Ass_2

Christoffersen J. & Haugland V.

Oktober 14. 2021

Del 1 - Innledning

Kort innledning

Kort litteraturgjennomgang på ca. 1 side

```
library(modelr)
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
```

```
library(ggplot2)
```

```
# Endrer navn på datasettet fra heights til hoyde
data("heights", package = "modelr")
hoyde <- heights
```

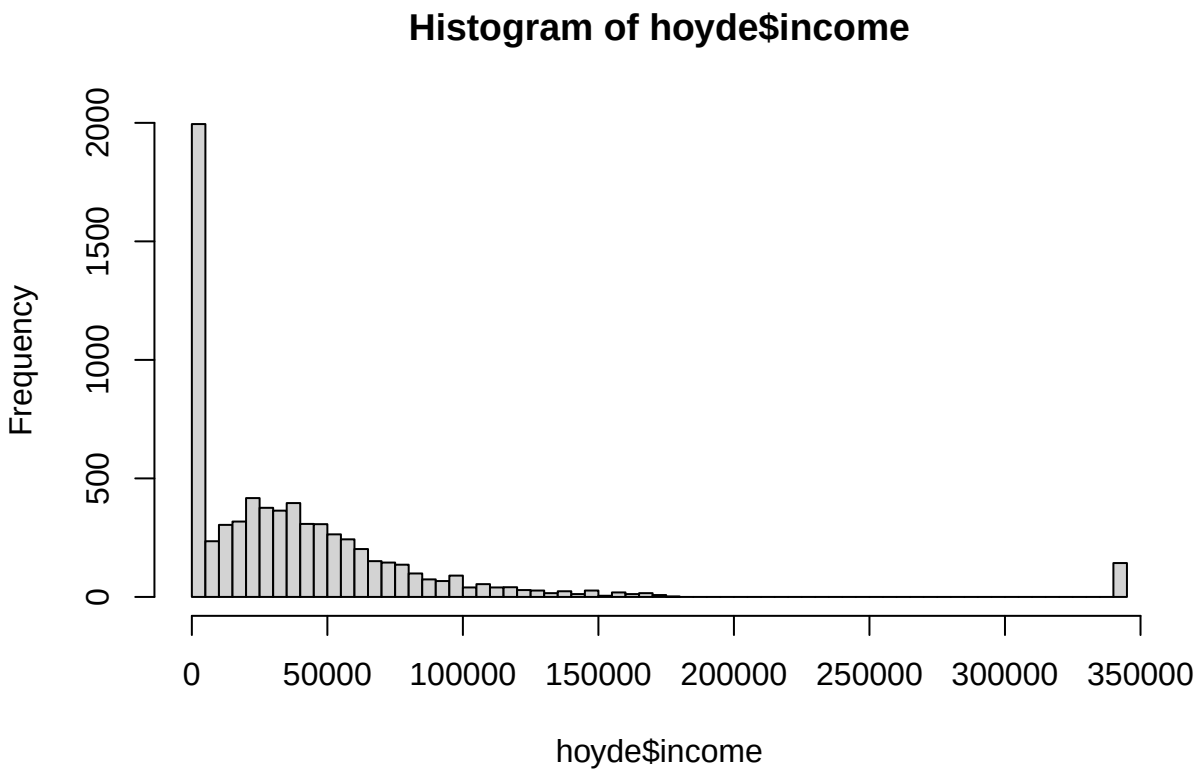
Beskrivende statistikk

I datasettet hoyde har vi $n = 7006$, som vil si at settet består av totalt 7006 ulike observasjoner. Videre har vi de 7 uavhengige variabelene *høyde*, *vekt*, *alder*, *sivilstatus*, *kjønn*, *utdannelse* og *afqt* (Score fra forvarets kvalifiseringstest), som bestemmer den avhengige variablen *inntekt* i dette settet. Hvorav variablene *sivilstatus* og *kjønn* er såkalte dummyvariabler. *Sivilstatus* varier med fem ulike faktorer som er *singel*, *gift*, *separert*, *skilt* og *enke* og dummyvariablen *kjønn* varier mellom tallet 1 og 2 som bestemmer om individet som er observert er enten mann eller kvinne.

EDA

- **Histogram**

```
hist(hoyde$income, breaks = 100)
```



- **Forklaring på utliggerne langt til høyre**

For å forklare utliggerne langt til høyere, kan vi begynne med å finne ut hva maksimumsverdien av variablen *income* er.

```
summary(hoyde)
```

```
##      income            height         weight           age
##  Min.   :     0.0   Min.   :52.0   Min.   : 76.0   Min.   :47.00
##  1st Qu.:   165.5   1st Qu.:64.0   1st Qu.:157.0   1st Qu.:49.00
##  Median : 29589.5   Median :67.0   Median :184.0   Median :51.00
##  Mean   : 41203.9   Mean   :67.1   Mean   :188.3   Mean   :51.33
##  3rd Qu.: 55000.0   3rd Qu.:70.0   3rd Qu.:212.0   3rd Qu.:53.00
##  Max.   :343830.0   Max.   :84.0   Max.   :524.0   Max.   :56.00
##                                    NA's   :95
##       marital         sex         education          afqt
##  single   :1124   male  :3402   Min.   : 1.00   Min.   :  0.00
##  married  :3806   female:3604   1st Qu.:12.00   1st Qu.: 15.12
```

```
##  separated: 366                Median :12.00   Median : 36.76
##  divorced :1549                Mean   :13.22   Mean   : 41.21
##  widowed  : 161                3rd Qu.:15.00   3rd Qu.: 65.24
##                                Max.   :20.00   Max.   :100.00
##                                NA's   :10      NA's   :262
```

