Ass_2

Christoffersen J. & Haugland V.

Oktober 14. 2021

Del 1 - Innledning

Kort innledning

Det eksisterer en mengde ulike faktorer som vil være med på å bestemme et individs inntekt. For eksempel, i tilskrivnings-kulturer som Japan er status noe som blir tildelt for hvem eller hva en person er, framfor hvordan de utfører sine oppgaver. Dette er et nokså bredt eksempel på tvers av kulturer, men forskning **Referanse?** har vist at det er en rekke biologiske faktorer som bestemmer hvordan vi mennesker tildeler status og anerkjennelse, ergo hvor mye individer kan forvente å tjene. Disse faktorene vil ikke være absolutt avgjørende, men de har tendenser til å være bestemmende for inntekt. I denne oppgaven skal det undersøkes om høyde er en bestemmende faktor for personers inntekt. I første del av oppgaven presenteres en litteraturgjennomgang og beskrivende statistikk. For å belyse denne problemstillingen skal vi benytte oss av datasettet *heights* fra R, (R Core Team 2021) vha. R-pakken *modelr,* (Wickham 2020), og gjennomføre en rekke ulike regresjonsanalyser med henhold til hvordan *høyde* og en rekke andre variabler påvirker inntekten. Resultatene fra disse analysene skal videre benyttes til lineære hypotesetester, hvor vi skal undersøke og konkludere variabelen *høyde* sin effekt på inntekt. Resultatet av oppgaven fremstilles avslutningsvis i en konklusjon.

Kort litteraturgjennomgang

For å anskaffe en bredere forståelse av dette fenomenet med høyde sin påvirkning på individets inntekt, skal vi kort gjennomgå artikkelen *The Effect of Physical Hieght on Workplace Sucess and Income: Prelimitary Test of a Theoretical Model*. Her undersøkte de forholdet mellom høyde og hvilken type karrièresuksess individene hadde. Resultatene fremviste at fysisk høyde har en signifikant påvirkning på individets sosiale aktelse, ledelse fremheving og ytelse(Judge and Cable 2004).

Det eksisterer mange underliggende faktorer som er med på å påvirke hvorfor høyde er så ønskelig. For det første er høye individer ansett som potensielt mer suksessfulle enn de som er lavere. En grunn til dette er at høyde er en sosialt ønskelig ressurs med mange goder. For eksempel, vil de som er høye bli betraktet som mer overbevisende, mer attraktiv som partner og ha en større sannsynlighet for å bli en fremtidig leder. For å illustrere sistnevnte, så kan vi se at det ikke har blitt utvalgt en president i USA som er lavere enn den gjennomsnittlige høyden helt siden 1896 (Judge and Cable 2004). Høydens viktighet er ikke bare observert blant mennesker, men vi ser den også i dyreriket. For dyrene er høyde noe som er svært enkelt å observere og dermed blir brukt som en indeks for dens makt og styrke, noe som spesielt fremvises i livstruende situasjoner. Vi kan dermed si fra et slikt sosiobiologisk perspektiv, at høyde vil tilegne seg makt og makt er noe som videre fører til respekt. I jobbsammenhenger kan høydens viktighet bli fremvist tydeligere, med henhold til at dette er en arena hvor faktorer som overtalelse og makt er ansett som uhyrlig viktige. Det er gjennomført flere ulike undersøkelser på høyde innen jobbsammenhenger og de viser blant annet betydningen når det kommer til hvordan høye salgspersoner blir betraktet i større grad og at lave politibetjenter kan forvente en større grad av klager enn høye politibetjenter (Judge and Cable 2004).

For å forstå disse underliggende faktorene til høyde kan vi benytte den teoretiske modellen for forholdet mellom høyde og karrière til å forklare dette sosiale fenomenet. For det første vil høyde ha en stor betydning for hvordan vi betrakter oss selv, altså vår selvtillit og selvfølelse. Med henhold til at høyde er sterkt korrelert med sosial makt, vil lave individer stå i fare for å være misfornøyde med sin fysiske form, og denne usikkerheten vil være med på å påvirke personligheten deres. Et ekstremt eksempel på dette vil være det som kalles *Napoleon komplekset*, hvor individets høyde har så stor negativ betydning at denne mangelen på høyde fører til at man føler seg utilstrekkelig som individ, som resulterer i en svært aggressiv opptreden (Just and Morris 2003).

