Navn: *Henrik Skaanes Steigård*
Brukernavn: *Henrisst* 17. oktober 2021

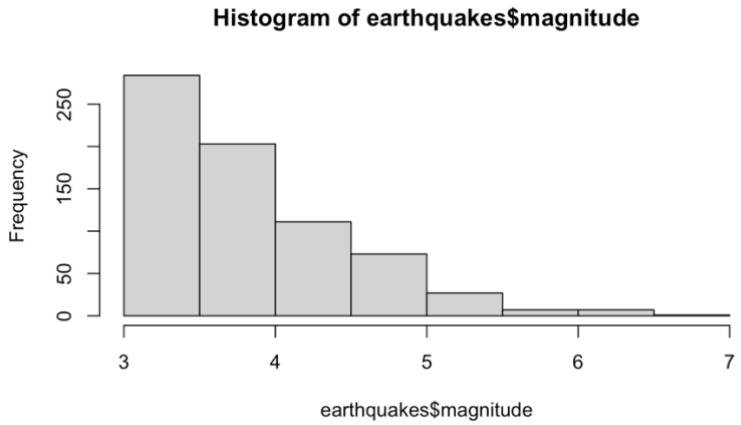
Obligatorisk innlevering 2 – STK1000

Oppgave 1

a) Laster inn nødvendige data:

```
1  data <- "https://www.uio.no/studier/emner/matnat/math/STK1000/data/obligdata/ +
2           oblig2/earthquakes.txt"
3  earthquakes <- read.table(data, header=TRUE)
4  hist(earthquakes$magnitude)
```

Histogrammet blir skrevet ut:



Av histogrammet ser vi at dataene ikke er normalfordelt, men heller sterkt høyreskjev.

b) Skriver nødvendige kommandoer:

```
5  mean(earthquakes$magnitude)
6  sd(earthquakes$magnitude)
```

Konsollen skriver ut følgende:

```
> sd(earthquakes$magnitude)
[1] 0.6623718
> mean(earthquakes$magnitude)
[1] 3.874334
```

Vi ser av konsollutskriften at jordskjelvene har en gjennomsnittlig verdi på ca. 3.87 på Richters skala og et standardavvik på ca. 0.66.

c) Skriver inn nødvendige kommandoer:

```
7  utvalg <- sample(earthquakes$magnitude, 50)
8  mean(utvalg)
```

Konsollen gir følgende utskrift:

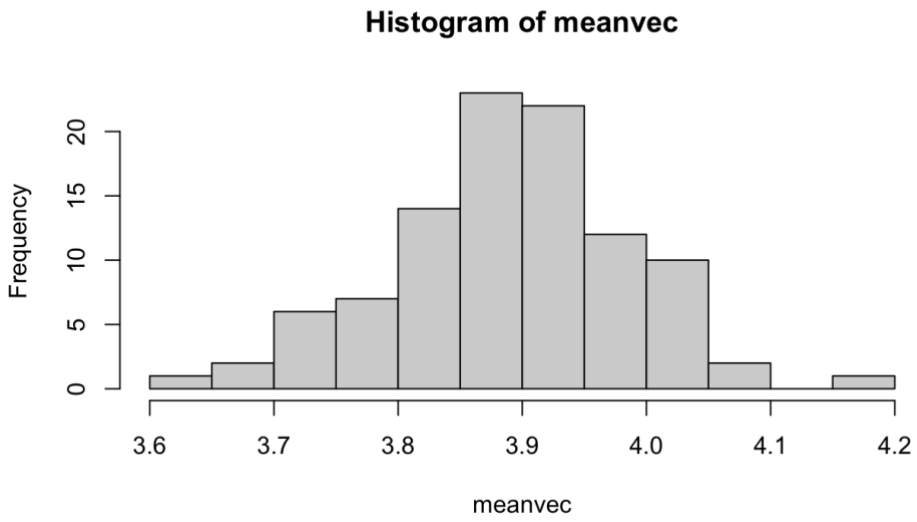
```
> mean(utvalg)
[1] 4.006
```

Vi ser at med et gjennomsnitt på ca. 4.0 på Richters skala, er gjennomsnittet i 50 tilfeldig utvalgte observasjoner høyere enn det som er gjennomsnittet for hele datasettet.

