STK1110
Statistiske metoder og dataanalyse

OBLIG 2

Egil Furnes
Student: 693784

Oppgave 1

(a)

Figur 1 viser et boksplott av vektene fordelt på de to fôrtypene. Vi ser at medianvekten for fisk som har fått fôr B ligger høyere enn for fisk som har fått fôr A. I tillegg virker fordelingen for fôr B generelt å ligge høyere, med unntak av noen få observasjoner. Dette gir en første indikasjon på at fôr B kan gi større vekt, men forskjellen må undersøkes statistisk.

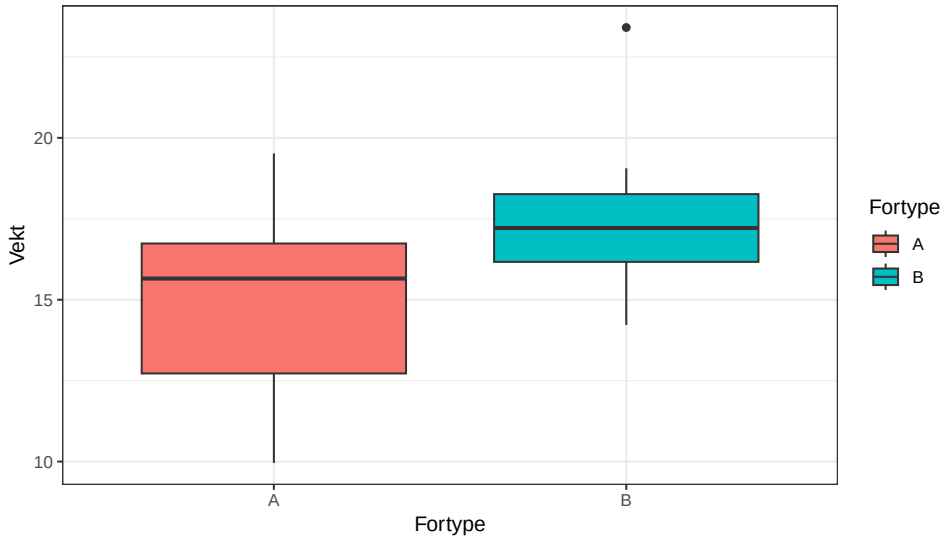

Figure 1: Boxplot av vekt fordelt på fôrtype

(b)

Normalfordelingsplottene i figur 2 viser at observasjonene i begge grupper ligger relativt tett langs referanselinjen. Det tyder på at antakelsen om normalfordelte data er rimelig. Ingen av gruppene viser klare tegn til outliers eller sterk skjevhet, og en parametrisk t-test er derfor passende.

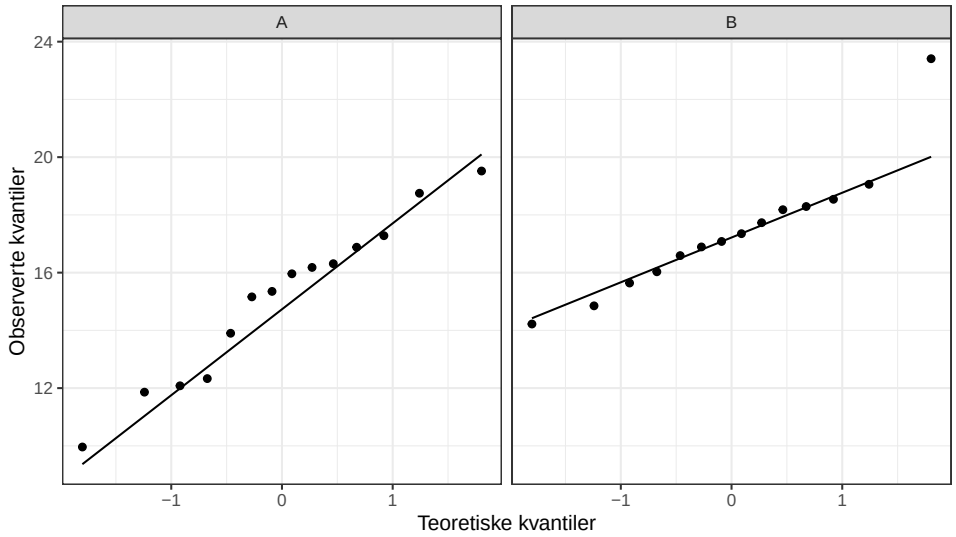

Figure 2: Normalfordelingsplot av vekt fordelt på fôrtype

(c)

