

# STK1110 Oblig 2

## Opppg 1

Leser inn dataen:

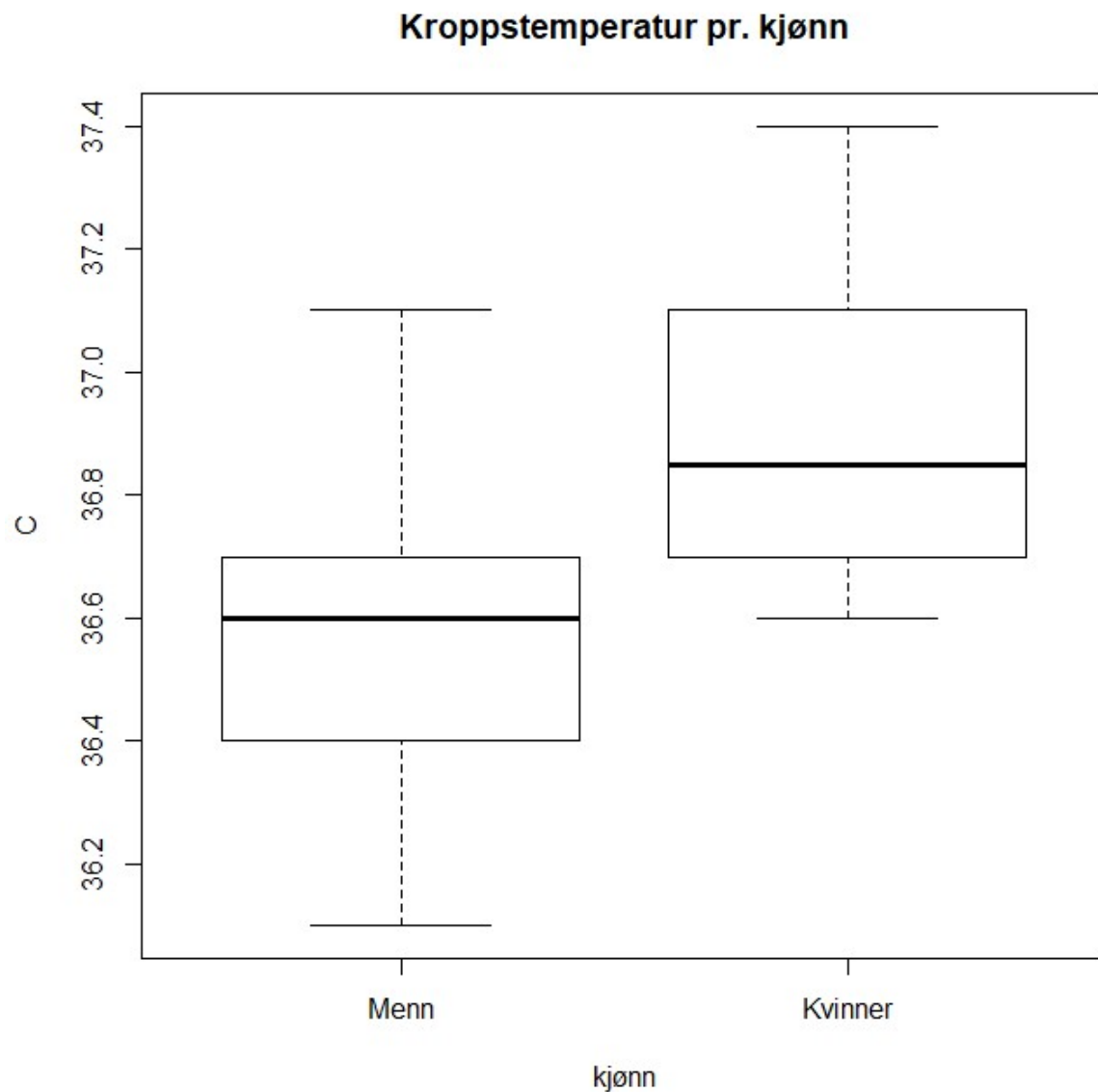
```
path="https://www.uio.no/studier/emner/matnat/math/STK1110/data/temp.txt"
data= read.table(path,header=T)

qqnorm(data$Menn, main="qqnorm Menn")
qqline(data$Menn)
qqnorm(data$Kvinner, main="qqnorm Kvinner")
qqline(data$Kvinner)
```

### a. Lag boksplot og normalfordelingsplott for observasjonene

Lager boxplot:

```
boxplot(data, xlab="kjønn", ylab="C", main="Kroppstemperatur pr. kjønn")
```