Høyde vil i likhet med selvtilliten påvirke hvordan individet vil bli betraktet av andre, altså den sosiale aktelsen. For eksempel, vil dette i stor grad føre til høyere objektiv ytelse, som er jobb eller oppgave utfall og resultater. Noe som kan spesielt observeres i situasjoner hvor sosiale interaksjoner er ansett som viktige. Dette kan fremvises ved at kunder tenderer til å gi mer betraktning mot høye individer, som gjør at de i større grad kommer til å kjøpe av en høy salgsperson. Denne beudringen vil også være essensiell i andre sammenhenger, slikt som å opprette tillit, motta informasjon og mer effektiv forhandling (Judge and Cable 2004), som vil resultere i økt effektivitet og ytelse. Individets egne selvtillit vil også ha en stor påvirkning på ytelsen. En positiv selvtillit er nært knyttet til høyere ytelse ved at slike individer gjerne blir ansett som mer gunstige og er dermed lettere likt blant andre.

Det siste steget i modellen forklarer hvordan denne ytelsen vil ha en påvirkning på videre suksess i karrièren. Med den økte sosiale aktelsen og selvtilliten det høye individet innehar, vil resultere i høyere ytelse. Denne ytelsen gjennom mer effektiv produktivitet vil dermed føre til organisatoriske belønninger som blant annet høyere lønninger og forfremmelser.

Pakker

```
# Bruk suppress så slipper en alle beskjeden fra pakkene som lastes
suppressPackageStartupMessages(
  c(
    library(modelr),
    library(dplyr),
    library(ggplot2),
    library(tinytex),
    library(tidyverse),
    library(ggpubr),
    library(huxtable),
    library(car)
    )
)
```

```
##   [1] "modelr"    "stats"     "graphics"  "grDevices" "utils"     "datasets"
##   [7] "methods"   "base"      "dplyr"     "modelr"    "stats"     "graphics"
##  [13] "grDevices" "utils"     "datasets"  "methods"   "base"      "ggplot2"
##  [19] "dplyr"     "modelr"    "stats"     "graphics"  "grDevices" "utils"
##  [25] "datasets"  "methods"   "base"      "tinytex"   "ggplot2"   "dplyr"
##  [31] "modelr"    "stats"     "graphics"  "grDevices" "utils"     "datasets"
##  [37] "methods"   "base"      "forcats"   "stringr"   "purrr"     "readr"
##  [43] "tidyr"     "tibble"    "tidyverse" "tinytex"   "ggplot2"   "dplyr"
##  [49] "modelr"    "stats"     "graphics"  "grDevices" "utils"     "datasets"
##  [55] "methods"   "base"      "ggpubr"    "forcats"   "stringr"   "purrr"
##  [61] "readr"     "tidyr"     "tibble"    "tidyverse" "tinytex"   "ggplot2"
##  [67] "dplyr"     "modelr"    "stats"     "graphics"  "grDevices" "utils"
##  [73] "datasets"  "methods"   "base"      "huxtable"  "ggpubr"    "forcats"
```

```
##  [79] "stringr"   "purrr"     "readr"     "tidyr"     "tibble"    "tidyverse"
##  [85] "tinytex"   "ggplot2"   "dplyr"     "modelr"    "stats"     "graphics"
##  [91] "grDevices" "utils"     "datasets"  "methods"   "base"      "car"
##  [97] "carData"   "huxtable"  "ggpubr"    "forcats"   "stringr"   "purrr"
## [103] "readr"     "tidyr"     "tibble"    "tidyverse" "tinytex"   "ggplot2"
## [109] "dplyr"     "modelr"    "stats"     "graphics"  "grDevices" "utils"
## [115] "datasets"  "methods"   "base"
```

```
# Jeg synes denne er for ekstrem. Gir uendelig mange nuller etter komma som
# ser litt rart ut
#options(scipen = 999)
options(scipen = 7)
```

