Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

TMA4245 Statistikk
Vår 2013

Øving nummer 13, blokk II
Løsningsskisse

Oppgave 1

a) $\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{-52.57}{60} = -0.876$
$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x} = \frac{360.37}{9} + 0.876 \cdot \frac{369}{9} = 75.96$

$\hat{\sigma^2} = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{\alpha} - \hat{\beta}x_i)^2$

b) $Var(\hat{\beta}) = Var(\frac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}) = \frac{1}{(\sum_{i=1}^n (x_i - \bar{x})^2)^2} \cdot Var(\sum_{i=1}^n (x_i - \bar{x})Y_i) =$
$\frac{1}{(\sum_{i=1}^n (x_i - \bar{x})^2)^2} \cdot \sum_{i=1}^n (x_i - \bar{x})^2 Var(Y_i) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$

$H_0 : \beta = 0 \qquad H_1 : \beta \neq 0$

Under H_0 har vi følgende observator og fordeling: $T = \frac{\hat{\beta}}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \sim t_{n-2}$

Vi forkaster H_0 dersom $T > t_{n-2,\alpha/2}$ eller om $T < -t_{n-2,\alpha/2}$. Med $\alpha = 0.01$, $n = 9$ har vi at $t_{7,0.005} = 3.5$.

$T = \frac{-0.876}{\frac{1.568}{\sqrt{60}}} = -4.33 < -3.50$
Vi forkaster H_0 på nivå $\alpha = 0.01$.

Tolkningen blir at alder har betydning for løypetiden.

c) Predikert tid er $\hat{Y}_0$.
$\hat{Y}_0 = \hat{\alpha} - \hat{\beta}x_0 = 75.96 - 0.876 \cdot 46 = 35.66$

$\hat{Y}_0$ er et estimat på sann verdi Y_0. Siden Y_i ene er uavhengige og $\hat{Y}_0$ er basert på andre Y_i er enn Y_0, så er $\hat{Y}_0$ og Y_0 uavhengige.
$E(\hat{Y}_0 - Y_0) = E(\hat{\alpha}) + E(\hat{\beta})x_0 - \alpha - \beta x_0 = 0$
$Var(\hat{Y}_0 - Y_0) = Var(\hat{Y}_0) + Var(Y_0) = \sigma^2(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}) + \sigma^2 = \sigma^2 \cdot v$

Vi har at observatoren $S = \frac{\hat{Y}_0 - Y_0}{\hat{\sigma} \cdot \sqrt{v}} \sim t_{n-2}$

$P(-t_{n-2,\alpha/2} < S < t_{n-2,\alpha/2}) = 1 - \alpha$
$P(\hat{Y}_0 - t_{n-2,\alpha/2}\hat{\sigma}\sqrt{v} < Y_0 < \hat{Y}_0 + t_{n-2,\alpha/2}\hat{\sigma}\sqrt{v}) = 1 - \alpha$

Innsatt verdier $n = 9$, $\alpha = 0.05$ og $t_{7,0.0025} = 2.36$, er et 0.95 prediksjonsintervall for Y_0 gitt ved:
$\{35.66 - 2.36 \cdot 1.568 \cdot \sqrt{1 + \frac{1}{9} + \frac{(46-41)^2}{60}}, 35.66 + 2.36 \cdot 1.568 \cdot \sqrt{1 + \frac{1}{9} + \frac{(46-41)^2}{60}}\} = \{31.09, 40.23\}$

Ekstrapolasjon så langt frem i tid bør ikke gjøres. Det er ikke sikkert at modellen holder utover de x verdiene hvor vi har data. Løperen vil neppe fortsette å forbedre seg i all fremtid.

Oppgave 2

1. Modellantagelser for en lineær regresjonsmodell $Y = \beta_0 + \beta_1 x + \epsilon$:
   - Antar at forventningsverdien til y er lineær mhp regresjonsvariabelen x, altså at $E(Y) = \beta_0 + \beta_1 x$
   - Antar at feilleddene, ϵ, er uavhengige og normalfordelte med forventning 0 og samme varians, altså $\epsilon_i \sim N(0, \sigma^2),\ i = 1, \ldots, n$
2. Dersom vi kjører matlabkoden `linearreg.m` for datasett 1 (endrer koden slik at det står `dataset = load('dataset1.txt');` i kodelinje 5) får vi figurer tilsvarende Figur 1 gitt i øvigen.
   - Plot 1 viser dataene (x_i, y_i) plottet mot den beregnede lineære regresjonslinja $y = \hat{\beta}_0 + \hat{\beta}_1 x$, og vi ser at den første modellantagelsen stemmer godt. En lineær regresjonsmodell passer dataene godt, datapunktene ligger jevnt spredt rundt regresjonslinja.
   - Plot 2 viser det samme som plot 1, men i tillegg et 95% prediksjonsinterval (PI) for responsene y_0 gitt ved

