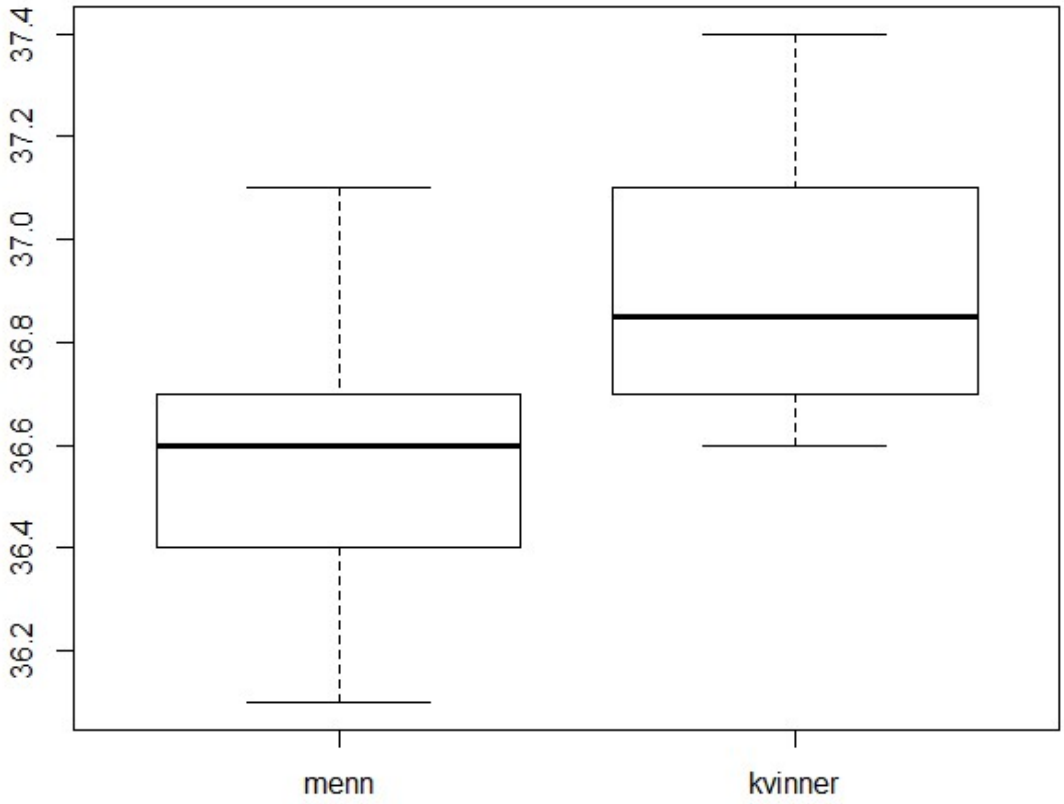
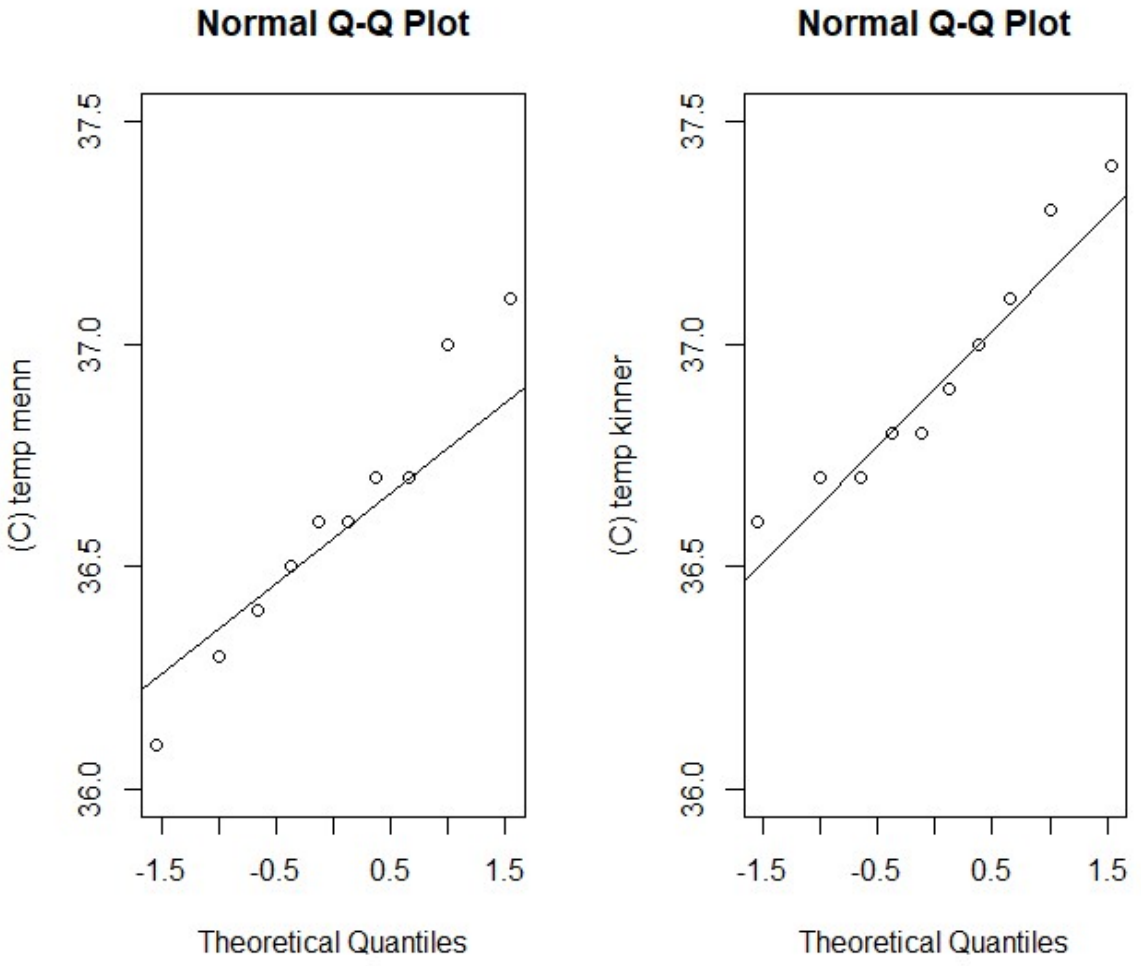
Oppg1