Vi ønsker å undersøke om fôr B gir høyere gjennomsnittsvekt enn fôr A. Hypotesene formuleres som:

$$H_0 : \mu_B = \mu_A \qquad \text{(ingen forskjell)}$$

$$H_a : \mu_B > \mu_A \qquad \text{(fôr B gir høyere vekt)}$$

Når vi antar lik varians, bruker vi testobservatoren

$$T = \frac{\bar{X}_B - \bar{X}_A}{s_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}}$$

der s_p^2 er det poolede variansestimatet. Under H_0 følger T en t-fordeling med $n_A + n_B - 2 = 26$ frihetsgrader.

(d)

Den observerte testverdien ble

$$t_{\text{obs}} = 2.44, \quad \text{df} = 26, \quad p = 0.011.$$

Siden $p < 0.05$ forkaster vi H_0, og konkluderer med at det er statistisk signifikant bevis for at fisk som får fôr B er tyngre enn fisk som får fôr A. Dette samsvarer med `t.test()` med antatt lik varians, som også gir $p = 0.011$.

```
> c(t_verdi = t_obs,
+   frihetsgrader = df,
```

```
+   p_ensidig = p_manual)
      t_verdi frihetsgrader     p_ensidig
   2.43602673   26.00000000    0.01100432
>
> t.test(B, A, var.equal = TRUE, alternative = "greater")

        Two Sample t-test

data:  B and A
t = 2.436, df = 26, p-value = 0.011
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.6926214       Inf
sample estimates:
mean of x mean of y
 17.41857  15.10857
```

(e)

Når vi ikke antar lik varians (Welch-test), får vi

$$t = 2.44, \quad p = 0.011.$$

Konklusjonen er dermed den samme: fôr B gir signifikant høyere vekt.

En F-test for forskjell i varians gir

$$p = 0.44,$$

som er langt over 0.05. Det finnes derfor ingen grunn til å anta ulike varianser. Dette forklarer hvorfor Welch-testen og testen med lik varians gir nesten identiske resultater.

```
> t.test(B, A, var.equal = FALSE, alternative = "greater")

        Welch Two Sample t-test

data:  B and A
t = 2.436, df = 24.858, p-value = 0.01118
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.6898736       Inf
sample estimates:
mean of x mean of y
 17.41857  15.10857

> var.test(B, A)

        F test to compare two variances
```

```
data:  B and A
F = 0.64701, num df = 13, denom df = 13, p-value = 0.4431
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.2077054 2.0154587
sample estimates:
ratio of variances
         0.6470098
```

(f)

Regresjonsmodellen

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad x_i = \begin{cases} 0 & \text{for fôr A} \\ 1 & \text{for fôr B} \end{cases}$$

gir $\hat{\beta}_1 = 2.31$ med $p = 0.022$. I regresjonsmodellen testes

$$H_0 : \beta_1 = 0 \quad \text{mot} \quad H_a : \beta_1 \neq 0,$$

som er en *tosidig* test. I del (c) brukte vi en *ensidig* test $H_a : \mu_B > \mu_A$. Derfor blir p-verdien i regresjonsmodellen omtrent dobbelt så stor som p-verdien i (c). Likevel er $p < 0.05$, og vi kommer til samme konklusjon: fôr B gir signifikant høyere vekt.

```
Call:
lm(formula = Vekt ~ x, data = f)

Residuals:
    Min      1Q  Median      3Q     Max
-5.1486 -1.4861  0.1464  1.1414  5.9914

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  15.1086     0.6705  22.532   <2e-16 ***
x             2.3100     0.9483   2.436    0.022 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.509 on 26 degrees of freedom
Multiple R-squared:  0.1858,        Adjusted R-squared:  0.1545
F-statistic: 5.934 on 1 and 26 DF,  p-value: 0.02201
```

