Does height decide income?

For Assignment 2 se etter innledende tekst.

Introduction

It has been claimed that the height of a person is one of the most deciding factors for the persons income (Judge and Cable 2004). In this short paper you will investigate this assertion by using a dataset from the *National Longitudinal Study* (U.S. Bureau of Labor Statistics). See `help(heights, package = modelr)` for details.

Summary statistics

We start by grouping the heights in 8 intervals and report summary statistics.

```
heights$heightInt <- cut(heights$height, breaks = 8)
kable(summary(heights[,1:4]))
```

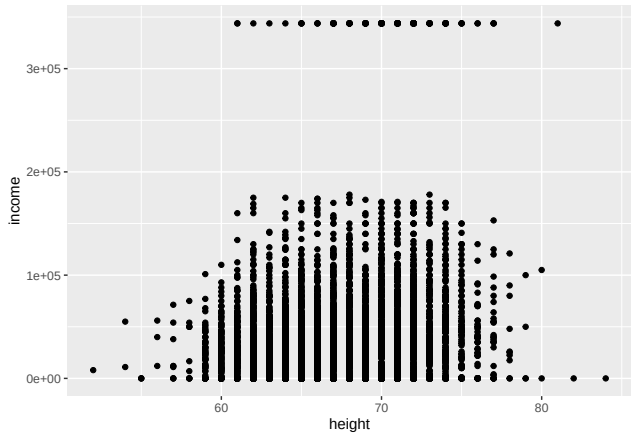
income	height	weight	age
Min. : 0.0	Min. :52.0	Min. : 76.0	Min. :47.00
1st Qu.: 165.5	1st Qu.:64.0	1st Qu.:157.0	1st Qu.:49.00
Median : 29589.5	Median :67.0	Median :184.0	Median :51.00
Mean : 41203.9	Mean :67.1	Mean :188.3	Mean :51.33
3rd Qu.: 55000.0	3rd Qu.:70.0	3rd Qu.:212.0	3rd Qu.:53.00
Max. :343830.0	Max. :84.0	Max. :524.0	Max. :56.00
NA	NA	NA's :95	NA

```
kable(summary(heights[,5:9]))
```

