TMA4245 Statistikk
Vår 2017

Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

Anbefalt øving 12
Løsningsskisse

Oppgave 1

a) Minste kvadraters metode tilpasser en linje til punktene ved å velge den linja som minimerer kvadratsummen

$$\mathrm{SSE} = \sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$$

av avstanden fra hvert punkt til linja. Derivasjon av SSE med hensyn på parametrene α og β gir

$$\frac{d\mathrm{SSE}}{d\alpha} = -2\sum_i (y_i - \alpha - \beta x_i) \quad \text{og} \quad \frac{d\mathrm{SSE}}{d\beta} = -2\sum_i x_i(y_i - \alpha - \beta x_i).$$

Setter vi de deriverte lik null, får vi

$$\sum_{i=1}^{n}(y_i - \alpha - \beta x_i) = 0 \quad \text{og} \quad \sum_{i=1}^{n} x_i(y_i - \alpha - \beta x_i) = 0,$$

og, når vi deler på n,

$$\bar{y} - \alpha - \beta\bar{x} = 0 \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \alpha\bar{x} - \beta\cdot\frac{1}{n}\sum_{i=1}^{n} x_i^2 = 0.$$

Løser den første likningen for α, og får

$$\alpha = \bar{y} - \beta\bar{x},$$

som innsatt i den andre likningen gir

$$\frac{1}{n}\sum_{i=1}^{n} x_i y_i - (\bar{y} - \beta\bar{x})\bar{x} - \beta\cdot\frac{1}{n}\sum_{i=1}^{n} x_i^2 = 0,$$

$$\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{y}\bar{x} + \beta\left(\bar{x}^2 - \frac{1}{n}\sum_{i=1}^{n} x_i^2\right) = 0 \Rightarrow \beta = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2}.$$

Ganger vi med n i teller og nevner i det siste uttrykket, får vi

$$\beta = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}.$$

For få de oppgitte estimatorene bytter vi ut y_i med den tilsvarende tilfeldige variabelen Y_i, altså

$$\widehat{\beta} = \frac{\sum_{i=1}^n x_i Y_i - n\bar{x}\bar{Y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \quad \text{og} \quad \widehat{\alpha} = \bar{Y} - \widehat{\beta}\bar{x}.$$

b) Utgangspunktet er

$$P\left(-t_{n-2,0.025} < \frac{(\hat{\beta}-\beta)}{s/\sqrt{\sum_{i=1}^n (x_i-\bar{x})^2}} < t_{n-2,0.025}\right) = 0.95$$

Løser hver av ulikhetene for β og får

$$-t_{n-2,0.025} < \frac{(\hat{\beta}-\beta)}{s/\sqrt{\sum_{i=1}^n (x_i-\bar{x})^2}} \Rightarrow -\frac{-st_{n-2,0.025}}{\sqrt{\sum_{i=1}^n (x_i-\bar{x})^2}} < \hat{\beta}-\beta$$

$$\Rightarrow \hat{\beta} + \frac{st_{n-2,0.025}}{\sqrt{\sum_{i=1}^n (x_i-\bar{x})^2}} > \beta$$

og

$$\frac{(\hat{\beta}-\beta)}{s/\sqrt{\sum_{i=1}^n (x_i-\bar{x})^2}} < t_{n-2,0.025} \Rightarrow \hat{\beta}-\beta < \frac{st_{n-2,0.025}}{\sqrt{\sum_{i=1}^n (x_i-\bar{x})^2}}$$

$$\Rightarrow \beta > \hat{\beta} - \frac{st_{n-2,0.025}}{\sqrt{\sum_{i=1}^n (x_i-\bar{x})^2}}.$$

Dermed har vi

$$P\left(\hat{\beta} - \frac{st_{n-2,0.025}}{\sqrt{\sum_{i=1}^n (x_i-\bar{x})^2}} < \beta < \hat{\beta} + \frac{st_{n-2,0.025}}{\sqrt{\sum_{i=1}^n (x_i-\bar{x})^2}}\right) = 0.95,$$

Og konfidensintervallet blir altså

$$\left(\hat{\beta} - \frac{st_{n-2,0.025}}{\sqrt{\sum_{i=1}^n (x_i-\bar{x})^2}}, \hat{\beta} + \frac{st_{n-2,0.025}}{\sqrt{\sum_{i=1}^n (x_i-\bar{x})^2}}\right).$$