d) Skriver inn oppgitte kommandoer:

```
 9  meanvec <- rep(0, 100)
10  for(i in 1:100) {
11    sample.now <- sample(earthquakes$magnitude, 50)
12    meanvec[i] <- mean(sample.now)
13  }
14  hist(meanvec)
```

Konsollen skriver ut følgende histogram:



Av histogrammet ser vi at når vi tar et større tilfeldig utvalg mange ganger og regner ut gjennomsnittet, vil gjennomsnittene danne en tilnærmet normalfordelt modell.

e) Vi har følgende formler for å finne forventningen og standardavviket til gjennomsnittet:

$$\mu_{\overline{X}} = \mu$$

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

Skriver nødvendige kommandoer inn i konsollvinduet:

```
10  meanvec <- rep(0, 100)
11  for(i in 1:100) {
12    sample.now <- sample(earthquakes$magnitude, 50)
13    meanvec[i] <- mean(sample.now)
14  }
15
16  hist(meanvec, breaks = 5)
17  sd(meanvec)
18  mean(meanvec)
19  mean(earthquakes$magnitude)
20  sd(earthquakes$magnitude)/sqrt(50)
```

```
> sd(meanvec)
[1] 0.1018701
> mean(meanvec)
[1] 3.87632
> mean(earthquakes$magnitude)
[1] 3.874334
> sd(earthquakes$magnitude)/sqrt(50)
[1] 0.09367352
```

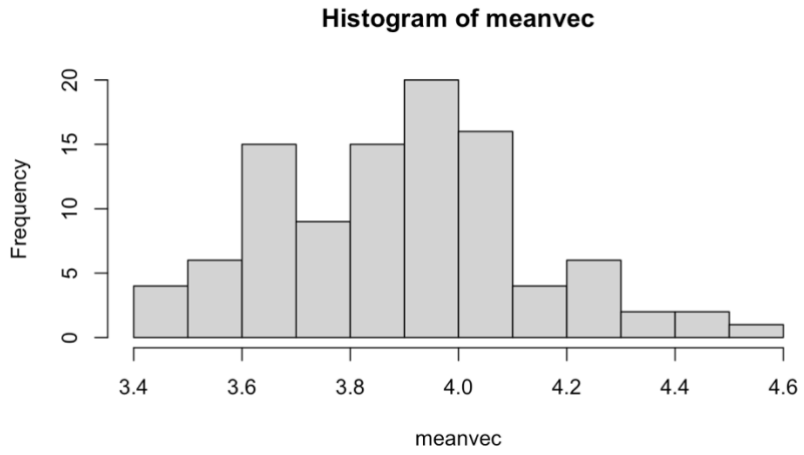
Vi ser at den empiriske forventningen på 3.874334 nærmer seg den teoretiske forventningen på 3.8826. Vi kan se at standardavviket til vektoren meanvec på ca. 0.1 nærmer seg det teoretiske standardavviket på ca. 0.09.

f) Begynner med 10 jordskjelv:

```
10 meanvec <- rep(0, 100)
11 for(i in 1:100) {
12   sample.now <- sample(earthquakes$magnitude, 10)
13   meanvec[i] <- mean(sample.now)
14 }
15
16 hist(meanvec)
17 sd(meanvec)
18 mean(meanvec)
19 mean(earthquakes$magnitude)
20 sd(earthquakes$magnitude)/sqrt(10)
```

```
> sd(meanvec)
[1] 0.1991127
> mean(meanvec)
[1] 3.8869
> mean(earthquakes$magnitude)
[1] 3.874334
> sd(earthquakes$magnitude)/sqrt(10)
[1] 0.2094604
```

Vi kan se at forventningen er tilnærmet lik den empiriske, men at standardavviket er samtidig mye større.



