Obligatorisk oppgave 2 i STK1110 – Høsten 2022

Oppgave 1

a) Boksplottet i Figur 1 antyder at verdiene er høyere for kvinner enn for menn.

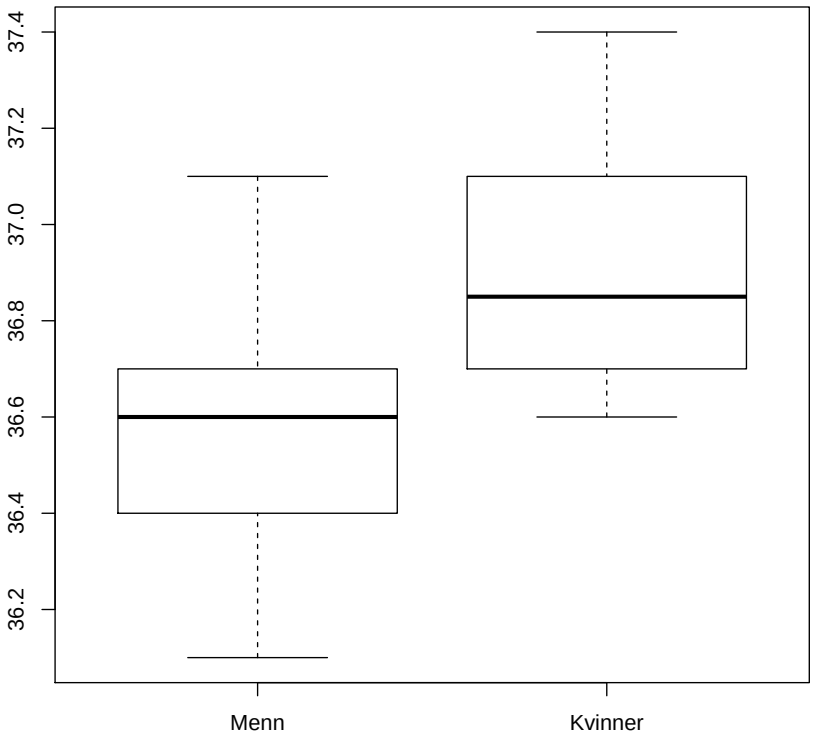

Figure 1: Boksplott over kroppstemperaturmålingene delt inn etter kjønn.

b) Kvantilplottene er vist i Figur 2 For menn passer den minste og de to største observasjonene så godt med den rette linja, men med bare ti observasjoner for hvert kjønn, kan vi ikke forvente å få en "perfekt" rett linje selv hvis dataene er normalfordelte.

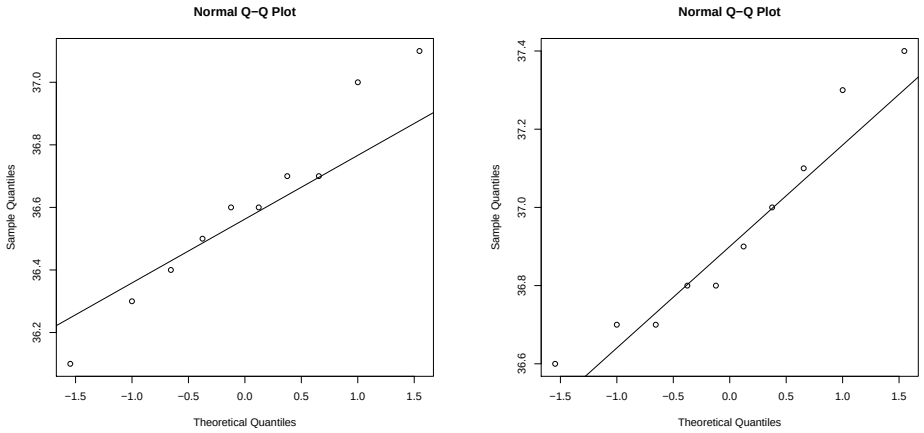

Figure 2: Kvantilplott over kroppstemperaturmålingene til menn (til venstre) og kvinner (til høyre).