Analyse

Vi lager først et histogram med inntekts-variabel.

```
# Endrer navn på datasettet fra heights til hoyde
# data("heights", package = "modelr")
# hoyde <- heights
# Bruker dere følgende slipper dere å ha både hoyde og heights lastet
hoyde <- modelr::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 variablene *høyde*, *vekt*, *alder*, *sivilstatus*, *kjønn*, *utdannelse* og *afqt* (score fra forsvarets kvalifiseringstest), som bestemmer den avhengige variabelen *inntekt* i dette settet. Hvorav variablene *sivilstatus* og *kjønn* er såkalte faktor-variabler (dummy-variabler). *Sivilstatus* varier med fem ulike faktorer som er *singel*, *gift*, *separert*, *skilt* og *enke* og dummy-variabelen *kjønn* varier mellom tallet 1 og 2 som bestemmer om individet som er observert er enten mann eller kvinne.

I datasettet er variablene oppgitt på engelsk:

Heights høyde oppgitt i tommer

Income årlig inntekt. Det er gjort et gjennomsnitt for å presentere topp 2% høyeste inntekt.

Weight Oppgir vekt i enheten pounds

Age Oppgir alder i antall år fra 47 til 56 år.

Martial Oppgir sivilstatus

Sex Kjønn. Mann/Kvinne

Education Oppgir antall år utdanning

AFGT Score fra forvarets kvalifiseringstest

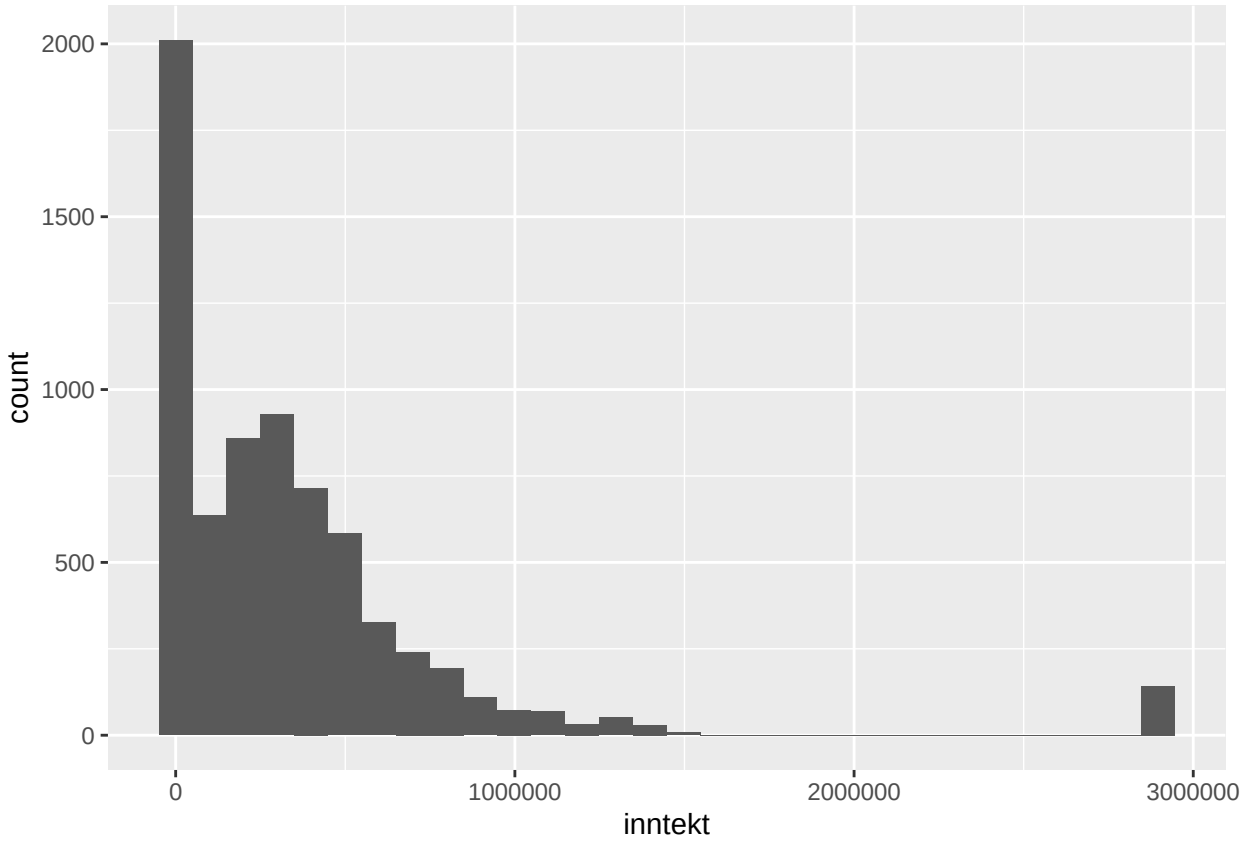
Forklarende dataanalyse

Til å begynne med lages tre nye variabler som settes til en metrisk standard.

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

Histogram til inntektsvariabelen

```r
ggplot(data = hoyde, aes(x = inntekt)) +
  geom_histogram(bins = 30)
```




```r
#eventuelt bruk binwidths, f.eks 100000
# geom_histogram(binwidth = 100000)
```

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 variabelen *inntekt*.

