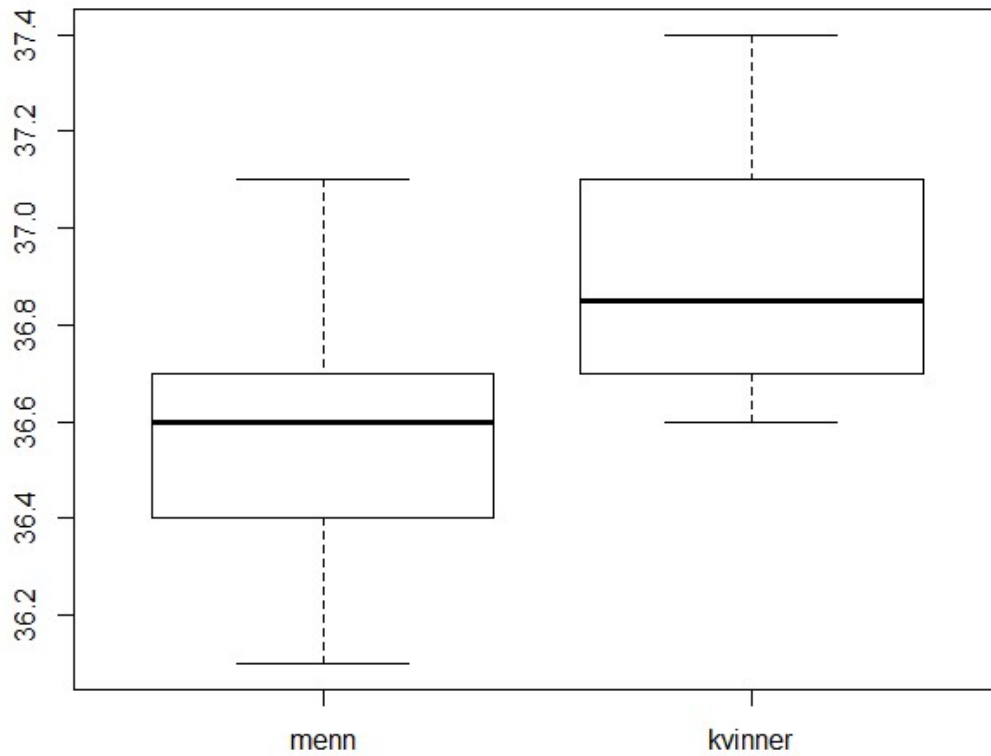


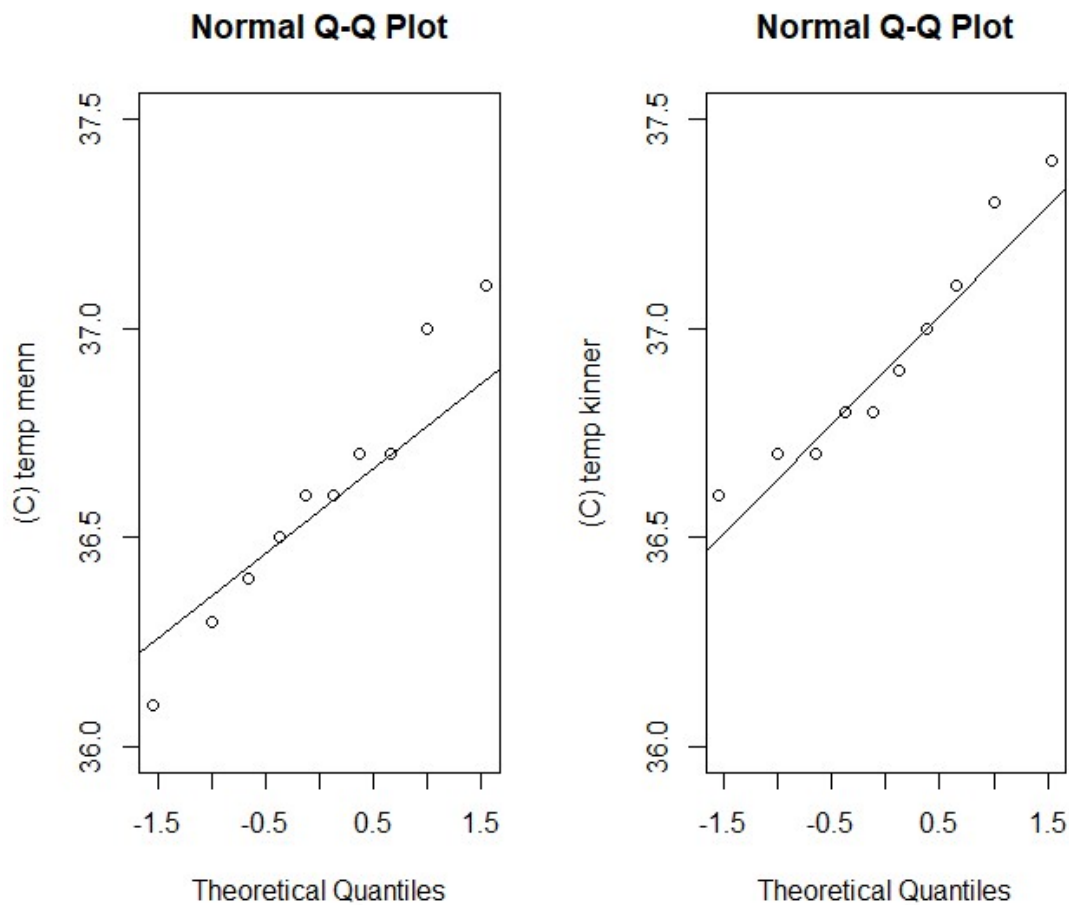
Oppg1

- a) Lag boksplott som viser fordelingen av observasjonene. Kommenter hva du finner.



Ser at kvinner sin snitt temperatur er høyere enn menns. Og at kvinner har et høymaksverdi, mens min verdien er nærmere snittet. Dermed er kvinners fordeling skjev. Menn fordelingen er mindre skjev med maks, og min verdier ganske jevnt rundt snittet. Dog snittet er litt positivt skjevt.

- b) Lag normalfordelingsplott for de to observasjonssettene. Kommenter hva du ser.



Ser at både kvinner og menn fordelingene er ganske normalt fordelte med få punkter lengre unna enn resten.

- c) Anta at variansen er den samme for de to utvalgene, og test med niv^o 5% om det er noen forskjell i forventet kroppstemperatur. Beregn ogs^a P-verdien, og lag et 95% konfidensintervall for forventet forskjell.

Vi sette $H_0: \mu_{\text{menn}} = \mu_{\text{kvinner}}$, $H_a: \mu_{\text{menn}} \neq \mu_{\text{kvinner}}$

For å gjøre testen trenger jeg t_{obs} , og t_{lim} :

b1C1

$$t_{obs} = \frac{|\bar{x}_1 - \bar{x}_2| - \overbrace{(\mu_1 - \mu_2)}^{=0, \text{ p.g.a } H_0}}{\sqrt{sp\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Setter vi inn tall får vi (Regnet i "R")
2,59

$$t_{lin} = t_{(1-\alpha/2, n-1)}$$

gir oss

$$t_{lin} = 2,26$$

Dvs $t_{lin} < t_{obs}$, som vil si at det er en signifikant forskjell. P.G.A vi forkaster H_0

For å finne P bruker jeg r koden $P = 2 * pt(t_{obs}, nx+ny-2)$.

Dvs. $P = 1.981519$

Vi kan finne CI ved
setter for ryddighetsdyld

$$t_{\alpha/2, n_x + n_y - 2} = t$$

$$1 - \alpha = P \left(-t < \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{S_p \left(\frac{1}{n_x} + \frac{1}{n_y} \right)}} < t \right)$$

