# Oblig 2-STAT111

Sigbjorn Fjelland

4/10/2020

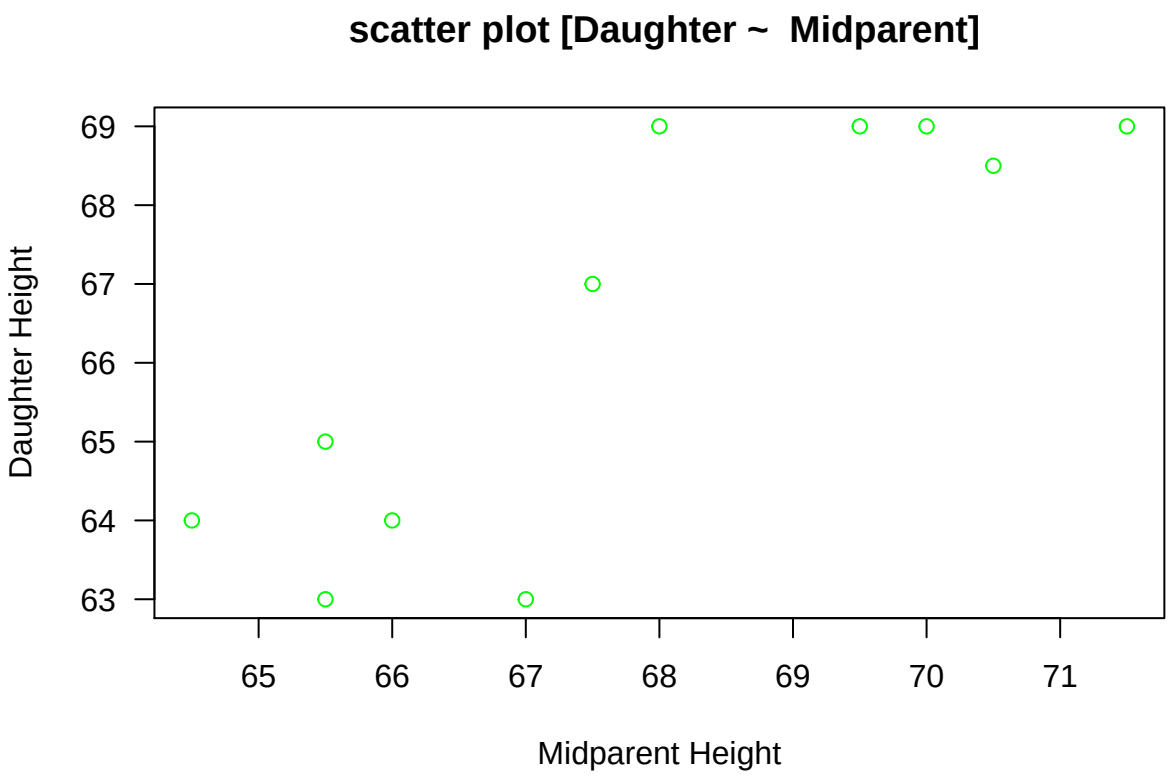
Oppgave 1

Galton brukte fedres høyde for å estimere høyden til sønnen. I oppgaven introduseres det at Galton også brukte også gjennomsnitts høyde på forreldre (altså gjennomsnitt mellom mor og far). Videre er det gitt en tabell med 11 snitthøyde til forreldre og høyde på tilhørende datter.

a)

I delspørsmål a skal vi lage ett spredningsplott for den gitte tabellen med snittøyde til forreldre sat

```
Midparent <- c(66.0, 65.5, 71.5, 68.0, 70.0, 65.5, 67.0, 70.5, 69.5, 64.5, 67.5)
Daughter <- c(64.0, 63.0, 69.0, 69.0, 69.0, 65.0, 63.0, 68.5, 69.0, 64.0, 67.0)


## Scatterplot
plot(x = Midparent, y = Daughter, type = "p", las = 1,
main = " scatter plot [Daughter ~  Midparent]", xlab = "Midparent Height", ylab = "Daughter Height", co
```




Konklusjon: Plottet viser at det samler seg nede i venstre og oppe i høyre hjørne. Det kan intuitivt se ut som om det er en positiv korrelasjon mellom snitt høyde for foreldre og døtres høyde.