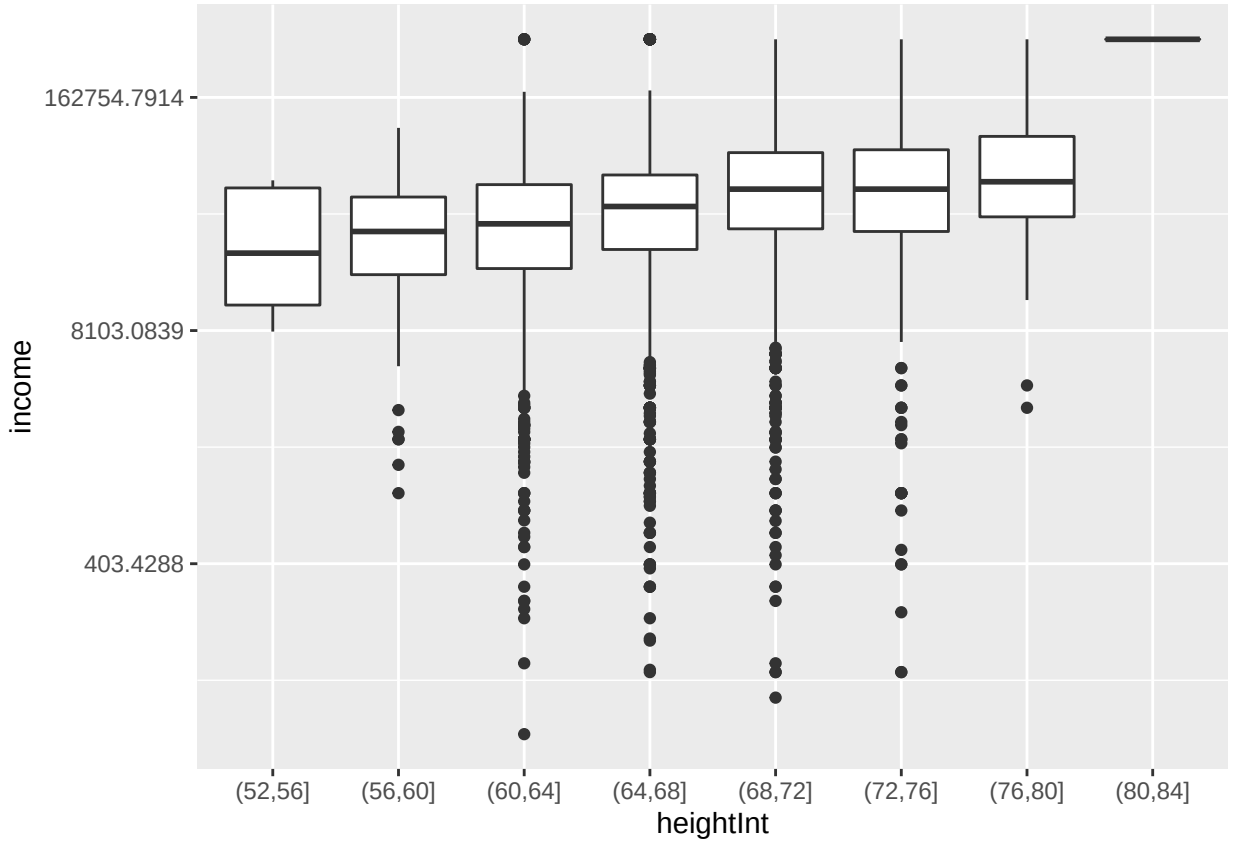
marital	sex	education	afqt	heightInt
single :1124	male :3402	Min. : 1.00	Min. : 0.00	(64,68]:2298
married :3806	female:3604	1st Qu.:12.00	1st Qu.: 15.12	(68,72]:1957
separated: 366	NA	Median :12.00	Median : 36.76	(60,64]:1778
divorced :1549	NA	Mean :13.22	Mean : 41.21	(72,76]: 628
widowed : 161	NA	3rd Qu.:15.00	3rd Qu.: 65.24	(56,60]: 285
NA	NA	Max. :20.00	Max. :100.00	(76,80]: 48
NA	NA	NA's :10	NA's :262	(Other): 12

Plots

```
ggplot(heights, mapping = aes(x = height, y = income)) +
  geom_point()
```