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



Ser at kvinner sin snitt temperatur er høyere en 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.



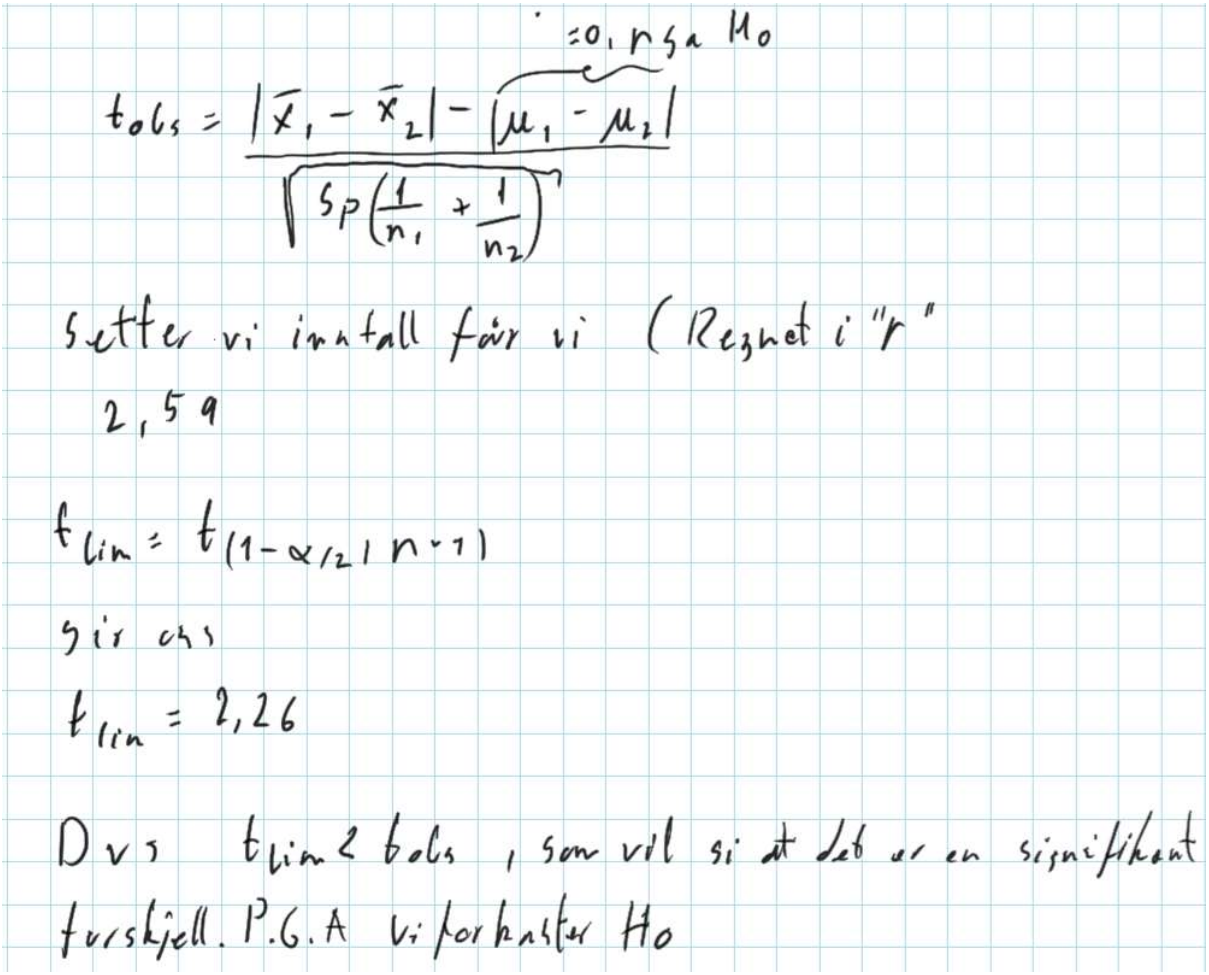
Ser at både kvinner og menn fordelingene er ganske normalt fordelte med få punkter lengere unna enn resten.

