Navn: *Henrik Skaanes Steigård*
Brukernavn: *Henrisst*
Oppgavespesifikasjon: *Geologi*

Torsdag 2. september 2021

OBLIGATORISK INNLEVERING 1 – STK1000

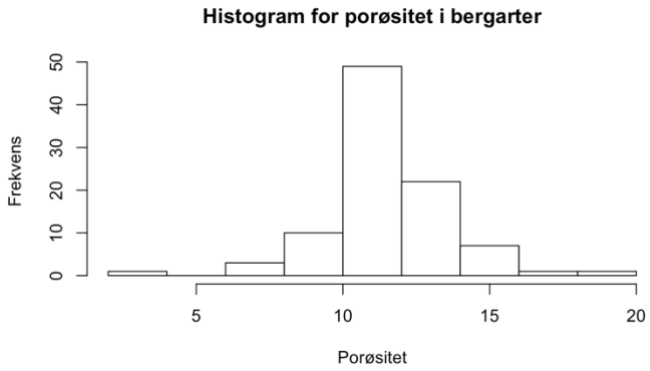
Oppgave 1

a)

R-kode for oppgaven:

```
#Inkluderer data som skal brukes i oppgaven:
data = "http://www.uio.no/studier/emner/matnat/math/STK1000/data +
/obligdata/oblig1/pennsylvania.txt"
myValues <- read.table(data,header=TRUE)
#Hadde problemer med R, noe som gjorde at
#Dataene måtte konverteres til numeriske data:
con = myValues[,1]su
con <- gsub(",", "", we)   # remove comma
con <- as.numeric(we)
#Histogrammet visualiseres med tilhørende spesifikasjoner:
hist(con, main = "Histogram for porøsitet i bergarter",
     xlab = "Porøsitet",
     ylab = "Frekvens",
     border = "black", col = "white")
```

R-koden er illustrert ved figur 1



Figur 1: *Figuren illustrerer histogrammet for 94 borreprøver Av ulike bergarter og deres porøsitet*

Av figur 1 ser man at mange av prøvene ser ut til å ligge mellom 10 % og 12 % i porøsitet. Man kan også se en mulig(e) uteligger(e) helt til venstre i histogrammet med porøsitet < 5 %.

b)

R-kode for oppgaven. Jeg fortsetter i samme script som i deloppgave a):

```
mean(con)
median(con)
```

Konsollen skriver ut følgende verdier:

```
> mean(con)
[1] 11.57128
> median(con)
[1] 11.4
```

Vi kan se av konsollutskriften at gjennomsnitt og median for borreprøvene er ganske nærme hverandre. Dette forteller oss at det ikke ser ut til å være ekstremer i porøsiteten til de ulike bergartene som er blitt undersøkt.

c)

R-kode for oppgaven. Jeg fortsetter i samme script som i deloppgave a) og b):

```
17  sd(con)
18  IQR(con)
```

Konsollen skriver ut følgende verdier:

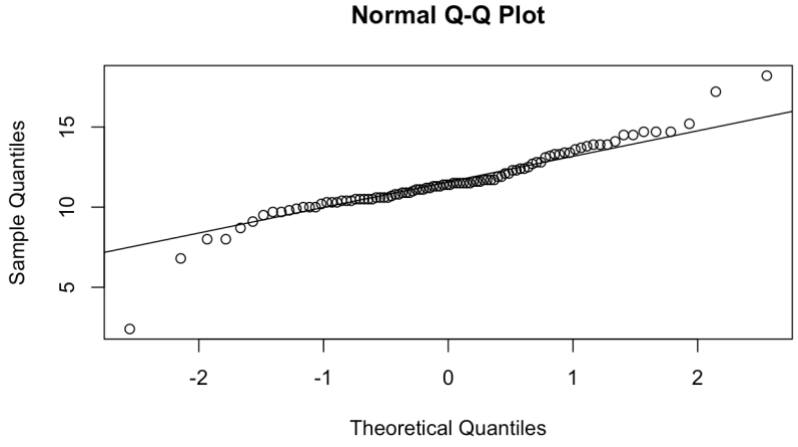
```
> sd(con)
[1] 2.059106
> IQR(con)
[1] 2.15
```