```r
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 variabelen *inntekt* er 343,830. Videre kan vi undersøke hvor mange av observasjonene som innehar denne verdien.

```
# FLOTT!
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 maksimumsinntekten og medianen til datasettet, kan vi si at utliggerne til høyere representerer et gjennomsnitt av ca. 2% av det totale utvalget.

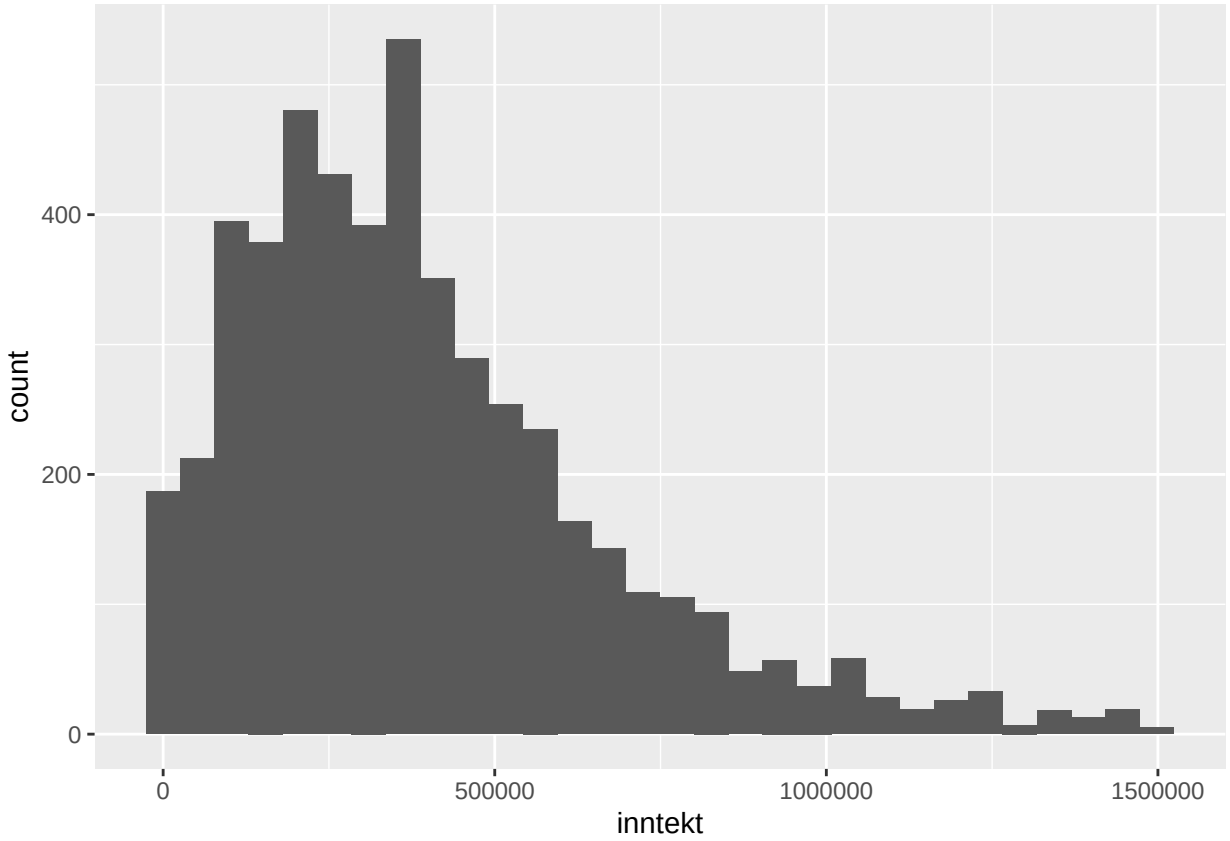
```
(143/7006)*100
```

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

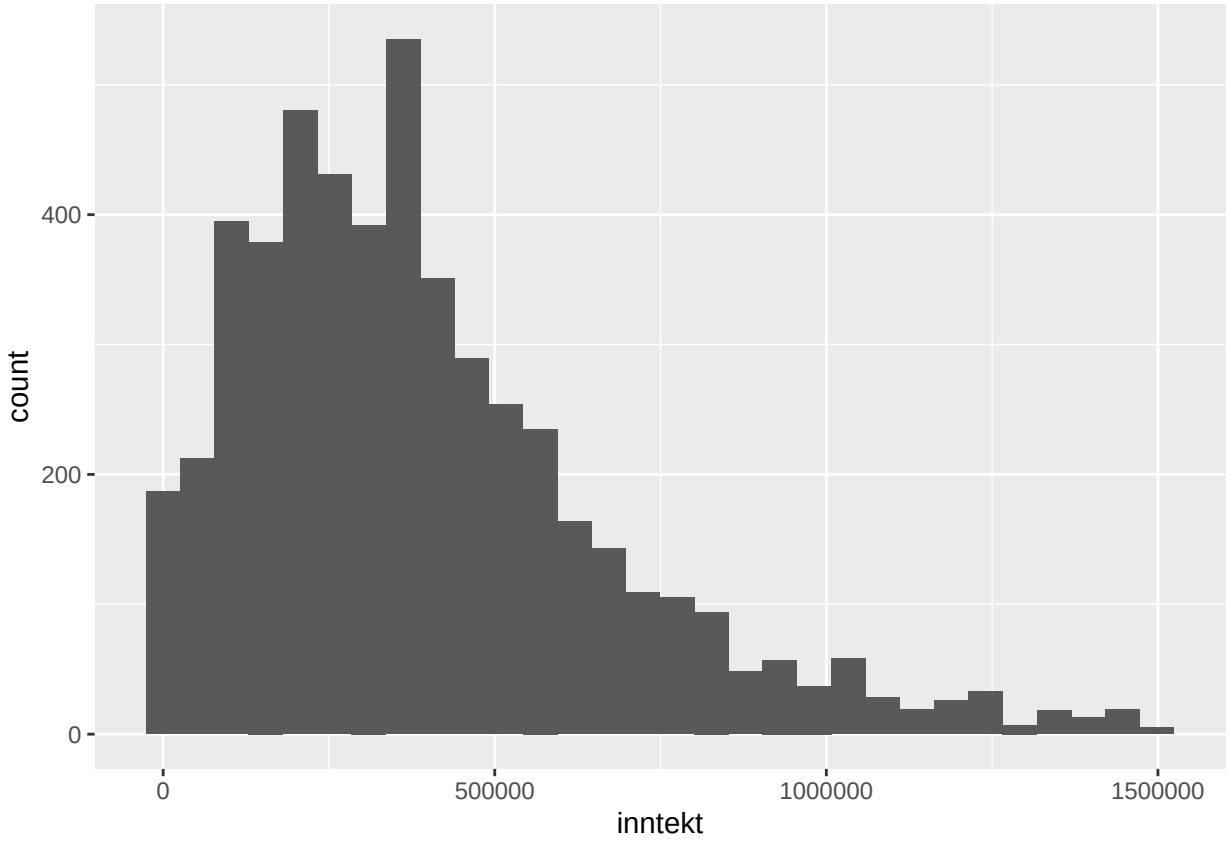
Dersom datasettet reduseres for topp 2% høyeste inntekt og inntekt lik 0, ser histogrammet slik ut:

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