R-kode

```r
library(tidyverse)
set.seed(1110)

d <- read.table("data/kveite_for.txt", sep = "", header = TRUE) %>%
  as_tibble() %>%
  mutate(Fortype = as.factor(Fortype))

# a)
plot_a <- d %>%
  ggplot(aes(x = Fortype, y = Vekt, fill = Fortype)) +
  geom_boxplot() +
  theme_bw()

ggsave("plots/oppg1_a.pdf", plot_a, width = 7, height = 4)

# b)
plot_b <- d %>%
  ggplot(aes(sample = Vekt)) +
  stat_qq() +
  stat_qq_line() +
  facet_wrap(~ Fortype) +
  theme_bw() +
  labs(x = "Teoretiske kvantiler", y = "Observerte kvantiler")

ggsave("plots/oppg1_b.pdf", plot_b, width = 7, height = 4)

# c)
A <- d$Vekt[d$Fortype == "A"]
B <- d$Vekt[d$Fortype == "B"]

# d)
sp2 <- ((var(A) * (length(A) - 1) + var(B) * (length(B) - 1)) /
          (length(A) + length(B) - 2))

t_obs <- (mean(B) - mean(A)) /
  sqrt(sp2 * (1/length(A) + 1/length(B)))

df <- length(A) + length(B) - 2
p_manual <- 1 - pt(t_obs, df)

c(t_verdi = t_obs,
  frihetsgrader = df,
  p_ensidig = p_manual)

t.test(B, A, var.equal = TRUE, alternative = "greater")

# e)
t.test(B, A, var.equal = FALSE, alternative = "greater")
var.test(B, A)

```

```r
# f)
f <- d %>% mutate(x = ifelse(Fortype == "B", 1, 0))
fit <- lm(Vekt ~ x, data = f)
summary(fit)
```

Oppgave 2

(a)

Vi sammenligner to behandlinger på samme individer, og ser på forskjellen i vekt før og etter. Siden hvert individ bidrar med to observasjoner som henger sammen, er dataene *avhengige*. Den riktige testen er derfor en *paret t-test*, ikke en to-utvalgs uavhengig t-test. Testvariabelen baseres på gjennomsnittet av differansene.

```r
# a) riktige forutsetninger
# Parete data --> bruk en paret t-test
```

(b)

Vi tester om gjennomsnittlig forskjell er lik null.

$$H_0 : \mu_D = 0 \qquad H_a : \mu_D \neq 0$$

Den observerte differansen var $\bar{D} = -3.26$. Standardfeilen blir

$$SE = \frac{sd_D}{\sqrt{n}} \approx \frac{8.81}{\sqrt{31}} \approx 1.58.$$

Dette gir testobservatoren

$$t = \frac{-3.26}{1.58} = -2.06, \qquad df = 30,$$

og

$$p = 0.048.$$

Siden $p < 0.05$, forkaster vi H_0. Det er altså signifikant evidens for at gjennomsnittlig forskjell ikke er null, og i dette tilfellet er differansen negativ, noe som betyr at de målte verdiene etter behandling er lavere i gjennomsnitt.

```
> c(t_verdi = t_obs,
+   frihetsgrader = df,
+   p_tosidig = pval)
      t_verdi frihetsgrader     p_tosidig
  -2.06026241   30.00000000   0.04813865
```

(c)

Et 95% konfidensintervall for middelverdien til differansene blir

$$(-6.49,\ -0.03).$$

Siden intervallet ligger helt under 0, støtter også konfidensintervallet konklusjonen om at verdiene etter behandling er lavere i gjennomsnitt. Beregnet standardfeil fra dataene blir

$$SE = \frac{8.81}{\sqrt{31}} = 1.582,$$

som stemmer godt med oppgitt verdi $SE = 1.58$.

```
> c(ci_95_lower = ci_low,
+   ci_95_upper = ci_high)
ci_95_lower ci_95_upper
-6.49153409 -0.02846591
>
> c(se_beregnet = sdD / sqrt(n),
+   se_gitt = 1.58)
se_beregnet     se_gitt
   1.582323    1.580000
```