Av histogrammet fra deloppgave a) ser vi at det er stor variasjon i porøsitet for de ulike bergartene. Standardavviket på ca. 2.06 forteller oss at det ikke er spesielt store avvik i porøsitet og at de fleste bergartene er samlet ganske nære medianen i porøsitet. En inter-kvartil avstand (IQR) på 2.15 forteller oss at 50% av observasjonene ligger innenfor +/- 1.075 i porøsitet i forhold til medianen på 11.4.

d)
R-kode for oppgaven. Jeg fortsetter i samme script som i deloppgave a) og b):

```
19  #Bruker Q-norm for aa distrubuere dataene
20  qqnorm(con)
21  #Bruker Q-line for å se hvordan datapunktene ligger i forhold til hverandre
22  qqline(con)
```

R plotter følgende i figur 2:



Figur 2

Av figur 2 ser vi at de fleste datapunktene ligger omtrent på den lineære linjen. Av figuren ser man at plottet har to haler på venstre og høyre side av plottet. Disse to halene ser ikke ut til å ha veldig stor betydning og det er rimelig å anta at *porosity* er normalfordelt.

e)

Den standardiserte verdien, eller Z-scoren, er gitt ved formelen:

$$Z = \frac{x - \mu}{\sigma}$$

Og ved å sette oppgavens oppgitte verdier inn i formelen for Z-score, får vi følgende

$$Z = \frac{x - \bar{x}}{s}$$

Den standardiserte verdien til *porosity* lik 14% kan vi finne ved hjelp av R og formelen vi har:

```
32 Z = (14 - mean(con))/sd(con)
```

Dette gir i konsollen:

```
Z                    1.17950364980287
```

Altså er den standardiserte verdien er ca. Z = 1.18. Z-scoren forteller oss hvor langt vi befinner oss fra gjennomsnittet med hensyn til antall standardavvik. Altså faller porøsitet på 14% over den 50-persentilen.

f)

Andel prosent av borreprøver med porøsitet lavere enn 8% er gitt ved R-koden:

```
25 #Finner andel ut fra datasettet vårt
26 pnorm(8,mean(con), sd(con))
```

Som konsollen beregner til følgende verdi:

```
> #Finner andel ut fra datasettet vårt
> pnorm(8,mean(con), sd(con))
[1] 0.04142518
```

Altså er det ca. 4.1% av borreprøvende som har porøsitet mindre enn 8%.

g)

Andel prosent av borreprøver med porøsitet høyere enn 15% er gitt ved følgende R-kode:

```
27 #Finner andel ut fra datasettet ved hjelp av subtraksjon:
28 1- pnorm(15, mean(con), sd(con))
```

Fra konsollen får vi beregnet prosentandel:

```
> 1- pnorm(15, mean(con), sd(con))
[1] 0.04794129
```

Dette forteller oss at ca. 4.8% av borreprøvene har høyere porøsitet enn 15%.

Oppgave 2

a)

Kategoriske variabler plasserer individer inn i ulike grupper. Kategoriske variabler kan for eksempel være et tilfeldig utvalg individer som blir plassert i én av to grupper: gifte og ugifte. I motsetning til kategoriske variabler, er kvantitative variabler brukt til å representere numeriske verdier. Dette kan for eksempel være alderen på et utvalg mennesker og deres body-mass-index (BMI). Her vil alder, vekt og høyde være de kvantitative variablene.

I oppgaven vil vekten og prosentandelen sand i kjerneprøven være kvantitative variable. Dybden hvor prøvene blir hentet fra, kan settes i to kategorier: grunt (dybde < 4m) og dypt (dybde > 4m). Dybden er en kategorisk variabel.

b)

Skriver følgende kode i R:

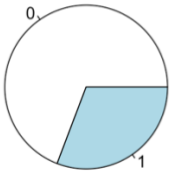
```
1 #Henter data
2 data="http://www.uio.no/studier/emner/matnat/math/STK1000/data/obligdata +
3 /oblig1/kjerneprover.txt"
4 kjerneprover <- read.table(data,header=TRUE)
5 #Får table og pie for dybde depth
6 table(kjerneprover$depth)
7 pie(table(kjerneprover$depth))
```

Konsollen skriver ut følgende verdier for table:

```
> table(kjerneprover$depth)

  0   1
137  61
```

Av utskriften ovenfor ser vi det er 137 prøver som er tatt på grunt vann og 61 prøver som er tatt på dypt vann.