Vi lar $X_1, \ldots, X_m$ med $m = 10$ betegne kroppstemperaturene for ti tilfeldig valgte menn og $Y_1, \ldots, Y_n$ med $n = 10$ betegne kroppstemperaturene ti tilfeldig valgte kvinner. Vi antar i resten av oppgaven at alle disse stokastiske variablene er uavhengige og at $X_i \sim N(\mu_1, \sigma_1^2)$ og $Y_j \sim N(\mu_2, \sigma_2^2)$. Altså er μ_1 forventet kroppstemperatur for mennene og μ_2 forventet kroppstemperatur for kvinnene. Videre antar vi i punkt c) at $\sigma_1 = \sigma_2 = \sigma$, mens vi ikke gjør denne antakelsen i punkt d).

c) La

$$\bar{X} = \frac{1}{m}\sum_{i=1}^{m} X_i \text{ og } \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

og

$$S_1^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \bar{X})^2 \text{ og } S_2^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2.$$

Vi har da at

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{m} + \frac{1}{n}}},$$

med $S_p^2 = [(n-1)S_1^2 + (m-1)S_2^2]/(n+m-2)$, er t_{n+m-2}-fordelt. Dermed kan vi bruke T med $\mu_1 - \mu_2 = 0$ for å teste $H_0 : \mu_1 = \mu_2$ mot $H_0 : \mu_1 \neq \mu_2$. Vi forkaster da H_0 dersom $T \leq -t_{0.025,m+n-2}$ eller $T \geq t_{0.025,m+n-2}$. I og med at t-fordelingen er symmetrisk om 0, er P-verdien da definert som

$$\mathrm{P}(T < -|t_{obs}|) + \mathrm{P}(T > |t_{obs}|) = 2\mathrm{P}(T > |t_{obs}|) = 0.0185.$$

Vi forkaster altså H_0 ved 5% signifikansnivå. For å konstruere konfidensinter-

vallet, benytter vi oss av at

$$\mathrm{P}\left(-t_{0.025,m+n-2} \leq \frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{m}+\frac{1}{n}}} \leq t_{0.025,m+n-2}\right) = 0.95.$$

Vi får da konfidensintervallet $\bar{x}-\bar{y} \pm t_{0.025,m+n-2}s_p\sqrt{\frac{1}{n}+\frac{1}{m}}$ for $\mu_1-\mu_2$, og når vi setter inn observerte verdier får vi $(-0.598, -0.0623)$, som ikke inneholder 0. Utskriften fra `t.test()` er:

```
Two Sample t-test

data:  x and y
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
```

som er samme resultater som over.
d) I dette tilfellet må vi bruke testobservatoren

$$T = \frac{\bar{X}-\bar{X}-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{m}+\frac{S_2^2}{n}}},$$

som er tilnærmet t_ν-fordelt med

$$\nu = \frac{\left(\frac{s_1^2}{m}+\frac{s_2^2}{n}\right)^2}{\frac{\left(\frac{s_1^2}{m}\right)^2}{m-1}+\frac{\left(\frac{s_2^2}{n}\right)^2}{n-1}}.$$

Dermed kan vi bruke T med $\mu_1-\mu_2=0$ for å teste $H_0:\mu_1=\mu_2$ mot $H_0:\mu_1 \neq \mu_2$. Vi forkaster da H_0 dersom $T \leq -t_{0.025,\nu}$ eller $T \geq t_{0.025,\nu}$. P-verdien er:

$$2\mathrm{P}(T > |t_{obs}|) = 0.0186,$$

som er omtrent den samme som vi fikk undr antakelsen om lik varians. Vi forkaster dermed H_0 ved 5% signifikansnivå også med denne testen. For å konstruere konfidensintervallet, benytter vi oss av at

$$\mathrm{P}\left(-t_{0.025,\nu} \leq \frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{m}+\frac{S_2^2}{n}}} \leq t_{0.025,\nu}\right) = 0.95,$$

