# Ass_2

Christoffersen J. & Haugland V.

Oktober 14. 2021

## Del 1 - Innledning

Bare en liten test.

### 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)
library(tinytex)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v tibble  3.1.3     v purrr   0.3.4
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggpubr)
library(huxtable)
```

```
##
## Attaching package: 'huxtable'

## The following object is masked from 'package:ggpubr':
##
##     font

## The following object is masked from 'package:ggplot2':
##
##     theme_grey

## The following object is masked from 'package:dplyr':
##
##     add_rownames
```

```
options(scipen = 999)
```

## Analyse

Vi lager først et histogram med inntekts-variabel.

```
# 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.

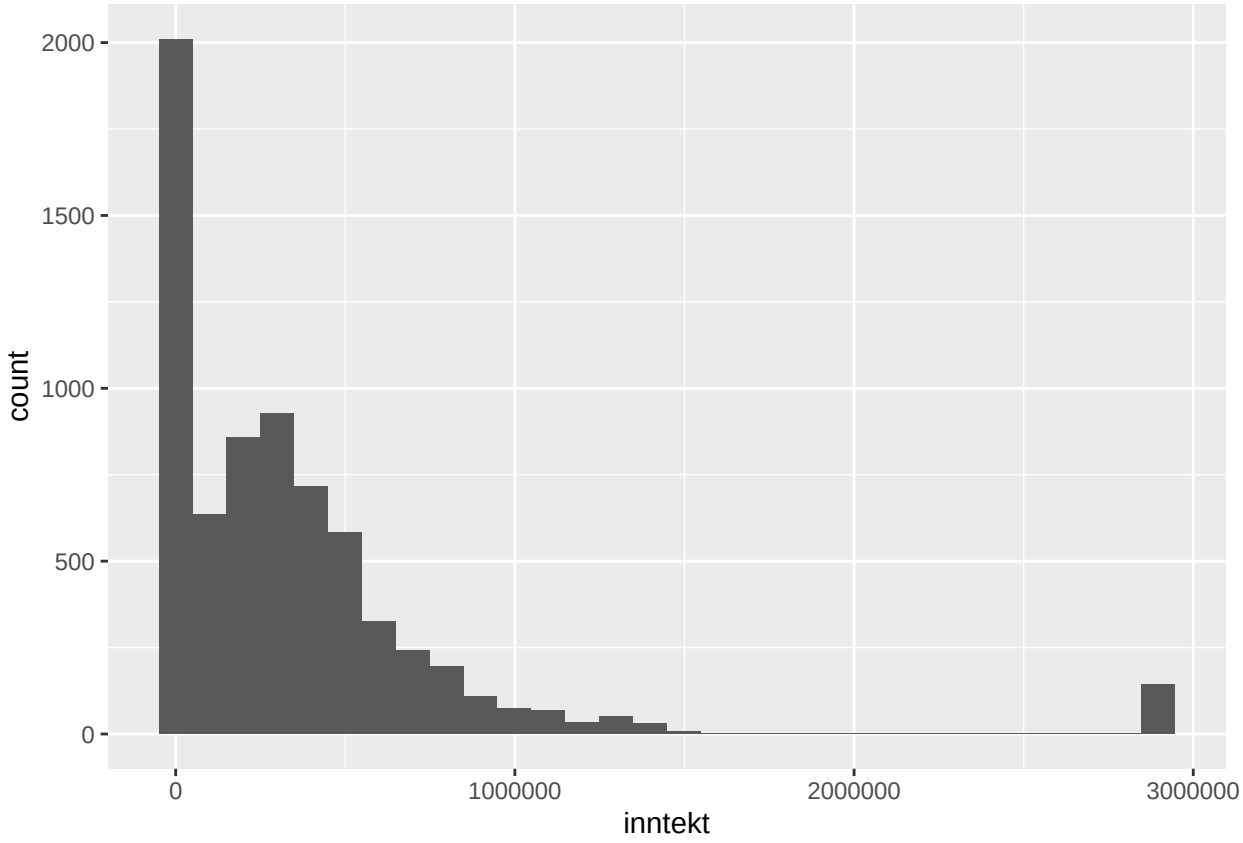
## EDA

- **Histogram til variabelen innekt** Til å begynne med lages tre nye variabler som settes til en metrisk standard.

```
hoyde <- hoyde %>%
  mutate(inntekt = income * 8.42,
         hoyde_cm = height * 2.54,
         vekt_kg = weight * 0.454,
         BMI = vekt_kg / (hoyde_cm/100)^2)
```

```
ggplot(data = hoyde, aes(x = inntekt)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
  geom_histogram(bins = 30)
```

```
## geom_bar: na.rm = FALSE, orientation = NA
## stat_bin: binwidth = NULL, bins = 30, na.rm = FALSE, orientation = NA, pad = FALSE
## position_stack
```

- **Forklaring på utliggerne langt til høyre**

Utliggeren til høyre viser gjennomsnittet av 2% topp inntekt. For å forklare dette bedre kan vi begynne med å finne maksimumsverdien av variablen *inntekt*.

```
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
##     inntekt           hoyde_cm        vekt_kg            BMI
##  Min.   :      0   Min.   :132.1   Min.   : 34.50   Min.   :12.89
##  1st Qu.:   1394   1st Qu.:162.6   1st Qu.: 71.28   1st Qu.:25.11
##  Median : 249144   Median :170.2   Median : 83.54   Median :28.35
##  Mean   : 346937   Mean   :170.4   Mean   : 85.49   Mean   :29.33
##  3rd Qu.: 463100   3rd Qu.:177.8   3rd Qu.: 96.25   3rd Qu.:32.31
##  Max.   :2895049   Max.   :213.4   Max.   :237.90   Max.   :75.06
##                                    NA's   :95       NA's   :95
```

Fra summeringstabellen ser vi at maksimumsverdien til variablen *inntekt* er 343,830. Videre kan vi undersøke hvor mange av observasjonene 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 et gjennomsnitt av ca. 2% av det totale utvalget.