Figur 3: *Over ser vi kakediagrammet for den kategoriske variabelen For dybde, depth. Diagrammet viser visuelt fordelingen av dype og Grunne borreprøver.*

c)

Skriver inn oppgitt R-kode:

```
10  kjerneprover.dypt <- kjerneprover[kjerneprover[, "depth"]==1,]
11  kjerneprover.grunt <- kjerneprover[kjerneprover[,"depth"]==0,]
```

Konsollen skriver ut følgende:

```
> kjerneprover.dypt <- kjerneprover[kjerneprover[, "depth"]==1,]
> kjerneprover.grunt <- kjerneprover[kjerneprover[,"depth"]==0,]
```

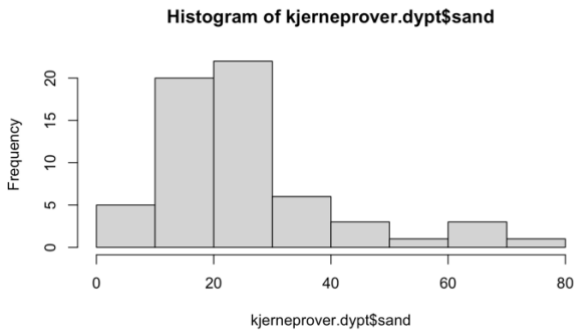
og dataene er dermed delt i to datasett.

d)

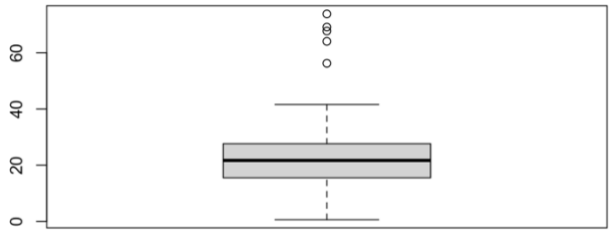
Fra forrige deloppgave har vi lagret de dype og grunne prøvene separat. Vi kan derfor kjøre følgende kommandoer:

```
12  #Lager et histogram for dype prøver og prosentandel sand:
13  hist(kjerneprover.dypt$sand)
14  #Lager et boxplot for dype prøver og prosentandel sand:
15  boxplot(kjerneprover.dypt$sand)
```

Figur 3 viser histogrammet for dype prøver, og figur 4 viser boksplottet for dype prøver. Begge figurene baserer seg på prosentandel sand i borreprøvene.



Figur 3: *histogram dype prøver*

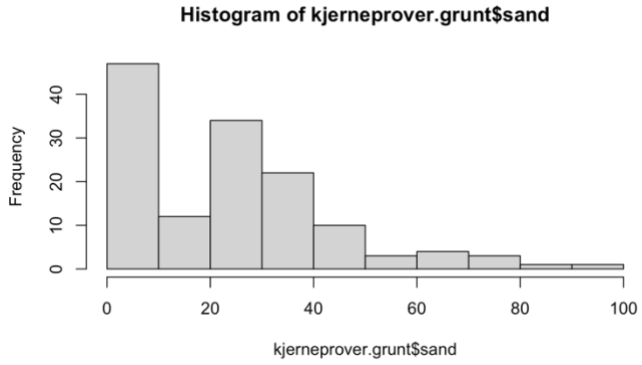


Figur 4: *boksplott av dype prøver*

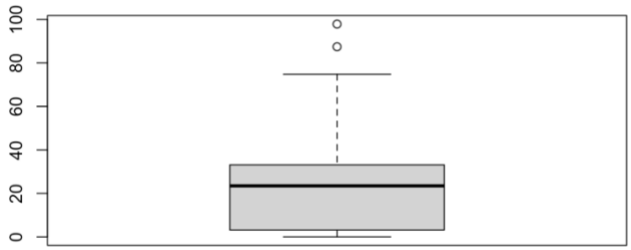
Vi kjører samme kode grunne prøver:

```
#Lager et histogram for grunne prøver og prosentandel sand:
hist(kjerneprover.grunt$sand)
#Lager et boxplot for grunne prøver og prosentandel sand:
boxplot(kjerneprover.grunt$sand)
```

Figur 5 viser histogrammet for grunne prøver, og figur 6 viser boksplottet for grunne prøver. Begge figurene baserer seg på prosentandel sand i borreprøvene.



