Ass_2

Christoffersen J. & Haugland V.

Oktober 14. 2021

Del 1 - Innledning

Bare en liten test.

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 har vist at det er en rekke biologiske faktorer som bestemmer hvordan vi mennesker tildeler status og annerkjennelse, ergo. hvor mye individer kan forvente å tjene. Disse faktorene vil ikke være absolutt avgjørende, men de har tendenser og er blant annet at blondiner tjener mer enn brunetter, skallete menn tjener mindre enn menn med fullt hår og at høye individer tjener mer enn lave individer. Det sistnevnde gjelder spesielt for menn og faktoren angående høyde kommer være problemstillingen vår for dette mini-paperet som er *Er det høyde som bestemmer inntekt?*. For å belyse denne problemstillingen skal vi benytte oss av datasettet *heights* fra R-pakken *modlr*, og gjennomføre en rekke ulike regresjonsanalyser med henhold til hvordan *høyde* og en rekke andre variabler påvirker inntekten. Resultatene fra disse analysene sklal videre benyttes til lineære hypotesetester, hvor vi skal undersøke og konkludere variablen *høyde* sin effekt på inntekt.

Kort litteraturgjennomgang på ca. 1 side

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 karrieresuksess individene hadde. Resultatene fremviste at fysisk høyde har en signifikant påvirkning på individets sosiale aktelse, ledelse fremheving og ytelse **(Judge et al., 2004)**

Det eksisterer mange underliggende faktorerer 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 et al., 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 tydligere, 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 betyningen 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 politbetjenter **(Judge et al., 2004)**.

For å forstå disse underliggende faktorene til høyde kan vi benytte den teoretiske modellen for forholdet mellom høyde og karriere 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 kompleks*, hvor individets høyde har så stor negativ betydning som fører til at denne mangelen på høyde fører til at man føler seg utilstrekkelig som individ, som resulterer i en svært aggresiv opptreden **(Just et al., 2003)**

Høyde vil i likhet med selvtillitten 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 kjøpe av en høy salgsperson. Denne berudringen vil også være essensiell i andre sammenehenger, slikt som å opprette tillit, motta informasjon og mer effektiv forhandling **(Judge et al., 2004)**, som vil resultere i økt effektivitet og ytelse. Individets egne selvtilitt vil også ha en stor påvirkning på ytelsen. En positiv selvtillitt 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 karrieren. Med den økte sosiale aktelsen og selvtilittien det høye individet innehar, vil resultere i høyere ytelse. Denne ytelsen gjennom mer effektiv produktiviet vil dermed føre til organisatoriske belønninger som blant annet høyere lønninger og forfremmelser.

```
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
```

