STK1110 - Statistiske metoder og dataanalyse høsten 2019
Løsningsforslag til oppgaver i læreboka for uke 43

Vinnie Ko & Ørnulf Borgan

October 23, 2019

Lærebok: *Modern Mathematical Statistics with Applications (2nd ed.)* - Devore & Berk

Exercise 10.41

a) Vi bruker formelen på side 513 i læreboka for å regne ut konfidensintervallet:

$$\left[\overline{d} - t_{\frac{\alpha}{2},n-1} \cdot \frac{s_D}{\sqrt{n}},\ \overline{d} + t_{\frac{\alpha}{2},n-1} \cdot \frac{s_D}{\sqrt{n}}\right] = \left[0.000246 - 2.18 \cdot \frac{0.000331}{\sqrt{13}},\ 0.000246 + 2.18 \cdot \frac{0.000331}{\sqrt{13}}\right]$$
$$= [0.000046, 0.000446]$$

Siden $0 \notin [0.000046, 0.000446]$, forkaster vi $H_0 : \mu_D = 0$ (mot $H_a : \mu_D \neq 0$) på 5% signifikansnivå og konkluderer at $\mu_D \neq 0$.

```
> # Leser data fra web:
> shovel.data = read.table("https://www.uio.no/studier/emner/matnat/math/STK1110/data/
    exe10-41.txt", header=T)
>
> # Regner ut differansene:
> shovel.data.diff = shovel.data[,"convent"] - shovel.data[,"perf"]
> n = length(shovel.data.diff)
>
> # Regner ut gjennomsnitt og standardavvik:
> diff.mean = mean(shovel.data.diff)
> print(diff.mean)
[1] 0.0002461538
> diff.sd = sd(shovel.data.diff)
> print(diff.sd)
[1] 0.0003306947
>
> # Bestemmer kritisk verdi:
> alpha = 0.05
> critical.value = qt(1-alpha/2, df=n-1)
> print(critical.value)
[1] 2.178813
>
> # Beregner konfidensintervallet:
> CI = c(
+   diff.mean - critical.value*diff.sd/sqrt(n),
+   diff.mean + critical.value*diff.sd/sqrt(n)
+ )
> print(round(CI,6))
[1] 0.000046 0.000446
```

b) Vi vil teste $H_0 : \mu_D = 0$ mot $H_a : \mu_D > 0$ med signifikansnivå $\alpha = 0.05$.

Basert på resultatene på side 511 i læreboka, bruker vi her testobservatoren (med $\Delta_0 = 0$)

$$t = \frac{\overline{d} - \Delta_0}{s_D/\sqrt{n}} = \frac{0.000246 - 0}{0.000331/\sqrt{13}} = 2.68$$

og forkaster H_0 så sant $t \geq t_{\alpha,n-1} = t_{0.05,12} = 1.78$. Her er $t = 2.68 > 1.78$ så vi forkaster nullhypotesen på nivå 5%.

P-verdier er her (der Φ_{12} er den kumulative t-fordelingen med 12 frihetsgrader)

$$p = 1 - \Phi_{12}(2.68) = 1 - 0.990 = 0.010$$

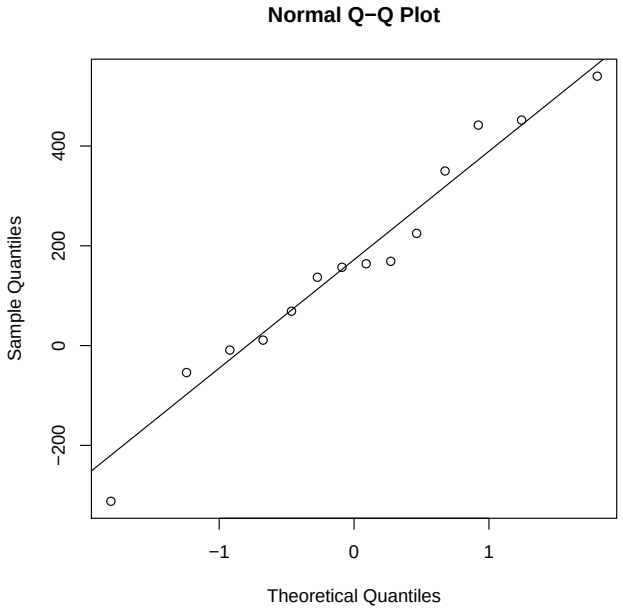
```
> # testobservator
> test.statistic = (diff.mean - 0)/(diff.sd/sqrt(n))
> print(test.statistic)
[1] 2.683806
> # p-verdi
> p.value = 1 - pt(q=test.statistic, df=n-1)
> print(p.value)
[1] 0.009948221
```