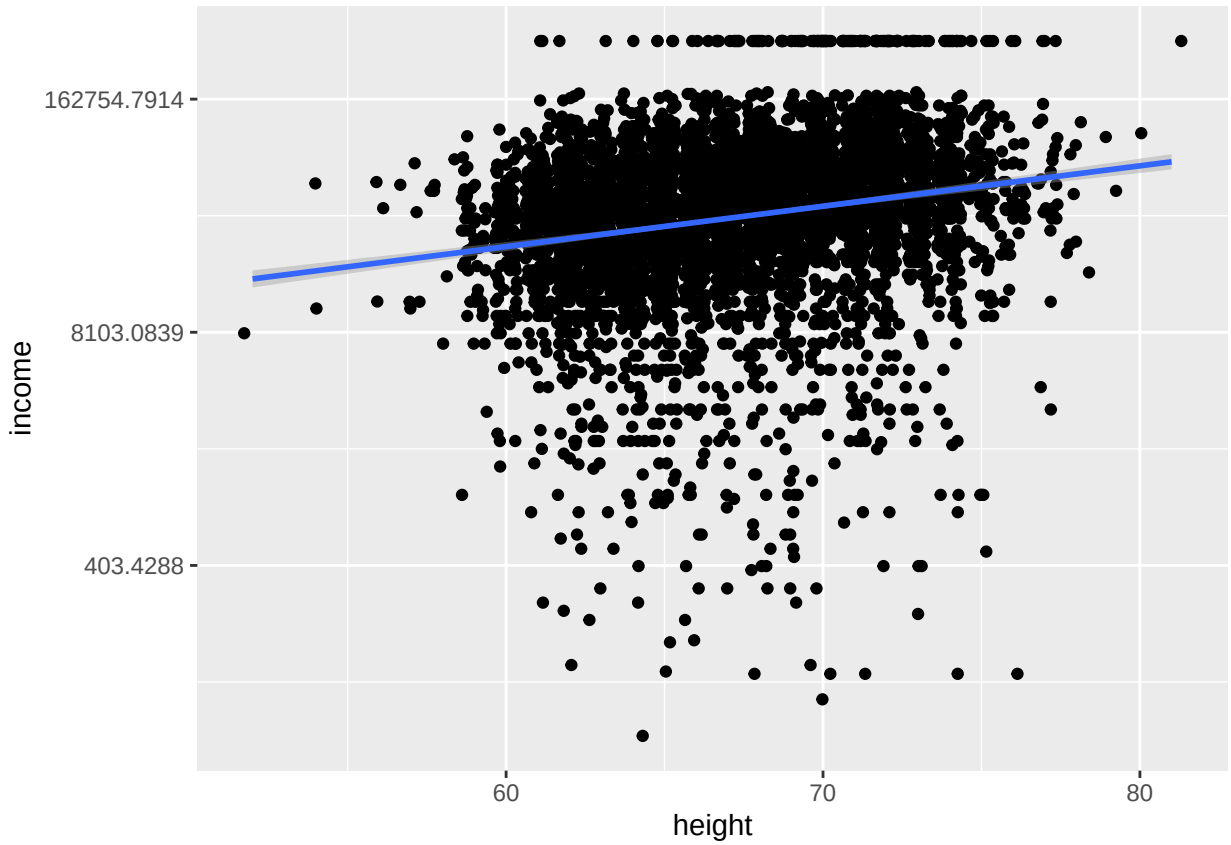



```
heightsPosInc <- subset(heights, income > 0)
ggplot(heightsPosInc, mapping = aes(x = heightInt, y = income)) +
  scale_y_continuous(trans = scales::log_trans()) +
  geom_boxplot()
```



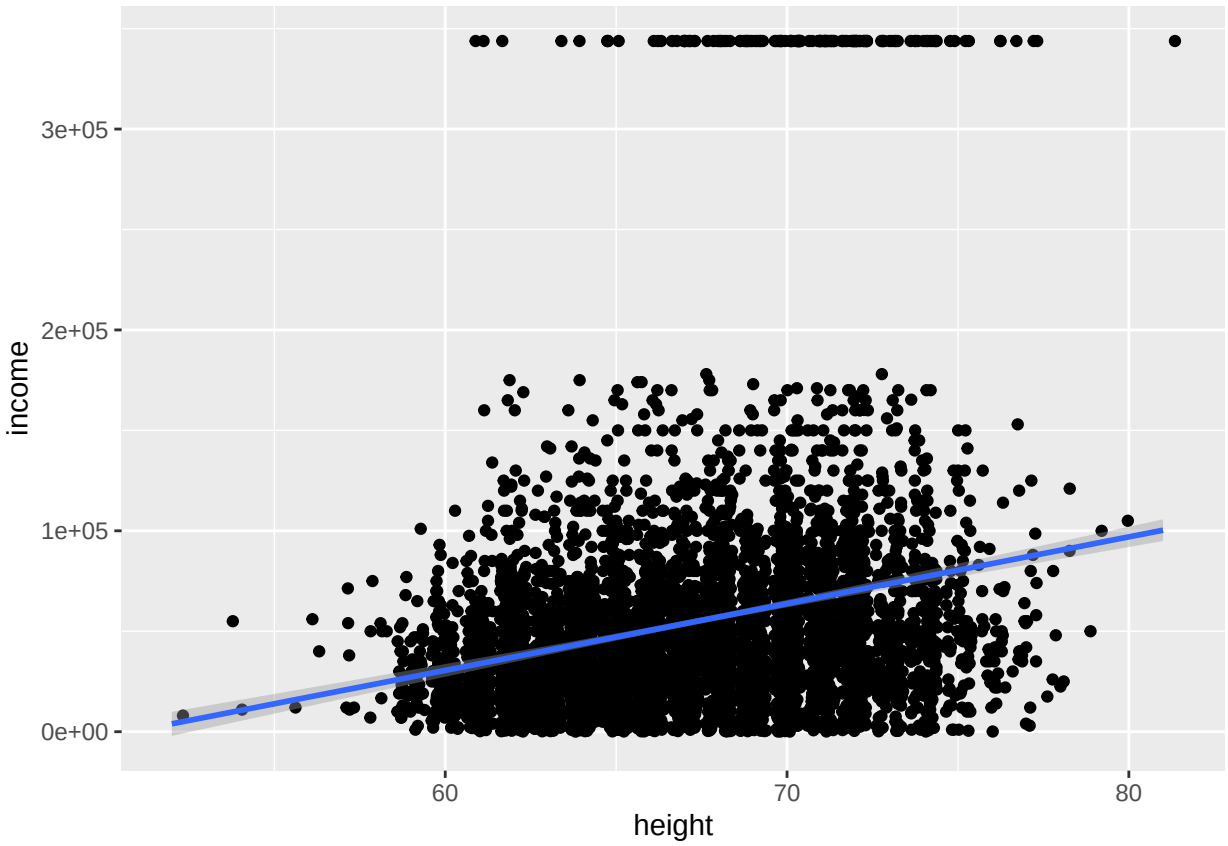

```
heightsPosInc <- subset(heights, income > 0)
ggplot(heightsPosInc, mapping = aes(x = height, y = income)) +
  scale_y_continuous(trans = scales::log_trans()) +
  geom_point(position = "jitter") +
  geom_smooth(method = 'lm')
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
heightsPosInc <- subset(heights, income > 0)
ggplot(heightsPosInc, mapping = aes(x = height, y = income)) +
  geom_point(position = "jitter") +
  geom_smooth(method = 'lm')
```

```
## `geom_smooth()` using formula 'y ~ x'
```