b) Vi ser av plottet at settene ikke er injektive. For "Daughter Hight" = 69 finnes flere "Midparent Height", og for "Midparent Height" = 65,5 er to kanditater kandidater av døttre. Dette sier oss at det ikke er perfekt korrelasjon, siden det da ville vert en til en. Dette sier oss ikke at det ikke er en korrelasjon, men at det også kan være andre faktorer som kan påvirker høyden.

c) Vi skal bruke dataene som står under oppgaven til å lage en regresjonslinje av plottet [f(x)=ax+b] (der jeg for min egen upraktiske del har satt x= midparent og y=daughters) og bruke funksjonen til å predikere hvilken høyde en datter som har forreldre med snitthøyde 70 har. Vi skal også drøfte om dette er en god approksimasjon dersom snithøyden til forreldre er 74.

```
## Gjennomsnitts høyde for døttre
mean_daughter <- mean(Daughter)

## Generer linjær regresjons model av vektorene Daughter~Midparent
lm_daughters_explained_by_midparent <- lm(formula = Daughter~Midparent)


print(summary(lm_daughters_explained_by_midparent))
```

```
##
## Call:
## lm(formula = Daughter ~ Midparent)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -2.6707 -0.8429  0.4627  0.8071  2.3737
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.6497    13.3632   0.123  0.90446
## Midparent     0.9555     0.1971   4.849  0.00091 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.451 on 9 degrees of freedom
## Multiple R-squared:  0.7232, Adjusted R-squared:  0.6924
## F-statistic: 23.51 on 1 and 9 DF,  p-value: 0.00091
```

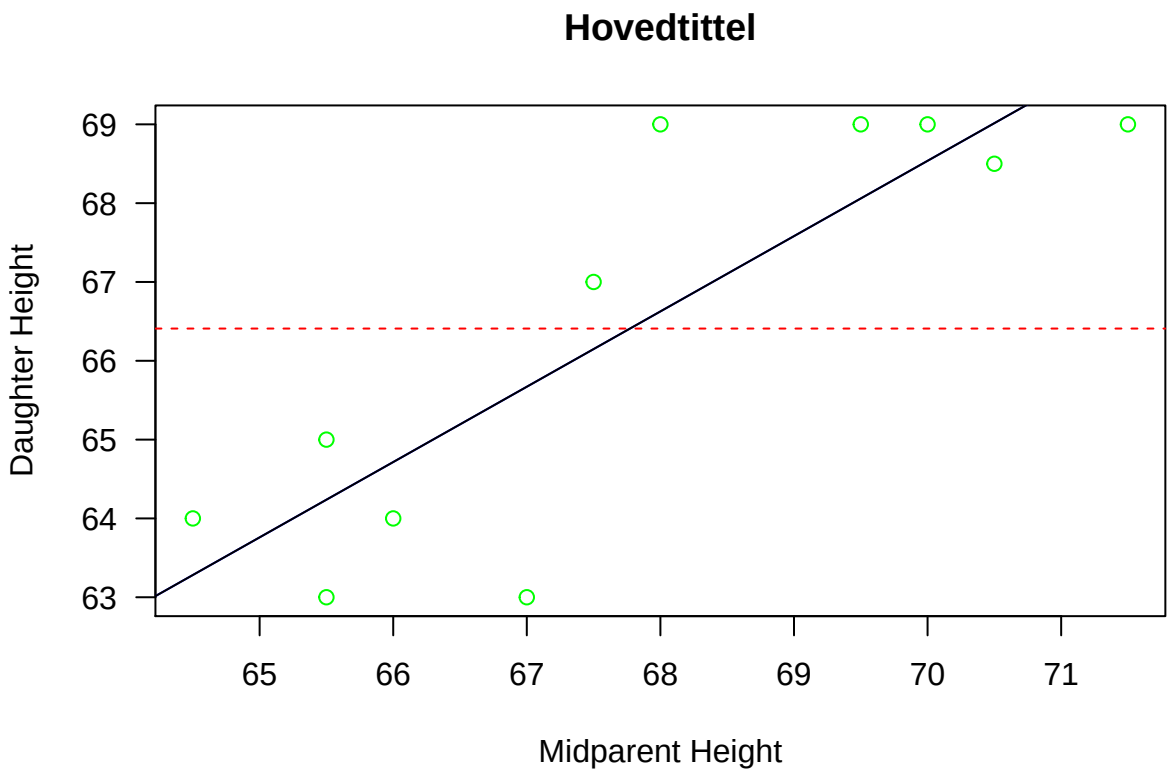
```
names(lm_daughters_explained_by_midparent)
```

```
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"
```

```
# Beregning av koefisienter til regresjonslinjen
a <- lm_daughters_explained_by_midparent$coefficients[2]
b <- lm_daughters_explained_by_midparent$coefficients[1]

# Det direkte måte å trekke ut koeefisientene a og b på
coef(lm_daughters_explained_by_midparent)
```

```
## (Intercept)   Midparent
##   1.6497483   0.9555369
```

```
# Funksjon for regresjonslinjen
regresjonslinje <- function(x){a*x+ b}


## Plott for å sammenligne målinger med modell og gjennomsnitt
plot(x = Midparent, y = Daughter, type = "p", las = 1,
main = "Hovedtittel", xlab = "Midparent Height", ylab = "Daughter Height", col = "green")
abline(h = mean_daughter, col = "red", lty =2)
abline(lm_daughters_explained_by_midparent, col = "blue")

# lagt over ser vi at regresjonslinje(x) korrelerer en til en med lm_daughters_explained
x = 64:72
y = regresjonslinje(x)
lines(x, y)
```




```
# Beregnet høyde for midparent = 70
inn_verdi_70 <- 70
verdi_70 <-regresjonslinje(inn_verdi_70)
print(verdi_70)
```

```
## Midparent
##  68.53733
```

```
# Beregnet høyde for midparent = 74
inn_verdi_74 <- 74
verdi_2 <-regresjonslinje(inn_verdi_74)
print(verdi_2)
```

```
## Midparent
##  72.35948
```

```
# Beregnet høyde for midparent = 5
inn_verdi_5 <- 5
verdi_3 <-regresjonslinje(inn_verdi_5)
print(verdi_3)
```

```
## Midparent
##  6.427433
```