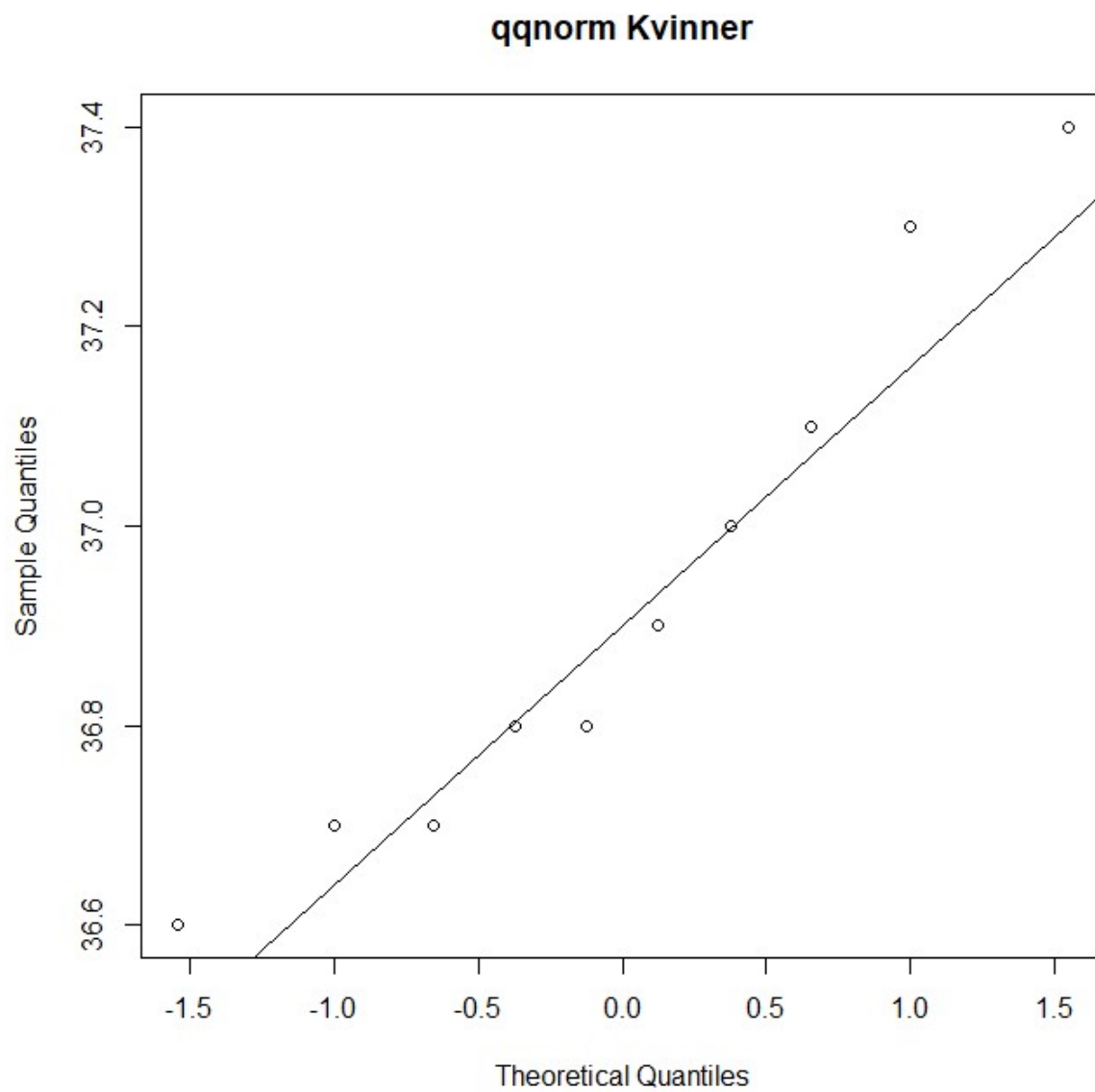


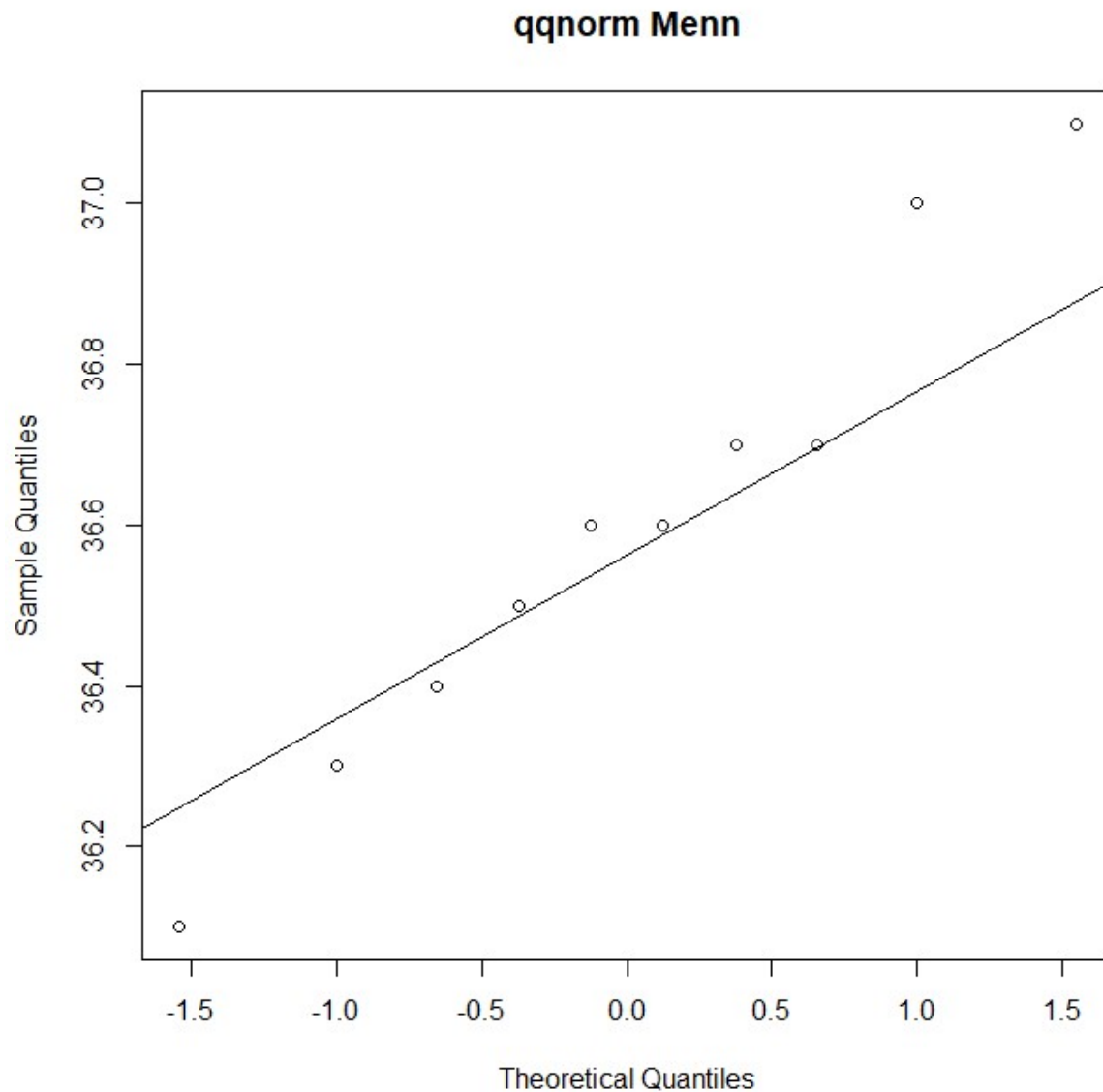
Fra boks plottet ser det ut som menn har lavere gjennomsnittstemperatur. Men med større ekstremverdier. Kvinner ser ut til å være betydelig varmere, men ingen klar signifikant forskjell. Dermed må vi ha en hypotesetest for å bestemme signifikansen.

For å gjøre hypotesetest med t-fordeling må vi anta at dataen er normalfordelt. Dette ser vi med et qqnorm og qqline plots.

```
qqnorm(data$Menn, main="qqnorm Menn")  
qqline(data$Menn)  
qqnorm(data$Kvinner, main="qqnorm Kvinner")  
qqline(data$Kvinner)
```

Som gir:





qqnorm Menn ser ut til å ha noe større ekstremverdier, enn en normalmodell ville frklart. Men siden det er såpass få verdier som er vel ekstreme, kan vi anta en normalfordeling. qqnorm Kvinner ser ut til å følge normalfordelingen tettere. Med noe avvik i de høyeste verdiene. Men vi antar her også at normalfordelingsmodellen holder.