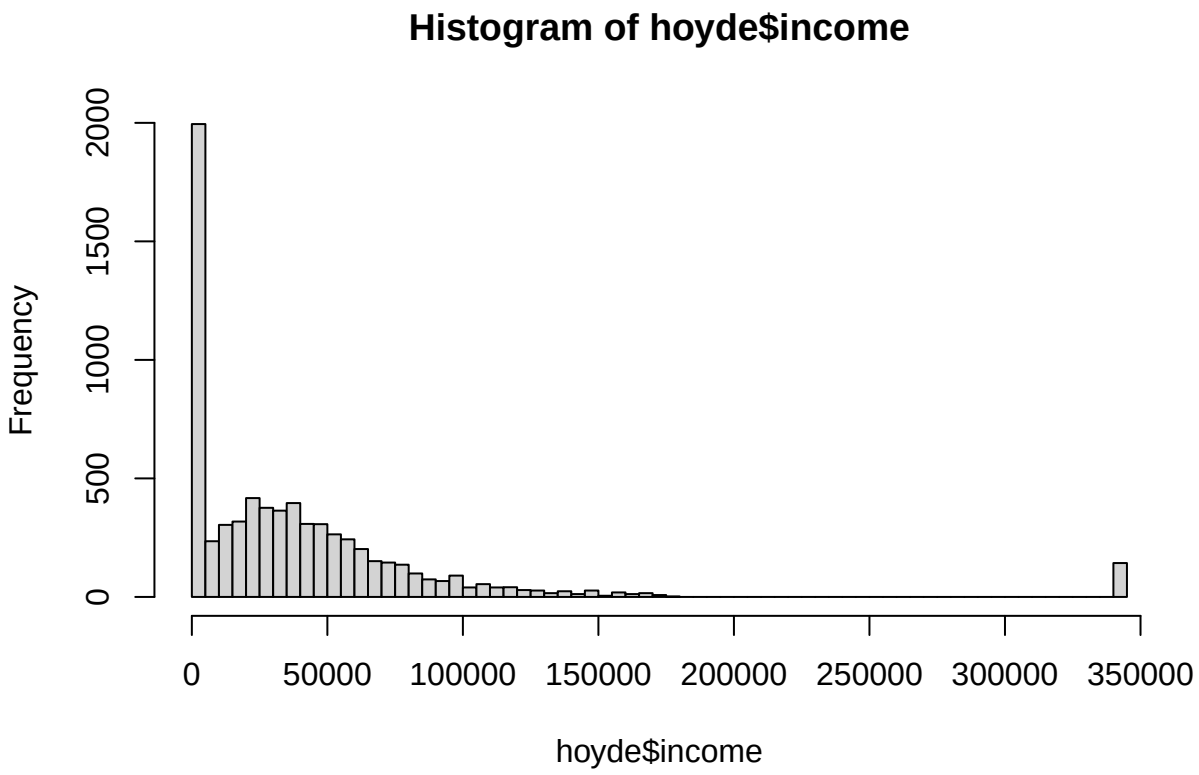
Beskrivende statistikk

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

```r
lm1 <- lm(income ~ height_cm, data = hoyde )
lm2 <- lm(income ~ height_cm + weight_kg + education, data = hoyde)
lm3 <- lm(income ~ height_cm + sex + age + education, data = hoyde)
```

Test interaksjon

```r
m_educsex_i <- "income ~ education*sex"
lm_educsex_i <- lm(m_educsex_i, data = hoyde)
```

```r
summary(lm_educsex_i)
```

```
##
## Call:
## lm(formula = m_educsex_i, data = hoyde)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -140850  -24337   -6104   14658  341178
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -110419.8     4339.8  -25.44   <2e-16 ***
## education             12563.5      326.2   38.51   <2e-16 ***
## sexfemale             68353.1     6056.2   11.29   <2e-16 ***
## education:sexfemale   -7203.4      450.0  -16.01   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 48760 on 6992 degrees of freedom
##   (10 observations deleted due to missingness)
## Multiple R-squared:  0.2396, Adjusted R-squared:  0.2393
## F-statistic: 734.5 on 3 and 6992 DF,  p-value: < 2.2e-16
```

```r
summary(lm2)
```

```
##
## Call:
## lm(formula = income ~ height_cm + weight_kg + education, data = hoyde)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -120347  -27261   -6696   14751  342030
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -250543.44   10525.58 -23.803   <2e-16 ***
## height_cm      1114.70      66.58  16.741   <2e-16 ***
## weight_kg       -75.30      34.48  -2.184    0.029 *
## education      8192.11     234.68  34.907   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 50470 on 6897 degrees of freedom
##   (105 observations deleted due to missingness)
## Multiple R-squared:  0.1932, Adjusted R-squared:  0.1928
## F-statistic: 550.5 on 3 and 6897 DF,  p-value: < 2.2e-16
```

```
library(huxtable)
```

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

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

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

```
huxreg(lm1, lm2, lm3)
```

Redusert datasett

```
# Legge til height_cm i redusert datasett (sample1)
sample1 <- sample1 %>%
  mutate(height_cm = height * 2.54)
```

```
# Legge til weight_kg i redusert datasett (sample1)
sample1 <- sample1 %>%
  mutate(weight_kg = weight *0.45)
```

```
# Legge til bmi i sample1
sample1 <- sample1 %>%
  mutate(sample1, bmi = weight_kg/(height_cm)^2)
```

```
# Legge til forenklet utgave av marrital i sample1
sample1 <- sample1 %>%
  mutate(
    married = factor(
      case_when(
        marital == "married" ~ TRUE,
        TRUE ~ FALSE
      )
    )
  )
```

	(1)	(2)	(3)
(Intercept)	-158888.057 ***	-250543.442 ***	-82298.523 ***
	(10733.752)	(10525.583)	(20011.844)
height_cm	1173.939 ***	1114.696 ***	238.326 **
	(62.859)	(66.584)	(82.719)
weight_kg		-75.300 *	
		(34.481)	
education		8192.113 ***	8673.503 ***
		(234.684)	(231.343)
sexfemale			-23169.494 ***
			(1715.513)
age			-385.839
			(265.470)
N	7006	6901	6996
R2	0.047	0.193	0.213
logLik	-86354.182	-84521.954	-85565.644
AIC	172714.364	169053.909	171143.288

*** p < 0.001; ** p < 0.01; * p < 0.05.

```
lm4 <- lm(income ~ height_cm, data = sample1)
lm5 <- lm(income ~ height_cm + weight_kg + education, data = sample1)
lm6 <- lm(income ~ height_cm + sex + age + education, data = sample1)
```

```
summary(lm5)
```

```
##
## Call:
## lm(formula = income ~ height_cm + weight_kg + education, data = sample1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -83925 -19943  -4176  14783 134541
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -143491.29    7545.87 -19.016   <2e-16 ***
## height_cm       730.37      47.46  15.389   <2e-16 ***
## weight_kg       -29.40      25.31  -1.161    0.246
```

```
## education      5048.21     166.66  30.290   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29910 on 5048 degrees of freedom
##   (71 observations deleted due to missingness)
## Multiple R-squared:  0.1969, Adjusted R-squared:  0.1964
## F-statistic: 412.6 on 3 and 5048 DF,  p-value: < 2.2e-16
```

```
huxreg(lm4, lm5, lm6)
```

