STK1110 - Statistiske metoder og dataanalyse
Obligatorisk innlevering 2

Rohullah Akbari[1]

[1] Matematisk institutt, Universitetet i Oslo (UiO)

January 2, 2021

Contents

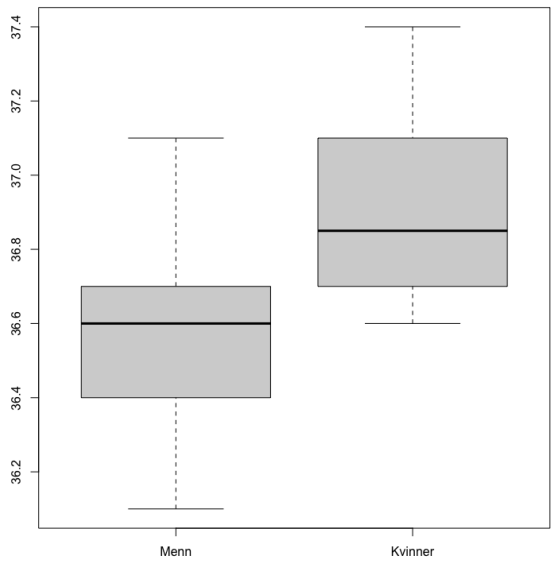


Figure 1: Viser ER-diagram fra oppgave 2.

Oppgave 1

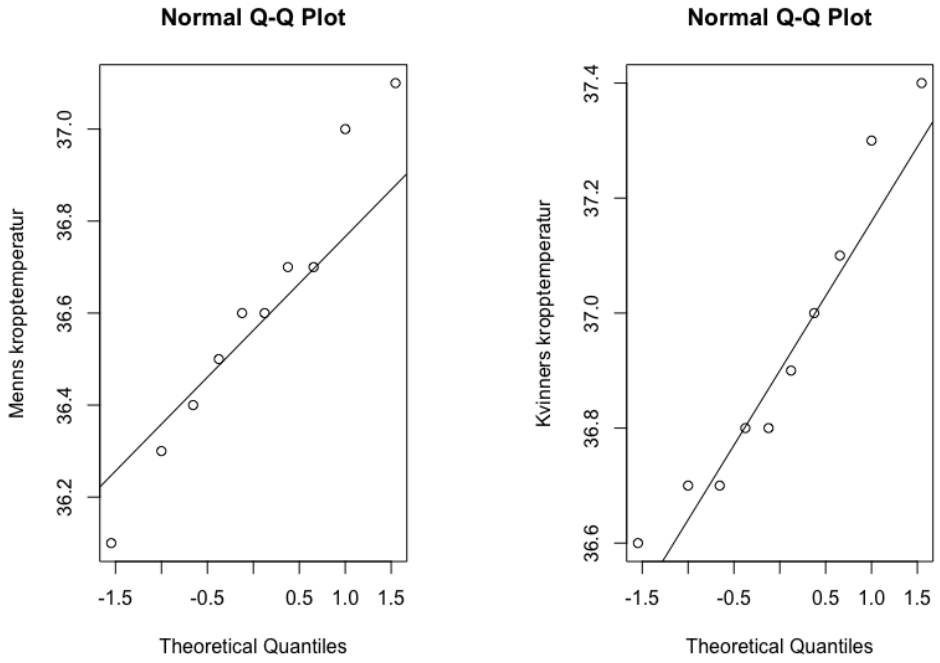
a) Lager ett boksplott, som viser fordelingen av observasjonene, i R med følgende kode:

```
data = read.table("https://www.uio.no/studier/emner/
    matnat/math/STK1110/data/temp.txt", header=T)
menn = data$Menn
kvinner = data$Kvinner
boxplot(menn, kvinner, names = c("Menn","Kvinner"))
```

Boksplottet er vist i figur(1). Her ser vi at menns kroppstemperatur er litt lavere enn kvinners. I tillegg til det er det observerbar at det er litt skjev fordeling blant kvinnenes kroppstemperatur i forhold til menn.

b) Vi lager normal-fordelingsplott for de to observasjonssettene ved å lage qq-plot. R-koden ble:

```
qqnorm(menn, ylab = "Menns kropptemperatur")
qqline(menn)
qqnorm(kvinner, ylab = "Kvinners kropptemperatur")
qqline(kvinner)
```



(a) Menn (b) Kvinner

Figure 2: Viser qq-plottene til menn og kvinner.