## b. Utled en H-test med alfa 0.05. for at forventet kroppstemperatur er den samme.

For å teste om forskjellen er betydelig antar vi at fordelingene er uavhengige og normalfordelte. Vi skal også ha antagelsen at variansen er lik.

Vi har to hypoteser:

$$H_0: \mu_0 = \mu_1 \text{ mot } H_a: \mu_0 \neq \mu_1$$

Og vi har Testobservator:

$$Z = \frac{\bar{x} - \bar{y}}{s \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

Og vi forkaster  $H_0$  hvis:

$$Z_{1-0.05/2, m+n-2} < Z < Z_{0.05/2, m+n-2}$$

Hvor:

$\mu_0$  er forventet temperatur for men,  $\mu_1$  er forventet temperatur for kvinner,  $s$  er den estimerte variansen,  $n$  er antall målinger for men og  $m$  er antall målinger for 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)
s = ((n-1)*sd.x + (m-1)*sd.y) / (m+n-2)
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)) )

z = (mean.y - mean.x)/ (s* sqrt(1/n + 1/m))
Z.CI = c(qt(0.05/2,m+n-2), qt(1-0.05/2,m+n-2))
```

Som gir resultatet:

$$Z = 2.59$$

$$CI = [-2.100922, 2.100922]$$

Dermed kan vi ikke forkaste  $H_0$  ved gitte antagelser.

Regner ut P verdi:

```
p.z = 1-pt(z, m+n-2)
```

Som gir:

$$P = 0.0091$$

Som sterkt forsvaret kastingen av  $H_0$ . Da vi måtte ha 0.025 for å ikke kunne forkaste  $H_0$ .

### c. F-test på variansene

Hypotesene:

$$H_0: s_{menn}^2 = s_{kvinner}^2 \text{ mot } H_a: s_{menn}^2 \neq s_{kvinner}^2 \text{ med signifikans } 0.05$$

Testobservator:

$$f = \frac{s_{menn}^2}{s_{kvinner}^2}$$

Kvintiler:

$$\left\{ f_{\frac{0.05}{2}, n-1, m-1}, f_{1-\frac{0.05}{2}, n-1, m-1} \right\}$$

R:

```
#c. F-test
f_obs = sd.x^2/sd.y^2

#kvantilene
f.CI = c( qf(alfa/2, n-1,m-1), qf(1-alfa/2, n-1,m-1) )
```

Som gir:

$$f_{obs} = 1.279251, \text{ kvantiler: } 0.2690492 \text{ } 3.7167919$$

Siden  $f_{obs}$  er imellom kvantilene kan vi ikke forkaste  $H_0$ . Dermed kan vi ikke si det er en betydelig forskjell mellom variansene.

## Opppg. 2

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. paret sammenligning.

Vi bruker paret sammenligning siden betingelsene mellom hvert par kan endres. I kontrast til tidligere hvor vi hadde en strand mot en annen. Her blir det som om vi ville se på forskjeller strender på sør- og nord-sidene av øyer. En øy i Sør-Afrika vil ikke ha de samme forutsetningene som en øy i Norge. Da må vi heller se lage en variabel som beskriver forskjellen på hver øy og så estimere en forventningsverdi på forskjellene.

Man kan tenke at de to strendene fra 1. er et par, mens nå har vi mange slike par vi skal undersøke.

I oppg2 så ser vi på tvillinger, men konseptet er det samme.

### b. Test forskjell.

Hypotese:

$H_0: \mu_D = 0$  mot  $H_a: \mu_D \neq 0$  hvor  $\mu_D$  er forskjell i IQ (diff.mean)

Testobservator:

$$t_{obs} = \frac{(\bar{x}_d - \Delta_0)}{s_d / \sqrt{n}}, \text{ hvor } \Delta_0 \text{ er fra } H_0$$

Kvantiler:

$$\left\{ t_{\frac{0.05}{2}, n-1}, t_{1-\frac{0.05}{2}, n-1} \right\}, \text{ med frihetsgrader} = df = n - 1$$

R:

```
mu_D = A.mean - B.mean
t_obs = mu_D / (diff.se_mean / sqrt(diff.N))
P.val = 1-pt(t_obs, diff.N-1)
```

Dette gir resultat

$$t_{obs} = -11.48792, \quad p_{val} = 1, \quad df = n-1 = 30$$

Dette er en urimelig høy p verdi, antar det er en feil i sd i test observatoren.

En slik p-verdi vil si at vi er veldig sikker på at vi ikke kan forkaste  $H_0$  for noen rimelige signifikansnivå.

### c. CI for $\mu_D$

CI for paret t for  $\mu_D$ :

$$\bar{d} \pm t_{\frac{\alpha}{2}, n-1} * \frac{SD}{\sqrt{n}}$$

R:

```
alfa = 0.05
df = diff.N - 1
diff.Ci = diff.mean + c(1,-1) * qt(alfa/2, df) * sqrt( (A.sd/A.n) +
(B.sd/A.n))
```

Dermed har vi et CI for  $\mu_D$ :

$$\{-5.24379, -1.27621\}$$

Det faktum at konfidensintervallet dekker bare negative verdier. Viser til en signifikant forskjell. Det twin A ser ut til å ha betydelig lavere IQ.

## Oppg. 3

Data:



```

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

```

## a. Lineær regresjon: Strength: respons, Temperature: forklaringsvariabel.

Estimerer  $\beta_0$  og  $\beta_1$ :

$$\widehat{\beta}_0 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \text{ og } \widehat{\beta}_1 = \frac{\sum y_i - \widehat{\beta}_1 \sum x_i}{n}$$

R

```

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

```

Som gir:

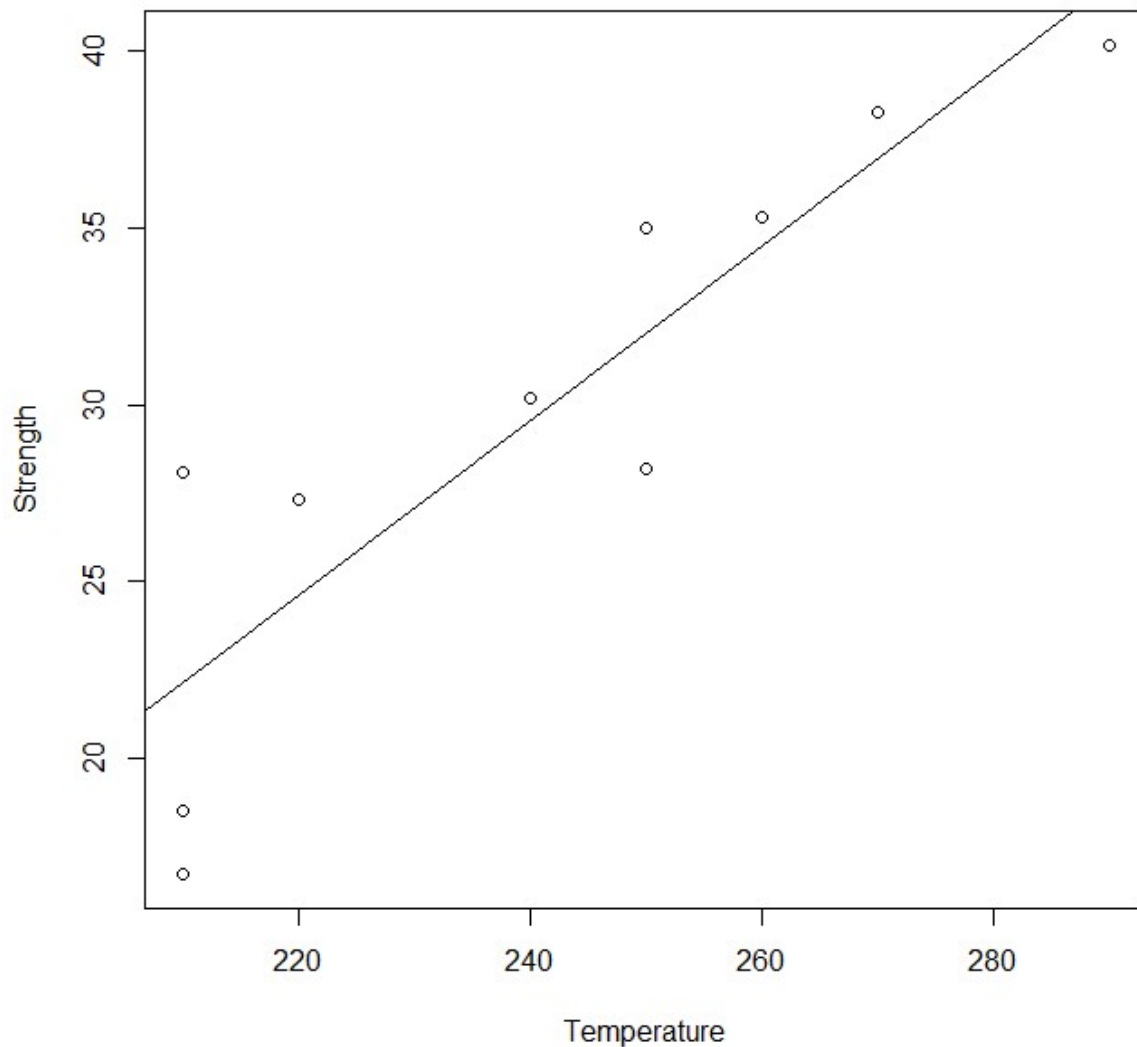
$$\widehat{\beta}_0 = -29.84795, \quad \widehat{\beta}_1 = 0.2474189$$

Plotter regresjonslinjen med dataen:

```

plot(X, Y, xlab="Temperature", ylab="Strength")
abline(b0, b1)

```



Det ser ut til at regresjonslinjen følger dataene. Men rundt Temp 210 ser det ut som det er andre faktorer som forklarer mer. Og det er noe variasjon i dataene, som kan være måleusikkerhet eller andre variabler som er med på å forklare styrken.

## **b. Konfidensintervall for regresjonskoeffisienten.**

0.95 konfidensintervall for  $b_1$  er gitt ved:

$$\widehat{\beta}_1 \pm t_{\frac{\alpha}{2}, n-2} * s_{\widehat{\beta}_1}$$

Hvor  $s_{\hat{\beta}_1}$  er:

$$s_{\hat{\beta}_1} = \frac{s}{\sqrt{S_{xx}}} \text{ og } S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

R:

```
alfa = 0.05
s = sqrt( (sum(Y^2) - b0*sum(Y) - b1*sum(X*Y)) / (n-2) )
Sxx = sum((X-X.mean)^2)
S.b1 = s / sqrt(Sxx)
b1.CI = b1 + c(1, -1) * qt(alfa/2,X.n-2) * S.b1
```

Som gir:

$$b1.CI = \{0.1436597, 0.3511781\}$$

Vi ser at 0 er ikke med i konfidensintervallet som tyder på en endring i Strength når temperaturen endrer seg.

## c. Konfidensintervall for forventningsverdien til Y

0.95 konfidensintervall for Y er gitt ved:

$$\widehat{\beta}_0 + \widehat{\beta}_1 x^* \pm t_{\frac{\alpha}{2}, n-2} * s_{\widehat{\beta}_0 + \widehat{\beta}_1 x^*} = \hat{y} \pm t_{\frac{\alpha}{2}, n-2} * s_{\hat{y}}$$

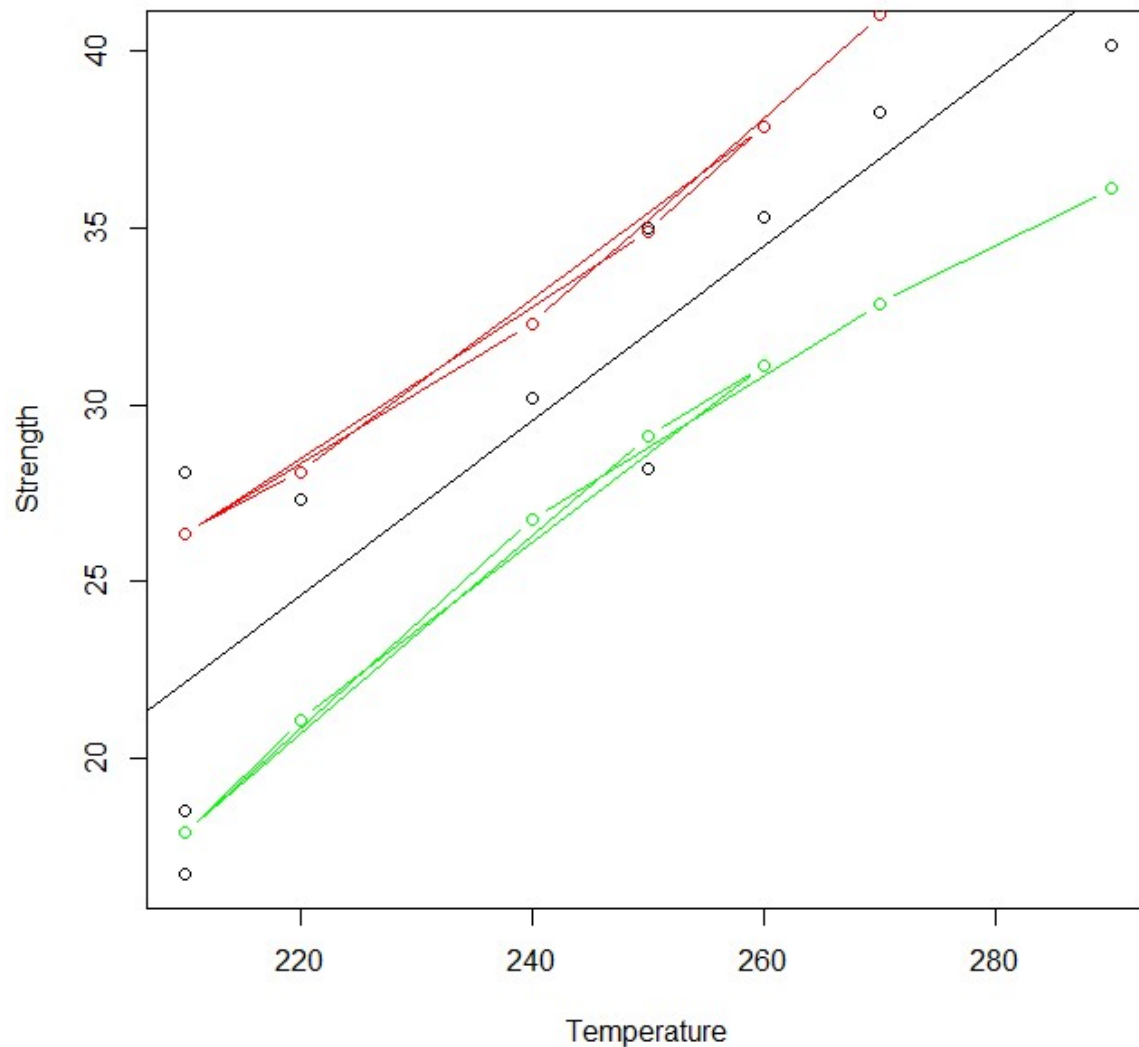
Hvor  $s_{\widehat{\beta}_0 + \widehat{\beta}_1 x^*} = s_{\hat{y}}$  er:

$$s_{\hat{y}} = s \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}, \quad s^2 = \frac{\sum Y_i^2 - \widehat{\beta}_0 \sum Y_i - \widehat{\beta}_1 \sum x_i Y_i}{n-2}, \quad S_{xx} = \sum (x_i - \bar{x})^2$$

R:

```
s = sqrt( (sum(Y^2) - b0*sum(Y) - b1*sum(X*Y)) / (n-2) )
s_xx = sum((X - X.mean)^2)
S_Y = s * sqrt( 1/n + ( X - X.mean )^2 / (s_xx) )
Y.hat = b0 + b1*X
Y.CI.p = Y.hat + qt(0.05/2, X.n-2) * S_Y
Y.CI.m = Y.hat - qt(0.05/2, X.n-2) * S_Y

lines(X, Y.CI.m, type="b")
lines(X,Y.CI.p, type="b")
```



Ut ifra plottet så ser konfidensintervallet ut til å stemme. Fleste punkter er inne i intervallet, intervallet følger regresjonslinjen, og intervallet er ikke mye videre enn punktene.

#### d. Gjennta a og b for Pressure. Hvilken ville jeg valgt.

Gjentar beregningene i R:

```

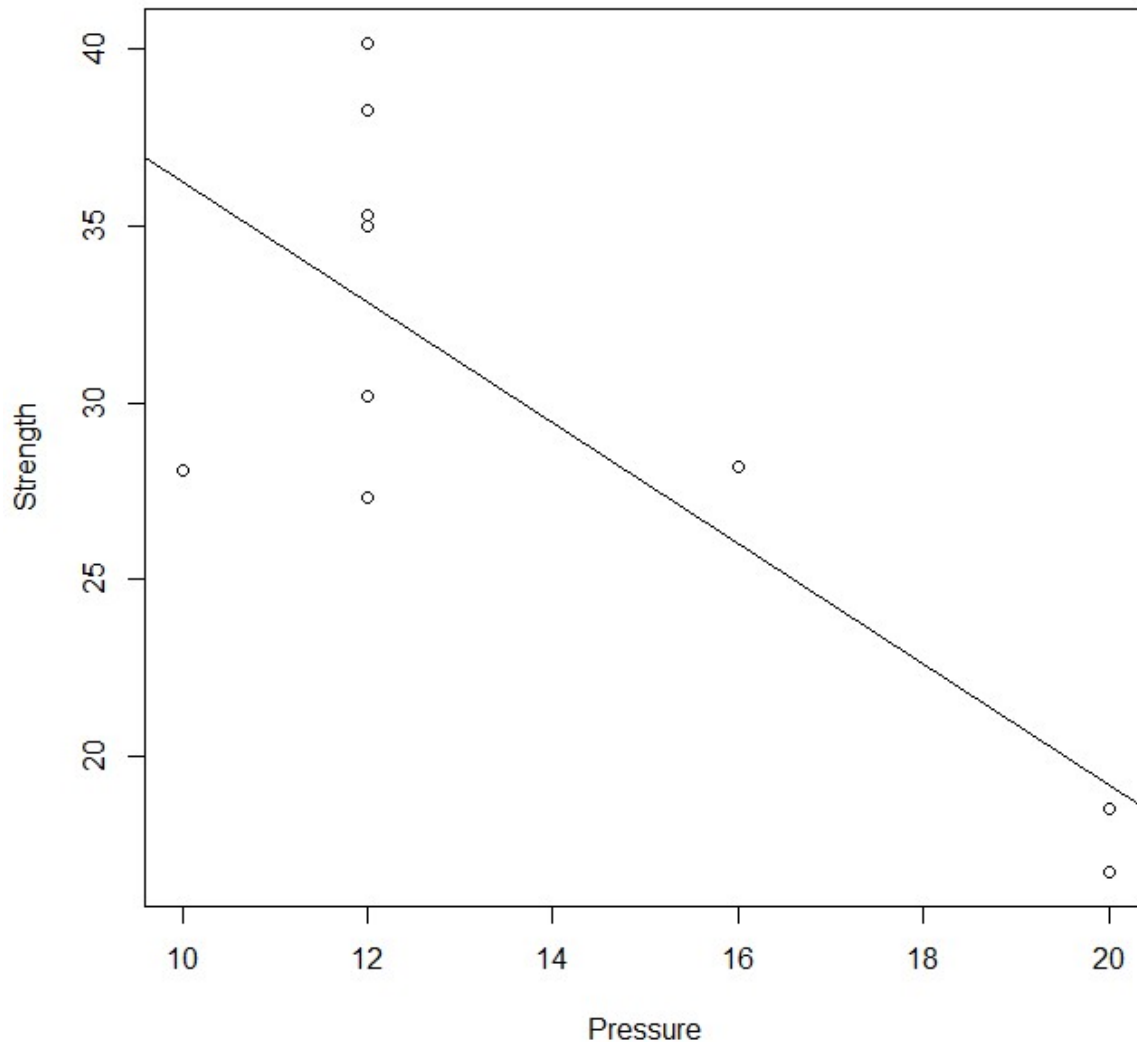
X = plast$Pressure
X.mean = mean(X)
X.n = length(X)

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="Pressure", ylab="Strength")
abline(b0, b1)

alfa = 0.05
s = sqrt( (sum(Y^2) - b0*sum(Y) - b1*sum(X*Y)) / (n-2) )
Sxx = sum((X-X.mean)^2)
S.b1 = s / sqrt(Sxx)
b1.CI = b1 + c(1, -1) * qt(alfa/2,X.n-2) * S.b1

```

Som gir plottet:



Punktene ser fortsatt ut til å følge linjen, men Pressure verdiene er lite spredd i X verdier. Og rundt pressure=12 så er det mange forskjellige styrkeverdier. Noe som tilsier at det nok er en annen faktor som spiller en stor rolle.

Konfidensintervallet blir:

$$b1.CI = \{-2.8109908 \ -0.6118466\}$$

Jeg ville valgt Temperaturen som forklaringsvariabel, hvis jeg bare kunne velge en av dem. Vi kan også se på forklaringsgraden  $r$ .

R:

```
SS_R = sum((Y.hat-Y.mean)^2)
SS_T = sum((Y - Y.mean)^2)
r = SS_R/SS_T
```

r for Temperature er 0.616896 og r for Pressure er 0.7907761. Som forteller at r for pressure forklarer mer av endringen. Dog jeg ville hvert påpasselig siden nesten halve veridene til pressure er 12.