Fra summeringstabellen ser vi at maksimumsverdien til variablen *income* er 343,830. Videre kan vi undersøke hvor mange av de observerte som innehar denne verdien.

```
sum(hoyde$income == 343830)
```

```
## [1] 143
```

Totalt var det 7006 observasjoner og med hensyn til at 143 hadde maksimumsverdien og at det er en relativt stor avstand mellom maksinntekten og medianen til datasettet, kan vi si at utliggerne til høyere representerer ca. 2% av det totale utvalget.

```
(143/7006)*100
```

```
## [1] 2.041108
```

- **Har vi med personer uten inntekt i datasettet?**

Ja. Datasettet inneholder observasjoner av en rekke individer uten inntekt. Dette kan vi enkelt observere ved å se på histogrammet på 0 langs x-aksen, hvor vi ser at frekvensen på denne verdien er < 1500. Vi kan også benytte summeringstabellen fra tidligere og se på minimumsverdien til variablen *income* som er på 0. Dersom vi ønsker å finne ut hvor mange av de observerte som innehar denne verdien, kan vi benytte følgende:

```
sum(hoyde$income == 0)
```

```
## [1] 1740
```

Del 2 - Regresjonsanalyse

Redusere datasettet

Med henhold til at vi har funnet ut at datasettet inneholder mange minimumsverdier på 0 og at maksimumsverdien på 343,830 utgjorde kun 2% av utvalget, skal vi fjerne disse verdiene for å kunne kjøre en endelig modell mot reduserte datasett. Dette gjør vi for å teste modellens robushet.

```
sample1=filter(hoyde,income!=0 & income!=343830)
```

Her har vi opprettet en ny model, under navnet *sample1*. For å sjekke om de riktige verdiene ble fjernet kan vi kjøre en ny summeringstabell.

```
summary(sample1)
```

```
##      income           height          weight           age
##  Min.   :    45   Min.   :52.00   Min.   : 78.0   Min.   :47.00
##  1st Qu.: 23000   1st Qu.:64.00   1st Qu.:159.0   1st Qu.:49.00
##  Median : 40000   Median :67.00   Median :185.0   Median :51.00
##  Mean   : 46751   Mean   :67.22   Mean   :188.4   Mean   :51.28
##  3rd Qu.: 62000   3rd Qu.:70.00   3rd Qu.:212.0   3rd Qu.:53.00
##  Max.   :178000   Max.   :80.00   Max.   :480.0   Max.   :56.00
##                                   NA's   :69
##       marital         sex         education          afqt
##  single   : 699   male  :2526   Min.   : 1.00   Min.   :  0.00
##  married  :2983   female:2597   1st Qu.:12.00   1st Qu.: 19.55
##  separated: 233                 Median :12.00   Median : 41.71
##  divorced :1102                 Mean   :13.48   Mean   : 44.40
##  widowed  : 106                 3rd Qu.:16.00   3rd Qu.: 67.89
##                                 Max.   :20.00   Max.   :100.00
##                                 NA's   :2       NA's   :184
```

Her kan vi se at både minimums- og makismumsverdien har endret seg. Det nye datasettet inneholder også færre observasjoner:

$$7006 - 143 - 1740 = 5123$$

Ny $n = 5123$ til datasettet *sample1*.

Mutering av eksisterende variabler fra imperial- til metric system

Siden vi benytter oss av det metriske systemet i Norge og datasettet har brukt det imperiske, så ønsker vi å endre disse variablene. Variablene dette omhandler er *height* og *weight* som vi skal gjøre om til *height_cm* og *weight_kg*. Har funnet ut på nettet at en inch er 2.54 cm og en pund er ca. 0.45 kg.

```
hoyde <- hoyde %>%
  mutate(height_cm = height * 2.54)
```

- **Vi kan teste om muteringen av *height* fungerer ved å:**

```
# Finne maksverdien av den nye variablen
max(hoyde$height_cm)
```

```
## [1] 213.36
```

```
# Vi kan sjekke at dette stemmer ved å se hva den høyeste i inches er
max(hoyde$height)
```

```
## [1] 84
```

```
# Og så gange med 2.54 som var overgangen fra inches til cm
84*2.54
```

```
## [1] 213.36
```

```r
# Bruker samme fremgangsmåte for pund til kg
hoyde <- hoyde %>%
  mutate(weight_kg = weight *0.45)
```

Fra summeringstabellen til datasettet *hoyde* kan vi se at variablen *weight* har flere NA-verdier, altså verdier som ikke er tilgjengelige. Dersom man ønsker å finne ut hva for eksempel maksimumsverdien til *weight_kg* er må man skrive følgende:

```r
max(hoyde$weight_kg, na.rm = TRUE)
```

```
## [1] 235.8
```

Vi putter inn $na.rm = TRUE$ slik at R utelater cellene med betegnelsen *NA* (Not available). Dette gjør at vi kun får cellene med tallverdier.

Dersom man ønsker å finne ut om andre variabler også innehar verdier som er NA, kan vi skrive:

```r
# For å sjekke hvilke celler som har NA
colSums(is.na(hoyde))
```

```
##    income    height    weight       age   marital       sex education      afqt
##         0         0        95         0         0         0        10       262
## height_cm weight_kg
##         0        95
```

Opprettelse av ny variabel BMI

```r
hoyde <- hoyde %>%
  mutate(hoyde, bmi = weight_kg/(height_cm)^2)
```

Forenklet utgave av variablen martial

Siden dummyvariabel *marital* innholdt hele 5 ulike faktorer, så ønsker vi å forenkle denne til enten "True" eller "False" for om individet er gift eller ikke-gift.

```r
hoyde <- hoyde %>%
  mutate(
    married = factor(
      case_when(
        marital == "married" ~ TRUE,
        TRUE ~ FALSE
      )
    )
  )
```

Del 3 - Estimere modeller