Plottene er vist i figure(2). På disse plottene kan vi se at det er mange punkter som ikke er på den rette linja og derfor er det vanskelig å avgjøre om disse observasjonene er normalfordelt.

I resten av oppgaven antar vi at observasjonene er realisasjoner av normal- fordelte variabler.

c) Vi antar at variansen er den samme for de to utvalgene, og tester med signifikansnivå 5% om det er noen forskjell i forventet kroppstemperatur, altså:

$$\sigma^2 = \sigma_1^2 = \sigma_2^2$$

Deretter lar vi $X_1...X_m$ være kroppstemperaturene hos menn og $Y_1...Y_n$ være kroppstemperaturene hos kvinner. Fra antagelsen om normalfordeling få vi:

$$X_1...X_m \sim N(\mu_X, \sigma_X^2)$$
$$Y_1...Y_m \sim N(\mu_Y, \sigma_Y^2)$$

Videre setter vi opp testen for om det er noen forskjell i forventet

kroppstemperatur hos menn og kvinner:

$$H_o : \mu_X = \mu_Y$$

$$H_a : \mu_X \neq \mu_Y$$

For å skjekke denne testen så regner vi ut Z. Fra seksjon 10.1 i læreboka har vi:

$$Z = \frac{\overline{X} - \overline{Y} - \Delta_o}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}} \sim N(0,1)$$

der $\Delta_o = \mu_X - \mu_Y$. Siden vi har lik varians så får vi:

$$Z = \frac{\overline{X} - \overline{Y} - \Delta_o}{\sqrt{\sigma^2(\frac{1}{m} + \frac{1}{n})}}$$

Under H_o så er $\Delta_o = 0$:

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\sigma^2(\frac{1}{m} + \frac{1}{n})}}$$

Bruker "Pooled t Procedures" fra s.504 i læreboka til å estimere σ^2:

$$\hat{\sigma^2} = S_P^2 = \frac{m-1}{m+n-2} S_X^2 + \frac{n-1}{m+n-2} S_Y^2$$

Videre har vi at S_X, S_Y er uavhengige og siden de har lik varians så får vi at summen av fordelingene til S_X og S_Y er to uavhengige chi-kvadratiske fordelinger med $m-1$ og $m-1$ frihetsgrader (eller $m-1+n-1 = m+n-2$ frihetsgrader). Altså:

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{S_P^2(\frac{1}{m} + \frac{1}{n})}}$$

Bruker R til å berenge denne t:

```
> m = length(menn)
> n = length(kvinner)
> sp2 = (m-1)/(m+n-2)*sd(menn)^2 + (n-1)/(m+n-2)*sd(
    kvinner)^2
> t = (mean(menn)-mean(kvinner))/sqrt(sp2 *(1/length(
    menn) + 1/length(kvinner)) )
> print(t)
[1] -2.590062
```

Altså $t = -2.59$. Bruker denne verdien til å berenge p-verdi:

```
> pvalue = 2*(1-pt(abs(t),18))
> print(pvalue)
[1] 0.01848131
```

Vi får p-verdi ca. lik 0.018. For å forkaste H_o så må enten $t \leq -t_{\alpha/2,v}$ eller $t \geq t_{\alpha/2,v}$. Siden

$$t = -2.59 < -t_{\alpha/2,v} = -2.01$$

så forkaster vi H_o. Deretter bruker vi T til å berenge 5% konfidensintervall:

$$P\left(-t_{\alpha/2,m+n-2} < \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{S_P^2(\frac{1}{m} + \frac{1}{n})}} < t_{\alpha/2,m+n-2} \right) = 0.95$$

$$P\left(-t_{\alpha/2,m+n-2}\sqrt{S_P^2(\frac{1}{m} + \frac{1}{n})} - \overline{X} + \overline{Y} < -(\mu_X - \mu_Y) < t_{\alpha/2,m+n-2}\sqrt{S_P^2(\frac{1}{m} + \frac{1}{n})} - \overline{X} + \overline{Y} \right)$$

$$P\left(-t_{\alpha/2,m+n-2}\sqrt{S_P^2(\frac{1}{m} + \frac{1}{n})} + \overline{X} - \overline{Y} < \mu_X - \mu_Y < t_{\alpha/2,m+n-2}\sqrt{S_P^2(\frac{1}{m} + \frac{1}{n})} + \overline{X} - \overline{Y} \right)$$