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

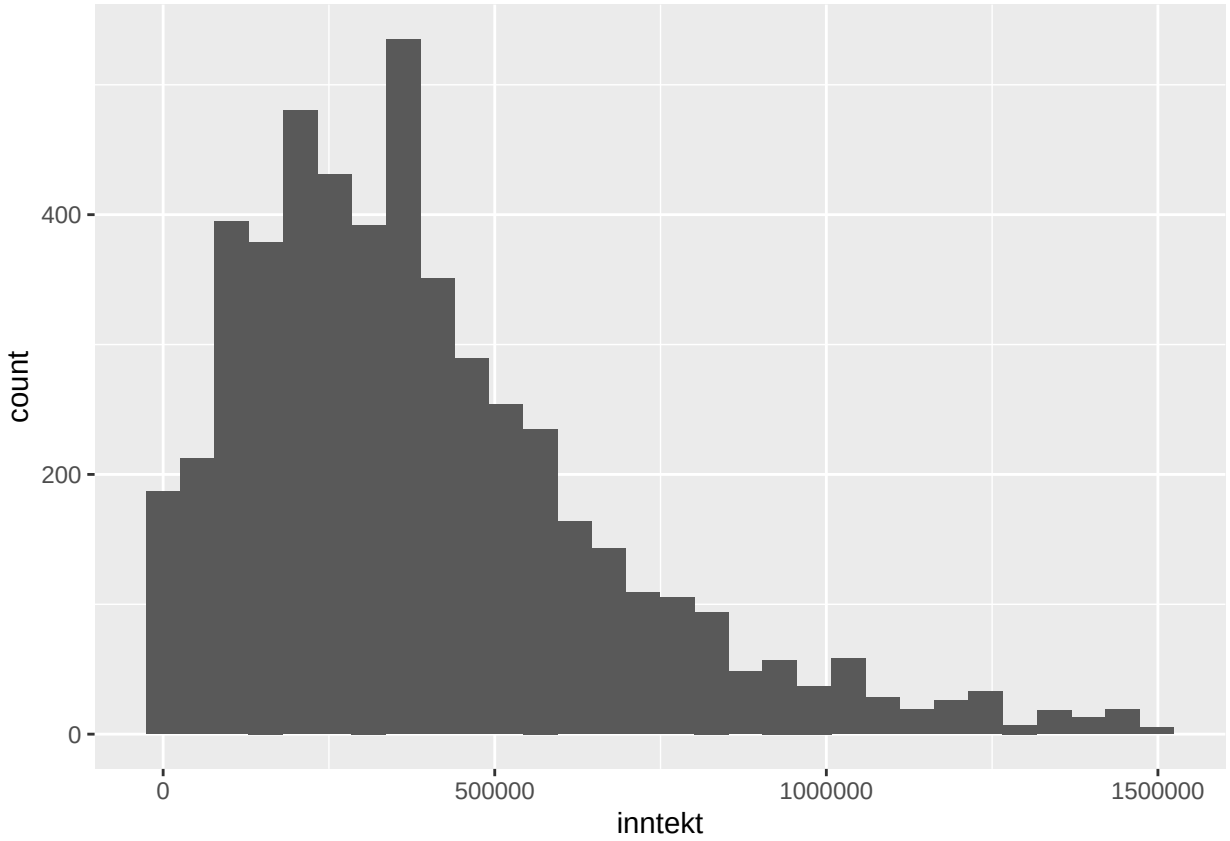
```
## [1] 2.041108
```

Dersom datasettet reduseres for topp 2% høyeste inntekt og inntekt lik 0, ser histogrammet slik ut:

```
hoyde_begr <- hoyde %>%
  filter(inntekt < 1500000,
         inntekt > 1)
```

```
ggplot(data = hoyde_begr, aes(x = inntekt)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




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

Ja. Datasettet inneholder observasjoner av en rekke individer uten inntekt. Dette fremkommer av histogrammet til variabelen inntekt hvor $< 1500$ er frekvensen til 0 langs x-aksen. Det fremkommer også av hoyde-summeringstabellen at minimumsverdien til inntekt-variabelen er 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

```
mod1 <- "inntekt ~ hoyde_cm"
lm1 <- lm(mod1, data = hoyde, subset = complete.cases(hoyde))
```

```
summary (lm1)
```

```
##
## Call:
## lm(formula = mod1, data = hoyde, subset = complete.cases(hoyde))
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -782810 -267359  -94513  123099 2699234
##
## Coefficients:
##               Estimate Std. Error t value            Pr(>|t|)
## (Intercept) -1361001.0    94430.0  -14.41 <0.0000000000000002 ***
## hoyde_cm       10047.9      552.8   18.18 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 467300 on 6643 degrees of freedom
## Multiple R-squared:  0.04737,    Adjusted R-squared:  0.04723
## F-statistic: 330.3 on 1 and 6643 DF,  p-value: < 0.00000000000000022
```