som gir konfidensintervallet $\bar{x} - \bar{y} \pm t_{0.025,\nu}\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}$ for $\mu_1 - \mu_2$, og når vi setter inn observerte verdier får vi $(-0.598, -0.0620)$, som er nesten det samme som i c) og som ikke inneholder 0. I og med at reusltatene er omtrent de samme enten vi antar like varianser eller ikke, ser altså antakelsen om like varianser ut til å være fornuftig. Utskriften fra `t.test()` er:

```
Welch Two Sample t-test

data:  x and y
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:
 -0.59796699 -0.06203301
sample estimates:
mean of x mean of y
    36.60     36.93
```

som er samme resultater som over.
e) Vi har at

$$\frac{(m-1)S_1^2}{\sigma_1^2} \sim \chi^2_{m-1} \text{ og } \frac{(n-1)S_2^2}{\sigma_2^2} \sim \chi^2_{2-1}$$

er uavhengige, slik at

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{\frac{(m-1)S_1^2/\sigma_1^2}{m-1}}{\frac{(n-1)S_2^2/\sigma_2^2}{n-1}}$$

er $F_{m-1,n-2}$-fordelt. Dette følger av proposisjon på s. 320 i boka, at de to utvalgene er uavhengige samt definisjonen av F -fordelingen. Vi forkaster da $H_0 : \sigma_1^2 = \sigma_2^2$ til fordel for $H_a : \sigma_1^2 \neq \sigma_2^2$ ved signifikansnivå 5% dersom $F < f_{0.975,m-1,n-1} = 0.25$ eller $F > f_{0.025,m-1,n-1} = 4.03$. Vi har $f_{obs} = 1.28$, som ikke gir noen grunn til å forkaste H_0. Utskriften fra `var.test()` er:

```
F test to compare two variances

data:  x and y
F = 1.2793, num df = 9, denom df = 9, p-value = 0.7197
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.3177479 5.1502577
sample estimates:
ratio of variances
          1.279251
```

Vi ser at P-verdien er på 0.72, som er langt fra forkasting av H_0, hvilket gir en forklaring på hvorfor vi får så like resultater i c) og d).
f) Vi har at

$$\mathrm{E}(X_{11} - Y_{11}) = \mu_1 - \mu_2,$$

som kan estimeres med $\bar{X} - \bar{X}$. Videre er

$$\begin{aligned}
\mathrm{E}(X_{11} - Y_{11} - (\bar{X} - \bar{Y})) =& \mathrm{E}(X_{11} - Y_{11}) - \mathrm{E}(\bar{X} - \bar{Y}) = \mu_1 - \mu_2 - (\mu_1 - \mu_2) = 0 \\
\mathrm{Var}(X_{11} - Y_{11} - (\bar{X} - \bar{Y})) \overset{uavh.}{=}& \mathrm{Var}(X_{11}) - \mathrm{Var}(Y_{11}) + \mathrm{Var}(\bar{X}) + \mathrm{Var}(\bar{Y}) \\
=& \sigma_1^2 + \sigma_2^2 + \sigma_1^2/m + \sigma_2^2/n \\
=& \sigma^2(2 + 1/m + 1/n),
\end{aligned}$$

da $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Dermed er $X_{11} - Y_{11} - (\bar{X} - \bar{Y}) \sim N(0, \sigma^2(2 + 1/m + 1/n))$. Vi har da at

$$T = \frac{X_{11} - Y_{11} - (\bar{X} - \bar{Y})}{S_p\sqrt{2 + \frac{1}{m} + \frac{1}{n}}}$$