Exercise 10.42

a) Med følgende kode lager vi et Q-Q plot, som er vist i Figur **??**.

```
> # Leser data fra web.
> infant.data = read.table("https://www.uio.no/studier/emner/matnat/math/STK1110/data/
    exe10-42.txt", header=T)
> infant.data.diff = infant.data[,"Diff"]
> n = length(infant.data.diff)
>
> # QQ plot
> qqnorm(infant.data.diff)
> qqline(infant.data.diff)
```

Fra Q-Q plotet kan vi se at punktene grovt sett ligger på en rett linje. Dermed konkluderer vi at det er rimelig å anta normalfordeling.



Figur 1: Q-Q plot for data fra Exercise 10.42.

b) Vi vil teste $H_0 : \mu_D = \mu_1 - \mu_2 = 0$ mot $H_a : \mu_D \neq 0$ ved å bruke testen beskrevet på side 511 i læreboka (med $\Delta_0 = 0$). Vi bruker testobservatoren

$$t = \frac{\overline{d} - \Delta_0}{s_D/\sqrt{n}} = \frac{167.21 - 0}{228.21/\sqrt{14}} = 2.74$$

og får P-verdien (der Φ_{13} er den kumulative t-fordelingen med 13 frihetsgrader)

$$p = 2(1 - \Phi_{13}(2.74)) = 2(1 - 0.9916) = 0.0168.$$

Siden P-verdien er mindre enn 0.05, forkaster vi H_0 på 5% nivå.

```
> # Regner ut gjennomsnitt og standardavvik:
> diff.mean = mean(infant.data.diff)
> print(diff.mean)
[1] 167.2143
> diff.sd = sd(infant.data.diff)
> print(diff.sd)
[1] 228.2108
>
> # testobservator
> test.statistic = (diff.mean - 0)/(diff.sd/sqrt(n))
> print(test.statistic)
[1] 2.741581
>
> # p-verdi
> p.value = 2*(1 - pt(test.statistic, df=n-1))
> print(p.value)
[1] 0.01680597
```

c) Hvis vi, ved en feiltagelse, hadde brukt en t-test for to utvalg, ville vi ha brukt testobservatoren beskrevet på sidene 504-505 i læreboka (hvis vi forutsetter lik varians i de to gruppene). Her er estimatet for det felles standardavviket gitt ved (siden vi her har $m = n = 14$)

$$s_p = \sqrt{\frac{m-1}{m+n-2}s_1^2 + \frac{n-1}{m+n-2}s_2^2} = \sqrt{\frac{1}{2}s_1^2 + \frac{1}{2}s_2^2} = \sqrt{\frac{1}{2}352.970^2 + \frac{1}{2}234.042^2} = 299.5$$

Testobservatoren ville dermed ha blitt

$$t = \frac{\overline{x} - \overline{y} - 0}{s_p\sqrt{\frac{1}{m} + \frac{1}{n}}} = \frac{167.21 - 0}{299.5\sqrt{\frac{1}{14} + \frac{1}{14}}} = 1.477$$

og vi ville ha fått P-verdien (der Φ_{26} er den kumulative t-fordelingen med $2 \cdot 14 - 2 = 26$ frihetsgrader)

$$p = 2(1 - \Phi_{26}(1.477)) = 2(1 - 0.924) = 0.152.$$

Siden P-verdien er større enn 0.05, ville vi ikke ha forkastet H_0 på 5% nivå hvis vi (feilaktig) hadde brukt en t-test for to utvalg.

```
> # Beregner estimat for standardavviket:
> sp=sqrt(var(infant.data[,"Isotopic"])/2+var(infant.data[,"Test"])/2)
> print(sp)
[1] 299.4689
>
> # Beregner testobservatoren
> test.statistic = (mean(infant.data[,"Isotopic"]) - mean(infant.data[,"Test"]) - 0)/
    sqrt(sp*sqrt(1/n+1/n))
> print(test.statistic)
[1] 1.477306
>
> # p-verdi
> p.value = 2*(1 - pt(test.statistic, df=2*n-2))
> print(p.value)
[1] 0.1516052
```

Exercise 11.4

Vi vil teste $H_0 : \mu_1 = \mu_2 = \mu_3$ mot H_a : minst to av μ_i-ene er ulike.