Dette gir intervallet:

$$\overline{x} - \overline{y} \pm t_{\alpha/2,m+n-2}\sqrt{S_P^2(\frac{1}{m} + \frac{1}{n})}$$

Bruker R til å regne intervallet:

```
> ovre = mean(menn) - mean(kvinner) + 2.101*sqrt(sp2 *
    (1/n + 1/m))
> nedre = mean(menn) - mean(kvinner) - 2.101*sqrt(sp2 *
    (1/n + 1/m))
> c(nedre,ovre)
[1] -0.59768862 -0.06231138
```

95% konfidensintervallet blir:

$$(-0.598, -0.062)$$

Bruker t.test() funksjonen i R til å sjekke svaret:

```
> t.test(menn, kvinner, var.equal = TRUE)

        Two Sample t-test

data:  menn and kvinner
t = -2.5901, df = 18, p-value = 0.01848
alternative hypothesis: true difference in means is not
    equal to 0
95 percent confidence interval:
 -0.59767869 -0.06232131
sample estimates:
mean of x mean of y
    36.60     36.93
```

Det ser ut som at vi har fått samme t, pverdien og konfidensintervallet som i test.t(), og konkluderer at vi har regnet riktig.

d) Utfører den samme testen, men i dette tilfellet så antar vi at det er ikke lik varians. Fra seksjon 10.2 i læreboka har vi:

$$T = \frac{\overline{X} - \overline{Y} - \Delta_o}{\sqrt{\frac{S_X^2}{m} + \frac{S_Y^2}{n}}} \sim t_v$$

der $\Delta_o = \mu_X - \mu_Y$. Under H_o har vi at:

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_X^2}{m} + \frac{S_Y^2}{n}}}$$

Bruker R til å beregne t:

```
> t = (mean(menn) - mean(kvinner))/sqrt(sd(menn)^2/m +
    sd(kvinner)^2/n)
> print(t)
[1] -2.590062
```

Så $t = -2.59$. Dette er den samme t-verdien vi fant i oppgaven over. Beregner v eller antall frihetsgrader:

$$v = \frac{(\frac{S_X^2}{m} + \frac{S_Y^2}{n})^2}{(\frac{S_X^2}{m})^2/(m-1) + (\frac{S_Y^2}{n})^2/(n-1)}$$

Setter denne i R og får:

```
> a = (sd(menn)^2 /m)^2 /(m-1)
> b = (sd(kvinner)^2 /n)^2 /(n-1)
> v = ((sd(menn)^2 / m) + (sd(kvinner)^2 / n))^2 / (a+b
    )
> print(v)
[1] 17.7338
```

Hittil har vi fått $t = -2.59$ og $v \simeq 18$. Disse verdiene er helt samme som oppgaven ovenfor og derfor beholder samme konklusjon, nemlig at vi forkaster H_o og beholder H_a.

Deretter bruker t.test() til å sjekke om jeg har regnet riktig:

```
> t.test(kvinner,menn)

        Welch Two Sample t-test

data:  kvinner and menn
t = 2.5901, df = 17.734, p-value = 0.01863
alternative hypothesis: true difference in means is not
    equal to 0
95 percent confidence interval:
 0.06203301 0.59796699
sample estimates:
mean of x mean of y
    36.93     36.60
```

Det ser ut som at det stemmer.

e) Vi tester opp og gjennomfører en F-test for å sjekke om det er noen grunn til å påstå at variansene er forskjellige. Slik blir hypotesetesten:

$$H_o : \sigma_X = \sigma_Y$$

$$H_a : \sigma_X \neq \sigma_Y$$

Fra seksjon 10.5 i læreboka har vi at:

$$F = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim F_{m-1,n-1}$$

Under H_o har vi:

$$F = \frac{S_X^2}{S_Y^2}$$

Bruker R til å regne ut f:

```
> f = sd(menn)^2/sd(kvinner)^2
> print(f)
[1] 1.279251
```

Så $f = 1.279$. Vi forkaster H_o ved signifikansnivå α dersom $f \leq F_{1-\alpha/2,m-1,n-1}$ eller $f \geq F_{\alpha/2,m-1,n-1}$. Vi får da:

$$f = 1.279 \not< F_{0,975/2,m-1,n-1} = 0.248$$

$$f = 1.279 \not> F_{0.025,m-1,n-1} = 4.026$$

Dermed konkluderer vi med at vi beholder H_o. Altså det var ingen god grunn til å påstå at variansene var forskjellige.

f) På grunn av dårlig tid og sykdom så rakk jeg ikke denne deloppgaven.