er t_{m+n-2}-fordelt. Det betyr at

$$\begin{aligned}
&\mathrm{P}\left(-t_{0.025,m+n-2} \leq \frac{X_{11} - Y_{11} - (\bar{X} - \bar{Y})}{S_p\sqrt{2 + \frac{1}{m} + \frac{1}{n}}} \leq t_{0.025,m+n-2}\right) = 0.95 \\
\to \mathrm{P}&\left(\bar{X} - \bar{Y} - t_{0.025,m+n-2}S_p\sqrt{2 + \frac{1}{m} + \frac{1}{n}} \leq X_{11} - Y_{11}\right. \\
&\qquad \left. \leq \bar{X} - \bar{Y} + t_{0.025,m+n-2}S_p\sqrt{2 + \frac{1}{m} + \frac{1}{n}}\right) = 0.95,
\end{aligned}$$

slik at intervallet $\bar{X} - \bar{Y} \pm t_{0.025,m+n-2}S_p\sqrt{2 + \frac{1}{m} + \frac{1}{n}}$ inneholder $X_{11} - Y_{11}$ med 95% sannsynlighet. Når vi setter inn observerte verdier får vi prediksjonsintervallet $(-1.220, 0.558)$. Vi får her et mye bredere intervall da vi både må ta hensyn til usikkerheten i parameterne, slik som i konfidensintervallet, og i de nye observasjonene.

Oppgave 2

a) De to tvillingene i et enegget tvillingpar har det samme genetiske materiale, så forskjeller mellom dem vil bare skyldes miljøet de har vokst opp i. Derfor vil observasjonene for tvillingene i et tvillingpar være positivt korrelert. Det vil resultere i at differansen mellom observasjonene for de to tvillingene i et tvillingpar har mindre varians (enn forskjellen på mellom to tilfeldige personer) og dermed gir mer nøyaktige resultater (enn det en ville fått med to uavhengige utvalg). Se s. 515-516 i læreboka for nærmere kommentarer.

La $D_1, \ldots, D_n$ være differansene i IQ-målinger for n tvillingpar (tvilling A minus tvilling B). Vi vil anta at disse differansene er uavhengige og $N(\mu_D, \sigma_D^2)$-fordelte.

b) Vi innfører

$$\bar{D} = \frac{1}{n}\sum_{i=1}^{n} D_i \text{ og } S_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2.$$

Av resultater for ett normalfordelt utvalg (se s. 401 i læreboka) har vi da at

$$T = \frac{\bar{D} - \mu_D}{S_D/\sqrt{n}},$$

t_{n-1}-fordelt. Vi vil teste nullhypotesen $H_0 : \mu_D = 0$ mot den alternative hypotesen $H_a : \mu_D \neq 0$. Vi får da en test med signifikansnivå 5% hvis vi forkaster H_0 så sant $T \leq -t_{0.025,n-1}$ eller $T \geq t_{0.025,n-1}$, med $\mu_D = 0$ satt inn i uttrykket for T. I dette tilfellet blir $t = 2.06$ og P-verdien er $2\mathrm{P}(T > |t_{obs}|) = 0.048$, dvs. at vi forkaster $H_0 : \mu_D = 0$ ved 5% signifikansnivå og konkluderer med at det er forskjell på IQ-en til tvilling A og tvilling B, men vi forkaster bare såvidt. Det er således en del usikkerhet rundt konklusjonen.
c) Ved å ta utgangspunkt i observatoren T fra b), får vi følgende 95% konfidensintervall for μ_D:

$$\bar{d} \pm t_{0.025,n-1} s_D/\sqrt{n}.$$

Når vi setter inn verdier fra dataene, får vi intervallet $(-6.49, -0.03)$. Vi ser at konfidensintervallet bare inneholder negative verdier. Det betyr at vi vil forvente at tvilling B har høyere IQ enn tvilling A. Merk videre at siden et 95% konfidensintevall ikke inneholder verdien 0, vil nullhypotesen $H_0 : \mu_D = 0$ forkastes til fordel for den alternative hypotesen $H_a : \mu_D \neq 0$ ved 5% signifikansnivå (slik vi alt har sett i punkt b).