	(1)	(2)	(3)
(Intercept)	-75562.510 ***	-143491.294 ***	-45055.441 **
	(7558.119)	(7545.871)	(14027.293)
height_cm	716.334 ***	730.370 ***	157.744 **
	(44.185)	(47.459)	(58.473)
weight_kg		-29.399	
		(25.312)	
education		5048.214 ***	5408.041 ***
		(166.662)	(164.762)
sexfemale			-15320.758 ***
			(1207.401)
age			-5.253
			(184.203)
N	5123	5052	5121
R2	0.049	0.197	0.221
logLik	-60485.803	-59232.213	-59951.106
AIC	120977.606	118474.426	119914.212

*** p < 0.001; ** p < 0.01; * p < 0.05.

```
huxreg(
  list("Modell 1" = lm1, "Modell 4" = lm4),
  error_format = "[{statistic}]",
  note = "Regresjonstabell 1: {stars}. T statistics in brackets."
)
```

	Modell 1	Modell 4
(Intercept)	-158888.057 ***	-75562.510 ***
	[-14.803]	[-9.998]
height_cm	1173.939 ***	716.334 ***
	[18.676]	[16.212]
N	7006	5123
R2	0.047	0.049
logLik	-86354.182	-60485.803
AIC	172714.364	120977.606

Regresjonstabell 1: *** p < 0.001; ** p < 0.01; * p < 0.05. T statistics in brackets.

```
huxreg(
  list("Modell 2" = lm2, "Modell 5" = lm5),
  error_format = "[{statistic}]",
  note = "Regresjonstabell 2: {stars}. T statistics in brackets."
)
```

```
huxreg(
  list("Modell 3" = lm3, "Modell 6" = lm6),
  error_format = "[{statistic}]",
  note = "Regresjonstabell 3: {stars}. T statistics in brackets."
)
```

Interaksjon på redusert datasett

```
m_educsex_i <- "income ~ education*sex"
lm_educsex_iI <- lm(m_educsex_i, data = sample1)
```

```
summary(lm_educsex_iI)
```

```
##
## Call:
## lm(formula = m_educsex_i, data = sample1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -96162 -19080  -4667  14305 128249
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
```

	Modell 2	Modell 5
(Intercept)	-250543.442 ***	-143491.294 ***
	[-23.803]	[-19.016]
height_cm	1114.696 ***	730.370 ***
	[16.741]	[15.389]
weight_kg	-75.300 *	-29.399
	[-2.184]	[-1.161]
education	8192.113 ***	5048.214 ***
	[34.907]	[30.290]
N	6901	5052
R2	0.193	0.197
logLik	-84521.954	-59232.213
AIC	169053.909	118474.426

Regresjonstabell 2: *** p < 0.001; ** p < 0.01; * p < 0.05. T statistics in brackets.

```
## (Intercept)         -27515.7     3173.0  -8.672  < 2e-16 ***
## education             6188.9      236.1  26.212  < 2e-16 ***
## sexfemale              917.9     4474.7   0.205    0.837
## education:sexfemale  -1382.4      326.6  -4.232 2.35e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29360 on 5117 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.2225, Adjusted R-squared:  0.222
## F-statistic:   488 on 3 and 5117 DF,  p-value: < 2.2e-16
```

```
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

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

```
summary(lm6)
```

	Modell 3	Modell 6
(Intercept)	-82298.523 ***	-45055.441 **
	[-4.112]	[-3.212]
height_cm	238.326 **	157.744 **
	[2.881]	[2.698]
sexfemale	-23169.494 ***	-15320.758 ***
	[-13.506]	[-12.689]
age	-385.839	-5.253
	[-1.453]	[-0.029]
education	8673.503 ***	5408.041 ***
	[37.492]	[32.823]
N	6996	5121
R2	0.213	0.221
logLik	-85565.644	-59951.106
AIC	171143.288	119914.212

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

```
##
## Call:
## lm(formula = income ~ height_cm + sex + age + education, data = sample1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -90790 -19227  -4129  14310 128176
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -45055.441  14027.293  -3.212  0.00133 **
## height_cm      157.744     58.473   2.698  0.00700 **
## sexfemale   -15320.758   1207.401 -12.689  < 2e-16 ***
## age             -5.253    184.203  -0.029  0.97725
## education     5408.041    164.762  32.823  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29390 on 5116 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.2209, Adjusted R-squared:  0.2202
## F-statistic: 362.5 on 4 and 5116 DF,  p-value: < 2.2e-16
```