Vi har $n = 29$ og tabelloppslag gir kvantilen $t_{n-2,0.025} = t_{27,0.025} = 2.0518$. Innsetting av tallverdier gir estimatet $\hat{\beta} = -6364.6/40169 = -0.1584$. Vinnertiden forventes å forkortes med $4 \cdot 0.1584 \approx 0.63$ sekunder mellom etterfølgende olympiske leker. Videre er 95%-konfidensintervallet for stigningstallet lik $(-0.1925, -0.1244)$.

c) Vi lar $x_0 = 2020$, og vi har $\hat{\alpha} = 109.0114 + 0.1584 \cdot 1956.6 = 418.9368$. Den predikerte tiden er $\hat{Y}_0 = \hat{\alpha} + 2020\hat{\beta} = 418.9368 - 0.1584 \cdot 2020 = 98.8768$, altså ca. 1 minutt og 39 sekunder. Vinnertiden i 2020 har 95%-prediksjonsintervall

$$\hat{Y}_0 \pm t_{n-2,0.025} s \sqrt{1 + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2} + 1/n}.$$

Med tallverdier innsatt blir det $(91.62, 106.14)$.

d) Vi har $\hat{Y}_0 = \hat{\alpha} + x_0\hat{\beta} = 90$, som betyr at

$$x_0 = \frac{90 - \hat{\alpha}}{\hat{\beta}} = \frac{90 - 418.9368}{-0.1584} = 2076.024$$

Siden $x_0 > 2076$ forventer vi strengt tatt ikke at 90-sekundersgrensen brytes under OL i 2076, men først under neste OL, altså i 2080. Tar man den store usikkerheten i betraktning, fremstår imidlertid 2076 som et like godt svar som 2080.

Modellantakelser: Det ser ut til at vinnertidene følger en ikkelineær trend i tid. Om vi bruker den tilpassede modellen til å ekstrapolere bakover i tid, ser vi at den tilsier at vinnertiden i år 0 ville vært 419 sekunder, hvilket er urimelig. Ekstrapolererer vi tilstrekkelig langt framover i tid, predikerer modellen dessuten negative vinnertider, hvilket er umulig.

Modellantakelsene kan kontrolleres ved hjelp av residualplott. Ser residualene $e_i = Y_i - \hat{Y}_i$ ut til å ha en trend? Ifølge modellen bør de være nærmest uavhengige og identisk normalfordelt.

Oppgave 2

a) $Y \sim n(y; 500, 80)$. Transformerer Y til standard $N(0,1)$-normalfordeling.

$$\begin{aligned}\text{Prob}(Y > 550) &= \text{Prob}(\frac{Y-500}{80} > \frac{550-500}{80}) = \text{Prob}(Z > \frac{5}{8}) \\ &= 1 - \text{Prob}(Z \leq \frac{5}{8}) = 1 - \Phi(0.625) = 1 - 0.734 = 0.266.\end{aligned}$$

$Y_1 - Y_2 \sim n(y; 0, \sqrt{2} \cdot 80)$. (Lineærkombinasjonen av to uavhengige normalfordelinger er normalfordelt, sjekk forventningsverdi og varians ved de vanlige regnereglene.)

Da kan vi regne ut sannsynligheten for at målingene avviker med mer enn 80 g/tonn.

$$\begin{aligned}\text{Prob}(|Y_1 - Y_2| > 80) &= 1 - \text{Prob}(-80 < Y_1 - Y_2 < 80) \\ &= 1 - \text{Prob}(\frac{-80}{80\sqrt{2}} < \frac{Y_1 - Y_2}{80\sqrt{2}} < \frac{80}{80\sqrt{2}}) \\ &= 1 - \text{Prob}(-\frac{\sqrt{2}}{2} < Z < \frac{\sqrt{2}}{2}) = 2\text{Prob}(Z \leq \frac{-\sqrt{2}}{2}) = 2\Phi(-0.707) \\ &= 2 \cdot 0.24 = 0.48.\end{aligned}$$

b) Setter inn $\overline{x} = 20$, $x_1 = \ldots = x_5 = 0$ og $x_6 = \ldots = x_{10} = 40$ i uttrykket for B.