Her ser man at inntekt øker med 10047.9 kr dersom høyde øker med 1 cm. Test:

```
-1361001.0 + (10047.9 * 173)
```

```
## [1] 377285.7
```

```
-1361001.0 + (10047.9 * 161)
```

```
## [1] 256710.9
```

```
mod2 <- "inntekt ~ hoyde_cm + vekt_kg"
lm2 <- lm(mod2, data = hoyde, subset = complete.cases(hoyde))
```

```
summary(lm2)
```

```
##
## Call:
## lm(formula = mod2, data = hoyde, subset = complete.cases(hoyde))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -843668 -263322  -92573  125798 2715000
##
## Coefficients:
##               Estimate Std. Error t value             Pr(>|t|)
## (Intercept) -1466873.6    96890.5 -15.139 < 0.0000000000000002 ***
## hoyde_cm       11430.3      624.3  18.308 < 0.0000000000000002 ***
## vekt_kg        -1518.4      320.5  -4.737           0.00000221 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 466600 on 6642 degrees of freedom
## Multiple R-squared:  0.05058,    Adjusted R-squared:  0.05029
## F-statistic: 176.9 on 2 and 6642 DF,  p-value: < 0.00000000000000022
```

Fra regresjonsanalysen over ser man inntekt øker ved økt høyde, men synker ved økning i vekt. Dersom høyde økes med 1 cm og vekt samtidig øker med 1 kg, vil også inntekten øke. Dette skyldes at inntekten øker betydelig mer ved økning i høyde, enn den syker ved tilsvarende økning i vekt.

