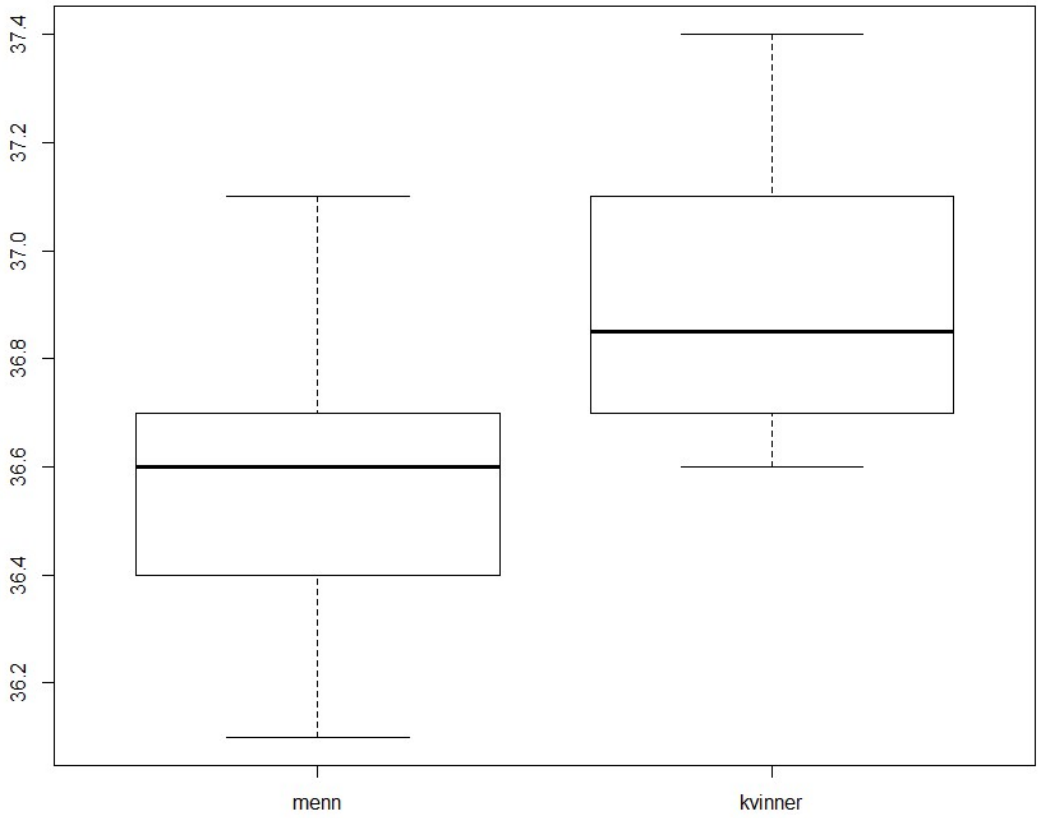
Oppg1

Inndata:

```
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.

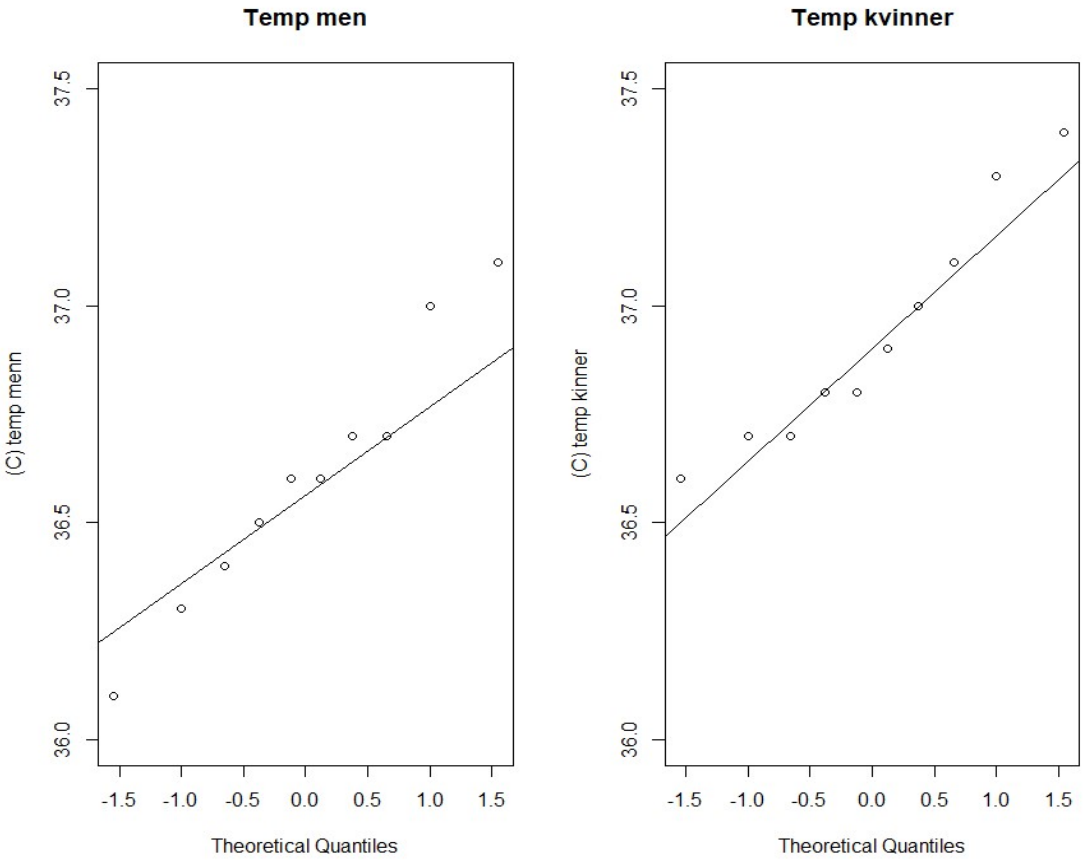
```
boxplot(menn, kvinner, names=c('menn', 'kvinner'))
```



Snittet til menn (36.6) er mindre enn snittet til kvinner (36.93). Maksimumsverdien til kvinner er også høyere enn menn sin maks. Kvinner sin minimum er den samme som men sitt minimum. Og menn sin minimum er en del mindre enn kvinners minimum. Ut i fra dette ser det ut til å være en betydelig forskjell. Men det må testes videre for å angi et sikkerhetsnivå (Konfidens nivå).

B: Lag normalfordelingsplott for de to observasjonssettene. Kommenter hva du ser.

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



Ser her at Temp menn kan ha en «Short Tail», altså en s-kurve. Med stor varianse på endene, i hver sin retning. Dette kan også bare være tilfeldig og med mer data så kan Temp men være mer normalfordelt, spesielt da det er bare 3 punkter som er langt unna normallinjen.

For Temp kvinner ser statistikken mer normalt fordelt ut, dog der ser ut til å noe «Right Skew» da det datapunktene i endene er litt over normallinjen.