R-kode

```r
library(tidyverse)
set.seed(1110)

n     <- 31
meanA <- 93.32
sdA   <- 15.41
meanB <- 96.58
sdB   <- 13.84
meanD <- -3.26
sdD   <- 8.81

# a)

# b)
t_obs <- meanD / (sdD / sqrt(n))
df    <- n - 1
pval  <- 2 * (1 - pt(abs(t_obs), df))

c(t_verdi = t_obs,
  frihetsgrader = df,
  p_tosidig = pval)
```

```r

# c)
t_crit <- qt(0.975, df)

ci_low  <- meanD - t_crit * sdD / sqrt(n)
ci_high <- meanD + t_crit * sdD / sqrt(n)

c(ci_95_lower = ci_low,
  ci_95_upper = ci_high)

c(se_beregnet = sdD / sqrt(n),
  se_gitt = 1.58)
```

Oppgave 3

(a)

Vi sammenligner andelen som ofte opplever tidsklemme hos fedre og mødre. Estimatene er

$$\hat{p}_{\text{menn}} = \tfrac{486}{3000} = 0.162, \qquad \hat{p}_{\text{kvinner}} = \tfrac{441}{3000} = 0.147, \qquad \hat{p}_{\text{menn}} - \hat{p}_{\text{kvinner}} = 0.015.$$

Vi tester

$$H_0 : p_{\text{menn}} = p_{\text{kvinner}} \quad \text{mot} \quad H_a : p_{\text{menn}} \neq p_{\text{kvinner}}.$$

Med pooled andel $\hat{p} = (486 + 441)/(3000 + 3000) = 0.1545$ blir

$$SE = \sqrt{\hat{p}(1-\hat{p})\left(\tfrac{1}{3000} + \tfrac{1}{3000}\right)} \approx 0.00933, \quad z = \frac{0.015}{SE} \approx 1.61,$$

og tosidig p-verdi ≈ 0.108. Vi forkaster *ikke* H_0 ved 5 %-nivå: forskjellen på ca. 1.5 prosentpoeng er ikke statistisk signifikant.

```
> c(z_verdi = z_obs,
+   p_tosidig = pval)
  z_verdi p_tosidig
1.6073699 0.1079732
```

(b)

Kontroll med `prop.test()` (uten kontinuitetskorreksjon) gir samme konklusjon: $p \approx 0.108$, og 95 % KI for $(p_{\text{menn}} - p_{\text{kvinner}})$ er $(-0.0033,\ 0.0333)$, som inkluderer 0. Dermed finner vi ikke signifikant forskjell mellom fedre og mødre i denne undersøkelsen.

```
> prop.test(c(x_menn, x_kvinner),
+           c(n_menn, n_kvinner),
+           alternative = "two.sided",
+           correct = FALSE)

        2-sample test for equality of proportions without continuity correction

data:  c(x_menn, x_kvinner) out of c(n_menn, n_kvinner)
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
```

R-kode

```r
library(tidyverse)
set.seed(1110)

n_menn    <- 3000
n_kvinner <- 3000
x_menn    <- 486
x_kvinner <- 441

p_menn    <- x_menn / n_menn
p_kvinner <- x_kvinner / n_kvinner
diff_hat  <- p_menn - p_kvinner

# a)
p_pooled <- (x_menn + x_kvinner) / (n_menn + n_kvinner)
se       <- sqrt(p_pooled * (1 - p_pooled) * (1/n_menn + 1/n_kvinner
   ))
z_obs    <- diff_hat / se
pval     <- 2 * (1 - pnorm(abs(z_obs)))

c(z_verdi = z_obs,
  p_tosidig = pval)

# b)
prop.test(c(x_menn, x_kvinner),
          c(n_menn, n_kvinner),
          alternative = "two.sided",
          correct = FALSE)
```

