Aflevering 8

Lucas Bagge

3.5

I denne opgave skal I bruge en lineær regressionsmodel til at sige noget om værdien af den forklarende variabel ud fra en målt responsværdi. I opgaven her er den forklarende variabel alderen af en løve, og respons er fraktion af sort pigment i løvens næsetip.

I artiklen Sustainable trophy hunting of African lions diskuteres hvordan trofæjagt af løver kan gøres bæredygtigt. Forfatternes konklusion er, at man skal sørge for, at de løver, der jages, er hanløver over en vis alder. Ofte bruger jægeren størelsen og farven af løvens manke til at vurdere alderen, men dette er en meget usikker metode. En mere sikker metode består i at bruge andelen af sort pigment i løvens næsetip. I opgaven her skal I se på, hvordan andelen af sort pigment afhænger af alderen, og hvor godt vi kan estimere alderen ud fra dette.

Data for 32 hanløver fra Serengeti og Ngorongoro nationalparkerne ligger i filen Loeve.csv, der har to søjler med henholdsvis alder (år) og fraktion af sort.

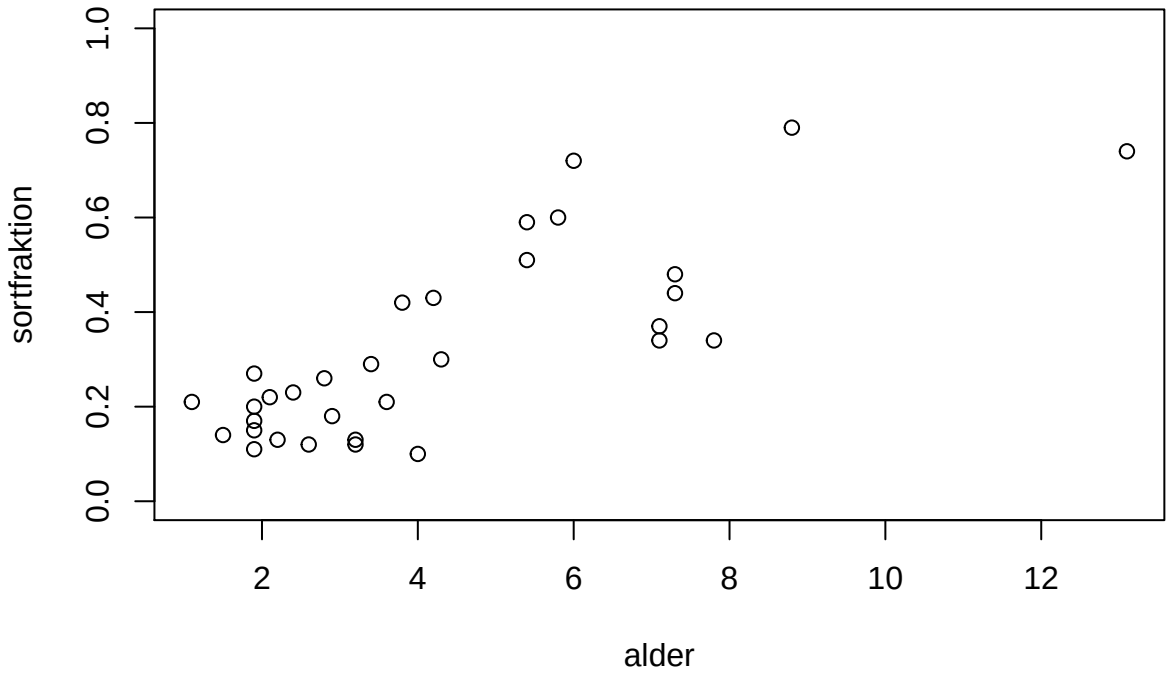
Indlæs data:

```
data <- read.csv("MatStat-R/data/JLJfiler/Loeve.csv") %>%
  janitor::clean_names()
head(data)
```

```
##   alder sortfraktion
## 1   1.1         0.21
## 2   1.5         0.14
## 3   1.9         0.11
## 4   2.2         0.13
## 5   2.6         0.12
## 6   3.2         0.13
```

a) Lav en figur, hvor fraktion af sort afsættes mod alder. Styr start og slut på andenaksen med tilføjelsen ylim=c(0,1) til plot. Synes du, at der er en lineær sammenhæng i data? Synes du, at sammenhængen er god med henblik på at estimere alder ud fra fraktion af sort i næstippen?
Jeg laver et scatter plot.

```
attach(data)
plot(alder, sortfraktion, ylim = c(0,1))
```