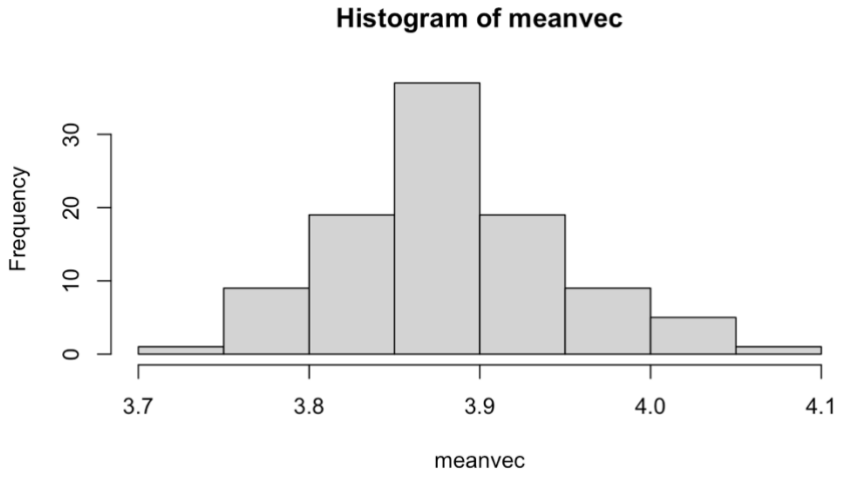
Dette gir en ikke-symmetrisk og heller ikke normalfordelt distribusjon.

For n lik 100:

```
10  meanvec <- rep(0, 100)
11  for(i in 1:100) {
12    sample.now <- sample(earthquakes$magnitude, 100)
13    meanvec[i] <- mean(sample.now)
14  }
15  hist(meanvec)
16  sd(meanvec)
17  mean(meanvec)
18  mean(earthquakes$magnitude)
19  sd(earthquakes$magnitude)/sqrt(100)
```

```
> hist(meanvec)
> sd(meanvec)
[1] 0.06787159
> mean(meanvec)
[1] 3.88355
> mean(earthquakes$magnitude)
[1] 3.874334
> sd(earthquakes$magnitude)/sqrt(100)
[1] 0.06623718
> mean()
```

Vi ser at det teoretiske standardavviket er tilnærmet likt det empiriske. Samtidig er også den teoretiske forventningen tilnærmet lik den empiriske. Dette har sammenheng med store talls lov.



Vi kan se av histogrammet at for de 100 utvalgene med 100 tilfeldige verdier, er symmetrisk og tilsynelatende normalfordelt.

g) Skriver inn nødvendige kommandoer i konsollen:

```
> meanvec <- rep(0, 100)
> for(i in 1:100) {
+   sample.now <- sample(earthquakes$magnitude, 100)
+   meanvec[i] <- mean(sample.now)
+ }
> 1 - pnorm(4.0, mean(meanvec), sd(meanvec))
[1] 0.008919555
```

For 100 utvalg av størrelse n lik 100 er sannsynligheten ca. 0.9% for at man trekker et utvalg med gjennomsnitt større enn 4.0.

```
> meanvec <- rep(0, 100)
> for(i in 1:100) {
+   sample.now <- sample(earthquakes$magnitude, 50)
+   meanvec[i] <- mean(sample.now)
+ }
> 1 - pnorm(4.0, mean(meanvec), sd(meanvec))
[1] 0.08468536
```

For 100 utvalg av størrelse n lik 50 er sannsynligheten ca. 8.5% for at man trekker et utvalg med gjennomsnitt større enn 4.0.

```
> meanvec <- rep(0, 100)
> for(i in 1:100) {
+   sample.now <- sample(earthquakes$magnitude, 10)
+   meanvec[i] <- mean(sample.now)
+ }
> 1 - pnorm(4.0, mean(meanvec), sd(meanvec))
[1] 0.3151439
```

For 100 utvalg av størrelse n lik 10 er sannsynligheten ca. 31.5% for at man trekker et utvalg med gjennomsnitt større enn 4.0.