$$\begin{aligned}B &= \frac{\sum_{j=1}^{5} -20Y_j + \sum_{j=6}^{10} 20Y_j}{\sum_{j=1}^{10} 20^2} = \frac{20\left(\sum_{j=6}^{10} Y_j - \sum_{j=1}^{5} Y_j\right)}{10 \cdot 20^2} \\ &= \frac{\sum_{j=6}^{10} Y_j - \sum_{j=1}^{5} Y_j}{200}, \text{som skulle vises.}\end{aligned}$$

$$A = \overline{Y} - B\overline{x} = \frac{1}{10}\sum_{j=1}^{10} Y_j - \frac{20}{200}\left(\sum_{j=6}^{10} Y_j - \sum_{j=1}^{5} Y_j\right) = \frac{1}{5}\sum_{j=1}^{5} Y_j.$$

A er skjæringspunktet regresjonslinja har med y-aksen. Det er kanskje ikke så rart at gjennomsnittet av målingene ved $x = 0$ er et estimat for denne verdien? (I hvert fall når målingene bare er gjort for to x-verdier.)

$$\mathrm{Var}(B) = \frac{1}{200^2}\left(\sum_{j=6}^{10}\mathrm{Var}(Y_j) + \sum_{j=1}^{5}\mathrm{Var}(Y_j)\right) = \frac{10\sigma^2}{200^2} = \frac{\sigma^2}{4000}.$$

c) Med bare to målepunkter, kan vi estimere variansen i hver ende for seg, dvs at vi beregner s_V^2 og s_E^2. (Husk at målingene ikke har samme forventningsverdi i de to endene av gruva, så vi kan ikke se på alle som ett datasett.) Ettersom vi antar samme varians i begge ender, er gjennomsnittet av s_V^2 og s_E^2 et godt estimat for σ^2.

Mer formelt, vi har en to-utvalgssituasjon, og kan da bruke s_p^2 fra pensum. Denne sikrer χ^2-fordeling og T-fordeling. Brukes estimatoren for variansen fra regresjonsanalysen, får en også samme resultat.

$$\begin{aligned} s^2 &= \frac{1}{2}\left(s_V^2 + s_E^2\right) = \frac{1}{2}\left(\frac{\sum_{j=1}^{5}(y_j - \overline{y}_V)^2}{5-1} + \frac{\sum_{j=6}^{10}(y_j - \overline{y}_E)^2}{5-1}\right) \\ &= \frac{1}{8}\left(\sum_{j=1}^{5}(y_j - \overline{y}_V)^2 + \sum_{j=6}^{10}(y_j - \overline{y}_E)^2\right) = \frac{26064 + 22720}{8} = 6098. \end{aligned}$$

Hypotesene blir: H_0: $\beta = 12$ mot H_1: $\beta > 12$.

Vi baserer testen på estimatoren B. Siden variansen til B er ukjent, bruker vi estimatet $S_B^2 = \frac{s^2}{4000} = 1.525$ i stedet for $\frac{\sigma^2}{4000}$.

Testobservatoren, $\frac{B-12}{S_B}$, er T-fordelt med 8 frihetsgrader. Det er $n-2$ frihetsgrader denne gangen, fordi vi bruker "pooled" varians, eller, som sagt, variansestimatoren fra regresjonsanalysen. (Estimert varians er basert på to gjennomsnitt, $\overline{y}_V$ og $\overline{y}_E$. Da er det ikke så urimelig at vi mister to frihetsgrader?) Med oppgitte data blir stigningstallet

$$b = \frac{\sum_{j=6}^{10} y_j - \sum_{j=1}^{5} y_j}{200} = \frac{\overline{y}_E - \overline{y}_V}{40} = 17.$$

Gjennomfører hypotesetesten.

$$\frac{b-12}{s_B} = \frac{17-12}{\sqrt{1.525}} = 4.05 > t_{0.05,8} = 1.86,$$

som betyr at vi forkaster nullhypotesen på signifikansnivå 5%.

d) Fra det første uttrykket for B får vi

$$\mathrm{Var}(B) = \frac{\sigma^2}{\sum_{j=1}^{n}(x_j - \overline{x})^2}.$$

Variansen er liten for $\sum_{j=1}^{n}(x_j - \overline{x})^2$ stor. Altså vil vi ha alle $|x_j - \overline{x}|$ så store som mulig. Når $\overline{x}$ er fast, bør x_j-ene legges til endene, som i denne oppgaven. (Det kan være andre grunner til å spre målepunktene, f.eks. for å vurdere om dataene tilnærmet følger en rett linje, her var det antatt kjent.)

$$\mathrm{Var}(Y_0 - \widehat{Y}_0) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{j=1}^{n}(x_j - \overline{x})^2}\right) = \frac{11}{10} \cdot \sigma^2$$

når $x_0 = \overline{x}$. Punktestimatet blir $\widehat{y}_0 = a + bx_0 = \overline{y}_V + 17 \cdot 20 = 470$.