```
mod3 <- "inntekt ~ hoyde_cm + vekt_kg + BMI"
lm3 <- lm(mod3, data = hoyde, subset = complete.cases(hoyde))
```

```
summary(lm3)
```

```
##
## Call:
## lm(formula = mod3, data = hoyde, subset = complete.cases(hoyde))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -886295 -261634  -93597  124905 2709981
##
## Coefficients:
##             Estimate Std. Error t value     Pr(>|t|)
## (Intercept) -2015890     447005  -4.510 0.0000066012 ***
## hoyde_cm       14669       2649   5.537 0.0000000319 ***
## vekt_kg        -4723       2567  -1.840       0.0658 .
## BMI             9224       7332   1.258       0.2084
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 466600 on 6641 degrees of freedom
## Multiple R-squared:  0.05081,    Adjusted R-squared:  0.05038
## F-statistic: 118.5 on 3 and 6641 DF,  p-value: < 0.00000000000000022
```

Videre fremkommer det at en kombinasjon av økning i høyde, vekt og BMI, også vil gi en økning i innekt.

## Forenklet utgave av variablen martial

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

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

## Resultat fra estimering rapporteres vha. huxreg

```
huxreg(list("mod1" = lm1, "mod2" = lm2, "mod3" = lm3),
       error_format = "[{statistic}]",
       note = "Regresjonstabell 3: {stars}. T statistics in brackets."
       )
```

| | mod1 | mod2 | mod3 |
|---|---|---|---|
| (Intercept) | -1361000.990 *** | -1466873.555 *** | -2015889.845 *** |
| | [-14.413] | [-15.139] | [-4.510] |
| hoyde_cm | 10047.860 *** | 11430.259 *** | 14669.413 *** |
| | [18.175] | [18.308] | [5.537] |
| vekt_kg | | -1518.381 *** | -4722.577 |
| | | [-4.737] | [-1.840] |
| BMI | | | 9224.408 |
| | | | [1.258] |
| N | 6645 | 6645 | 6645 |
| R2 | 0.047 | 0.051 | 0.051 |
| logLik | -96177.211 | -96166.004 | -96165.212 |
| AIC | 192360.423 | 192340.008 | 192340.424 |

Regresjonstabell 3: *** p < 0.001; ** p < 0.01; * p < 0.05. T statistics in brackets.

## Modell med interaksjonsvariabel

```
mod4 <- "inntekt ~ sex*hoyde_cm + vekt_kg + I(vekt_kg^2) + BMI + I(BMI^2)"
lm4 <- lm(mod4, data = hoyde)
summary(lm4)
```

```
##
## Call:
## lm(formula = mod4, data = hoyde)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -786022 -247378  -90398  126933 2685039
##
## Coefficients:
##                      Estimate Std. Error t value  Pr(>|t|)
## (Intercept)          23856.53 1214202.31   0.020  0.984325
## sexfemale          1018826.81  266432.35   3.824  0.000132 ***
```

```
## hoyde_cm              1982.64    7194.15   0.276  0.782871
## vekt_kg              18018.56   12853.89   1.402  0.161020
## I(vekt_kg^2)           -61.12      32.37  -1.888  0.059068 .
## BMI                 -47099.77   36660.01  -1.285  0.198915
## I(BMI^2)              369.41     268.16   1.378  0.168384
## sexfemale:hoyde_cm   -6640.80    1562.43  -4.250 0.0000216 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 458400 on 6903 degrees of freedom
##   (95 observations deleted due to missingness)
## Multiple R-squared:  0.06105,    Adjusted R-squared:  0.0601
## F-statistic: 64.12 on 7 and 6903 DF,  p-value: < 0.00000000000000022
```

### Modell med flere variasjonsvariabler