h) Bias handler om forventningsskjevheten til observatoren. En observator brukes til å estimere parameter. En observator er forventningsrett hvis forventninga til utvalgsfordelingen har lik verdi som den sanne parameterverdien. Variansen til en observator er beskrevet av spredninga til utvalgsfordelingen. Spredningen blir bestemt av utvalgsdesignet og utvalgsstørrelsen n. Ved å bruke større utvalg får man også mindre spredning på dataene. Hvis man velger utvalgsstørrelsen n tilstrekkelig stor og samtidig benytter seg av tilfeldige utvalg, vil man kunne få lav bias (forventningsskjevhet) og lav spredning (variasjon) av observatoren. Dette betyr igjen at forventningen og standardavviket til observatoren vil nærme seg, og ofte være tilnærmet lik, den faktiske forventningen og standardavviket til parameteren.

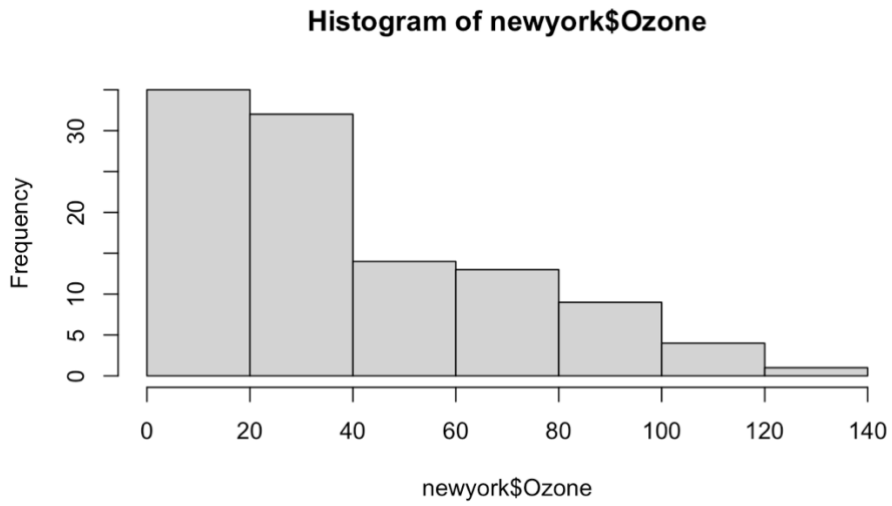
Oppgave 2)

a) Skriver inn følgende kommandoer i r:

```
1 data <- "https://www.uio.no/studier/emner/matnat/math/STK1000/data/ +
2          obligdata/oblig2/ozone.txt"
3 newyork <- read.table(data,header=TRUE)
4 summary(newyork$Ozone)
5 hist(newyork$Ozone)
6
```

Konsollen skriver da ut følgende:

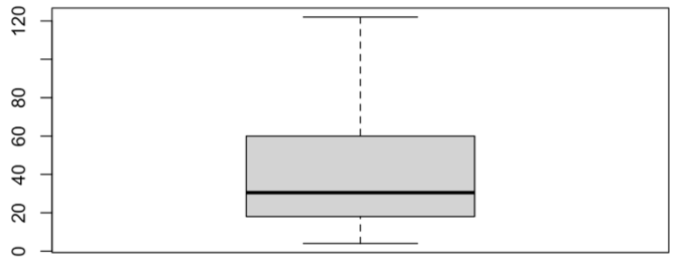
```
> summary(newyork$Ozone)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   4.00   18.00   30.50   40.45   59.50  122.00
```



Av utskriften i R ser vi at ozonnivået har en forventning på 40.45 på syttitallet. Av histogrammet ser vi at fordelingen er sterkt høyreskjev, og at fordelingen derfor ikke er normalfordelt. For å finne eventuelle uteliggere finner vi Q1 - 1.5xIQR og Q3 + 1.5xIQR:

```
> 59.5 + 1.5*IQR(newyork$Ozone)
[1] 121.75
> 18.0 - 1.5*IQR(newyork$Ozone)
[1] -44.25
```

Av utskriften i konsollvinduet over, ser vi at mistenkte uteliggere vil ha en verdi større enn 121.75.