Figur 5: *histogram av grunne prøver*



Figur 6: *boksplott av grunne prøver*

Ved å sammenligne histogrammene og boksplottene for dype og grunne prøver, ser vi flere opplysninger om prøvene: Prøvene tatt på dypt og grunt vann ser ut til å ha omtrentlig samme median for prøvene. Vi ser at den interkvartile avstanden for dype prøver er mye lavere hos de dype prøvene, noe som tilsier at omtrent halvparten av prøvene har en sandandel mellom ca. 18% til 30%. Sammenligner vi dette med boksplottet for grunne prøver, ser vi at den interkvartile avstanden har større spredning når det kommer til prosentandel sand i prøvene. I de grunne prøvene varierer ca. Halvparten av prøvene mellom større enn 0% til ca. 38%. Vi ser også at maksimum og minimum av de dype prøvene ligger omtrent like langt fra medianen til prøvene. Ved å sammenligne histogrammene og boksplottene, kan man konkludere med at sannsynligheten for å finne prøver med relativ høy prosentandel sand ser ut til å være mye høyere hos dype prøver.

e)

Skriver inn følgende R-kode for å finne median og gjennomsnitt:

```
#Bruker median og mean for median og gjennomsnitt av grunne prøver:
median(kjerneprover.grunt$sand)
mean(kjerneprover.grunt$sand)
#Bruker median og mean for median og gjennomsnitt av dype prøver:
median(kjerneprover.dypt$sand)
mean(kjerneprover.dypt$sand)
```

Når man kjører koden, blir følgende skrevet ut av konsollen:

```
> median(kjerneprover.grunt$sand)
[1] 23.46
> mean(kjerneprover.grunt$sand)
[1] 22.97161
```

Over ser vi at medianen og gjennomsnitt er 23.46 og ca. 22.97 for de grunne borreprøvene.

```
> median(kjerneprover.dypt$sand)
[1] 21.71
> mean(kjerneprover.dypt$sand)
[1] 25.01574
```

Over ser vi at medianen og gjennomsnitt er 21.71 og ca. 25.02 for de dype borreprøvene.

Vi ser at medianen og gjennomsnittet for dype og grunne prøver ikke har veldig store variasjoner. Gjennomsnittet er litt lavere enn medianen hos de grunne prøvene, noe som kan tyde på at vi har prøver med lav andel sand. I de dype prøvene er gjennomsnittet en større enn medianen, noe som tyder på at vi har uteliggere med høy andel sand.

f)

Ved å se på boksplottet til de dype borreprøvene, ser vi at det er mange uteliggere, og fra deloppgave e) observerer vi at medianen og gjennomsnittet har en betydelig differanse. Dette gir en god indikasjon på at femtallsoppsummering er et godt mål for å få bred oversikt over spredningen til dataene.

For de grunne borreprøvene vil gjennomsnitt og standardavvik fungere bra, men ikke som eneste mulige fremstilling, ettersom dataene inneholder bare noen få uteliggere.

f og g)

Ved å sammenligne histogrammene fra deloppgave e), ser vi at begge histogrammene ikke er symmetriske og det er derfor heller ikke rimelig å anta at de er normalfordelte. Ut ifra det vi kan observere, vil det være et godt valg å bruke femtallsoppsummering for å beskrive de to datasettene. Femtallsoppsummeringen vil gi oss en god oversikt over spredningen til datasettene, og uten at de to er normalfordelte, vil ikke gjennomsnitt og standardavvik si mye om spredningen av dataene.