Vi benytter fortsatt estimatet S^2 for σ^2, derfor fortsatt T-fordeling med $n-2$ frihetsgrader. Prediksjonsintervallet blir derfor

$$(\widehat{y}_0 \pm t_{0.025,8} \cdot s\sqrt{\frac{11}{10}}) = (470 \pm 2.306 \cdot \sqrt{6098} \cdot \sqrt{1.1}) = (281.1, 658.9).$$

Den nye målingen, 600 g/tonn, ligger innenfor prediksjonsintervallet, så vi kan ikke konkludere med at den eller modellen er urimelig.

Oppgave 3

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

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

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

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

Forventning og varians blir

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

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

b) Vi laster inn dataene i Matlab og tilpasser den oppgitte modellen på følgende måte:

```
% Skriv inn observasjoner
x = [22 68 108 137 255 315 390 405 685 700 1100];
y = [1.2 3.8 5.1 7.5 14.9 19.2 21.4 23 39.2 41.6 60.8];


% Tilpass modell
mdl = fitlm(x, y,'Intercept', false);
```

```
% Skriv ut modell
mdl

mdl =


Linear regression model:
    y ~ x1

Estimated Coefficients:
          Estimate       SE         tStat       pValue
          --------    ----------    ------    ----------

    x1    0.056691    0.00060617    93.523    4.7825e-16


Number of observations: 11, Error degrees of freedom: 10
Root Mean Squared Error: 0.993
```

Merk at kommandoen `mdl` skriver ut den tilpassede modellen. Fra utskriften leser vi at $\hat{\beta} = 0.056691$ og at den tilhørende p-verdien er 4.7825×10^{-16}. Vi vil derfor forkaste H_0. Den lineære modellen er plottet i Figur 3, og vi ser at den lineære tilpasningen samsvarer godt med observasjonene.

c) Vi plotter den tilpassede modellen, og lager normalsannsynlighetsplott og residualplott som følger:

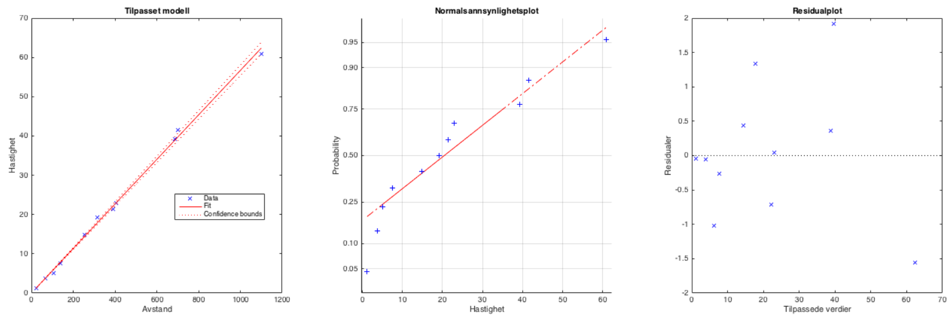
```
% Plott modellen og observasjoner
figure;
subplot(1,3,1)
plot(mdl);
xlabel('Avstand')
ylabel('Hastighet')
title('Tilpasset modell')


% Normalsannsynlighetsplott
subplot(1,3,2)
normplot(y)
xlabel('Hastighet')
title('Normalsannsynlighetsplott');

% Plott residualer
subplot(1,3,3)
plotResiduals(mdl,'fitted');
xlabel('Tilpassede verdier')
ylabel('Residualer')
```

```
title('Residualplott')
```

I Figur 3 ser vi fra normalsannsynlighetsplottet at normalantakelsen er rimelig. I tillegg virker residualene å være tilfeldig fordelt, men det kan være en svak indikasjon på at variansen øker med x.



Figur 1: Tilpasset lineær modell, normalsannsynlighetsplott og residualplott.