Lösen p.h.m $\mu_x - \mu_y$

setter $\mu_x - \mu_y = \mu_D$,

$$1 - \alpha = P \left(-t < \frac{\bar{X} - \bar{Y} - \mu_D}{\left(S_p \left(\frac{1}{n_x} + \frac{1}{n_y} \right) \right)^{1/2}} < t \right)$$

$$= P \left(-t \sqrt{S_p \left(\frac{1}{n_x} + \frac{1}{n_y} \right)} < \bar{X} - \bar{Y} - \mu_D < t \sqrt{S_p \left(\frac{1}{n_x} + \frac{1}{n_y} \right)} \right)$$

$$= P \left(\bar{X} - \bar{Y} - t \sqrt{S_p \left(\frac{1}{n_x} + \frac{1}{n_y} \right)} < \mu_D < \bar{X} - \bar{Y} + t \sqrt{S_p \left(\frac{1}{n_x} + \frac{1}{n_y} \right)} \right)$$

Dermed har vi et CI for $\mu_D = \mu_x - \mu_y$

$$\bar{X} - \bar{Y} \pm t_{\alpha/2, n_x + n_y - 2} \sqrt{S_p \left(\frac{1}{n_x} + \frac{1}{n_y} \right)}$$

Setter vi inn verdiene får vi ("r")

$$[0.062, 0.598]$$

- d) Gjennomfør testen og beregn P-verdien også i det tilfellet der man ikke antar felles varians. Diskuter og forklar resultatene.

Hvis det ikke er samme varians bruker vi

$$t_{obs} = \frac{|\bar{x}_1 - \bar{x}_2| - |\mu_1 - \mu_2|}{\sqrt{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}}$$

0 Det gir oss en t_{obs} verdi på 3.36483, som er en noe større verdi enn det vi hadde når vi antok lik varians. De er fortsatt ganske lik noe vi ville antatt da variansen for de to fordelingene er rimelig like.

Vi estimerer frihetsgraden med:

$$n - 1 = \frac{(s_x^2/n_x + s_y^2/n_y)^2}{\frac{(s_x^2/n_x)^2}{n_x - 1} + \frac{(s_y^2/n_y)^2}{n_y - 1}}$$

Så runder vi $n-1$ til nærmeste hele tall. Da har vi estimerte frihetsgrader.

Estimatet er $v = 18$ som stemmer overens med den tidligere utregningen da vi regnet med at Variansene var det samme.

Oppg2

b)) Kall forventet forskjell mellom Twin A og Twin B for μ_D . Sett opp nullhypotese og alternativ hypotese for å besvare spørsmålet om forskjell i IQ. Finn en egnet testobservator, og beregn dennes numeriske verdi. Beregn så tilhørende P-verdi. Spesifiser antall frihetsgrader i fordelingen du bruker. Formuler din konklusjon på testen

	N	Mean	StDev	SE Mean
Twin A	31	93.32	15.41	2.77
Twin B	31	96.58	13.84	2.49
Difference	31	-3.26	8.81	1.58

$$t_{obs} = \frac{\mu_D - A_D}{\sigma_D / \sqrt{n}}$$

Setter inn verdiene

$$t_{obs} = \frac{-3,26 - 0}{8,81 / \sqrt{31}} = -2,06$$

Dermed trenger vi P til å sammenligne med $\alpha = 0,05$

$$P = 2 P(t_{n-1} \geq t_{obs})$$

Setter inn verdiene

$$P = 2 P(t_{30} \geq -2,06)$$

$$\stackrel{\text{"r"}}{=} 0,046$$

Dermed har vi $P < \alpha$ så vi forkaster H_0

Vi har $n-1$ frihetsgrader.

c) Finn et 95% konfidensintervall for μ_D . Hva betyr det at dette intervallet dekker kun negative verdier? Forklar kort om sammenhengen mellom tosidig testing og konfidensintervaller.