Oppgave 2

a) Grunnen til at vi bruker en paret sammenligning er fordi vi ønsker å se forskjellen på de to metodene. Eggede tvillinger er "bygd" av samme materialer men kan kun ha forskjellige IQ pga forskjellige metoder. Antagelser vi gjør i denne oppgaven er at vi antar data settet består av n uavhengige valgte par $(X_1, Y_1)...(X_n, Y_n)$ med $E(X_i) = \mu_X$ og $E(Y_i) = \mu_Y$. Vi kaller "Twin A" for X og "Twin B" for Y. Differansen mellom X og Y kaller vi for D slik at vi har $D_1...D_n$. Videre antar vi disse D-verdiene er normalfordelt.

b) Vi setter testen for å å besvare spørsmålet om forskjell i I:

$$H_o : \mu_D = 0$$

$$H_a : \mu_D \neq 0$$

Har følgende testobservator:

$$T = \frac{\overline{D} - \Delta_o}{\hat{\sigma}/\sqrt{n}}$$

Under H_o har vi at:

$$t = \frac{\overline{D}}{\hat{\sigma}/\sqrt{n}} \sim t_{n-1}$$

Regner ut t:

$$t = \frac{-3.26}{1.58} = -2.06.$$

Bruker det til å berenge p-verdien:

$$p = 2[1 - \Phi(|t|)]$$

Bruker R til å beregne p-verdien:

```
p_verdi = 2*(1-pt(abs(T),31))
> print(p_verdi)
[1] 0.04754466
```

P-verdien ble funnet til å være omtrent lik 0.0475. Fra side 457 har vi at for en signifikansnivå α så kan vi forkaste H_o dersom p-verdi er mindre eller lik α. I dette tilfellet så kan vi velge et signifikansnivå α lik 5%, da kunne vi forkaste H_o. Altså vi forventer at det er forskjell i IQ-en med signifikansnivå lik 5%.

c) Setter opp konfidensintervallet for μ_D:

$$P\left(-t_{\alpha/2,n-1} < \frac{\overline{d} - \mu_D}{s_D/\sqrt{n}} < t_{\alpha/2,n-1} \right) = 0.95$$

$$P\left(-t_{\alpha/2,n-1}s_D/\sqrt{n} - d < -\mu_D < t_{\alpha/2,n-1}s_D/\sqrt{n} - d \right) = 0.95$$

$$P\left(-t_{\alpha/2,n-1}s_D/\sqrt{n} + d < \mu_D < t_{\alpha/2,n-1}s_D/\sqrt{n} + d \right) = 0.95$$

Vi har intervallet:

$$\left(d - t_{\alpha/2,n-1}s_D/\sqrt{n}, d + t_{\alpha/2,n-1}s_D/\sqrt{n} \right)$$

Bruker R til å regne ut intervallet:

```
> nedre = -3.26 - qt(0.025,df=30,lower.tail = F)*1.58
> ovre = -3.26 + qt(0.025,df=30,lower.tail = F)*1.58
> c(nedre,ovre)
[1] -6.48679048 -0.03320952
```

Så intervallet blir: (-6.487, -0.033). Med andre ord så dekker intervallet kun negative verdier. I dette tilfellet så betyr det at vi forkaster H_o siden $\mu_D = 0$ er ikke engang inneholdt i dette intervallet.

Oppgave 3

Det er 3000 menn og 3000 kvinner. Vi har p1 donere 16% av fedre og p1 donere 14.7% av mødre som opplever tidsklemma.

a) Vi setter opp en hypotesetest for å undersøke om forskjellen mellom mødre og fedre er signifikant. Slik blir hypotesetesten:

$$H_o : p_1 = p_2$$

$$H_a : p_1 \neq p_2$$

Fra seksjon 10.4 i læreboka har vi følgende testobservator:

$$Z = \frac{\hat{p_1} - \hat{p_2} - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{m} + \frac{p_2 q_2}{n}}}$$

Der $q_1 = 1 - p_1$ og $q_2 = 1 - p_2$. Under H_o har vi at:

$$Z = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{pq\left(\frac{1}{m} + \frac{1}{n}\right)}}$$