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

Vi sette H0: mu_menn = mu_kvinner, Ha: mu_menn != mu_kvinner

For å gjøre testen trenger jeg t_obs, og t_lim:

b1CI



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 nyttighetsskyld

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

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

Løsen 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 at CI for $\mu_D = \mu_x - \mu_y$

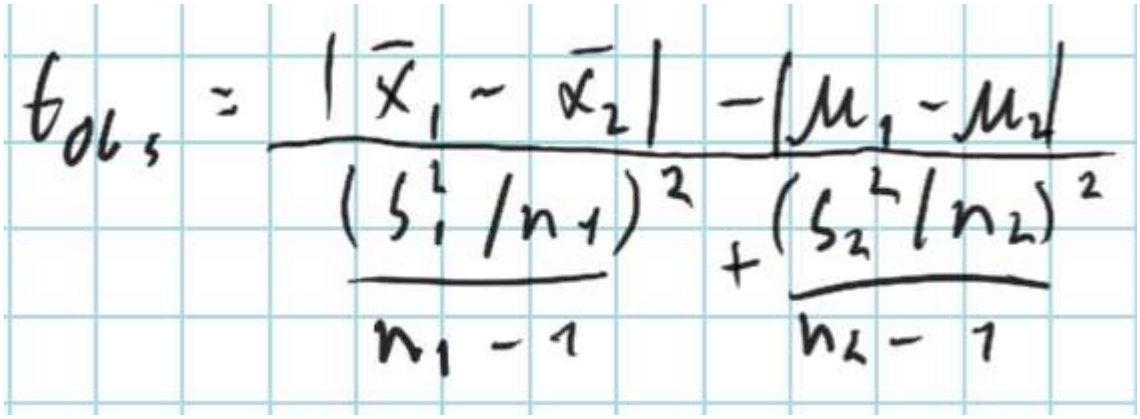
$$\bar{X}-\bar{Y} \pm t_{\alpha/2,\, n_x+n_y-2}\sqrt{S_p\,\frac{1}{n_x}+\frac{1}{n_y}}$$

Setter vi inn verdiene får vi ("r")

$$[0.062,\ 0,598]$$

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

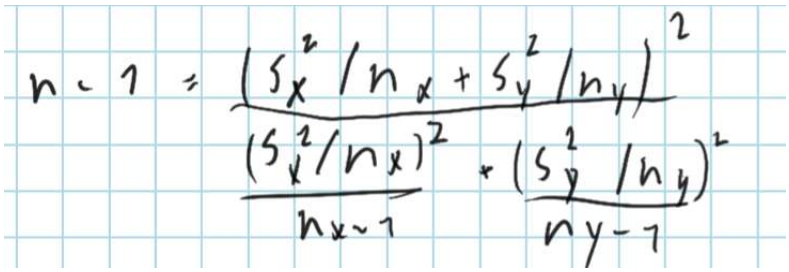
Hvis det ikke er samme varians bruker vi