```
mod5 <- "inntekt ~ sex*(hoyde_cm + vekt_kg + I(vekt_kg^2)) + BMI + I(BMI^2)"
lm5 <- lm(mod5, data = hoyde)
summary(lm5)
```

```
##
## Call:
## lm(formula = mod5, data = hoyde)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -864444 -245100  -91019  126362 2681172
##
## Coefficients:
##                            Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)            -2821666.91  1904365.52  -1.482   0.13847
## sexfemale               1181398.44   293082.63   4.031 0.0000562 ***
## hoyde_cm                  17091.78    10627.73   1.608   0.10783
## vekt_kg                   -4749.34    17977.28  -0.264   0.79164
## I(vekt_kg^2)                -17.95       42.26  -0.425   0.67109
## BMI                       34177.41    57584.98   0.594   0.55286
## I(BMI^2)                   -190.52      435.11  -0.438   0.66150
## sexfemale:hoyde_cm        -4729.20     1812.91  -2.609   0.00911 **
## sexfemale:vekt_kg         -9825.85     5200.88  -1.889   0.05890 .
## sexfemale:I(vekt_kg^2)       45.96       27.06   1.699   0.08941 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 458300 on 6901 degrees of freedom
##   (95 observations deleted due to missingness)
## Multiple R-squared:  0.06165,    Adjusted R-squared:  0.06043
## F-statistic: 50.38 on 9 and 6901 DF,  p-value: < 0.00000000000000022
```

### Tester modellen med redusert datasett

```
mod5 <- "inntekt ~ hoyde_cm + vekt_kg + BMI"
lm5 <- lm(mod3, data = hoyde_begr, subset = complete.cases(hoyde))
summary(lm5)
```

```
##
## Call:
## lm(formula = mod3, data = hoyde_begr, subset = complete.cases(hoyde))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -537793 -189174  -56803  135653 1139990
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -343501     328879  -1.044   0.2963
## hoyde_cm        4681       1940   2.413   0.0159 *
## vekt_kg         1393       1879   0.741   0.4586
## BMI            -6193       5430  -1.141   0.2541
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 274100 on 4800 degrees of freedom
##   (1841 observations deleted due to missingness)
## Multiple R-squared:  0.05215,    Adjusted R-squared:  0.05156
## F-statistic: 88.03 on 3 and 4800 DF,  p-value: < 0.00000000000000022
```

### Legger til residualer

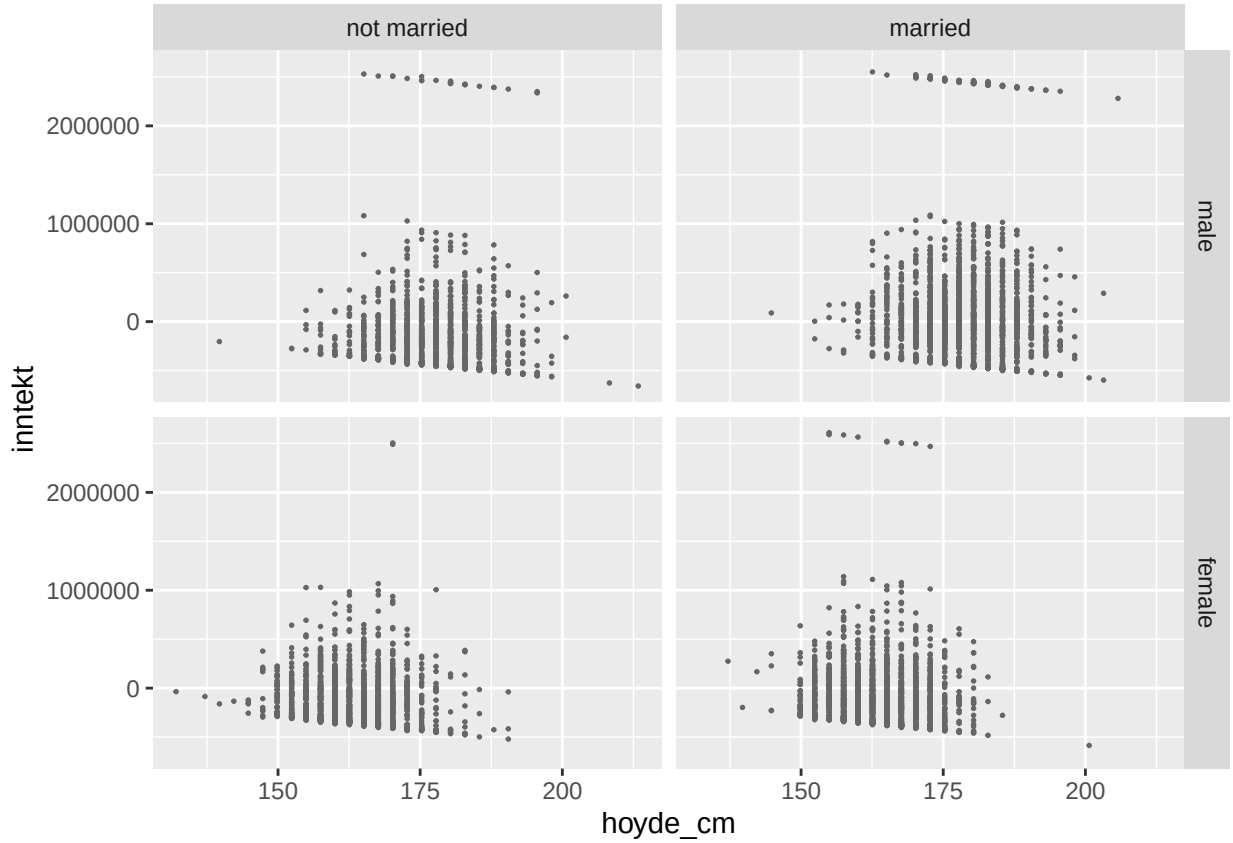
```
hoyde_begr <- hoyde %>%
  add_residuals(lm5)
hoyde_begr %>%
  head(n=10)
```

### Samtlige observasjoner

