Oblig 2 - STK1110

Oppg. 1

```
path="https://www.uio.no/studier/emner/matnat/math/STK1110/data/temp.txt"
data =read.table(path,header=T)


# a. boxplot og normalfordelingsplot.
boxplot(data, xlab="kjønn", ylab="C", main="Kroppstemperatur pr. kjønn")

qqnorm(data$Menn, main="qqnorm Menn")
qqline(data$Menn)
qqnorm(data$Kvinner, main="qqnorm Kvinner")
qqline(data$Kvinner)
#ser nogen lunde normalt ut. Dog men har litt større ekstremaliteter

# b. alfa = 0.05, H0: mu_menn == mu_kvinner altså del_0 = 0, Ha: mu_menn != mu_kvinner
# x:menn, x:kvinner
alfa =0.05
x = data$Menn
y = data$Kvinner
#mellomregning
mean.x = mean(x)
mean.y = mean(y)
sd.x = sd(x)
sd.y = sd(y)
m = length(x)
n = length(y)
se.x = sd.x/sqrt(m)
se.y = sd.y/sqrt(n)

df = ( (se.x)^2 + (se.y)^2 )^2 / ( (((se.x)^4)/(m-1)) + ((se.y)^4/(n-1)) )
t_obs = (mean.x - mean.y) / sqrt( (sd.x^2)/m + (sd.y^2)/n )

p.t_obs = pt(t_obs, df) # = 0.015 > alfa/2 så vi kan forkaste null hypotesen med stor margin
# todo: ser ut til av vi skal ha sd.x = sd.y, må finne mer ut om dette.

#c. F-test
f = Sd.x^2/sd.x^2
```

Oppg 2

```
# oppg 2

# twin A: at biological parents, twin B: not
# Q: is there difference in IQ

#data
A.n = 31
A.mean = 93.32
A.sd = 15.41
A.se_mean = 2.77

B.n = 31
B.mean = 96.58
B.sd = 13.84
B.se_mean = 2.49

diff.N = 31
diff.mean = -3.26
diff.sd = 8.81
diff.se_mean = 1.58

#a
# Vi burde ha en paret sammenligning siden:
  # Vi sammenligner to resulater, og ikke et nytt resultat mot et tidligere.
# Vi maa annta at:
  # X1, X2, ... , Xm er uavhenige og identisk fordelte, med forventingsverdi mu_x
og varians sigma_x
  # tillsvarende for Y.
  # og vi må ha at Xi og Yi er uavhengige.

#b
#Hypotesetest: H0: A - B = mu_D ==  0, mot Ha: A - B = mu_D = 0

mu_D = A.mean - B.mean
#df
#se.x = sd.x/sqrt(m)
#se.y = sd.y/sqrt(n)

df = ( (A.se_mean)^2 + (B.se_mean)^2 )^2 / ( (((A.se_mean)^4)/(A.n-1)) +
((B.se_mean)^4/(B.n-1)) ) # frihetsgrader
t_obs = (A.mean - B.mean - 0) / sqrt( (A.sd/A.n) + (B.sd/B.n))

P.val = pt(t_obs, df) # hvis denne er regnet rett er det stor grunn til å forkaste
H0

#c
alfa = 0.05
diff.Ci = diff.mean + c(1,-1) * qt(alfa/2, df) * sqrt( (A.sd/A.n) + (B.sd/A.n))
# det at diff.Ci bare dekker negative verdier forteller at vi er mer en .95 sikker
på at den sanne verdien er negative, siden vi er .95 sikker på at den sanne
verdien ligger innenfor vaart CI.
# En slik tosidig test, tester om vi er lengere unna, enn an størrelse beskrevet
av alfa/2, mu fra H0.
```

```
# Mens en ensidig test tester om test observatoren er en sørrelse beskrevet av
alfa, mindre enn mu fra H0.
```