```
ggplot(data = hoyde_begr, aes(x = inntekt)) +
  geom_histogram(bins=30)
```




```
# Koden ovenfor er helt ok, men jeg vil likevel foreslå følgende fordi jeg
# mener løsningen er mer fleksibel og holder oss til ett datasett
hoyde %>%
  # Her lærte jeg noe nytt. Jeg ville skrevet (inntekt < 1500000 & inntekt > 0)
  # men dokumentasjonen avslører at filter vil bytte ut , med & når det er flere
  # betingelser. Så å skille dem med , fungerer helt fint
  filter(inntekt < 1500000, inntekt > 0) %>%
  ggplot(aes(x = inntekt)) +
  geom_histogram(bins=30)
```




```
# ser at i koden ovenfor er det lett å legge inn ytterligere filter(), mutate() etc.
# og vi har fremdeles bare ett datasett å bry oss om
```

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:

```
# Flott!
sum(hoyde$income == 0)
```

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

Del 2 - Regresjonsanalyser

Regresjonsanalyse: inntekt og høyde

```
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   <2e-16 ***
## hoyde_cm       10047.9      552.8   18.18   <2e-16 ***
## ---
## 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: < 2.2e-16
```

```
# Igjen vil jeg argumentere for
lm1 <- hoyde %>%
  # punktumene nedfor er for å vis ehvor vi vil ha dataene i fra pipen
  filter(complete.cases(.)) %>%
  lm(mod1, data = .)
```

```
summary(lm1)
```

```
##
## Call:
## lm(formula = mod1, data = .)
##
## 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   <2e-16 ***
## hoyde_cm       10047.9      552.8   18.18   <2e-16 ***
## ---
## 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: < 2.2e-16
```

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

```r
-1361001.0 + (10047.9 * 185)
```

```
## [1] 497860.5
```

```r
-1361001.0 + (10047.9 * 177)
```

```
## [1] 417477.3
```

Regresjonsanalyse: inntekt, høyde & vekt

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

```r
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    < 2e-16 ***
## hoyde_cm       11430.3      624.3  18.308    < 2e-16 ***
## 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: < 2.2e-16
```

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 synker ved tilsvarende økning i vekt.

Regresjonsanalyse: inntekt, høyde, vekt & BMI

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

```r
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: < 2.2e-16
```

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."
       )
```

Huxreg modellen over viser en samlet fremstilling av de presenterte modellene.

Modell med interaksjonsvariabel

Videre legges det til en interaksjon for variabelen "sex."

	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.

```
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: < 2.2e-16
```

Fra analysen fremkommer det at dummy-variabelen for sexfemale og interaksjonsvariablene er signifikante.

Modell med flere interaksjonsvariabler

```
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: < 2.2e-16
```

Test av koeffisienter vha. linearHypothesis

```
linearHypothesis(lm4, c("sexfemale = 0", "sexfemale:hoyde_cm = 0"))
```

Tester endelig modell med redusert datasett

Videre utformes den endelige modellen ved datasett redusert for topp 2% inntekt og inntekt lik 0.

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
6.90e+03	1.46e+15				
6.9e+03	1.45e+15	2	1.42e+13	33.8	2.43e-15

```
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: < 2.2e-16
```

```
# Koden kan altså også skrives som
hoyde %>%
  filter(inntekt > 0, inntekt < 1500000) %>%
  filter(complete.cases(.)) %>%
  lm(mod5, data = .) %>%
  # summary()
  # eller for robust standard error, trenger pakkene lmtest og sandwich installert
  lmtest::coeftest(vcov = sandwich::vcovHC, type = "HC3")
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -446168.93  323594.21 -1.3788 0.168022
## hoyde_cm       5309.18    1932.06  2.7479 0.006019 **
## vekt_kg         669.45    1819.42  0.3679 0.712931
## BMI           -4183.07    5114.34 -0.8179 0.413448
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
hoyde %>%
  filter(inntekt > 0, inntekt < 1500000) %>%
  filter(complete.cases(.)) %>%
  lm(mod5, data = .) %>%
  linearHypothesis(c("vekt_kg = 0", "BMI = 0"),
                   white.adjust = "hc3")
```

Res.Df	Df	F	Pr(>F)
4.87e+03			
4.87e+03	2	7.23	0.000735

Robust, dvs. med robuste se, simultan test av om den begrensede («restricted») modellen der koeffisientene for vekt_kg og BMI er begrenset til 0 (dvs. utelatt) er den beste. Ser at H0 klart kan forkastes så vi bør beholde vekt_kg og BMI i modellen.