$$95\%PI\ :\ \hat{y}_0 \pm t_{0.025,n-1} \cdot s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

     hvor $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$ og $s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}}$. Plottet viser dette beregnede prediksjonsintervalet for alle verdier x_0. Vi ser at alle de observerte datapunktene ligger innenfor prediksjonsintervalet, og at intervalet er ganske smalt.
   - Plot 3 viser to forskjellige plot av residualene. Det første plottet, residualplottet, plotter regresjonsvariablene x_i mot residualene $e_i = y_i - \hat{y}_i$. Dersom modellantagelen om uavhengige residualer stemmer, vil punktene i dette plottet (x_i, e_i) ligge tilfeldig spredt omkring 0 som vi ser stemmer bra for datasett 1. Det andre plottet, q-q-plottet (Normal Probability Plot), er et plot av residualene i stigende rekkefølge plottet mot den tilsvarende kvantilen for en normalfordeling. Dersom

modellantagelsen om at residualene er normalfordelte med samme varians stemmer, vil datapunktene ligge langs en lineær linje markert som en rød strek i plottet. For datasett 1 ser vi at dette stemmer greit, punktene ligger jevnt langs linja.

Konklusjonen for datasett 1 fra de tre plottene er dermed at dette datasettet oppfyller modellantagelsene for en lineær regresjonsmodell godt.

3. Dersom vi kjører matlabkoden `linearreg.m` for datasett 2 (endrer koden slik at det står `dataset = load('dataset2.txt');` i kodelinje 5) får vi figurer tilsvarende Figur 2 gitt i øvigen. Fra plot 1 og 2 ser vi nå at datapunktene (x_i, y_i) ikke ligger godt langs den beregnede lineære regresjonslinja, her ser dataen mer ut som om de kommer fra en kvadratisk modell (ser formen på en annengradsligning). Prediksjonsintervalet er nå mye bredere, som tilsier at de predikerte verdiene er veldig usikre. Fra residualplottene ser ikke residualene uavhengige ut, men ligger systematisk omkring en kvadratisk linje. Halene i q-q-plottet ligger ikke langs den lineære linja, som tilsier at residualene ikke oppfyller antagelsen om at de er normalfordelte godt. Konklusjonen for datasett 2 fra de tre plottene er dermed at dette datasettet oppfyller modellantagelsene for en lineær regresjonsmodell dårlig.
Kommentar: Dataene for datasett 2 ble simulert fra modellen $y = \beta_0 + \beta_1 x^2 + \epsilon$ *hvor* $\epsilon \sim N(0, \sigma^2)$*, altså fra en kvadratisk modell.*

4. Dersom vi kjører matlabkoden `linearreg.m` for datasett 3 (endrer koden slik at det står `dataset = load('dataset3.txt');` i kodelinje 5) får vi figurer tilsvarende Figur 3 gitt i øvigen. Fra plot 1 og 2 ser datapunktene (x_i, y_i) ut til å ligge ganske greit spredt rundt den beregnede lineære regresjonslinja, men at avviket fra linja øker for økende verdier av x. Prediksjonsintervalet er som for datasett 2 igjen bredt. Fra residualplottene ser det ikke ut som residualene har samme varians, de ligger greit spredt rundt 0, men det ser ut til at variansen øker for økende verdier av x. Datapunktene i q-q-plottet overlapper den lineære linja dårlig, som tilsier at residualene ikke oppfyller antagelsen om at de er normalfordelte godt. Konklusjonen for datasett 3 fra de tre plottene er dermed at dette datasettet oppfyller modellantagelse 1 for lineær forventningsverdi rimelig greit, men modellantagelse 2 for residualene oppfylles dårlig.
Kommentar: Dataene for datasett 3 ble simulert fra modellen $y = \beta_0 + \beta_1 x + \epsilon_x$ *hvor* $\epsilon_x \sim N(0, (x\sigma)^2)$*, altså fra en lineær modell hvor variansen er avhengig av* x*.*

Oppgave 3

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

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

Dette oppnås ved å velge

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

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 4

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 $\hat{y}_0 = x_0\hat{\beta} = 900 \cdot 0.0567 = 51.03$.

Vi har at $\hat{y}_0 - Y_0 = \hat{\beta}x_o - \beta x_0 - \epsilon_0 = x_0(\hat{\beta} - \beta) - \epsilon_0$, dvs $E[\hat{y}_0 - Y_0] = E[x_0(\hat{\beta} - \beta)] = 0$, og

$$Var[\hat{y}_0 - Y_0] = Var[x_0(\hat{\beta} - \beta) - \epsilon_0] = \frac{x_0^2\sigma^2}{\sum_{i=1}^{11} x_i^2} + \sigma^2.$$

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 observasjon Y_0 gitt ved

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

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