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

0Det gir oss en t_obs verdi på 3.36483. som er en noe større verdi enn det vi hadde når vi anntok lik varianse. De er fortsatt ganske lik noe vi ville anntat da vaniansen 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 nermeste 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 µD. Sett opp nullhypotese og alternativ hypotese for ˚a besvare spørsm˚alet om forskjell i IQ. Finn en egnet testobservator, og beregn dennes numeriske verdi. Beregn s˚a tilhørende P-verdi. Spesifiser antall frihetsgrader i fordelingen du bruker. Formuler din konklusjon p˚a 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 for å sammenligne med $\alpha = 0{,}05$

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

Setter inn verdiene

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

"r"

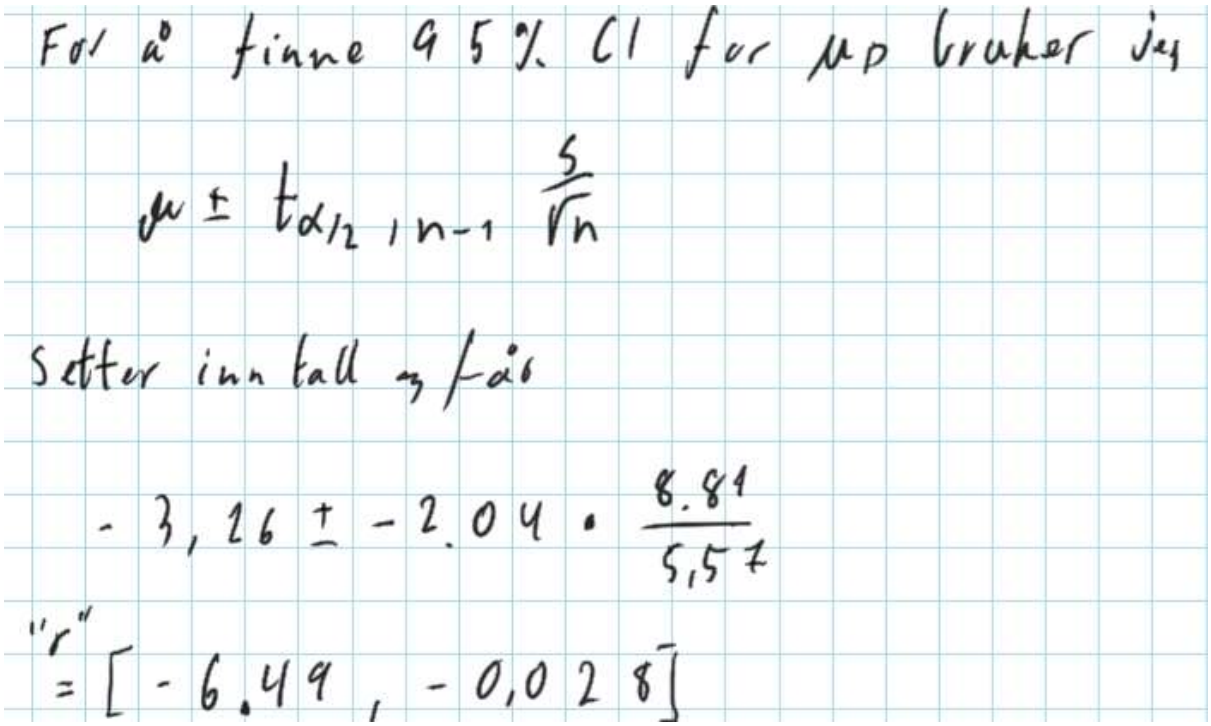
$$= 0{,}046$$

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

II

Vi har n-1 frihetsgrader.