Hvis vi ser fra vores bog *Mathematical statistic with resampling and R* på side 39, så ser der ud til at være en positiv lineær sammenhæng og vi kan dermed godt estimere alder ud fra fraktion af sort i næsetippen.

b) Opskriv den lineære regressionsmodel, hvor respons er fraktion af sort i næsetippen, og den forklarende variabel er alder. Find skøn og 95%-konfidensinterval for hældning og skæring, og indtegn den skønnede linje i figuren fra foregående spørgsmål. Angiv også et skøn over spredningen σ omkring den lineære sammenhæng. Lav figurer, der kan bruges til modelkontrol, og kommenter på disse figurer. Den lineære regressions model vil være følgende:

$$sortfraktion \sim N(\alpha + \beta \cdot alder_i)$$

Implementerin i R:

```
model <- lm(sortfraktion ~ alder, data = data)
sumUD <- summary(model)
sumUD
```

```
##
## Call:
## lm(formula = sortfraktion ~ alder, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.20406 -0.07758 -0.01750  0.07913  0.29876
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.069696   0.041956   1.661    0.107
## alder       0.058591   0.008307   7.053 7.68e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1238 on 30 degrees of freedom
## Multiple R-squared:  0.6238, Adjusted R-squared:  0.6113
## F-statistic: 49.75 on 1 and 30 DF,  p-value: 7.677e-08
```

Hvor skøn for skæring er: 0.069696 mens den for hældning er 0.5891.

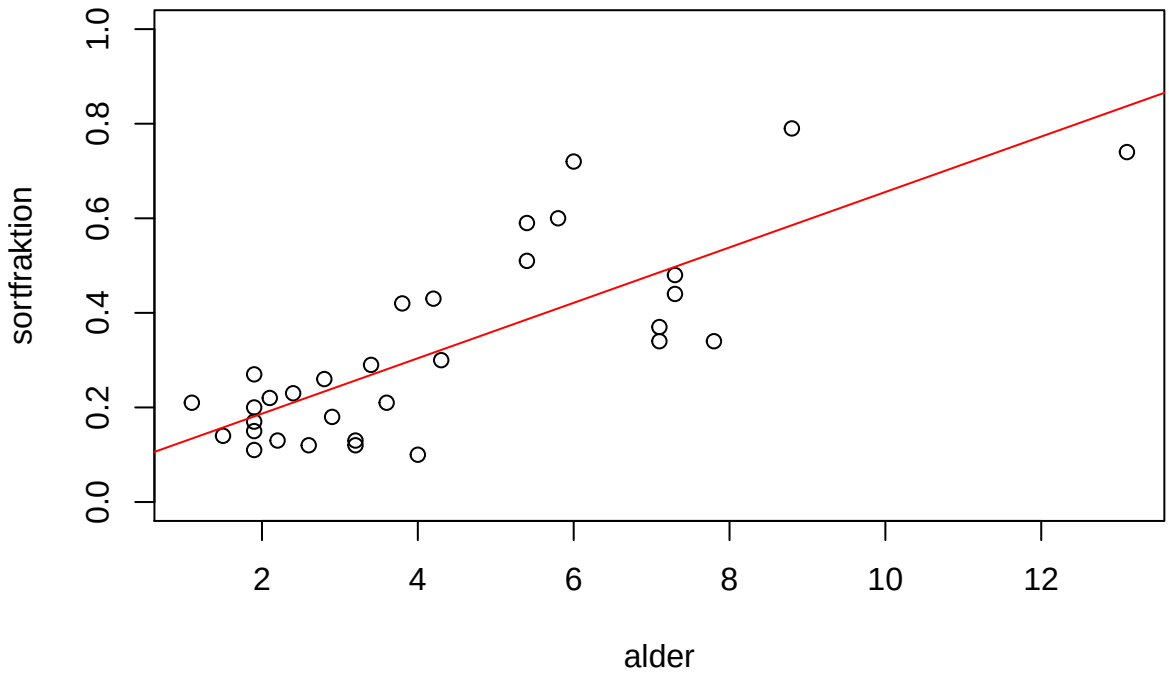
Vi kan burge `confint` til at finde konfidens intervallet.

```
confint(model)
```

```
##                   2.5 %     97.5 %
## (Intercept) -0.01598977 0.15538230
## alder        0.04162643 0.07555588
```