Konklusjon: Dersom man bruker modellen til å estimere innenfor den rangen av snithøyder som er gitt i oppgaven gir den ett plausibelt resultat. Når vi ekstrapolerer til 74 tommer som er like utenfor chartet gir dette en høyde på datteren som er ca 72,36, hvilket også er realistisk. Dersom man har forreldre som tommelite på med snitthøyde på 5 tommer begynner det å bli verre. Da har datteren en høyde på 6,43 tommer. Litt domenekunskap tilsier at slike mennesker ikke finnes (gitt at man ikke har tatt Terry Pratchets

"The Wee Free Men" bokstavlig). Vi er her offer for "the danger of extrapolation". Det vil si at ekstrapolering i linjær regresjon bør skje innen ett rimelig spenn av modellen.

d) Vi skal her finne Sum of Squared Errors (SSE), Sum of Squared Total (SST) og coefficient of determination, bedre kjent som $r^2$.

```
## Utregning av SSE, SST og r.squared kan gjøres direkte ved å hente ut
## residualene og bruke formlene på side 631, 633 og 634.
residualer <- residuals(lm_daughters_explained_by_midparent)
SSE <- sum(residualer^2)
SST <- sum((Daughter - mean(Daughter))^2)
r.squared <- 1 - SSE/SST

cat("rsq =  ",r.squared, "   SST =  ", SST, "   SSE =  ",SSE)
```

```
## rsq =   0.7231605    SST =   68.40909    SSE =   18.93834
```

```
## Alternativt kan dere finne SSE ved å se på resultatet fra `anova` og
## dere kan finne r.squared ved å se på resultatet fra `summary`.
summary <- summary(object = lm_daughters_explained_by_midparent)
anova <- anova(object = lm_daughters_explained_by_midparent)
```

Konklusjon: Sum of Squared Errors (SSE) er summen av avvik mellom reel høyde på, i dette tilfellet, døttres høyde og estimert høyde i regresjonslinjen kvadrert. Sum of Squared Total (SST) er samme prinsipp, men spenner mellom målt høyde og middelverdi, også dette kvadrert for å ungå at de forskjellige målingene kanselerer segt selv ut. $r^2$ er ens sammenheng som kommer ut av disse to komponentene $r^2 = \frac{SST-SSE}{SST} = 1 - \frac{SSE}{SST}$. Mer intuitivt kan det forklares som forskjellen mellom varians rundt regresjonslinjen og varians rund middelverdi relativ til varians til middelverdi. Dette forklarer oss hvor mye av variansen som kan forklares av korellasjon og hvor mye som er tilfeldig. I dette tilfellet er $r^2 \approx 68,41$, sammenheng mellom høyden på døttre og forreldres snitt høyde.