For å finne 95% CI for μ_D bruker jeg

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

Setter inn tall og får

$$-3,26 \pm -2,04 \cdot \frac{8,81}{\sqrt{5,57}}$$

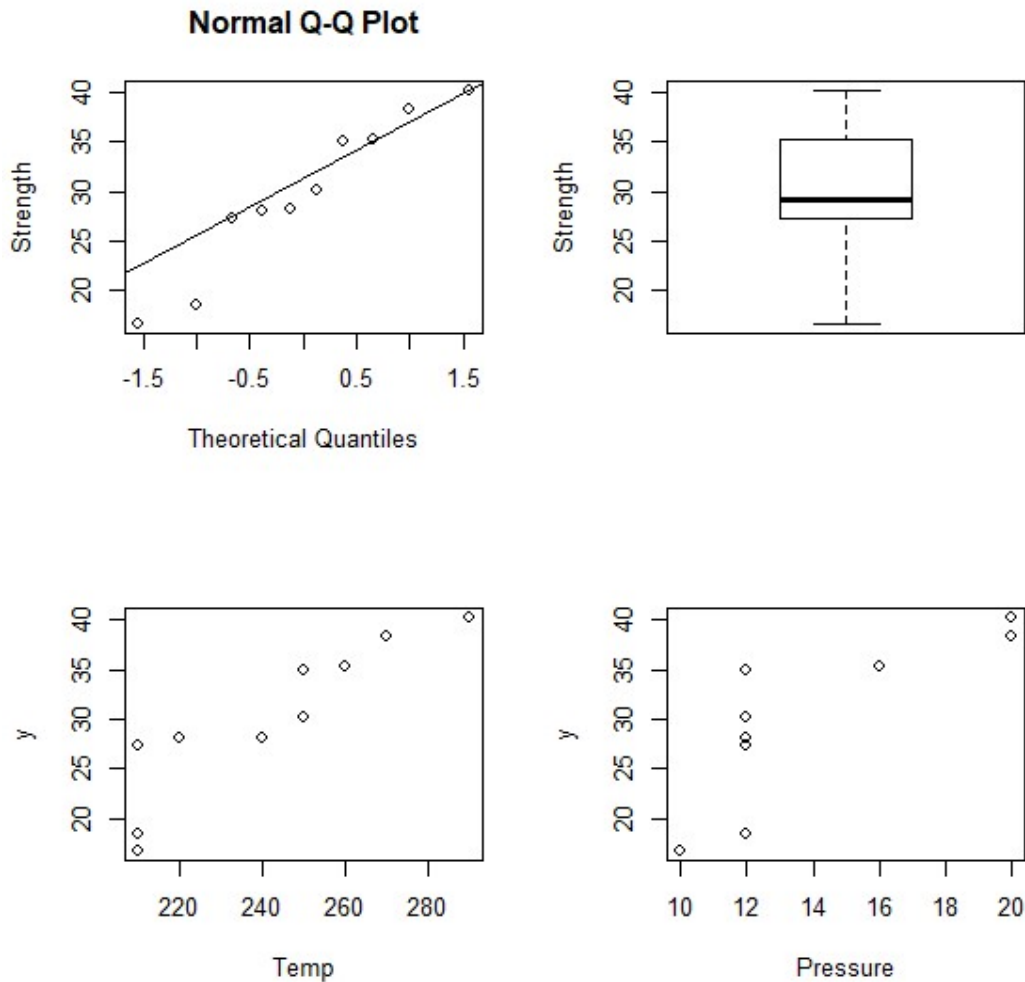
"r"

$$= [-6,49, -0,028]$$

Vår forventningsverdi er negativ da får vi et konfidensintervall som er sentrert rundt forventningsverdien.

Oppg3

- a) Utfør ulike typer eksplorativ data analyse (e.g. histogram, qqplot (probability plot), spredningsplott, dvs. scatterplot) p`a datasettet og beskriv hva du ser.