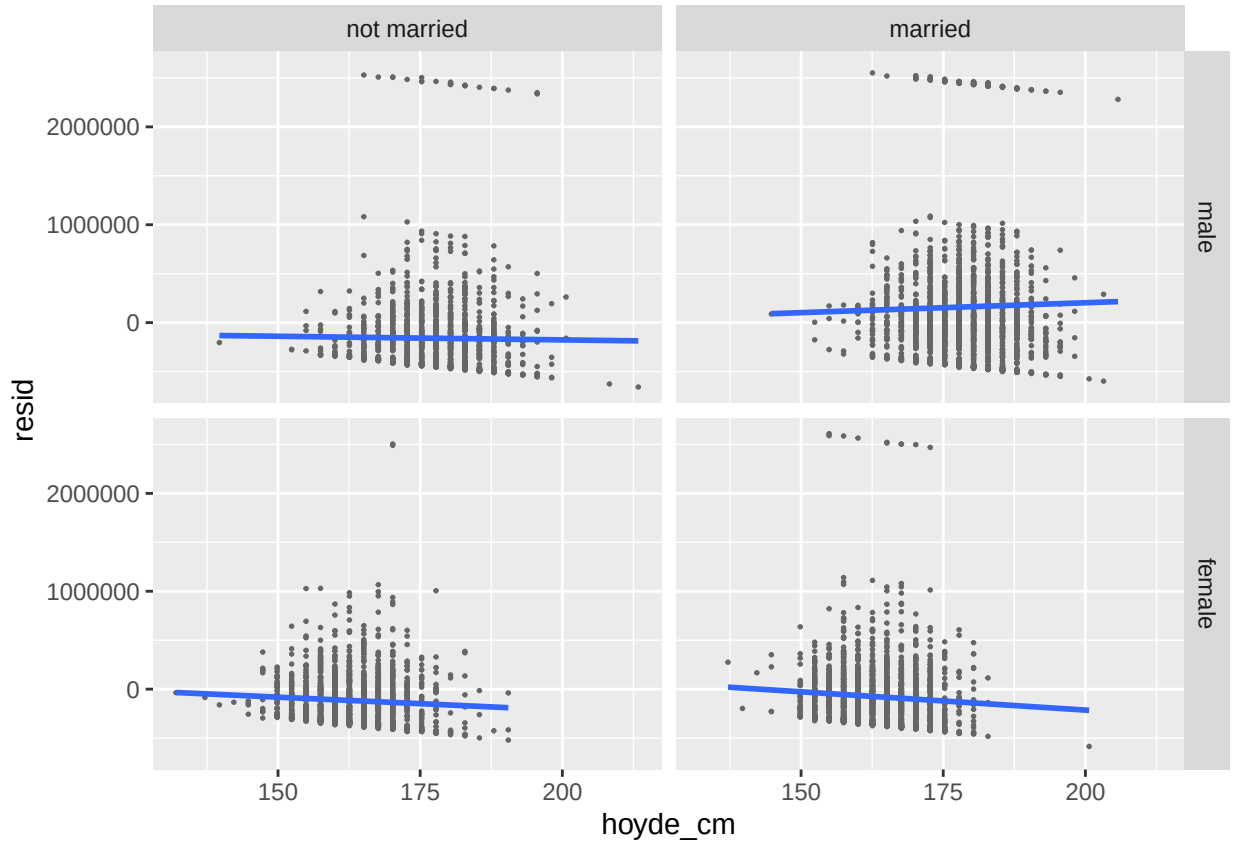
Legger til residualer

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

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

Samtlige observasjoner

```
hoyde_begr %>%
  filter(complete.cases(.) == TRUE) %>%
ggplot(mapping = aes(x = hoyde_cm, y = resid)) +
  facet_grid(sex ~ factor(married, labels = c("not married", "married"))) +
  geom_point(
    colour = "grey40",
    size = 0.3
  ) +
  # formula = er bare for å bli kvitt warning fra ggplot
  geom_smooth(formula='y ~ x', method = "lm", se =FALSE)
```