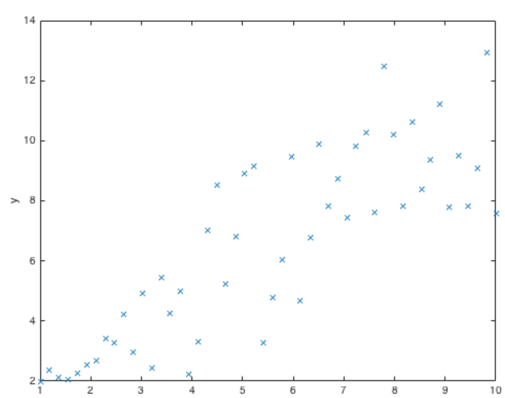
d) Vi predikerer en ny observasjon som følger i Matlab:

```
% Prediker ny observasjon
xpred = 900;
[ypred, yci] = predict(mdl, xpred, 'Prediction','observation');
```

Vi får predikert verdi 51.0220 med 95% prediksjonsintervall: $(48.4969, 53.5471)$.

Oppgave 4

a) Fra Figur 2 ser vi at y øker når x øker, derfor er korrelasjonen positiv.



Figur 2: Spredningsplott av (x_i, y_i) for $i = 1, \ldots, 50$.

```
% Les inn data
```

```
A = load('anb12.txt');
x = A(:, 1);
y = A(:, 2);

% Plott x mot y
figure
plot(x,y, 'x')
xlabel('x')
ylabel('y')
```

Vi tilpasser den lineære modellen på følgende måte i Matlab:

```
% Tilpass modell
mdl = fitlm(x, y);

% Skriv ut modell
mdl

mdl =


Linear regression model:
    y ~ 1 + x1

Estimated Coefficients:
                   Estimate       SE        tStat       pValue
                   ________    ________    ______    __________

    (Intercept)     1.1549      0.55791      2.07      0.043855
    x1             0.97084     0.091379    10.624    3.3607e-14


Number of observations: 50, Error degrees of freedom: 48
Root Mean Squared Error: 1.71
R-squared: 0.702,  Adjusted R-Squared 0.695
F-statistic vs. constant model: 113, p-value = 3.36e-14
```

Fra utskriften i Matlab ser vi at $\hat{\alpha} = 1.1549$ og $\hat{\beta} = 0.97084$. Siden p-verdien er 0.043855 vil vi ikke forkaste H_0 ved 5% signifikansnivå.

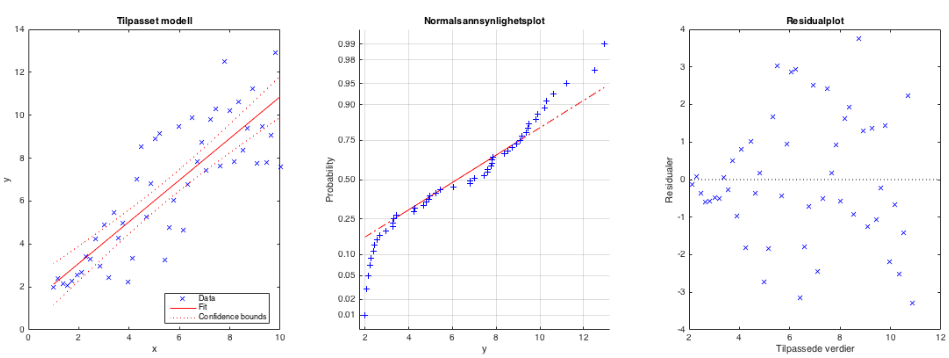
Vi kan plotte den tilpassede modellen, normalsannsynlighetsplottet og residualplottet som følger:

```
% Plott modellen og observasjoner
figure;
subplot(1,3,1)
plot(mdl);
xlabel('x')
ylabel('y')
title('Tilpasset modell')
```

```
% Normalsannsynlighetsplott
subplot(1,3,2)
normplot(y)
xlabel('y')
title('Normalsannsynlighetsplott');

% Plott residualer
subplot(1,3,3)
plotResiduals(mdl,'fitted');
xlabel('Tilpassede verdier')
ylabel('Residualer')
title('Residualplott')
```

Fra observasjonene ser vi at en lineær modell passer nokså godt siden forventningen til Y er lineær i x. Fra plottet av den tilpassede modellen ser vi at støyleddene ser ut til å være normalfordelte siden observasjonene er jevnt fordelt over og under regresjonslinjen. Fra residualplottet ser vi variansen øker med x. Altså er *ikke* kravet om konstant varians oppfylt.



Figur 3: Tilpasset lineær modell, normalsannsynlighetsplott og residualplott.