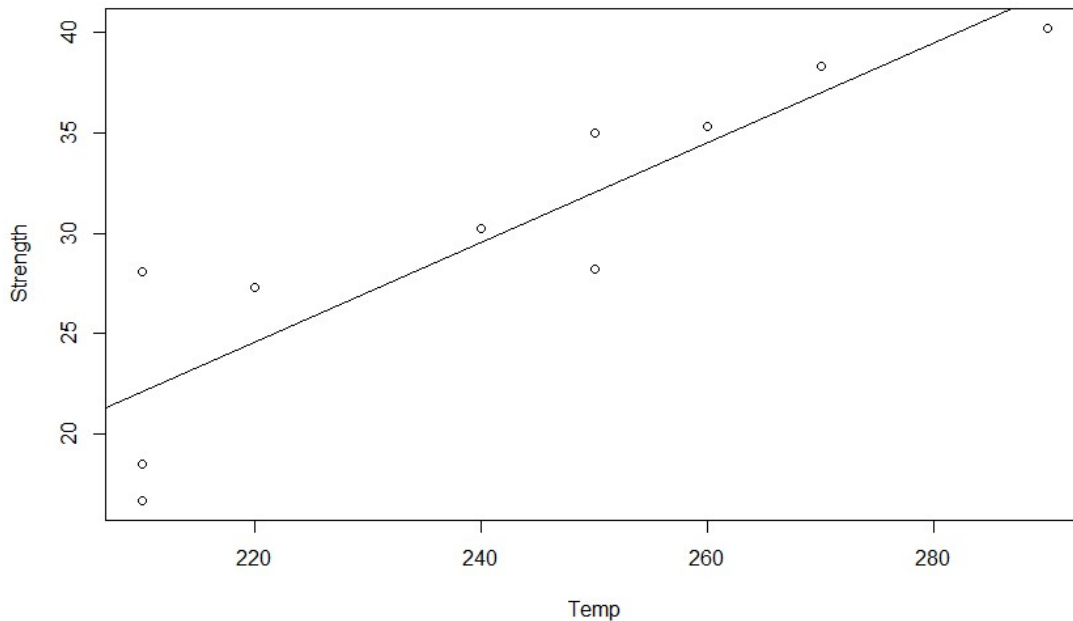


Fra Normal qq Plot ser vi at Strength er tilnærmet normalfordelt. Siden det ser ut som punktene følger normalen.

Fra boxplottet ser vi at Strength er positivt skjevt fordelt. Og vi har noen veldig lave verdier sammenlignet med forventingsverdien.

Fra de to siste plottene, er Strength(y) plottet mot Temp(x1) og Pressure(x2). Nå kan vi se hvilke som ser ut til å være mest linjer. Vi ser at y(temp) ser mer linjer ut, enn y(Pressure), som har et diskontinuerlig hop rundt $x_2 = 12$.

- b) Vi vil i første omgang konsentrere oss om forklaringsvariablen Temperature. Utfør en enkel lineær regresjon i dette tilfellet. Plott data sammen med den tilpassende regresjonslinjen og komment'er resultatene.



Ved å bruke minste kvadraters problem kan vi finne regresjonslinjen.

Vi skal finne $\hat{\beta}_0 + \hat{\beta}_1 x$

Vi finner β_1 ved

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} \quad (12.2)$$

$$\hat{\beta}_0 = \frac{\sum y_i - \hat{\beta}_1 \sum x_i}{n} = \bar{y} - \hat{\beta}_1 \bar{x} \quad (12.3)$$

Ved r kan vi bruke `lm(x1, y, xlab='Temp', ylab='Strength')` for å finne β_{1_hat} og β_{0_hat}

(kalt $\hat{\beta}_1$ og $\hat{\alpha}_1$ i koden). Får da verdiene $\hat{\beta}_1 = 0.2474189$, $\hat{\beta}_0 = -29.8479549$. Fra plottet ser regresjonen rimelig ut, og det virker som regresjonen er en ok estimator for $y(x_1)$.

C) Lag konfidensintervall for regresjonskoeffisienten for Temperature. Gir intervallet indikasjon på om temperatur er en viktig forklaringsvariabel?