```
ggplot(data = hoyde_begr, mapping = aes(x = hoyde_cm, y = inntekt)) +
  geom_point(
    data = hoyde_begr,
    mapping = aes(x = hoyde_cm, y = resid),
    colour = "grey40",
    size = 0.3
  ) +
  facet_grid(sex ~ factor(married, labels = c("not married", "married")))
```

```
## Warning: Removed 95 rows containing missing values (geom_point).
```

| ght | weight | age | marital | sex | education | afqt | inntekt | hoyde_cm | vekt_kg | BMI | married |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 60 | 155 | 53 | married | female | 13 | 6.84 | 1.6e+05 | 152 | 70.4 | 30.3 | TRUE |
| 70 | 156 | 51 | married | female | 10 | 49.4 | 2.95e+05 | 178 | 70.8 | 22.4 | TRUE |
| 65 | 195 | 52 | married | male | 16 | 99.4 | 8.84e+05 | 165 | 88.5 | 32.5 | TRUE |
| 63 | 197 | 54 | married | female | 14 | 44 | 3.37e+05 | 160 | 89.4 | 34.9 | TRUE |
| 66 | 190 | 49 | married | male | 14 | 59.7 | 6.32e+05 | 168 | 86.3 | 30.7 | TRUE |
| 68 | 200 | 49 | divorced | female | 18 | 98.8 | 8.59e+05 | 173 | 90.8 | 30.4 | FALSE |
| 74 | 225 | 48 | married | male | 16 | 82.3 | 0 | 188 | 102 | 28.9 | TRUE |
| 64 | 160 | 54 | divorced | female | 12 | 50.3 | 5.89e+05 | 163 | 72.6 | 27.5 | FALSE |
| 69 | 162 | 55 | divorced | male | 12 | 89.7 | 5.05e+05 | 175 | 73.5 | 23.9 | FALSE |
| 69 | 194 | 54 | divorced | male | 13 | 96 | 1.26e+06 | 175 | 88.1 | 28.7 | FALSE |



## Konklusjon

Fra testene fremkommer det at høyde ikke bestemmer inntekt. Det fremkommer dog at menn tjener mer enn kvinner.