Oppgave 3

a) La X og Y være antall henholdsvis fedre og mødre som ofte opplever problemer med tidsklemma, og la p_1 og p_2 være de tilsvarende populasjonsandelene. Vi antar da at $X \sim Binomisk(3000, p_1)$ og $Y \sim Binomisk(3000, p_2)$ er uavhengige. De to populasjonsandelene kan estimeres med $\hat{p}_1 = X/3000$ og $\hat{p}_2 = Y/3000$. Da $3000\hat{p}_1 = 486$, $3000(1 - \hat{p}_1) = 2514$, $3000\hat{p}_2 = 441$ og $3000(1 - \hat{p}_2) = 2559$ er såpass store, er $\hat{p}_1$ og $\hat{p}_2$ tilnærmet normalfordelt, slik at $\hat{p}_1 - \hat{p}_2$ også er det. Vi har

$$\begin{aligned}
\mathrm{E}(\hat{p}_1 - \hat{p}_2) =& \mathrm{E}(\hat{p}_1) - \mathrm{E}(\hat{p}_2) = p_1 - p_2 \\
\mathrm{Var}(\hat{p}_1 - \hat{p}_2) \overset{uavh.}{=}& \mathrm{Var}(\hat{p}_1) + \mathrm{Var}(\hat{p}_2) = \frac{p_1(1-p_1)}{3000} + \frac{p_2(1-p_2)}{3000},
\end{aligned}$$

slik at

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)+\hat{p}_2(1-\hat{p}_2)}{3000}}} \overset{tiln.}{\sim} N(0,1).$$

Vi ønsker å teste å teste om det er en signifikant forskjell mellom de to andelene, altså $H_0 : p_1 = p_2 = p$ mot $H_a : p_1 \neq p_2$. Vi kan bruke testobservatoren Z med $p_1 = p_2 = p$ i testen, som er

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\frac{2}{3000}}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\frac{1}{1500}}},$$

der $\hat{p} = \frac{m}{m+n}\hat{p}_1 + \frac{n}{m+n}\hat{p}_2 = \frac{1}{2}\hat{p}_1 + \frac{1}{2}\hat{p}_2$. Vi forkaster da H_0 ved 5% signifikansnivå dersom $Z < -z_{0.025} = -1.96$ eller $Z > z_{0.025} = 1.96$. Når vi setter inn, får vi $z_{obs} = 1.61$, altså forkaster vi ikke nullhypotesen om at andelene menn og kvinner som ofte opplever problemer med tidsklemma er like ved 5% signifikansnivå . P-verdien er gitt ved:

$$2\mathrm{P}(Z > |z_{obs}|) = 0.108,$$

hvilket er et godt stykke fra å forkaste H_0.

b)

```
Utskriften fra \texttt{prop.test()} i R er:
2-sample test for equality of proportions
without continuity correction

data:  c(p1.hat * m, p2.hat * n) out of c(m, n)
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 får samme P-verdi som i a). I R-funksjonen brukes testobservatoren Z^2 i stedet for Z, og denne er tilnærmet χ^2_1-fordelt. Vi har $z^2_{obs} = 2.58$, altså det samme som i utskriften fra R.

Oppgave 4

a) La Y_i og x_i være henholdsvis vannstanden og snøinnholdet målt i sesong i, $i = 1, \ldots, 18$. Da antar vi følgende lineære regresjonsmodell:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

med $\epsilon_i \overset{uif}{\sim} N(0, \sigma^2)$, $i = 1, \ldots, 18$. Utskriften fra R-funksjonen `lm()` er:

```
Call:
lm(formula = vann ~ snoe, data = snoe.vann)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7341 -1.4207 -0.1391  1.5444  3.3584