Vi vil bruke F-testen beskrevet på side 559 i læreboka. Testobservatoren er

$$F = \frac{\text{MSTr}}{\text{MSE}} = \frac{\text{SSTr}/(I-1)}{\text{SSE}/I(J-1)}$$

der $\text{SSTr} = J\sum_{i=1}^{I}\left(\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot}\right)^2$ og $\text{SSE} = \sum_{i=1}^{I}\sum_{j=1}^{J}\left(X_{ij} - \overline{X}_{i\cdot}\right)^2$.

Hvis nullhypotesen er riktig, er testobservatoren F-fordelt med $I - 1 = 3 - 1 = 2$ og $I(J-1) = 3(10-1) = 27$ frihetsgrader. Så med signifikansnivå $\alpha = 0.01$, forkaster vi nullhypotesen så sant $F > F_{0.01,2,27} = 5.49$.

Ved å bruke dataene gitt i oppgaven får vi

$$\begin{aligned}
\text{MSTr} &= \frac{\text{SSTr}}{I-1} \\
&= \frac{J}{I-1}\sum_{i=1}^{I}\left(\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot}\right)^2 \\
&= \frac{10}{3-1}\left[(1.63-1.5367)^2 + (1.56-1.5367)^2 + (1.42-1.5367)^2\right] \\
&= 0.1143
\end{aligned}$$

og

$$\begin{aligned}
\text{MSE} &= \frac{\text{SSE}}{I(J-1)} \\
&= \frac{1}{I(J-1)}\sum_{i=1}^{I}\sum_{j=1}^{J}\left(X_{ij} - \overline{X}_{i\cdot}\right)^2 \\
&= \frac{1}{I}\sum_{i=1}^{I}\frac{1}{J-1}\sum_{j=1}^{J}\left(X_{ij} - \overline{X}_{i\cdot}\right)^2 \\
&= \frac{1}{I}\sum_{i=1}^{I} S_i^2 \\
&= \frac{1}{3}\left[0.27^2 + 0.24^2 + 0.26^2\right] \\
&= 0.066
\end{aligned}$$

som gir

$$f = \frac{\text{MSTr}}{\text{MSE}} = \frac{0.1143}{0.066} = 1.73.$$

Siden vi har $f = 1.73 < 5.49$, forkaster vi ikke H_0 på nivå 1%, så det er ikke signifikant forskjell på de tre gruppene.

```
> mean.vec = c(1.63,1.56,1.42)
> sd.vec = c(0.27,0.24,0.26)
>
> MSTr = 10/(3-1)*sum((mean.vec - mean(mean.vec))^2)
> print(MSTr)
[1] 0.1143333
>
> MSE = (1/3)*sum(sd.vec^2)
> print(MSE)
```

```
[1] 0.06603333
>
> test.statistic = MSTr/MSE
> print(test.statistic)
[1] 1.731449
>
> # critical value
> alpha = 0.01
> critical.value = qf(1-alpha, df1=2, df2=27)
> print(critical.value)
[1] 5.488118
>
> # p-value
> p.value = 1 - pf(test.statistic, df1=2, df2=27)
> print(p.value)
[1] 0.1961099
```

Exercise 11.5

Vi vil teste $H_0 : \mu_1 = \mu_2 = \mu_3$ mot H_a : minst to av μ_i-ene er ulike.

Vi vil bruke F-testen beskrevet på side 559 i læreboka. Testobservatoren er

$$F = \frac{\text{MSTr}}{\text{MSE}} = \frac{\text{SSTr}/(I-1)}{\text{SSE}/I(J-1)}$$

der $\text{SSTr} = J\sum_{i=1}^{I}\left(\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot}\right)^2$ og $\text{SSE} = \sum_{i=1}^{I}\sum_{j=1}^{J}\left(X_{ij} - \overline{X}_{i\cdot}\right)^2$.

Hvis nullhypotesen er riktig, er testobservatoren F-fordelt med $4-1=3$ og $4(10-1)=36$ frihetsgrader. Så med signifikansnivå $\alpha = 0.01$, forkaster vi nullhypotesen så sant $F > F_{0.01,3,36} = 4.38$.

Ved å bruke data gitt i oppgaven får vi følgende varianalysetabell; se R-kommandoene nedenfor.