Oppgave 4

(a)

Vi tilpasser en enkel lineær regresjonsmodell

$$Vannstand = \beta_0 + \beta_1 \cdot Snoinnhold + \varepsilon.$$

Resultatene fra `lm()` er vist under. Estimatet $\hat{\beta}_1 = 0.506$ betyr at vannstanden øker med omtrent 0.51 enheter for hver økning på 1 i snøinnhold. Koeffisienten er svært signifikant ($p < 0.000001$), og modellen forklarer omtrent 84% av variasjonen i vannstanden ($R^2 = 0.84$). Det tyder på en klar lineær sammenheng mellom snøinnhold og vannstand.

```
Call:
lm(formula = Vannstand ~ Snoinnhold, data = d)

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
Snoinnhold   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
```

Figur 3 viser regresjonslinjen sammen med datapunktene. Punktene ligger tett langs linjen, noe som samsvarer med høy forklaringsgrad.

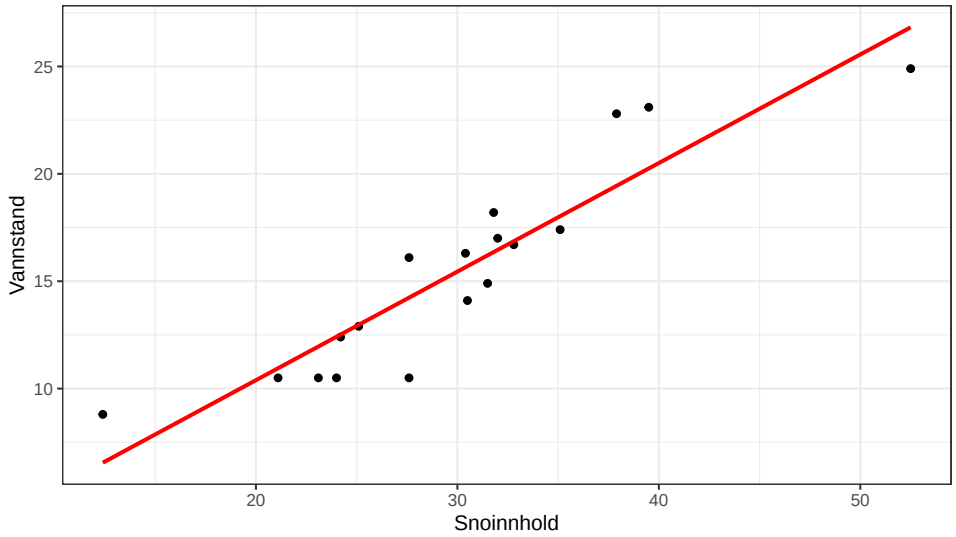

Figure 3: Spredningsplott med tilpasset regresjonslinje

(b)

Residualplottet (figur 4) viser ingen tydelig systematikk: residualene ligger tilfeldig rundt null-linjen. Det tyder på at en lineær modell passer dataene godt. QQ-plottet (figur 5) viser at residualene følger normalfordeling rimelig godt. Det er ingen sterke avvik eller skjevheter. Det virker derfor forsvarlig å bruke vanlige t- og F-tester i analysen.