CI for regresjonsintervallet finner vi ved:

$$CI = \hat{\beta}_1 \pm t_{\alpha/2, n-2} \cdot s_b$$

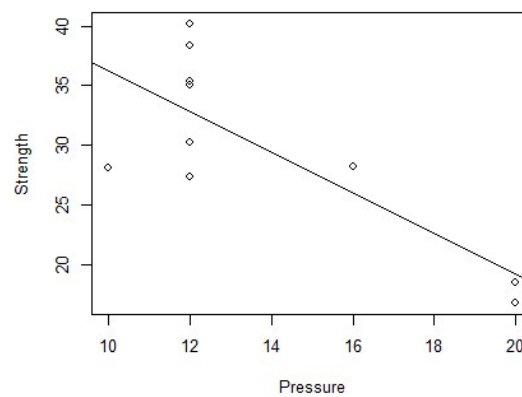
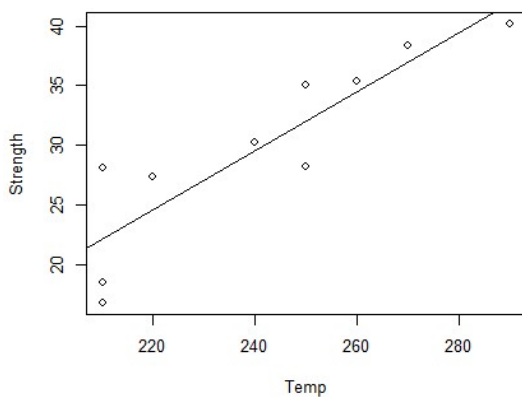
Setter vi i inn tall får vi

$$= 0,247 \pm 2,306 \cdot 0,098$$

$$= [0,020, 0,4743]$$

Vi ser at min tallet 0.02 ikke er veldig nærme -1 og maksimum er 0.47 som da er nærmere 1, dette avkrefter ikke at vi kan bruke dette som en modell, men det er ikke veldig bra.

f)



Fra bildet ser vi at regresjonslinjen til Strength(Pressure) går nedover, og punktene til linjen ser ikke ut til å følge linjen i det hele tatt. Dermed konkluderer jeg med at Strength(Temp) er et mye bedre estimert avhengighet.

Kode

Oppg1

```
y = menn <- c(36.1, 36.3, 36.4, 36.6, 36.6, 36.7, 36.7, 37.0, 36.5, 37.1)
```

```
x = kvinner <- c(36.6, 36.7, 36.8, 36.8, 36.7, 37.0, 37.1, 37.3, 36.9, 37.4)
```

#a) Lag boksplott som viser fordelingen av observasjonene. Kommenter hva du finner.

```
boxplot(menn, kvinner, names=c('menn', 'kvinner'))
```

```
,
```

Vi ser fra plottet at menn har litt høyere kroppstemperatur enn menn.

Vi ser også at mennene sin fordeling er sjevt mot lavere verdier,

menns kvinners fordeling er har ingen skjevhet

```
,
```

#b) Lag normalfordelingsplott for de to observasjonssettene. Kommenter hva du ser.

```
limy = c(36,37.5)
```

```
par(mfrow=c(1,2))
```

```
qqnorm(menn, ylab='(C) temp menn', ylim=limy)
```

```
qqline(menn)
```

```
qqnorm(kvinner, ylab='(C) temp kinner', ylim=limy)
```

```
qqline(kvinner)
```

#c)

```
alfa = 0.05
```

```
nx = length(kvinner)
```

```
ny = length(menn)
```

```
Sx = sd(x)
```

```
Sy = sd(y)
```

```
sp_square = ((nx-1)*Sx**2 + (ny-1)*Sy**2)/(nx+ny-2)
```

```
t_obs = abs((mean(kvinner)-mean(menn))/sqrt(sp_square*(1/nx + 1/ny)))
```

```
t_lim = qt(1-alfa/2, nx+ny-2)
```

```
P = 2 * pt(t_obs, nx+ny-2)
```

```
t_CI = mean(x) - mean(y) + c(1,-1)*qt(alfa/2, nx+ny-2)*sqrt(sp_square*(1/nx + 1/ny))
```

#d)

```
t_obs_NonEqualVar = abs(mean(kvinner)-mean(menn))/sqrt(Sx**2**2/nx + Sy**2/ny)
v = round((Sx**2/nx + Sy**2/ny)**2/((Sx**2/nx)**2/(nx-1) + (Sy**2/ny)**2/(ny-1)))
P = 2* pt(t_obs_NonEqualVar, v)

#e) Utled og gjennomfør en F-test for å sjekke om det er noen grunn til å påstå at
#variansene er forskjellige. Sjekk mot var.test() i R.

f = Sx**2/Sy**2
Clf = qf(c(alfa/2, 1-alfa/2),nx-1,ny-1)
```