Source of Variation	df	Sum of Squares	Mean Square	f
Treaments	3	509.1	169.71	10.85
Error	36	563.1	15.64	
Total	39	1072.2		

Av variansanalysetabellen får vi at verdien av testobservatoren er $f = 10.85$

Siden $f = 10.85 > 4.38 = F_{0.01,3,36}$ forkaster vi nullhypotesen på 1% signifikansnivå.

```
> # Leser data fra web.
> Fe.data = read.table("https://www.uio.no/studier/emner/matnat/math/STK1110/data/exe11
    -05.txt", header=T)
>
> # Utfører enveis variansanalyse:
> aov.fit=aov(Fe~factor(form),data=Fe.data)
> summary(aov.fit)
             Df Sum Sq Mean Sq F value  Pr(>F)
factor(form)  3  509.1  169.71   10.85 3.2e-05
Residuals    36  563.1   15.64
```

Exercise 11.6

Det er gitt i oppgaven at det er $I = 4$ grupper og $J = 5$ observasjoner i hver gruppe.

Ved å bruke Table 11.2 fra side 561 av læreboka får vi følgende variansanalysetabell:

Source of Variation	df	Sum of Squares	Mean Square	f
Treaments	3	75081.72	25027.24	1.70
Error	16	235419.04	14713.69	
Total	19	310500.76		

Hvis nullhypotesen er riktig, er testobservatoren F-fordelt med $4-1=3$ og $4(5-1)=16$ frihetsgrader.

Av variansanalysetabellen ser vi at testobservatoren får verdien $f = 1.70$.

Vi får P-verdien (der $\Phi_{3,16}$ er den kumulative F-fordelingen med 3 og 16 frihetsgrader)

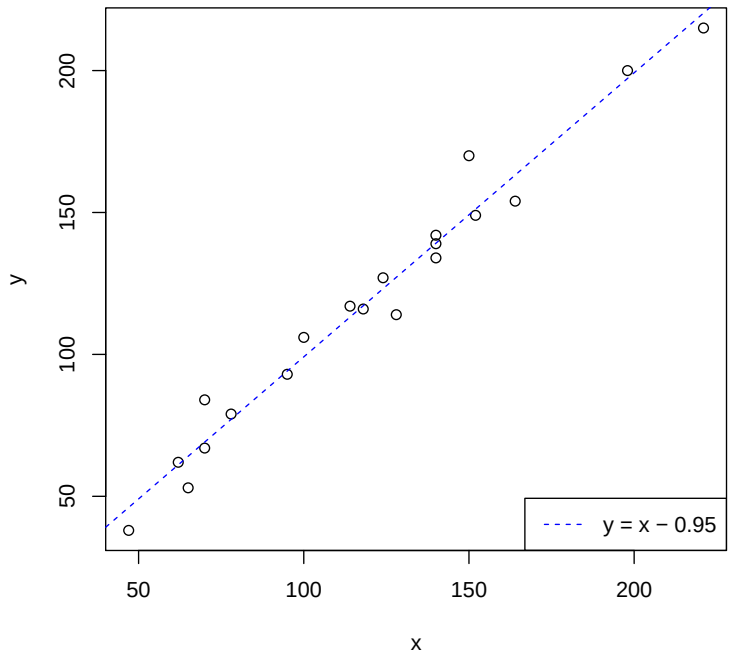
$$p = 1 - \Phi_{3,16}(1.70) = 1 - 0.793 = 0.207.$$

Siden P-verdien såpass stor, forkaster vi ikke nullhypotesen.

```
> p.value = 1 - pf(1.70, df1=3, df2=16)
> print(p.value)
[1] 0.2071208
```

Exercise 12.3

Et spredningsplott med en tilpasset regresjonslinje er vist i Figur **??**. Fra plottet kan vi se at det er en klar lineær sammenheng mellom `x` og `y`.



Figur 2: Spredningsplott for dataene i Exercise 12.3.