Herefter skal vi tegne vores skøn ind:

```
plot(alder, sortfraktion, ylim = c(0,1))
abline(model, col = "red")
```



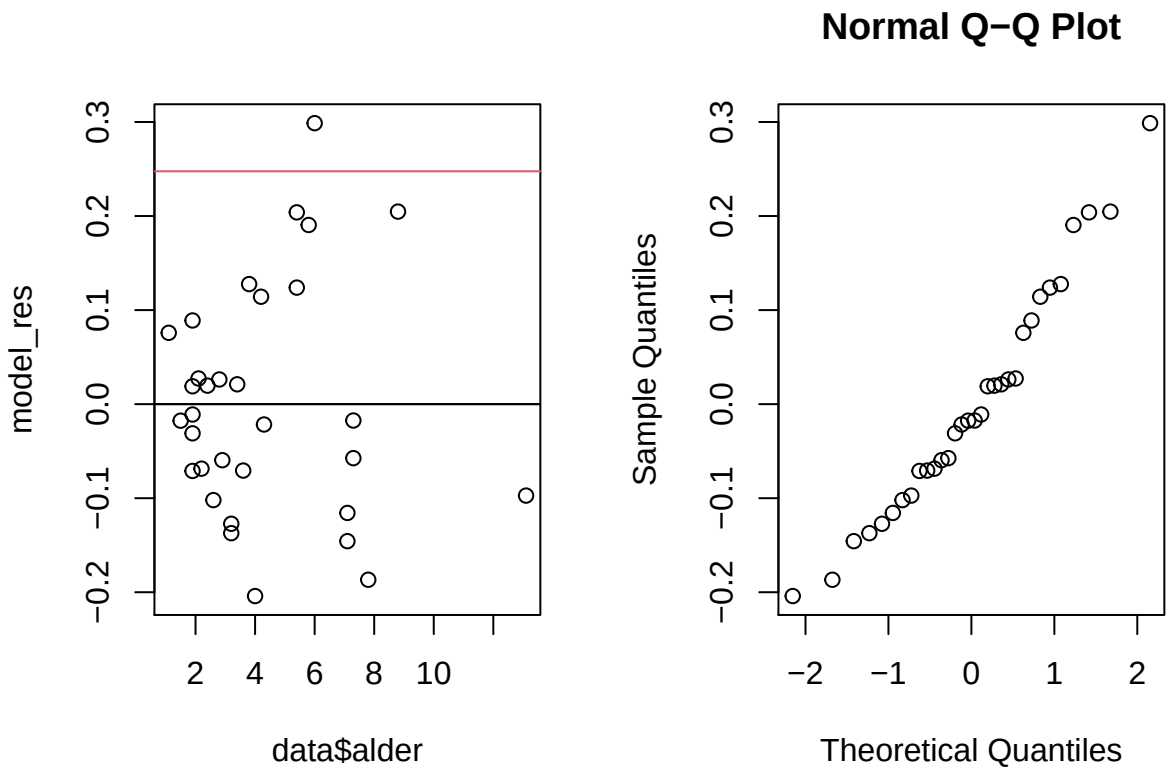
Så skal vi angive skønnet over spredning, s_r:

```
sumUD$sigma
```

```
## [1] 0.1237927
```