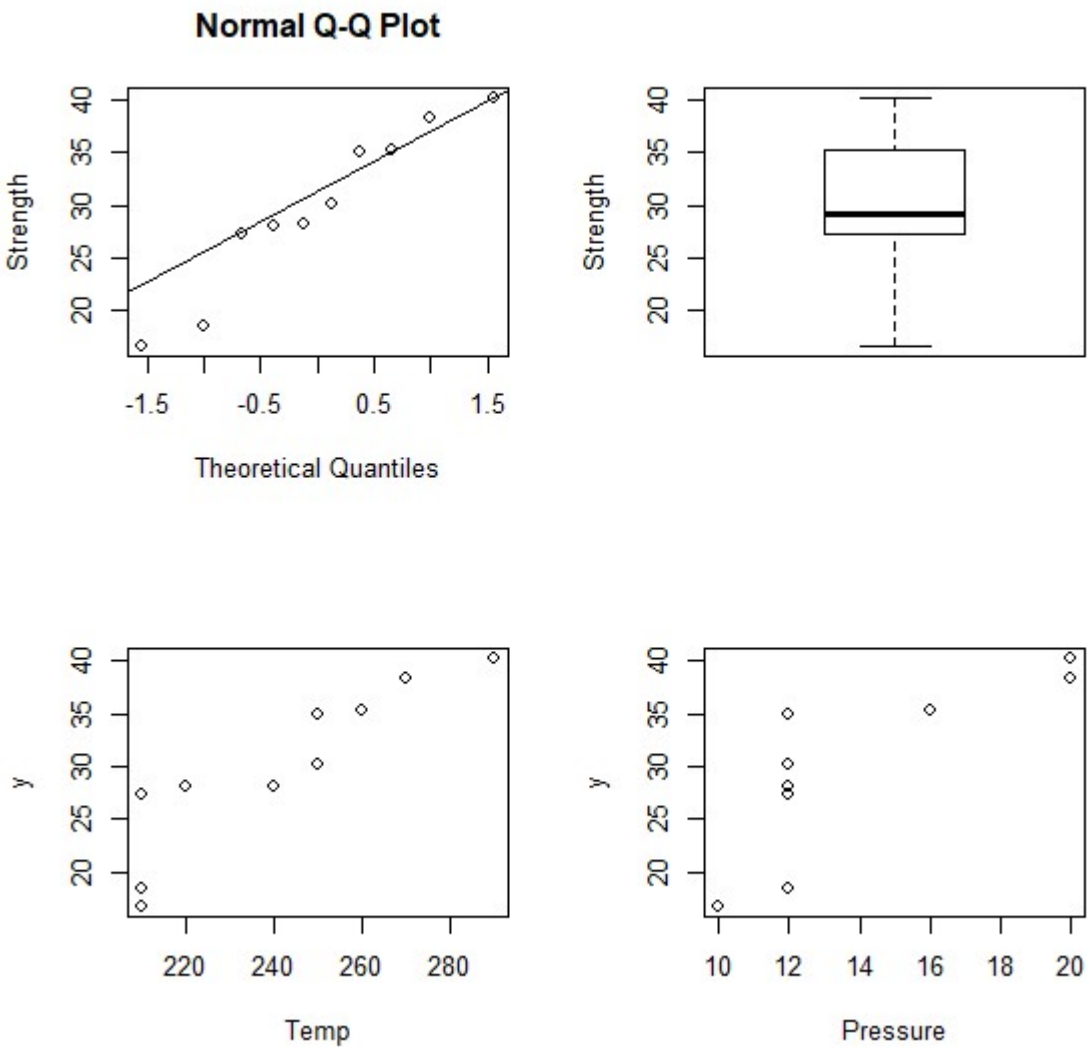
c) Finn et 95% konfidensintervall for µD. Hva betyr det at dette intervallet dekker kun negative verdier? Forklar kort om sammenhengen mellom tosidig testing og konfidensintervaller.



Vår forventningsverdi er negativ da får vi et konfidensintervall som er sentrer 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.