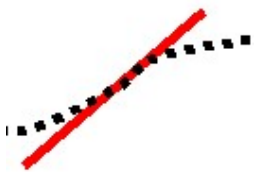


Figure 1: Short Tails

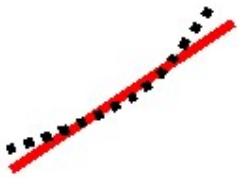


Figure 2: Right Skew

Figure 1, 2: http://www.skymark.com/resources/tools/normal_test_plot.asp, 2018.11.23

C:

Oppg2

Inndata:

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

A: Begrunn hvorfor en paret sammenligning er best egnet i denne situasjonen. Beskriv kort hvilke antakelser vi må legge til grunn for videre analyse.

Siden vi har 2 og 2 avhengige test par. Kan vi ikke bruke vanlig sammenligning siden vi der antar uavhengighet mellom de to gruppene vi skal sjekke om der er noe forskjell.

F.eks. Hvis vi skal måle en medisins effekt på blodtrykk. Så blir det galt å sammenligne gjennomsnittet av blodtrykket før med blodtrykket etter. Da burde vi heller ha en paret test der vi istedenfor ser på reduksjonen, som utgangspunkt for en test.

Antagelser:

Betingelsene endrer seg fra test til test. (fra tvillingpar til tvillingpar)

Uavhengighet mellom parene.

Antar at variasjonen av differansen kan beskrives som en normalfordeling.

Oppg3

3c