Oppgave 3

a)

Skriver inn følgende data i R-kode:

```
#Henter data
data = "http://www.uio.no/studier/emner/matnat/math/STK1000/data +
/obligdata/oblig1/vitruvisk.txt"
vitruvisk <- read.table(data,header=TRUE)
#Får en oppsummering av dataene:
summary(vitruvisk)
#Regner ut hvor mange menn og kvinner som deltok:
table(vitruvisk$kjonn)
```

Og konsollen skriver ut følgende:

```
> table(vitruvisk$kjonn)

  K   M
150  73
```

```
> summary(vitruvisk)
    kjonn            kroppslengde     fot.navle       navle.isse
 Length:223         Min.   :152.0   Min.   : 87.0   Min.   :52.00
 Class :character   1st Qu.:166.0   1st Qu.:101.0   1st Qu.:65.00
 Mode  :character   Median :172.0   Median :104.0   Median :67.00
                    Mean   :172.3   Mean   :104.8   Mean   :67.34
                    3rd Qu.:178.0   3rd Qu.:109.0   3rd Qu.:70.00
                    Max.   :196.0   Max.   :125.0   Max.   :81.00
     favn
 Min.   :146.0
 1st Qu.:165.0
 Median :171.0
 Mean   :172.4
 3rd Qu.:180.0
 Max.   :202.0
```

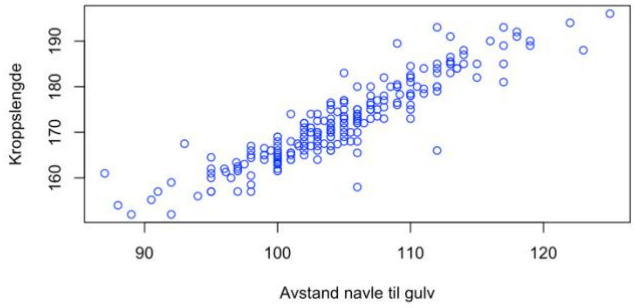
Av konsollen ser vi at en overvekt på 150 individer i undersøkelsen var kvinner og 73 var menn. Fra fempunktsoppsummeringene ser vi at kroppslengde har minimum lik 152, første kvartil er 166 (den 25 persentilen), medianen er 172 (den 50 persentilen), tredje kvartil er 178 (75 persentilen) og maksimum er 196. For fot.navle er femtallsoppsummeringen slik: minimum lik 87, første kvartil er 101 (den 25 persentilen), medianen er 104 (den 50 persentilen), tredje kvartil er 109 (75 persentilen) og maksimum er 125.

b)

Skriver inn følgende kode i R for å få spesifisert plott:

```
#Plotter x- og y-verdiene for plottet etter oppgavespesifikasjonen:
plot(vitruvisk$fot.navle, vitruvisk$kroppslengde, col = 'blue',
     xlab = "Avstand navle til gulv", ylab = "Kroppslengde")
```

I figur 9 ser vi plottet gitt av R-koden over:



Figur 9

Ved å se på det grafiske plottet for fot.navle på x-aksen og kroppslengde på y-aksen, ser vi at det er et ganske lineært proporsjonalt forhold mellom fot.navle-størrelsen og kroppslengden på en person. Altså, jo høyere man er, desto større er sannsynligheten for at man også har større fot.navle-avstand.

c)

Skriver inn følgende kode i R for å få korrelasjonen:

```
#Bruker koorelasjonsfunksjon:
cor(vitruvisk$fot.navle, vitruvisk$kroppslengde)
```

Konsollen skriver ut følgende verdi:

```
> cor(vitruvisk$fot.navle, vitruvisk$kroppslengde)
[1] 0.9140397
```

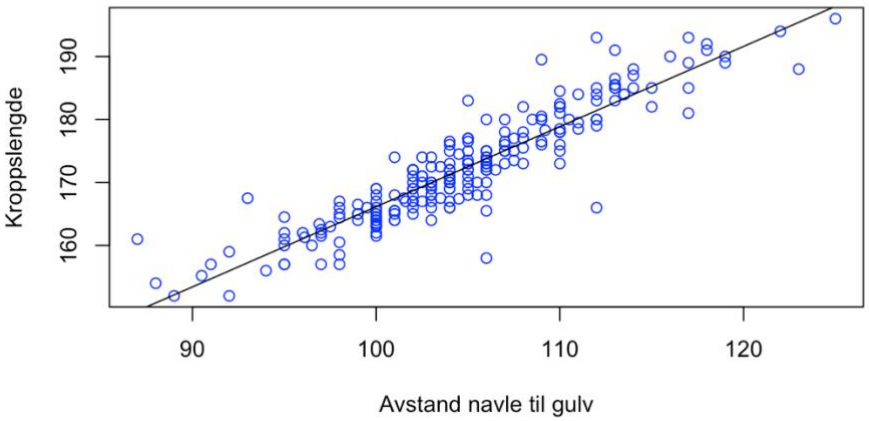
Av utskriften ser vi at korrelasjonen mellom navlehøyde og kroppshøyde har en positiv, sterk lineær sammenheng.

