n}(y_i - \beta_g - \beta_1 x_i)^2$$

We differentiating this expression and set the derivative equal to 0.

$$\frac{dSSE}{d\beta_b} = -\sum_{i=1}^{n_b}(y_i - \beta_b - \beta_1 x_i) = n_b\beta_b + \beta_1\sum_{i=1}^{n_b}x_i + \sum_{i=1}^{n_b}y_i = 0$$

$$\frac{dSSE}{d\beta_g} = -\sum_{i=n_b+1}^{n}(y_i - \beta_g - \beta_1 x_i) = n_g\beta_g + \beta_1\sum_{i=n_b+1}^{n}x_i + \sum_{i=n_b+1}^{n}y_i = 0$$

From these two first expressions we have:

$$\hat{\beta}_b = (1/n_b)\sum_{i=1}^{n_b}y_i - \hat{\beta}_b(1/n_b)\sum_{i=1}^{n_b}x_i$$

$$\hat{\beta}_g = (1/n_g)\sum_{i=n_b+1}^{n}y_i - \hat{\beta}_1(1/n_g)\sum_{i=n_b+1}^{n}x_i$$

Here, $n_g = n - n_b = 12$.

We have to insert these into the expression for β_1 to solve for the slope:

$$\frac{dSSE}{d\beta_1} = -\sum_{i=1}^{n_b}x_i(y_i - \beta_b - \beta_1 x_i) - \sum_{i=n_b+1}^{n}x_i(y_i - \beta_g - \beta_1 x_i)$$

$$\frac{dSSE}{d\beta_1} = -\sum_{i=1}^{n}x_i y_i + \beta_b\sum_{i=1}^{n_b}x_i + \beta_g\sum_{i=n_b+1}^{n}x_i + \beta_1\sum_{i=1}^{n}x_i^2$$

Now, inserting for the solution to the slopes we get:

$$\frac{dSSE}{d\beta_1} = -a + \beta_1 b = 0, \hat{\beta}_1 = a/b$$

where

$$a = \sum_{i=1}^{n}x_i y_i - (1/n_b)\sum_{i=1}^{n_b}y_i\sum_{i=1}^{n_b}x_i - (1/n_g)\sum_{i=n_b+1}^{n}y_i\sum_{i=n_b+1}^{n}x_i$$

$$b = \sum_{i=1}^{n} x_i^2 - (1/n_b)(\sum_{i=1}^{n_b} x_i)^2 - (1/n_g)(\sum_{i=n_b+1}^{n} x_i)^2$$

Inserting numbers we have: $a = 2\,752\,667 - (1/12) \cdot 36\,258 \cdot 460 - (1/12) \cdot 34\,936 \cdot 465 = 9\,007$, $b = 35\,737 - (1/12) \cdot 460^2 - (1/12) \cdot 465^2 = 75$.

Finally, $\hat{\beta}_1 = 9\,007/75 = \underline{\underline{120.22}}$.

$\hat{\beta}_b = 36\,258/12 - 120.22 \cdot (1/12) \cdot 460 = -\underline{\underline{1\,587}}$.

$\hat{\beta}_g = 34\,936/12 - 120.22 \cdot (1/12) \cdot 465 = -\underline{\underline{1\,747}}$.