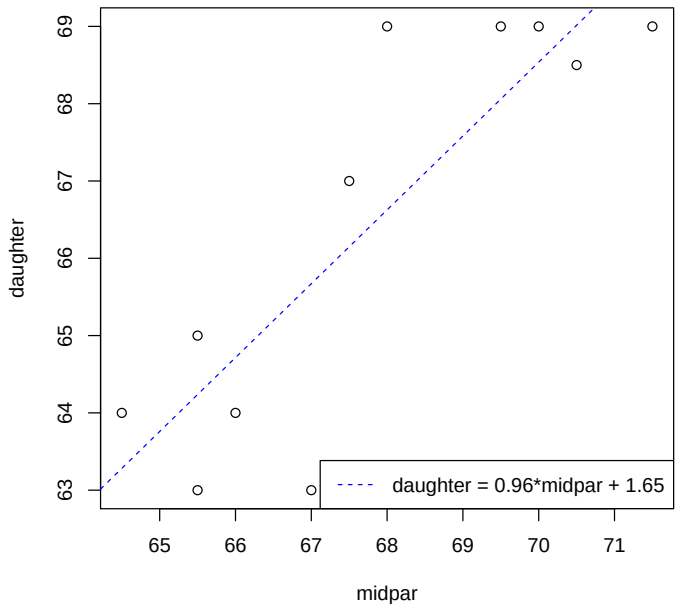
```
> # Leser data fra web.
> hydrogen.data = read.table("https://www.uio.no/studier/emner/matnat/math/STK1110/data/
    exe12-03.txt", header=T)
>
> # Minste kvadraters linje
```

```
> lm.fit = lm(y~x, data=hydrogen.data)
> print(lm.fit)
>
Call:
lm(formula = y ~ x, data = hydrogen.data)

Coefficients:
(Intercept)            x
    -0.9504       1.0008
>
> # Plot
> plot(x=hydrogen.data[,"x"], y=hydrogen.data[,"y"], xlab="x", ylab="y", main="")
> abline(lm.fit, lty=2, col="blue")
> legend("bottomright", legend=c("y = x - 0.95"), col=c("blue"), lty=c(2,1,1), cex=0.8)
```

Exercise 12.16

a) Et spredningsplott med en tilpasset regresjonslinje er vist i Figur **??**. Fra plottet kan vi se at det er en klar lineær sammenheng mellom `daughter` og `midpar`.



Figur 3: Spredningsplott for dataene i Exercise 12.16.

```
> # Leser data fra web.
> height.data = read.table("https://www.uio.no/studier/emner/matnat/math/STK1110/data/
    exe12-16.txt", header=T)
>
> # Minste kvadraters linje
> lm.fit = lm(daughter~midpar, data=height.data)
> print(lm.fit)

Call:
lm(formula = daughter ~ midpar, data = height.data)

Coefficients:
(Intercept)       midpar
```

```
      1.6497        0.9555
>
> # Plot
> plot(x=height.data[,"midpar"], y=height.data[,"daughter"], xlab="midpar", ylab="
    daughter", main="")
> abline(lm.fit, lty=2, col="blue")
> legend("bottomright", legend=c("daughter = 0.96*midpar + 1.65"), col=c("blue"), lty=c
    (2,1,1))
```

b) Nei. Hvis det skulle vært slik, måtte vi ha hatt en perfekt lineær sammenheng mellom `daughter` og `midpar`.

c) Den tilpassede lineær linja er gitt ved

$$y = 1.65 + 0.96x.$$

Når $x = 70$, har vi $y = 1.65 + 0.96 \cdot 70 + 1.65 = 68.5$.

Verdien $x = 74$ er et godt stykke utenfor det området av x-verdier som ble brukt til å tilpasse den lineære linja. Vi kan ikke vite om den lineære sammenhengen holder for slike `x`-verdier.

d) Fra utskriften, kan vi se at SSE $= 18.938$ og SST $= 68.409$. Det følger at

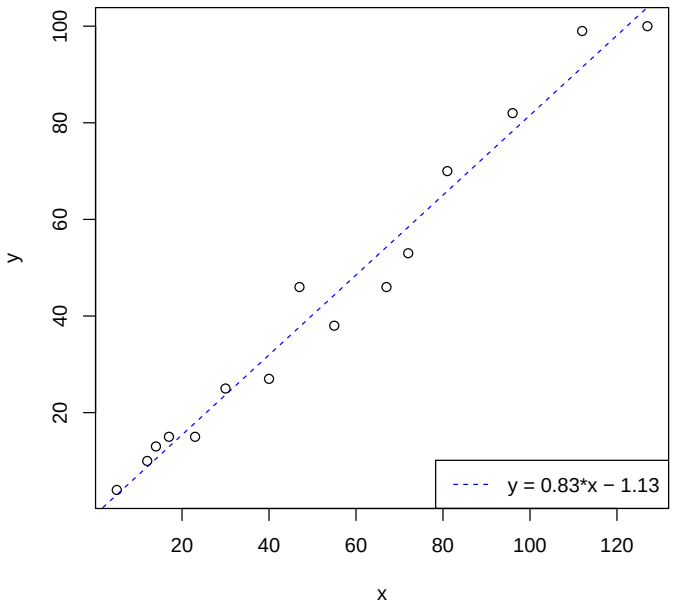
$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} = 1 - \frac{18.938}{68.409} = 0.723$$