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.28001    1.71191   0.164    0.872
snoe         0.50558    0.05508   9.180 8.91e-08 ***
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.943 on 16 degrees of freedom
Multiple R-squared:  0.8404,Adjusted R-squared:  0.8305
F-statistic: 84.27 on 1 and 16 DF,  p-value: 8.913e-08
```

Vi ser at de estimerte koeffisientene er $\hat{\beta}_0 = 0.28$ og $\hat{\beta}_1 = 0.51$. Disse virker rimelige, vannstanden øker med snøinnholdet, og konstantleddet er ikke svært stort. Tilpasningen ser også bra ut i Figur 3.

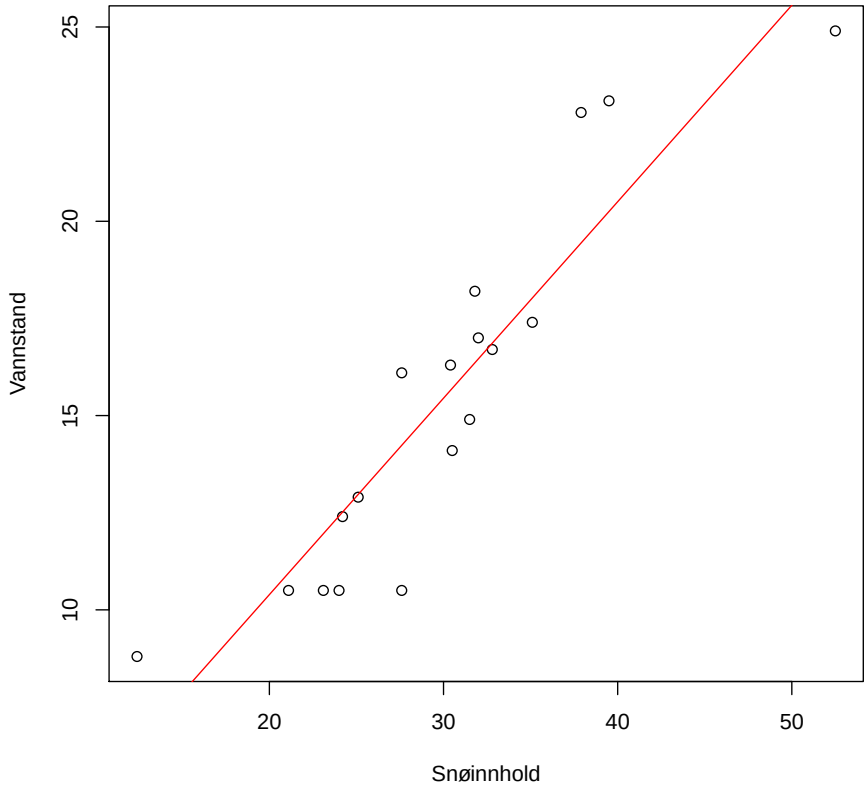

Figure 3: Vannstand mot snøinnhold sammen med tilpasset regresjonslinje.

b) Figur 4 viser residualene plottet mot forklaringsvariabelen (til venstre). Denne viser ikke noe tydelig mønster eller andre tegn til spesielle avvik fra modellantakelsene. Til høyre i figuren ser vi et normalfordelingsplott av residualene. Det gir heller ingen indikasjoner på at antakelsen om normalfordelte

residualer er feil. Ut fra disse resultatene, samt de fra a) ser det altså ut til at modellen passer ganske bra til dataene.

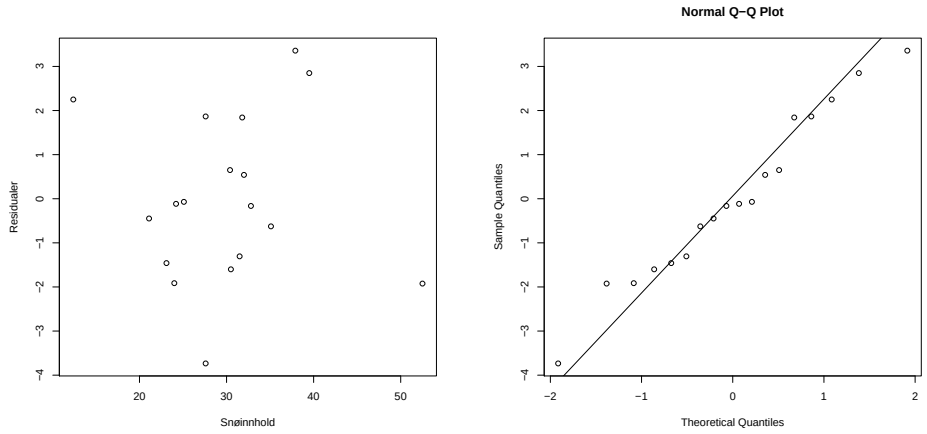

Figure 4: Residualer plottet mot forklaringsvariabelen (til venstre) og normalfordelingsplott av residualene (til høyre)

c) Vi vet at $\hat{beta}_1 \sim N(\beta_1, \mathrm{Var}(\hat{\beta}_1))$, der

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Residualene ϵ_i kan estimeres med $\hat{\epsilon}_i = y_i - \hat{\beta}_o - \hat{\beta}_1 x_i$, slik at en estimator for variansen σ^2 er

$$S^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

Dermed kan $\mathrm{Var}(\hat{\beta}_1)$ estimeres med

$$S^2_{\hat{\beta}_1} = \frac{S^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Vi får:

$$\mathrm{P}\left(-t_{0.025,n-2} \leq \frac{\hat{\beta}_1 - \beta_1}{S/\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \leq t_{0.025,n-2}\right) = 0.95$$

$$\rightarrow \mathrm{P}\left(\hat{\beta}_1 - t_{0.025,n-2}S/\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \leq \beta_1 \leq \hat{\beta}_1 + t_{0.025,n-2}S/\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) = 0.95.$$

Et 95% konfidensintervall er derfor $(\hat{\beta}_1 - t_{0.975,n-2}s_{\hat{\beta}_1}, \hat{\beta}_1 - t_{0.975,n-2}s_{\hat{\beta}_1})$, og når vi setter inn verdier fra dataene får vi $(0.389, 0.622)$

d) Vi vet at $\hat{beta}_0 \sim N(\beta_0, \text{Var}(\hat{\beta}_0))$, der

$$\text{Var}(\hat{\beta}_0) = \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \sigma^2,$$

som kan estimeres med

$$S^2_{\hat{\beta}_0} = \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) S^2$$

med S^2 som i c). Videre vet vi at S^2 og $\hat{\beta}_0$ er uavhengige, og at $(n-2)S^2/\sigma^2 \sim \chi^2_{n-2}$, slik at

$$\frac{\hat{\beta}_0 - \beta_0}{S_{\hat{\beta}_0}} \sim t_{n-2}.$$

vi får dermed en test for $H_0 : \beta_0 = 0$ mot $H_a : \beta_0 \neq 0$ med signifikansnivå 5% dersom vi forkaster H_0 hvis $T = \hat{\beta}_0 / S_{\hat{\beta}_0} \leq -t_{0.025,n-2}$ eller $T \geq t_{0.025,n-2}$. Når vi setter inne verdier fra dataene, får vi $t_{obs} = 0.164$, mens $t_{0.025,n-2} = t_{0.025,16} = 2.12$. Altså forkaster vi ikke nullhypotesen om at $\beta_0 = 0$, altså at modellen for vannstand ikke trenger noe konstantledd. Det betyr i praksis at vanstanden er omtrent proporsjonal med snøinnholdet. I og med at t-fordelingen er symmetrisk om 0, er P-verdien gitt ved

$$2\text{P}(T \geq |t_{obs}|) = 0.872,$$

som er veldig høy. Vi er altså ikke i nærheten av å forkaste H_0. Vi ser at disse resultatene er helt i overensstemmelse med utskriften fra funksjonen `lm()` i a).

R-kode