d)

Skriver inn følgende R-kode:

```
plot(vitruvisk$fot.navle, vitruvisk$kroppslengde, col = 'blue',
     xlab = "Avstand navle til gulv", ylab = "Kroppslengde")
#Legger til nødvendig tilleggskode
fit <- lm(vitruvisk$kroppslengde ~ vitruvisk$fot.navle)
abline(fit)
```

Koden gir oss den lineære regresjonslinjen i spredningsplottet for navlehøyde og kroppshøyde i figur 10.



Figur 10

Vi ser av figuren at det er en sterk, positiv lineær sammenheng mellom navlehøyde og kroppslengde.

e)

Skriver inn følgende kode i R:

```
cor(vitruvisk$fot.navle, vitruvisk$kroppslengde)
plot(vitruvisk$fot.navle, vitruvisk$kroppslengde, col = 'blue',
     xlab = "Avstand navle til gulv", ylab = "Kroppslengde")
#Legger til nødvendig tilleggskode
fit <- lm(vitruvisk$kroppslengde ~ vitruvisk$fot.navle)
abline(fit)
summary(fit)
```

Og konsollen skriver ut følgende:

```
> summary(fit)

Call:
lm(formula = vitruvisk$kroppslengde ~ vitruvisk$fot.navle)

Residuals:
     Min       1Q   Median       3Q      Max
-15.7843  -1.9667  -0.0568   2.1695  11.8981

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         38.89675    3.98888   9.751   <2e-16 ***
vitruvisk$fot.navle  1.27252    0.03799  33.499   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.605 on 221 degrees of freedom
Multiple R-squared:  0.8355,    Adjusted R-squared:  0.8347
F-statistic:  1122 on 1 and 221 DF,  p-value: < 2.2e-16
```

Av utskriften over ser vi at hvis navlehøyde øker med én cm, vil den predikerte kroppslengden øke med ca. 1.273 cm. Vi kan også se at skjæringspunktet er gitt ved ca. 38.9 ved y-aksen. Vi har altså den tilnærmede lineære modellen

$$y = 1.273x + 38.9$$

Der x er navlehøyden, 1.273 er stigningstallet og 38.9 er konstantleddet.

f)

Ved å bruke modellen fra d), få vi følgende for den predikerte kroppslengden til en person som er 121 cm over bakken:

$$y = 1.273x + 38.9$$

$$y = (1.273 * 121)\,cm + 38.9\,cm$$

$$y \approx 192.9\,cm$$

Altså vil en person med navlehøyde på 121 cm ha en predikert kroppslengde på ca. 192. 9 cm.

g)

Henter utskrift fra deloppgave e)

```
> summary(fit)

Call:
lm(formula = vitruvisk$kroppslengde ~ vitruvisk$fot.navle)

Residuals:
     Min       1Q   Median       3Q      Max
-15.7843  -1.9667  -0.0568   2.1695  11.8981

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         38.89675    3.98888   9.751   <2e-16 ***
vitruvisk$fot.navle  1.27252    0.03799  33.499   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.605 on 221 degrees of freedom
Multiple R-squared:  0.8355,    Adjusted R-squared:  0.8347
F-statistic:  1122 on 1 and 221 DF,  p-value: < 2.2e-16
```

Vi leser av konsollutskriften at multippel r-kvadrert er 0.[redacted]. Dette betyr at 83.55% av variasjonen i kroppslengde kan forklares av navlehøyden.

h)