Vi kan se at boksplottet ikke har noen uteliggere.

b) En t-test brukes for å sammenligne forventningen til to utvalg. Bruk av t-test forutsier at dataene er distribuert ved en normalfordeling, men dette kan omgås dersom utvalgsstørrelsen n av dataene våre er større enn 15. T-test er også følsom for uteliggere, derfor forutsier bruk av t-test at man ikke har uteliggere. T-test forutsier også et enkelt tilfeldig utvalg. Selv om modellen for ozon i New York på syttitallet er høyreskjev (og derfor ikke normalfordelt), oppfyller datasettet likevel kriteriet for å kunne bruke t-test.

c) Vi begynner med å formulere en nullhypoteste for ozonnivået i New York for perioden mai til september på syttitallet sammenlignet med forventet ozonnivå i en skandinavisk by på årsbasis:

$$H_0: \quad \mu_{NY:\,mai-sep} - \mu_{Skan} \leq 0$$

Vi formulerer så en alternativ hypotese:

$$H_1: \quad \mu_{NY:\,mai-sep} - \mu_{Skan} > 0$$

Vi velger deretter signifikantnivået *alpha*:

$$\alpha = 0.05$$

Vi kan nå utføre en t-test i R:

```
12  t.test(newyork$Ozone, alternative = "greater")
```

Vi får følgende utskrift i konsollvinduet:

```
data:  newyork$Ozone
t = 14.084, df = 107, p-value < 2.2e-16
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
 35.68791      Inf
sample estimates:
mean of x
  40.4537
```

Av utskriften over kan vi se at p-verdien er tilnærmet lik 0 (veldig liten). Dette betyr at vi kan forkaste nullhypotesen, for det er sterkere bevis i dataene som er i favør for den alternative hypotesen gitt ved:

$$H_1: \quad \mu_{NY:\,mai-sep} - \mu_{Skan} > 0$$

Vi kan også se at forventningen for ozonnivået i New York på syttitallet i månedene mai til september ligger på ca. 40.5 ppm.

d) Skriver inn nødvendig kommando i R for konfidensintervall på 90%:

```
14  t.test(newyork$Ozone, conf.level = 0.9)
```

I konsollen får vi følgende utskrift:

```
data:  newyork$Ozone
t = 14.084, df = 107, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
90 percent confidence interval:
 35.68791 45.21949
sample estimates:
mean of x
  40.4537
```

Vi ser at et konfidensintervall på 90% ligger mellom ca. 35.7 og ca. 42.2.

Skriver inn nødvendig kommando i R for konfidensintervall på 95%:

```
14  t.test(newyork$Ozone, conf.level = 0.95)
```

I konsollen får vi følgende utskrift:

```
t = 14.084, df = 107, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 34.75969 46.14772
sample estimates:
mean of x
  40.4537
```

Vi ser at et konfidensintervall på 95% ligger mellom ca. 34.8 og ca. 46.1.

Skriver inn nødvendig kommando i R for konfidensintervall på 99%:

```
14  t.test(newyork$Ozone, conf.level = 0.99)
```

I konsollen får vi følgende utskrift:

```
data:  newyork$Ozone
t = 14.084, df = 107, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
99 percent confidence interval:
 32.9209 47.9865
sample estimates:
mean of x
  40.4537
```

Vi ser at et konfidensintervall på 99% ligger mellom ca. 33.0 og ca. 48.0.