Oppg 3

```
# notation: Yi = b0 + b1*Xi + ei
# b1: slope coefficient,
# Y: responsvariabel
# xi forklaringsvariabel.
path="https://www.uio.no/studier/emner/matnat/math/STK1110/data/plastic.txt"
plast=read.table(path,header=T)

# a
# Strenght = respons = Y
# Temperature = forklaringsvariable = Xi
Y = plast$Strength
X = plast$Temperature
X.mean = mean(X)
Y.mean = mean(Y)
X.n = length(X)
Y.n = length(Y)

b1 = (sum((X-X.mean)*(Y-Y.mean))) / sum((X-X.mean)**2)# = Sxy/Sxx
b0 = (sum(Y) - b1*sum(X))/X.n # Y.mean - b1*X.mean

# plot
plot(X, Y, xlab="Temperature", ylab="Strength")
abline(b0, b1)
# kommentar:
    # Regresjonslinjen ser ut til å stemme. Standardfeilen er dog stor rundt Temp
= 200

# b) CI for regresjonskoeffisienten for Temperature. Git intervallet indikasjon på
om temperatur er en viktig forklaringsvariabel
# CI for the slope B1:
    # b1 +- t_alfa/2,n-2 * Sb1
alfa = 0.05
S2 = (sum(Y^2) - b0*sum(Y) - b1*sum(X*Y)) / (X.n-1)
S = sqrt(S2)
Sxx =  sum((X- X.mean)^2)
Sb1 = S/sqrt(Sxx)

b1.CI = b1 + c(1,-1) * qt(alfa/2, X.n-2) * Sb1
t_ratio = b1 / Sb1

# t_ratio er betydelig større en max(b1.CI) så dermed forkaster vi b1.CI

# c)
Sy = S * sqrt( (1/Y.n) + (X - X.mean)^2/Sxx ) # dont know wich N to use
Y.CI.m = b0+b1*X - qt(alfa/2, Y.n-2) * Sy # same here
```

```
Y.CI.p = b0+b1*X + qt(alfa/2, Y.n-2) * Sy # same here

# PI
  # b0 + b1*x +- t(alfa/2, n-2) * S * sqrt( 1 + 1/n + (X - X.mean)/sXX )
x_star = c(210, 240, 270)
y_hat = b0 + b1 * x_star
PI <- function(b0, b1, x, x_mean, S, Sxx, n) {
    p = b0 + b1*x + qt(alfa/2, n-2) * S * sqrt(1+1/n+(x-x_mean)^2/Sxx)
    m = b0 + b1*x - qt(alfa/2, n-2) * S * sqrt(1+1/n+(x-x_mean)^2/Sxx)
    res = data.frame(p, m)
    colnames(res) <- c("p", "m")

    return(res)
}
# Y.PI.x_star.p = ( b0 + b1*x_star ) + qt(alfa/2, Y.n-2) * S * sqrt( 1 + 1/Y.n +
(X - X.mean)/Sxx )
# Y.PI.x_star.m = ( b0 + b1*x_star ) - qt(alfa/2, Y.n-2) * S * sqrt( 1 + 1/Y.n +
(X - X.mean)/Sxx )
Y.PI.1 = PI(b0, b1, x_star[1], X.mean, S, Sxx, Y.n)
Y.PI.2 = PI(b0, b1, x_star[2], X.mean, S, Sxx, Y.n)
Y.PI.3 = PI(b0, b1, x_star[3], X.mean, S, Sxx, Y.n)

# d)
X2 = plast$Pressure
S2 = (sum(Y^2) - b0*sum(Y) - b1*sum(X2*Y)) / (X.n-1)
S = sqrt(S2)
sum((X2- X.mean)^2)
Sy = S * sqrt( (1/Y.n) + (X - X.mean)^2/Sxx )
reg_estimate <- function(Y, X) {
    n = length(Y)
    b1 = sum((X - mean(X)) * (Y - mean(Y))) / sum((X - mean(X))^2)
    b0 = (sum(Y - b1) * sum(X)) / n
    coeffs = data.frame(b0 = b0, b1 = b1)
    return(coeffs)
}

CI <- function(b0, b1, x, n, alfa, Sy) {
    Sy = S
    p = b0+b1*x + qt(alfa/2, n-2) * Sy
    m = b0+b1*x + qt(alfa/2, n-2) * Sy
    return data.frame(
        p = p,
        m = m
    )
}

estimate = reg_estimate(Y, X2)
b0 = estimate$b0
b1 = estimate$b1
Y.CI = CI(b0, b1, X2, Y.n, alfa, Sy2)
```

4 a) Vis at minste kvadraters estimator for β er $\hat{\beta} = \dfrac{\sum x_i Y_i}{\sum x_i^2}$

Vi skal minimere funksjonen

$$f(\beta) = \sum (Y_i - \beta x_i)^2 = \sum Y_i^2 - 2Y_i \beta x_i + \beta^2 x_i^2$$

Dermed deriverer jeg og setter lik 0

$$\frac{\partial f \beta}{\partial \beta} = \sum -2Y_i x_i + 2\beta x_i^2 = 0$$

Løser for β

$$0 = \sum -2Y_i x_i + \sum 2\beta x_i^2 \quad / \cdot -1, \; -\sum 2\beta x_i^2, \; \frac{1}{2}$$

$$\beta \sum x_i^2 = \sum \cancel{2} Y_i x_i \quad / \frac{1}{\sum x_i^2}$$

$$\hat{\beta} = \frac{\sum Y_i x_i}{\sum x_i^2}$$

$\square$

4 b)

$$T = \frac{\hat{\beta} - \beta}{S} \sqrt{\sum x^2}$$

Det kan skrives som

$$T = \frac{\hat{\beta} - \beta}{S / \sqrt{\sum x^2}}$$

Som kan skrives som

$$T = \frac{\dfrac{\hat{\beta} - \beta}{\sigma / \sqrt{S_{xx}}}}{\sqrt{\dfrac{(n-1)S^2/\sigma^2}{(n-1)}}}$$

Vi vet at $\dfrac{\hat{\beta} - \beta}{\sigma / \sqrt{S_{xx}}}$ er normalfordelt $\sim N(0,1)$

og at $(n-1)S^2/\sigma^2$ er kji-kvadratfordelt

og at en standard normal stokastisk variabel delt på roten av en kji-kvadrat er en t-fordeling

#

4 d)

Setter opp Hypotesetest

$$H_0: \beta = \beta_0 \quad , \quad H_a: \beta \neq \beta_0$$

CI for $\hat{\beta}$ blir

$$\hat{\beta} \pm t(t_{\alpha/2,\, n-1})\, s_{\beta}$$

$n-1$: siden vi har en parameter

$\alpha/2$ siden vi har tosidig test

$s_{\beta} =$

4 d) Setter opp Hypotesetest

$H_0: \beta = \beta_0$, $H_a: \beta \neq \beta_0$

Testverdi

$$t = \frac{\hat{\beta} - \beta_0}{S}\sqrt{\sum x_i^2}$$

Forkaster H_0 hvis

$t \geq t_{\alpha/2, n-1}$ eller $t \leq -t_{\alpha/2, n-1}$

Der α er signifikansverdi

$\square$

e) Vis at

$$\Sigma(Y_i - \hat{\beta} x_i)^2 = \Sigma Y_i^2 - \beta^2 \Sigma x^2$$

Løser kvadratet

$$\Sigma Y_i^2 - 2 \cancel{Y_i} \hat{\beta} x_i + \hat{\beta}^2 x_i^2$$

$$\Sigma Y_i^2 + \Sigma \hat{\beta}^2 x_i^2$$