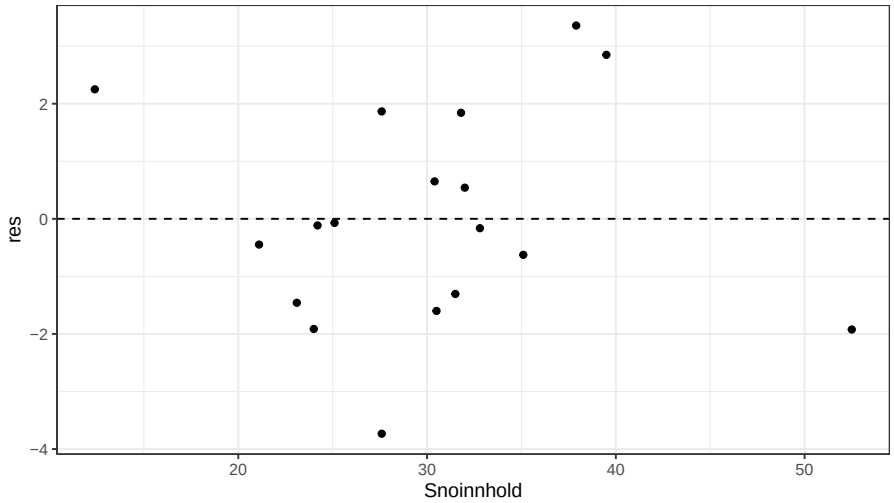

Figure 4: Residualer mot snøinnhold

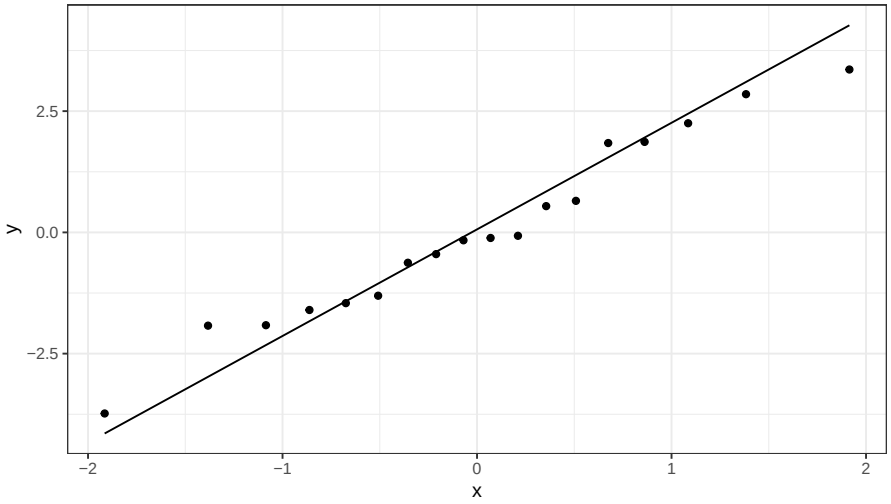

Figure 5: QQ-plott av residualer

(c)

Estimert feilvarians ble

$$\hat{\sigma}^2 = 3.77.$$

En 95% konfidensintervall for β_1 ble

$$0.389 \leq \beta_1 \leq 0.622.$$

Siden hele intervallet er positivt, støtter dette at vannstanden øker når snøinnholdet øker.

```
> confint(fit, "Snoinnhold", level = 0.95)
               2.5 %    97.5 %
Snoinnhold 0.3888275 0.6223388
```

(d)

Vi tester $H_0 : \beta_0 = 0$. Estimatet er $\hat{\beta}_0 = 0.28$ og p-verdien er $p = 0.87$. Det er dermed *ingen* statistisk støtte for at konstantleddet er forskjellig fra null, noe som ikke har noen praktisk betydning for modellens tolkning: det viktige poenget er at β_1 er klart positiv og signifikant.

```
> c(beta0 = beta0, SE_beta0 = se_b0, t_verdi = t_b0, p_tosidig = p_b0)
    beta0.(Intercept) SE_beta0    t_verdi.(Intercept)   p_tosidig.(Intercept)
            0.2800063 1.7119069             0.1635640             0.8721226
```

(e)

Vi tester om en modell med polynom av 2. og 3. grad gir bedre forklaring enn den enkle lineære modellen. Resultatene viser:

- Andregradspolynom gir *ingen* signifikant forbedring. Koeffisienten for x^2 har p-verdi 0.99 og R^2 blir ikke bedre. - Tredjegradspolynom gir høyere R^2 (fra 0.84 til 0.91), og alle leddene blir signifikante.