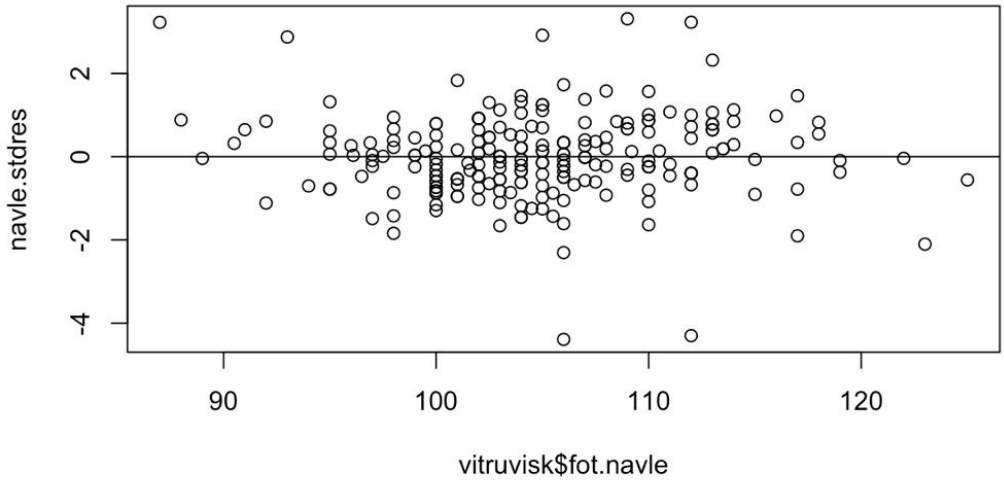
Skriver inn følgende R-kode:

```
19 #Skriver inn oppgitt kode i oppgaven:
20 plot(vitruvisk$fot.navle,residuals(fit))
21 abline(h=0)
```

Figur 11 viser plottingen av residualene til datasettet i oppgave:

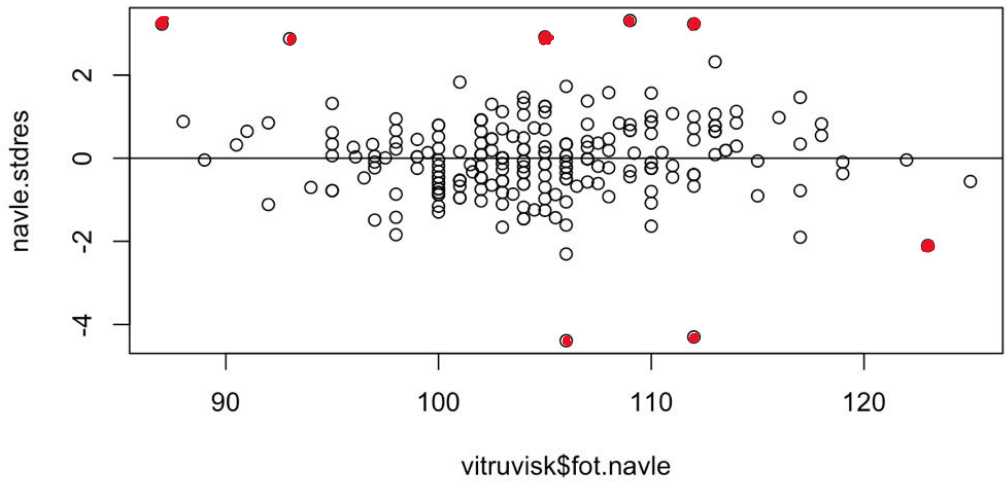
```
23 #Standardiserer dataene for aa se dataene i forhold til standardaavvik.
24 navle.lm = lm(vitruvisk$kroppslengde ~ vitruvisk$fot.navle, data=vitruvisk)
25 navle.stdres = rstandard(navle.lm)
26 plot(vitruvisk$fot.navle, navle.stdres)
27 abline(0,0)
```

Konsollen skriver ut følgende residualplott som er angitt som figur 11:



Figur 11

Fra figuren med residualplottet, ser vi at spredningen av dataene er veldig tilfeldig i residualplottet: vi ser ingen mønstre. Dette er en god indikasjon på at modellen passer veldig godt til å beskrive dataene våre. Y-aksen viser antall standardavvik fra forventet verdi, og eventuelle uteliggere kan vi definere slik: En uteligger ligger mer enn 2.5 standardavvik under og over forventet verdi. Vi får denne modellen med eventuelle uteliggere i figur 12:



Figur 12: *Figuren viser mulige uteliggere i residualplottet*
Som har et standardavvik på +/- 2.5 standardavvik
Fra forventet verdi.