Opppg2

```
AN = 31
BN = 31
AMean = 93.32
BMean = 96.58
AStDev = 15.41
BStDev = 13.84
ASE_Mean = 2.77
BSE_Mean = 2.49

DN = 31
DMean = -3.26
DStDev = 8.81
DSE_Mean = 1.58

alfa = 0.05
```

```
t_obs = DMean/(DStDev/sqrt(DN ))
P = 2* pt(t_obs, DN-1)

#c)
t_CI = DMean + c(1,-1)*qt(alfa/2, DN-1)*DStDev/sqrt(DN)
```

oppg3

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

```
y = Strength = plast[, 'Strength']
x1 = Temp = plast[, 'Temperature']
x2 = Pressure = plast[, 'Pressure']
```

```
par(mfrow=c(2,2))
qqnorm(Strength, ylab='Strength')
qqline(Strength)
#Plottet ser tilnærmet normalfordelt ut.
```

```
boxplot(Strength, xlab='Strength')
#Boxplotet viser en poseti sjevhet
```

```
qqplot(x1,y, xlab='Temp')
#qqline(y)
qqplot(x2,y, xlab='Pressure')
```

#Det ser ut som y er mer jevnt avhengig av Temp, mens det er en stor endring når Pressure er 12 som kan vise til at Styrken til plasten øker betydelig når trykket er 12.

#Dette kan ytterligere vises med en korrelasjonskoeffisient.

```
#b) Regresjon, y(x1)
```

```
x_bar = c(mean(x1), mean(x2))
y_bar = mean(y)
```

```
#b1_hat = sum((x1-x_bar[1])*(y-mean(y)))/sum((x1-x_bar[1])**2) s
#a1_hat = y_bar-b1_hat*x_bar[1]
```

```
#b2 = sum((x2-x_bar[2])*(y-mean(y)))/ sum((x-x_bar[1])**2)
```

```
l.lm = lm(y~x1)
plot(x1, y, xlab='Temp', ylab='Strength')
abline(l.lm)
# ser tilsynelatende ut til å ha en korrelasjon
```

```
#c)
betaX1hat = coefficients(l.lm)[2]
```

```
alfax1hat = coefficients(l.lm)[1]

alfa = 0.05

n = length(y)

#s= sqrt((sum(y**2-alfax1hat)*sum(y-betaX1hat)*sum(x1*y))/(n-2))

s = sqrt(sum((y-mean(y))**2) /(n-2))

Sxx = sum((x1-x_bar[1])**2)

sb = s/sqrt(Sxx)

b1CI = betaX1hat+ c(1,-1)*qt(alfa/2, n-2) * sb

#intervallet virker lite og viser at variasjoen er liten i betaX1hat

#d)

CI_Strength = y_bar + c(1,-1)*qt(alfa/2, n-2)*sd(y)/sqrt(n)

x_star = c(210, 240, 270)

y_hat = alfax1hat+betaX1hat*x_star

CI_pred_pluss = y_hat +qt(alfa/2,n-2)*sqrt(s**2+sd(y)**2)

CI_pred_minus = y_hat -qt(alfa/2,n-2)*sqrt(s**2+sd(y)**2)

CI_pred_x_sar = cbind(CI_pred_pluss, CI_pred_minus)

#prediksjons intervallet er betydelig større en CI_Strength siden prediksjonen basserer seg på vår estimerte regresjon

#og CI_Strength er bassert på våre observasjoner.

#s.659

#f)

l.lm2 = lm(y~x2)

plot(x2, y, xlab='Pressure',ylab='Strength')

abline(l.lm2)

betaX2hat = coefficients(l.lm2)[2]

alfax2hat = coefficients(l.lm2)[1]

Sxx2 = sum((x2-x_bar[2])**2)

sb2 = s/sqrt(Sxx2)

b2CI = betaX2hat+ c(1,-1)*qt(alfa/2, n-2) * sb2
```