Men når vi sammenligner modeller med AIC og BIC ser vi at: - AIC er lavest for tredjegradspolynom (72.5), - BIC er også lavest for tredjegradspolynom (76.9),

som betyr at fit3 er statistisk best. Likevel må man være forsiktig: tredjegradspolynom gir mer kurving og høyere risiko for overtilpasning, og det er få observasjoner ($n = 18$). Lineær modell kan være enklere å tolke og kan fortsatt være tilstrekkelig.

```
> summary(fit2)

Call:
lm(formula = Vannstand ~ poly(Snoinnhold, 2, raw = TRUE), data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7366 -1.4212 -0.1411  1.5409  3.3563

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      2.338e-01  4.204e+00   0.056   0.9564
poly(Snoinnhold, 2, raw = TRUE)1  5.087e-01  2.600e-01   1.956   0.0693 .
poly(Snoinnhold, 2, raw = TRUE)2 -4.741e-05  3.918e-03  -0.012   0.9905
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.007 on 15 degrees of freedom
Multiple R-squared:  0.8404,        Adjusted R-squared:  0.8192
F-statistic:  39.5 on 2 and 15 DF,  p-value: 1.052e-06
```

```
> summary(fit3)

Call:
lm(formula = Vannstand ~ poly(Snoinnhold, 3, raw = TRUE), data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-3.04811 -0.73756 -0.01764  0.97302  2.55189

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)                       23.4791445  7.7538782   3.028  0.00903 **
poly(Snoinnhold, 3, raw = TRUE)1 -2.1245867  0.8220027  -2.585  0.02161 *
poly(Snoinnhold, 3, raw = TRUE)2  0.0893035  0.0272093   3.282  0.00545 **
poly(Snoinnhold, 3, raw = TRUE)3 -0.0009189  0.0002781  -3.305  0.00522 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.557 on 14 degrees of freedom
Multiple R-squared:  0.9104,        Adjusted R-squared:  0.8911
F-statistic: 47.39 on 3 and 14 DF,  p-value: 1.404e-07
```

```
> AIC(fit, fit2, fit3)
     df      AIC
fit   3 78.87098
fit2  4 80.87081
fit3  5 72.49189
> BIC(fit, fit2, fit3)
     df      BIC
fit   3 81.54210
fit2  4 84.43229
fit3  5 76.94375
```

R-kode

```r
library(tidyverse)
set.seed(1110)

# problem 4

d <- read.table("data/snoe_vann.txt", header = FALSE) %>%
  as_tibble() %>%
  rename(Snoinnhold = V1, Vannstand = V2)


# a)

fit <- lm(Vannstand ~ Snoinnhold, data = d)
summary(fit)

plot_a <- d %>%
  ggplot(aes(x = Snoinnhold, y = Vannstand)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  theme_bw()

ggsave("plots/oppg4_a.pdf", plot_a, width = 7, height = 4)

```

```r

# b)

res <- residuals(fit)

plot_resid <- ggplot(data.frame(Snoinnhold = d$Snoinnhold, res = res),
                     aes(x = Snoinnhold, y = res)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  theme_bw()

ggsave("plots/oppg4_residualer.pdf", plot_resid, width = 7, height = 4)

plot_qq <- ggplot(data.frame(res), aes(sample = res)) +
  stat_qq() +
  stat_qq_line() +
  theme_bw()

ggsave("plots/oppg4_residual_qq.pdf", plot_qq, width = 7, height = 4)


# c)

sigma2_hat <- sum(res^2) / (nrow(d) - 2)
sigma2_hat

confint(fit, "Snoinnhold", level = 0.95)


# d)

beta0 <- coef(fit)[1]
se_b0 <- summary(fit)$coefficients[1, 2]
t_b0 <- beta0 / se_b0
p_b0 <- 2 * (1 - pt(abs(t_b0), df = nrow(d) - 2))

c(beta0 = beta0,
  SE_beta0 = se_b0,
  t_verdi = t_b0,
  p_tosidig = p_b0)


# e)

fit2 <- lm(Vannstand ~ poly(Snoinnhold, 2, raw = TRUE), data = d)
fit3 <- lm(Vannstand ~ poly(Snoinnhold, 3, raw = TRUE), data = d)

summary(fit2)
```

```r
summary(fit3)

AIC(fit, fit2, fit3)
BIC(fit, fit2, fit3)
```