```
alfa = 0.05
n = length(y)
Sxx1 = sd(x1)
#Sxx1 = sum(x1^2)- (sum(x1)^2)/n #vet ikke hvorfor dette blir ulikt linjen over.
SSE = sum((y - (beta0x1hat + beta1X1hat*x1))^2)
s_square = SSE/(n-2)
s = sqrt(s_square)
S_beta1x1hat = s/sqrt(Sxx1)
CI_beta1x1 = beta1X1hat + c(1,-1)*qt(alfa/2,n-2) * S_beta1x1hat
```

```
> beta1X1hat
       x1
0.2474189
> CI_beta1x1
[1] 0.1436597 0.3511781
```

Konfidensintervallet gir en indikasjon på at Temperature er en viktig forklaringsvariabel. Da intervallet ikke er velig stor. Den øvre grensen ligger på 1.72 og den nedre grensen ligger på 0,70 av den estimerte verdien beta1x1hat.

3d: Lag s˚a konfidensintervall for forventet verdi av Strength og dessuten prediksjonsintervall for nye verdier av Strength ved de følgende 3 verdiene av Temperature: 210, 240 og 270. Sammenlign intervallene og diskuter forskjellene mellom konfidens og prediksjonsintervall.

Kommentar:

Jeg tolker oppgaven slik at jeg skal lage konfidens intervall og prediksjonsintervall for de nye verdiene. Beskrevet i boken sec. 12.4

```
x1_star = c(210, 240, 270)
y_hat = beta0x1hat+beta1X1hat*x1_star
S_y_hat = s*sqrt(1/n + (x1_star - mean(x1))^2/Sxx1)

CI_y_hat_pluss = y_hat + qt(alfa/2, n-2)*S_y_hat
CI_y_hat_minus = y_hat - qt(alfa/2, n-2)*S_y_hat
CI_y_hat = cbind(CI_y_hat_pluss, CI_y_hat_minus)

PI_y_hat_pluss = y_hat + qt(alfa/2, n-2)*sqrt(s_square +S_y_hat^2)
PI_y_hat_minus = y_hat - qt(alfa/2, n-2)*sqrt(s_square +S_y_hat^2)
PI_y_hat = cbind(PI_y_hat_pluss, PI_y_hat_minus)
```

```
> CI_y_hat
     CI_y_hat_pluss CI_y_hat_minus
[1,]       17.86983       26.35020
[2,]       26.76783       32.29733
[3,]       32.87014       41.04015
> PI_y_hat
     PI_y_hat_pluss PI_y_hat_minus
[1,]       12.39868       31.82135
[2,]       20.36881       38.69635
[3,]       27.31056       46.59974
```

Et konfidensintervall av en forventningsverdi forteller hvor vi forventer at 1-alfa (95%) av verdiene vi har samlet ligger. Mens et prediksjonsintervall forteller hvor du kan forvente det neste datapunktet. Så hvis man samler inn data, så finner man prediksjonsintervallet gjerne for 95%, så samler inn ny data. Da vil du få 95% av de nye verdiene innenfor prediksjonsintervallet.. hvis vi antar at fordelingen på verdiene passer vår valgte fordeling. Prediksjonsintervallet må ta til høye for usikkerheten i vår forventningsverdi, og spredningen på vår data. Det medfører at prediksjonsintervallet alltid er videre enn konfidensintervallet. Noe som gjenspeiles i vår data over.

3e:

```
x1_star = c(210, 240, 270)
y_hat = beta0x1hat+beta1X1hat*x1_star
S_y_hat = s*sqrt(1/n + (x1_star - mean(x1))^2/Sxx1)

CI_y_hat_pluss = y_hat + qt(alfa/2, n-2)*S_y_hat
CI_y_hat_minus = y_hat - qt(alfa/2, n-2)*S_y_hat
CI_y_hat = cbind(CI_y_hat_pluss, CI_y_hat_minus)

PI_y_hat_pluss = y_hat + qt(alfa/2, n-2)*sqrt(s_square +S_y_hat^2)
PI_y_hat_minus = y_hat - qt(alfa/2, n-2)*sqrt(s_square +S_y_hat^2)
PI_y_hat = cbind(PI_y_hat_pluss, PI_y_hat_minus)
```