Så for modelkontrol

```
model_res <- resid(model)
sigma <- summary(model)$sigma
par(mfrow=c(1,2))
plot(data$alder, model_res)
abline(0,0)
abline(h = 2 * sigma, col = 2)
abline(h = -(2 * sigma), col = 2)
abline(0,0)
qqnorm(model_res)
```



Residualplotter tyder ikke på afvigelser fra en lineære sammenhæng. Kun vi en outlier måske med den ydersste observation. QQplottet giver ikke anledning til bekymring med hensyn til normalfordelingsantagelsen.

c) Betragt situationen, hvor en ny løve registreres, og fraktion af sort i næsen for denne løve er 0.2. Beregn et 95%-konfidensinterval for løvens alder i dette tilfælde. Gentag beregningen i tre andre tilfælde, hvor en løve er observeret med henholdsvis 0.4, 0.6 og 0.8 for fraktionen af sort i næsen. Lav en tabel med resultaterne for de fire tilfælde. Hvis vi kun ønsker at skyde løver, der er mindst 5 år gamle, hvor stor synes du så fraktionen af sort i næsen skal være, før

du skyder? Nu lader vi som om vi observer en ny løve der har en sort fraktion på 0.2. Hertil laver jeg et nyt datasæt.

```
new_data <- data.frame(
  sortfraktion = 0.2
)
```

Nu skal vi altså forudsige hvad alderen på den nye løve er når den har en sortfraktion på 0.2. Hertil laver vi en ny model hvor vi regresser alder på sortfraktion og bruger R prediction til at lave denne analyse:

```
source("MatStat-R/source/Rfunktioner.R")
inversReg(model, new_data$sortfraktion)
```

```
## $estimat
## [1] 2.223949
##
## $konfidensinterval
## [1] -2.592041  6.658276
```

Her ser vi at løven burde være omkring 2 år indenfor -2.59 og 6.65.

Nu får vi yderligere tre nye løver hvor vi skal lave samme analyse. Her kobler jeg bare alle fire på, da vi bliver bedt om at lave en tabel:

```
new_data <- data.frame(
  sortfraktion = c(0.2, 0.4, 0.6, 0.8)
)
inversReg(model, 0.4)
```

```
## $estimat
## [1] 5.637434
##
## $konfidensinterval
## [1]  1.161802 10.356120
```

```
inversReg(model, 0.6)
```

```
## $estimat
## [1] 9.050919
##
## $konfidensinterval
## [1]  4.667829 14.301781
```

```
inversReg(model, 0.8)
```

```
## $estimat
## [1] 12.4644
##
## $konfidensinterval
## [1]  7.957056 18.464243
```

Her til sidst skal vi så svarer på hvad fraktionen af sort skal være før vi vil skyde en løve på baggrund af den er mindst 5 år. Ud fra vores prediction så skal skal en løve have omkring 0.4 fraktion så vil den være omkring 5 år.

Kigger vi dog på selve data:

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
data %>%
  filter(alder > 5.0)
```

```
##    alder sortfraktion
## 1    5.4         0.59
## 2    5.8         0.60
## 3    6.0         0.72
## 4    7.3         0.48
## 5    7.3         0.44
## 6    7.8         0.34
## 7    7.1         0.37
## 8    7.1         0.34
## 9   13.1         0.74
## 10   8.8         0.79
## 11   5.4         0.51
```

Foroven laver jeg et filter på mit data og ser at mange løver som er omkring 5 til 6 år har en sortfraktion på 0.5 til 0.7, som giver god mening ud fra vores model prediction, men ser vi på alderen 7, så har den en sortfraktion på under 0.5, og det er misvisende for så tyder det på at vi har nogle løver som er gammel men ikke har meget fraktion på næsen.

Derfor er vores model ikke særlig god og jeg tør ikke udtale mig om hvad fraktionen skal være for vi vil skyde en løve som er mindst 5 år.