Altså er 72.3% av variasjonen i `daughter` er forklart av `midpar`.

e) Nei. Det er simpelheten fordi menn generelt er høyere enn kvinner.

Exercise 12.17

a) Et spredningsplott med en tilpasset regresjonslinje er vist i Figur **??**. Fra plottet kan vi se at x og y har en klar lineær sammenheng.



Figur 4: Spredningsplott for dataene i Exercise 12.17.

```
> # Leser data fra web.
> highway.data = read.table("https://www.uio.no/studier/emner/matnat/math/STK1110/data/
    exe12-17.txt", header=T)
>
> # Tilpasser minste kvadraters linje
> lm.fit = lm(y~x, data=highway.data)
> summary(lm.fit)

Call:
lm(formula = y ~ x, data = highway.data)

Residuals:
   Min     1Q Median     3Q    Max
-8.279 -4.424  1.205  3.145  8.261

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.12830    2.36778  -0.477    0.642
x            0.82697    0.03652  22.642  7.9e-12 ***
---
Signif. codes:  0  '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.24 on 13 degrees of freedom
Multiple R-squared:  0.9753,  Adjusted R-squared:  0.9734
F-statistic: 512.7 on 1 and 13 DF,  p-value: 7.896e-12


>
> # Plot
> plot(x=highway.data[,"x"], y=highway.data[,"y"], xlab="x", ylab="y", main="")
> abline(lm.fit, lty=2, col="blue")
> legend("bottomright", legend=c("y = 0.83*x - 1.13"), col=c("blue"), lty=c(2))
```

b) Fra resultatene på sidene 626 og 627 i læreboka har vi at

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2} = \frac{S_{xy}}{S_{xx}}, \qquad \widehat{\beta}_0 = \widehat{y} - \widehat{\beta}_1 \overline{x}$$

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

Ved å bruke disse forlene får vi $\widehat{\beta}_1 = 0.83$ og $\widehat{\beta}_0 = -1.13$. Se R-kommandoene nedenfor. Svarene blir (selvfølgelig!) de samme som vi fikk ved `summary`-kommandoen i punkt a.

```
> x = highway.data[,"x"]
> y = highway.data[,"y"]
>
> S.xy = sum((x-mean(x))*(y-mean(y)))
> S.xx = sum((x-mean(x))^2)
>
> beta.1.hat = S.xy/S.xx
> print(beta.1.hat)
[1] 0.8269731
>
> beta.0.hat = mean(y) - beta.1.hat*mean(x)
> print(beta.0.hat)
[1] -1.128305
```

c) For $x = 50$ får vi estimatet

$$y = 0.83 \cdot 50 - 1.13 = 40.2$$

d) Et esimate for standardavviket er (jf. side 631 i læreboka)

$$s = \sqrt{\frac{\text{SSE}}{n-2}} = \sqrt{\frac{1}{n-2} \sum_{i=1}^n (y_i - \widehat{y}_i)^2}$$

Ved å bruke denne formelen får vi $s = 5.24$. Se R-kommandoene nedenfor. Dette svaret finner vi også som "residual standard error" ved `summary`-kommandoen i punkt a.

```
> y.hat.vec = predict(object=lm.fit, newdata=highway.data)
> sqrt(sum((y - y.hat.vec)^2)/(length(y)-2))
[1] 5.240462
```

e) Fra side 634 har vi

$$r^2 = 1 - \frac{\text{SSE}}{\text{SST}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

Ved å bruke denne formelen får vi $r^2 = 0.9753$. Se R-kommandoene nedenfor. Dette svaret finner vi også som "multiple R-squared:" ved `summary`-kommandoen i punkt a.

Det betyr at 97.53% av variasjonen i y er forklart av x.

```
> 1 - sum((y - y.hat.vec)^2)/((length(y)-1)*var(y))
[1] 0.9752689
```