```
> PI_pred
[1] 11.25206 48.30794
```

Siden det prediksjonsintervallet vi lagde i oppgave 3d er avhengig av Temp så blir prediksjonen smalere. Men om vi skal predikere uten en avhengighet blir det tilsvarende med at vi ikke hat kontroll over temperaturen og da vil variasjonen bli større. Dermed får vi et stort intervall, som her omslutter hele intervallet.

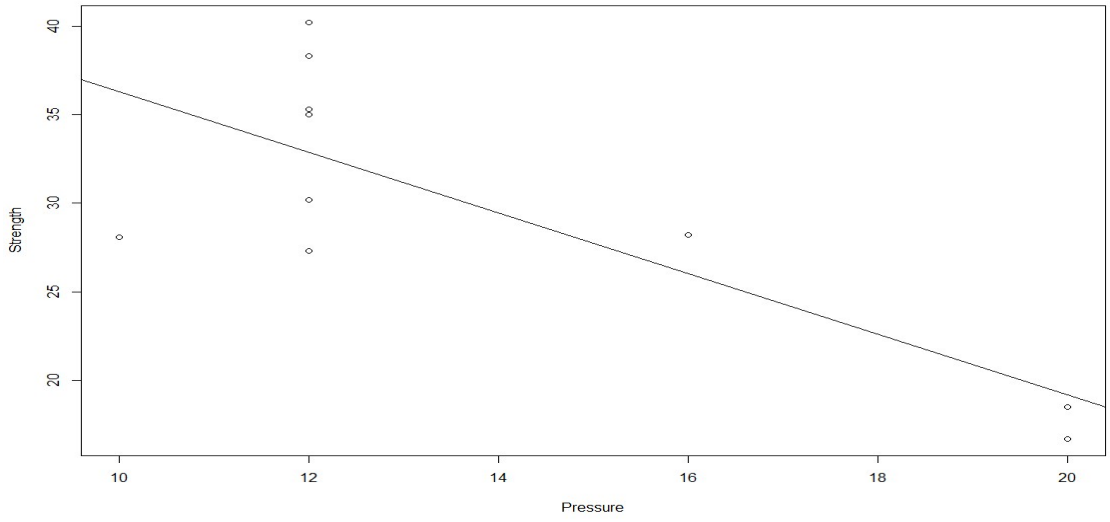
3f:

```
#f) b)

l.lm = lm(y~x2)
plot(x2, y, xlab='Temp',ylab='Strength')
abline(l.lm)

beta1X2hat = coefficients(l.lm)[2]
beta0x2hat = coefficients(l.lm)[1]

#f) c)
alfa = 0.05
n = length(y)
#Sxx2 = sd(x2)
Sxx2 = sum(x2^2)- (sum(x2)^2)/n #vet ikke hvorfor dette blir ulikt linjen over.
SSE2 = sum((y - (beta0x2hat + beta1X2hat*x2))^2)
s_square2 = SSE2/(n-2)
s2 = sqrt(s_square2)
S_beta1x2hat = s2/sqrt(Sxx2)
CI_beta1x2 = beta1X2hat + c(1,-1)*qt(alfa/2,n-2) * S_beta1x2hat
```



```
> c(beta1X2hat, beta0x2hat)
         x2 (Intercept)
  -1.711419   53.397578
> CI_beta1x2
[1] -2.8109908 -0.6118466
> abs(-2.8109908 - (-0.6118466))
[1] 2.199144

> c(beta1X1hat, beta0x1hat)
         x1 (Intercept)
  0.2474189 -29.8479549
> CI_beta1x1
[1] 0.1436597 0.3511781
> abs(0.1436597 - 0.3511781)
[1] 0.2075184
```

Vi ser at vidden på CI_beta1x2 er en del større enn vidden til CI_beta1x1. Og vi ser i plottet at det kan være en korrelasjon men ikke en lineær en. Kan spekulere i om det er en logistisk vekst kurve, da vi ser en stor endring før den muligens flater ut, men det trenger vi flere datapunkter for å avgjøre siden vi har noen anomalier (hvis vi antar logistisk vekst).

Dermed konkluder jeg med at Temperaturen er en bedre forklarende variabel.