Her er verdien til $\hat{p}$ ukjent og vi må derfor estimere det:

$$\hat{p} = \frac{X + Y}{m + n} = \frac{m}{m + n}\hat{p_1} + \frac{n}{m + n}\hat{p_2}$$

$$= \frac{3000}{6000}0.162 + \frac{3000}{6000}0.147 = 0.1545$$

Bruker denne til å beregne:

$$Z = \frac{0.162 - 0.147}{\sqrt{0.1545(1 - 0.1545)\left(\frac{1}{3000} + \frac{1}{3000}\right)}} \simeq= 1.607$$

Beregner p-verdi med R:

```
> p_verdi = 2*(1-pnorm(1.607))
> print(p_verdi)
[1] 0.1073979
```

Altså,

$$p = 2[1 - \Phi(|1.607|)] = 0.107$$

Fra side 457 har vi at for en signifikansnivå α så kan vi forkaste H_o dersom p-verdi er mindre eller lik α. For å beholde H_o så må signifikansnivået være $\alpha = 0.1$, slik at $p_{verdi} > \alpha$. Med dette signifikansnivået så kan vi konkludere med at det er ikke noen forskjell i tidsklemma mellom småbarnsfedre og småbarnsfedre.

b) Vi bruker prop.test() i R for å kontrollere svaret:

```
> prop.test(c(486,441),c(3000,3000) , correct=F)

        2-sample test for equality of proportions
            without continuity correction

data:  c(486, 441) out of c(3000, 3000)
X-squared = 2.5836, df = 1, p-value = 0.108
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.003286474  0.033286474
sample estimates:
prop 1 prop 2
 0.162  0.147
```

Vi ser at vi har fått samme resultat for p-verdien og dermed konkluderer med at vi har regnet riktig.

Oppgave 4

a) I denne oppgaven skal vi lage en regresjonsmodell for sammenhengen mellom snømengden og vannstaden. Vi lar Y donere vannstanden og x være snømengden, slik at vi får følgende relasjon:

$$Y = \beta_0 + \beta_1 x + \epsilon$$

der $\epsilon \sim N(0, \sigma^2)$. Her er β_0 og β_1 ukjente parametere og vi bruker funksjonen lm() i R til å finne de parametere:

```
> data = read.table("https://www.uio.no/studier/emner/
    matnat/math/STK1110/data/snoe_vann.txt", header=T)
> sno =  c(23.1,data$X23.1)
> vann = c(10.5,data$X10.5)
> lm(vann~sno)

Call:
lm(formula = vann ~ sno)

Coefficients:
(Intercept)          sno
     0.2800       0.5056
```

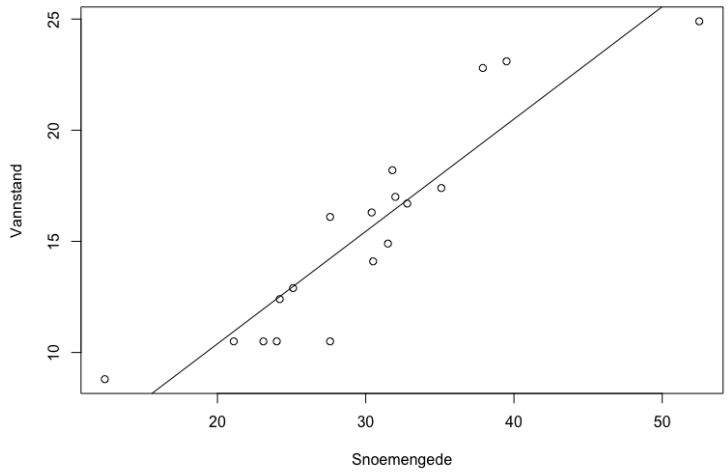


Figure 3: Viser plott av observasjonene og den tilpassede regresjonslinja.

Fra dette får vi at $\beta_0 = 0.28$ og $\beta_1 = 0.5056$. Så dette fører til:

$$Y = 0.28 + 0.5056x + \epsilon$$

Videre plotter vi observasjonene og den tilpassede regresjonslinja:

```
plot(sno,vann,xlab="Snoemengede",ylab = "Vannstand")
abline(lm(vann~sno))
```

Plottet er vist i figur(3).

b) I denne oppgaven plotter vi residualene mot forklaringsvariabelen. Slik ble koden:

```
vann_res = resid(lm(vann~sno))
plot(sno,vann_res, ylab="Residualer", xlab="Sno")
```

Deretter lager vi normal-fordelingsplott av residualene ved å lage et qq-plot:

```
vann_standard = rstandard(lm(vann~sno))
qqnorm(vann_standard)
qqline(vann_standard)
```

Resultatetene er vist i figur(4) og (5).