Vi ser av de forskjellige utskriftene for henholdsvis 90-, 95- og 99-konfidensintervaller at de er forskjellige. Et konfidensintervall kan fortelle oss ved en prosentandel hvor sikre man kan være på at den sanne populasjonsforventningen ligger innenfor et gitt intervall. Dette forklarer hvorfor et 99- er mindre enn et 95- som igjen er mindre enn et 90-konfidensialtietsintervall.

e) Vi begynner med å formulere en nullhypoteste:

$$H_0: \quad \mu_{NY:\, juli,august} - \mu_{NY:\, mai,juni,september} = 0$$

Vi formulerer så en alternativ hypotese:

$$H_1: \quad \mu_{NY:\, juli,august} - \mu_{NY:\, mai,juni,september} > 0$$

Vi velger deretter signifikantnivået *alpha*:

$$\alpha = 0.05$$

Vi kan nå utføre en paret t-test i R:

```
17  t.test(oz.juli.august, oz.mai.juni.sept, alternative = "greater")
```

Konsollen skriver ut følgende:

```
data:  oz.juli.august and oz.mai.juni.sept
t = 4.9526, df = 80.631, p-value = 1.978e-06
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 17.82669      Inf
sample estimates:
mean of x mean of y
 55.61702  28.77049
```

Vi kan se at p-verdien er tilnærmet 0 og er mindre enn signifikantnivået. Det er da rimelig å anta at nullhypoteste er gal og dataene er i favør for den alternative hypotesen.

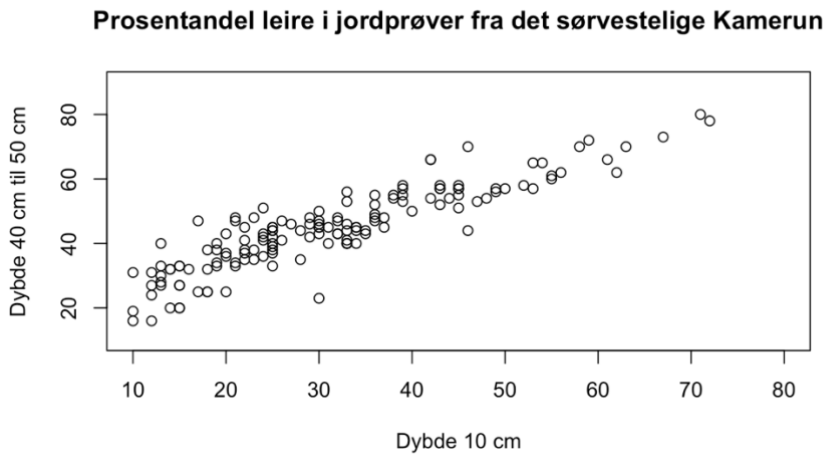
Oppgave 3)

a) Skriver inn nødvendig kode i R:

```
data <- "https://www.uio.no/studier/emner/matnat/math/STK1000/data/ +
         obligdata/oblig2/cameroonclay.txt"
cameroon <- read.table(data,header=TRUE)

clay1 <- cameroon$clay1
clay5 <- cameroon$clay5
plot(clay1, clay5, xlim = c(10, 80), ylim = c(10, 90),
     xlab = "Dybde 10 cm",
     ylab = "Dybde 40 cm til 50 cm",
     main = "Prosentandel leire i jordprøver fra det sørvestelige Kamerun")
```

Konsollen skriver ut følgende plott:

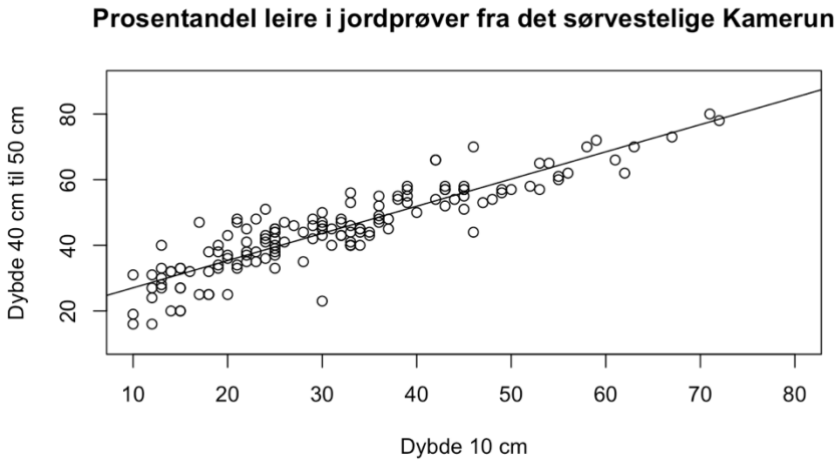


Fra plottet ovenfor ser det ut til å være en tilnærmet lineær sammenheng mellom prosentandel leire i det første laget (p <= 10 cm) og det femte laget (40 cm <= p <= 50 cm). Sammenhengen er ikke veldig sterk, men vi kan se av datapunktene at det danner seg en tilnærmet rett linje gjennom punktene.

b) Skriver inn nødvendie kommandoer i R:

```
fit <- lm(clay5 ~ clay1)
abline(fit)
```

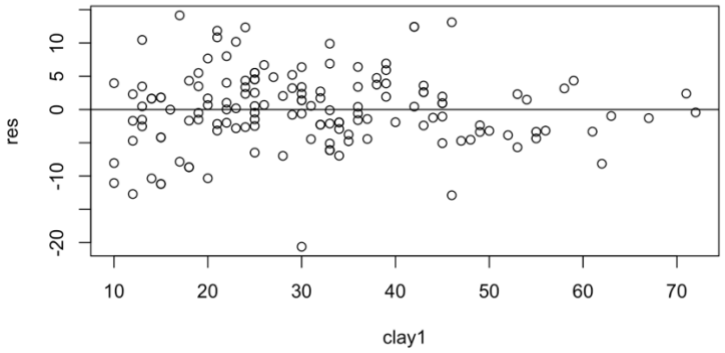
Konsollen skriver ut følgende plott:



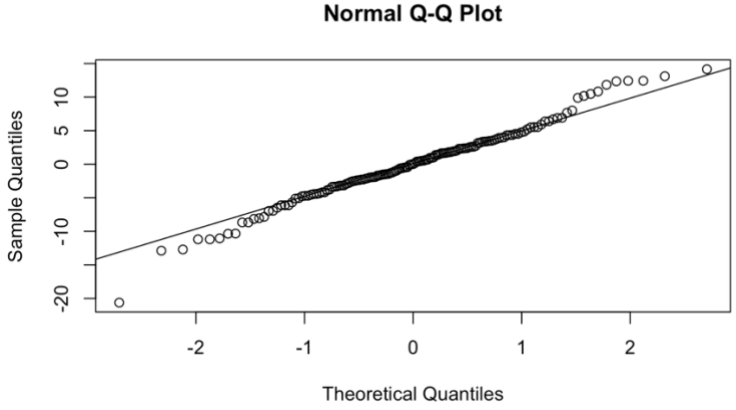
c) Skriver inn følgende kommandoer i R:

```
13 res <- residuals(fit)
14 plot(clay1, res)
15 abline(h = 0)
16 qqnorm(res)
17 qqline(res)
```

Konsollen skriver ut følgende utskrifter:



Vi ser av utskriften ovenfor at residualene ser ut til å være uavhengige av hverandre på grunn av den store spredningen.



Av utskriften ovenfor ser vi at begge endene av regresjonslinjen har en «hale» av datapunkter. Resten av datapunktene ligger tilnærmet på regresjonslinjen. Med denne informasjonen tyder det på at vi har en normalfordeling for x-variabelen.

d) Skriver inn nødvendig kommando i R:

```
18  summary(fit)
```

Konsollen skriver ut følgende utskrift:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 18.75856    1.15561   16.23   <2e-16 ***
clay1        0.82891    0.03377   24.54   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.687 on 145 degrees of freedom
Multiple R-squared:  0.806,     Adjusted R-squared:  0.8047
F-statistic: 602.4 on 1 and 145 DF,  p-value: < 2.2e-16
```

Lar R regne ut en prosentvis økning på 1% for leire i det første laget:

```
19  y_1 = 18.75856 + 0.01*0.82891
20  y_2 = 18.75856
21  y_1 - y_2
```

Konsollen skriver ut følgende:

```
> y_1 - y_2
[1] 0.0082891
```

Vi kan se at en prosentvis økning på 1% med leire i det første laget, gir en predikert prosentvis økning i det femte laget på ca. 0.83%.

e) Fra forrige summary-utskrift i R har vi følgende:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 18.75856    1.15561   16.23   <2e-16 ***
clay1        0.82891    0.03377   24.54   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.687 on 145 degrees of freedom
Multiple R-squared:  0.806,     Adjusted R-squared:  0.8047
F-statistic: 602.4 on 1 and 145 DF,  p-value: < 2.2e-16
```

Vi kan se at p-verdien er tilnærmet 0. Dette forteller oss at vi kan se bort i fra nullhypotesen og deklarere denne som mer eller mindre feil. Vi ser altså på den alternative hypotesen gitt som at stigningstallet er større enn 0 som sann.

f) Skriver inn nødvendige kommandoer i R:

```
24  b1 <- summary(fit)$coefficients[2, 1]
25  se.b1 <- summary(fit)$coefficients[2, 2]
26  df <- fit$df.residual
27  lower <- b1 + qt(0.025, df) * se.b1
28  upper <- b1 + qt(0.975, df) * se.b1
29  upper
30  lower
```

Konsollen skriver ut følgende:

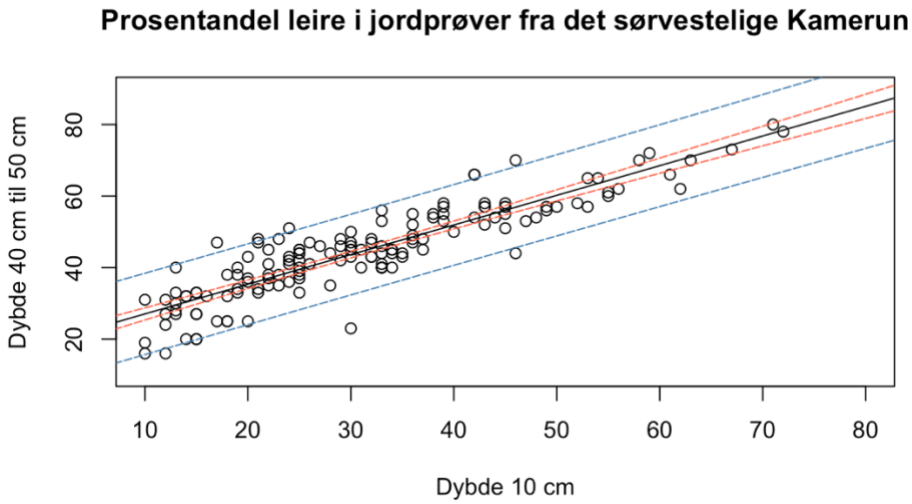
```
> upper
[1] 0.8956582
> lower
[1] 0.7621583
```

Vi ser av utskriften over at 95-konfidensintervallet er fra og med ca. 0.896 til og med ca. 0.762.

g) Skriver inn nødvendige kommandoer i R:

```
plot(clay1, clay5, xlim = c(10, 80), ylim = c(10, 90),
     xlab = "Dybde 10 cm", ylab = "Dybde 40 cm til 50 cm" ,
     main = "Prosentandel leire i jordprøver fra det sørvestelige Kamerun")
abline(fit)
xval <- seq(0, 100, by = 0.01)
new <- data.frame(clay1 = xval)
pred.int <- predict(fit, newdata = new, interval = "prediction")
mean.int <- predict(fit, newdata = new, interval = "confidence")
matlines(xval, cbind(pred.int[, 2], pred.int[, 3]), lty = 2,
         col = "steelblue")
matlines(xval, cbind(mean.int[, 2], mean.int[, 3]), lty = 2,
         col = "tomato")
```

R skriver ut følgende utskrift:



h) Fra forrige utskrift ser vi at for x = 60 for 95-prediksjonsintervallet er fra og med ca. 58 til og med ca. 80 ved å lese av grafen.