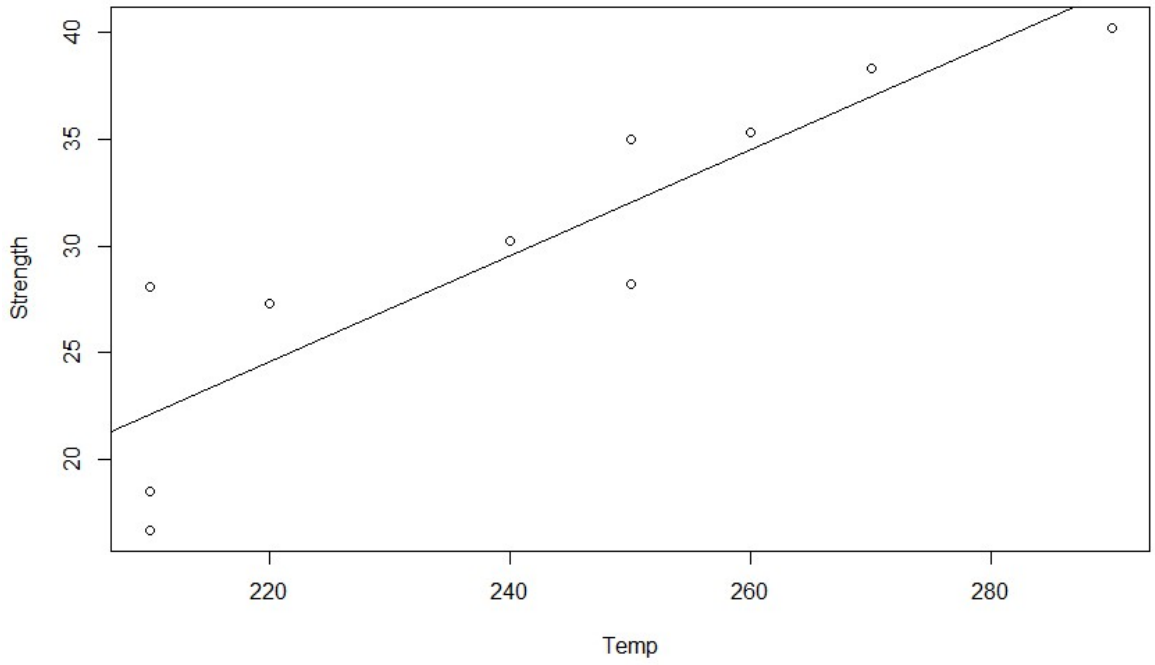


Fra Normal qq Plot ser vi at Strength er tilnærmet normalfordelt. Siden det ser ut som punktene følger normalen.

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 linijer ut,enn y(Pressure), som har et diskontinuerlig hop rundt x2 = 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

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} \qquad (12,2)$$

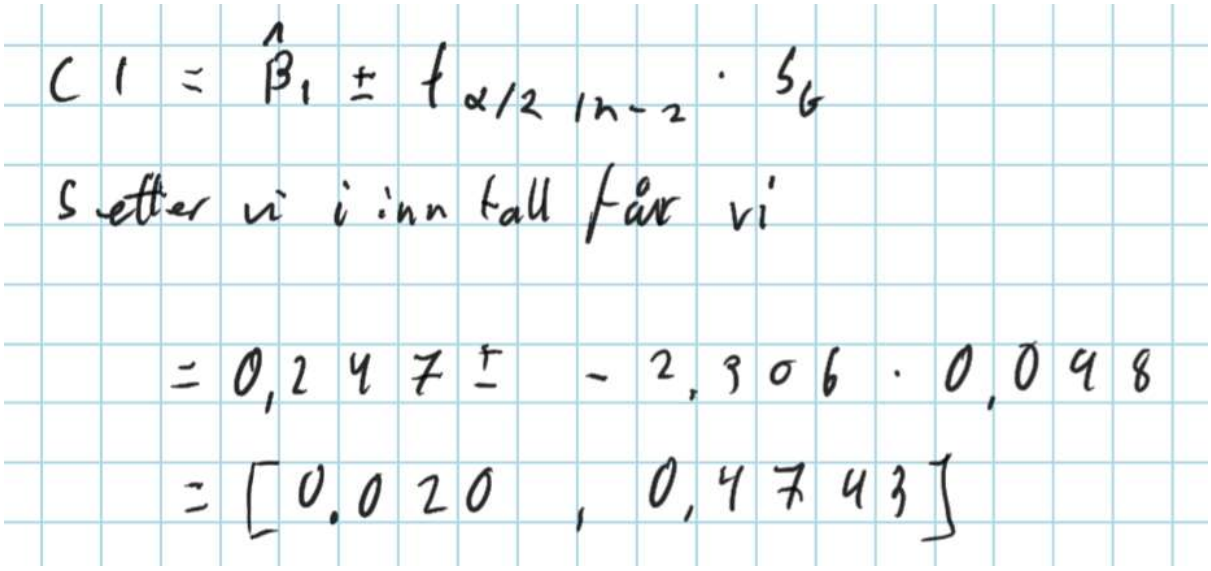
$$\hat{\beta}_0 = \frac{\sum y_i - \hat{\beta}_1 \sum x_i}{n} = \bar{y} - \beta_1 \bar{x} \qquad (12,3)$$

Ved r kan vi bruke lm(x1, y, xlab='Temp',ylab='Strength') for å finne beta1_hat og beta0_hat

(kalt betaX1hat og alfaX1hat i koden). Får da verdiene beta1_hat=0.2474189, beta0_hat =-29.8479549. Fra plottet ser regresjonen rimelig ut, og det virker som regresjonen er en ok estimator for y(x1).

C) Lag konfidensintervall for regresjonskoeffsienten for Temperature. Gir intervallet indikasjon p˚a om temperatur er en viktig forklaringsvariabel?

CI for regresjonsintervallet finner vi ved:



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