```
### Oppgave 1

##Lesedata

d = read.table("https://www.uio.no/studier/emner/matnat/math/STK1110/data/temp.txt",header =
x = d$Menn
m = length(x)
y = d$Kvinner
n = length(y)

##a)

boxplot(d)

##b)

qqnorm(x)
qqline(x)
```

```
qqnorm(y)
qqline(y)

##c)

s2.p  =  ((m-1)*var(x)+(n-1)*var(y))/(m+n-2)
t = (mean(x)-mean(y))/sqrt(s2.p*(1/m+1/n))
nu = m+n-2
p.value = 2*(1-pt(abs(t),nu))
signif(p.value,3)

alpha = 0.05
signif(mean(x)-mean(y)+qt(c(alpha/2,1-alpha/2),nu)*sqrt(s2.p*(1/m+1/n)),3)

t.test(x,y,var.equal=TRUE)

## d)

se1 = sd(x)/sqrt(m)
se2 = sd(y)/sqrt(n)
t = (mean(x)-mean(y))/sqrt(se1^2+se2^2)
nu = (se1^2+se2^2)^2/(se1^4/(m-1)+se2^4/(n-1))
p.value = 2*(1-pt(abs(t),nu))
signif(p.value,3)

alpha = 0.05
signif(mean(x)-mean(y)+qt(c(alpha/2,1-alpha/2),nu)*sqrt(se1^2+se2^2),3)

t.test(x,y,var.equal = FALSE)

## e)

f = var(x)/var(y)
alpha = 0.05
f.crit.low = qf(1-alpha/2,m-1,n-1,lower.tail=FALSE)
f.crit.high = qf(alpha/2,m-1,n-1,lower.tail=FALSE)
signif(c(f,f.crit.low,f.crit.high),3)

var.test(x,y,alternative="two.sided",conf.level=1-alpha)

## f)

v = s2.p*(2+1/m+1/n)
signif(mean(x)-mean(y)+qt(c(alpha/2,1-alpha/2),nu)*sqrt(v),3)
```

```
### Oppgave 2

## b)

m = -3.26
se = 1.58
n = 31
t = -m/se
p.value = 2*(1-pt(abs(t),n-1))
signif(p.value,3)

## c)

alpha = 0.05
signif(m+qt(c(alpha/2,1-alpha/2),n-1)*se,3)

### Oppgave 3

## a)

p1.hat = 0.162
p2.hat = 0.147
m = n = 3000
p.hat = 0.5*p1.hat+0.5*p2.hat
z.obs = (p1.hat-p2.hat)/sqrt(p.hat*(1-p.hat)/1500)
p.value = 2*(1-pnorm(z.obs))
signif(c(z.obs,p.value),3)

## b)

alpha = 0.05
prop.test(c(p1.hat*m,p2.hat*n),c(m,n),alternative="two.sided",conf.level=1-alpha,correct=FAI


### Oppgave 4

## Lese dataene

snoe.vann <-read.table("http://www.uio.no/studier/emner/matnat/math/STK1110/data/snoe_vann.t
colnames(snoe.vann) = c("snoe","vann")

## a)

n = nrow(snoe.vann)
x = snoe.vann$snoe # kovariat
y = snoe.vann$vann #   respons
```

```
mod = lm(vann~snoe,data=snoe.vann)
summary(mod)

plot(x,y,xlab="Snøinnhold",ylab="Vannstand")
abline(mod$coeff[1],mod$coeff[2],col=2)

## b)

plot(x,mod$res,xlab="Snøinnhold",ylab="Residualer")

qqnorm(mod$res)
qqline(mod$res)

## c)

alpha = 0.05
res = mod$res
s.2 = sum(res^2)/(n-2)
s.beta1hat.2 = s.2/sum((x-mean(x))^2)
beta.1.hat = mod$coeff[2]
signif(beta.1.hat + qt(c(0.975,0.025),df=n-2,lower.tail=FALSE)*sqrt(s.beta1hat.2),3)

## d)

alpha = 0.05
s.beta0hat.2 = (1/n+mean(x)^2/sum((x-mean(x))^2))*s.2
beta.0.hat = mod$coeff[1]
t.obs = beta.0.hat/sqrt(s.beta0hat.2)
t.crit = qt(alpha/2,n-2,lower.tail = FALSE)
p.value = 2*(1-pt(abs(t.obs),n-2))
signif(c(t.obs,t.crit,p.value),3)
```