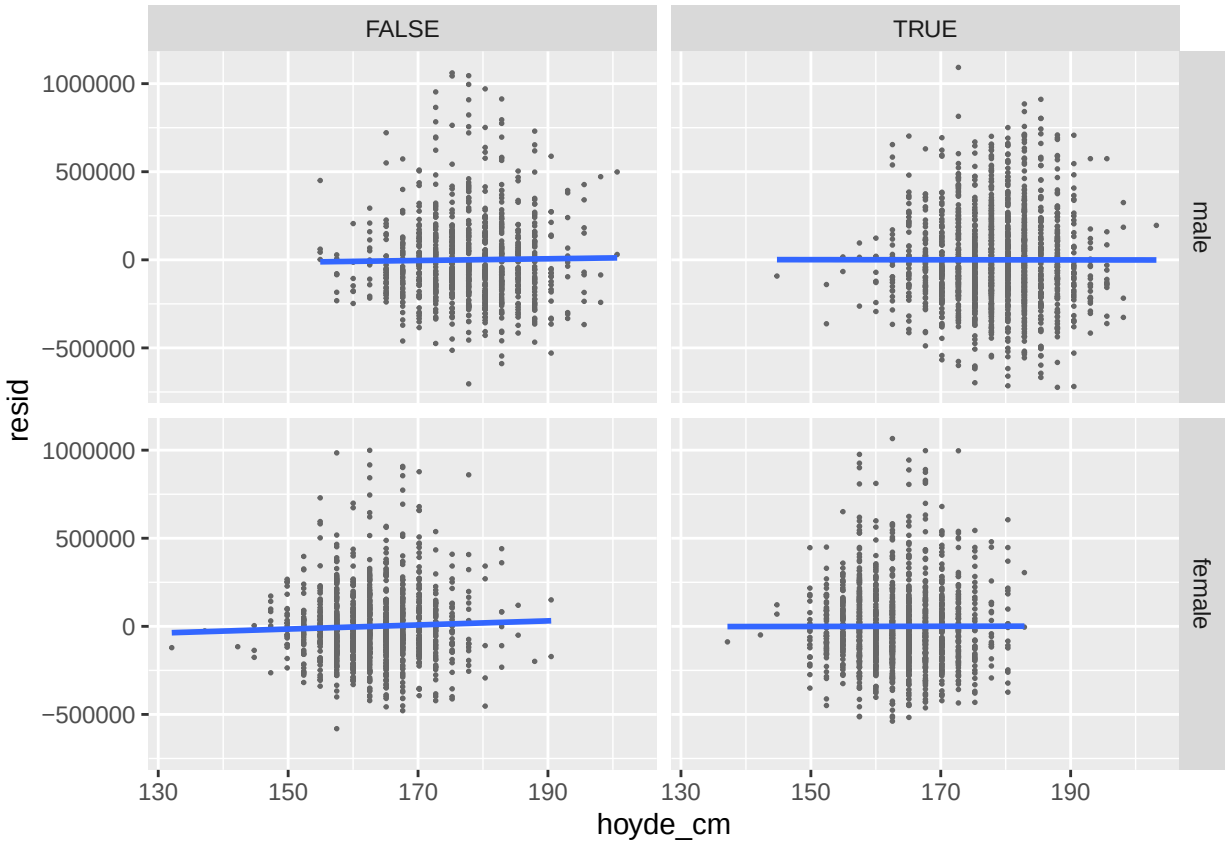
Videre fremkommer det at en total av residualene motsier den endelige modellen, og at høyde ikke nødvendigvis har en reell påvirkning på inntekt.

Merk! modell 5 inneholder høyde som en forklaringsvariabel så ikke noen overraskelse hvis residualene ikke viser noen sammenheng med høyde. Kan imidlertid lage en modell der høyde ikke inngår og så lage plottet.

```
mod6 <- "inntekt ~ sex*(education + afqt + married + age)"
hoyde_begr_ag <- hoyde %>%
  filter(inntekt > 0, inntekt < 1500000) %>%
  filter(complete.cases(.) == TRUE) %>%
  add_residuals(lm(mod6, data = .))


hoyde_begr_ag %>%
```

```
ggplot(mapping = aes(x = hoyde_cm, y = resid)) +
  facet_grid(sex ~ married) +
  geom_point(
    colour = "grey40",
    size = 0.3
  ) +
  # formula = er bare for å bli kvitt warning fra ggplot
  geom_smooth(formula='y ~ x', method = "lm", se =FALSE)
```



Når vi korrigere for kjønn, utdannelse, evner, alder og sivil status (gift/ugift) ser det ut til å være en svak positiv sammenheng mellom høyde og inntekt for ugifte kvinner. Ellers ser inntekt ut til å være uavhengig av høyde.

Konklusjon

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

Referanser

Judge, Timothy A., and Daniel M. Cable. 2004. "The Effect of Physical Height on Workplace Success and Income: Preliminary Test of a Theoretical Model." *Journal of Applied Psychology* 89 (3): 428–41. https://doi.org/10.1037/0021-9010.89.3.428.

Just, Winfried, and Molly R. Morris. 2003. "The Napoleon Complex: Why Smaller Males Pick Fights." *Evolutionary Ecology* 17 (5-6): 509–22. https://doi.org/10.1023/B:EVEC.0000005629.54152.83.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2020. *Modelr: Modelling Functions That Work with the Pipe*. https://CRAN.R-project.org/package=modelr.