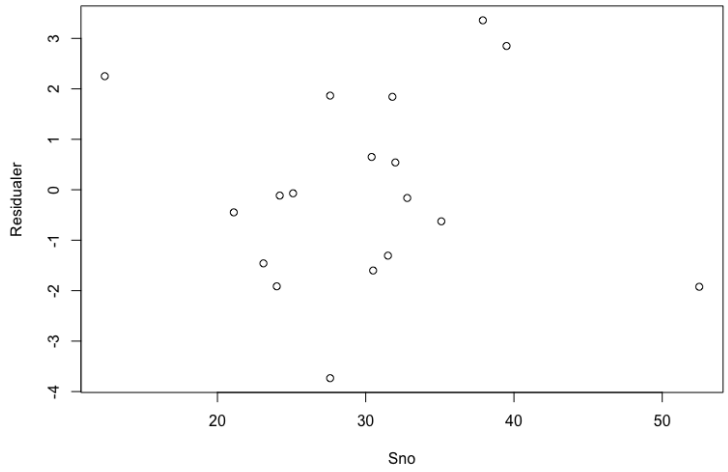


Figure 4: Viser plott av residualene mot forklaringsvariabelen.

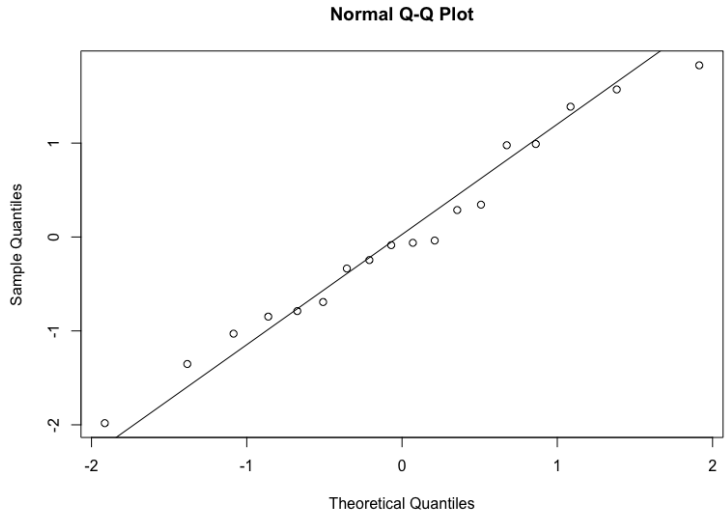


Figure 5: Viser plott av normalfordelingsplott av residualene.

c) Et estimat for variansen til feileddene, fra s. 631, kan beregnes ved:

$$\hat{\sigma^2} = s^2 = \frac{SSE}{n-2} = \frac{\sum y_i - (\hat{\beta_0} + \hat{\beta_1 x_i})}{n-2}$$

Bruker R til å berenge det:

```
> B0 = 0.2800
> B1 = 0.5056
> n = length(sno)
> SSE = sum((vann - (B0+B1*sno))^2)
> s2 = SSE/(n-2)
> print(s2)
[1] 3.774599
```

Vi konstruerer et 95 % konfidensintervall for stigningstallet β_1. Fra seksjon 10.3 har vi:

$$P\left(-t_{\alpha/2,n-2} < \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} < t_{\alpha/2,n-1} \right) = 0.95$$

$$P\left(\hat{\beta}_1 - t_{\alpha/2,n-2} S_{\hat{\beta}_1} < \beta_1 < \hat{\beta}_1 + t_{\alpha/2,n-2} S_{\hat{\beta}_1} \right) = 0.95$$

Dette gir intervallet:

$$\left(\hat{\beta}_1 - t_{\alpha/2,n-2} S_{\hat{\beta}_1}, \hat{\beta}_1 + t_{\alpha/2,n-2} S_{\hat{\beta}_1} \right)$$

der $S_{\hat{\beta}_1} = \frac{S}{\sqrt{S_{XX}}} = \frac{n^2}{n\sum x_i^2 - (\sum x_i)^2}$.

Bruker R til å regne ut intervallet:

```
> S_B = n*s2 / ( n*sum(sno^2) - ((sum(sno))^2) )
> ovre = B1 + S_B * 2.120
> nedre = B1 - S_B * 2.120
> c(nedre,ovre)
[1] 0.4991693 0.5120307
```

95 % konfidensintervall ble (0.499,0.512).

d På grunn av dårlig tid og sykdom så rakk jeg ikke denne deloppgaven.