How much do we earn extra per inch

```
summary(lm(income ~ height, data = heights))
```

```
##
## Call:
## lm(formula = income ~ height, data = heights)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -91583 -31511 -10893  14882 320828
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -158888.1    10733.8  -14.80   <2e-16 ***
## height         2981.8      159.7   18.68   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54550 on 7004 degrees of freedom
## Multiple R-squared:  0.04744,    Adjusted R-squared:  0.0473
## F-statistic: 348.8 on 1 and 7004 DF,  p-value: < 2.2e-16
```

Conclusion

One extra inch (2,72 cm) gives you $2982 extra income per year. **You should have eaten your vegetables**, as your mother said.

Is there more to this story? That is what you are going to investigate in this assignment. We will write it in Norwegian this time.

Oppgave (assignment) 2

«Mini-paperet» ovenfor er et nokså ubehjelpelig forsøk på en empirisk undersøkelse. Vi klarer bedre enn det!

Skriv en liten artikkel på norsk (bruk lang: no-NB eller lang: no-NN for hhv. bokmål og nynorsk i YAML header-en, se dokumentet «Debriefing assignment 1» for detaljer) der du benytter datasettet `heights` fra pakken `modelr` til å undersøke problemstillingen *Er det høyde som bestemmer inntekt?* Artikkelen skal inneholde:

- En kort innledning
- En kort litteraturgjennomgang på ca. 1 side (se dokumentet «Debriefing assignment 1» for hensiktsmessig bruk av Zotero når flere forfattere jobber sammen).
- Start analyse med å lage din egen versjon av datasettet. Kall det for hoyde og jobb med dette.

```
# Fjern eval = FALSE for å utføre
data('heights', package = 'modelr')
hoyde <- heights
```

- Beskrivende statistikk, dvs. kort beskrivelse av dataene
- EDA (vha. ggplot) av datasettet.
  - Lag et histogram av variabelen income
  - Hva er forklaring på utliggerne langt til høyre? (se help for datasettet)
  - Har vi med personer uten inntekt i datasettet?
- Regresjonsanalyse (dokumentet *Liten introduksjon til å kjøre regresjonsanalyser i R* kan være til hjelp)
  - Vi benytter hele datasettet, men vil kjøre endelig modell også mot reduserte datasett (uten 2% topp inntekt og uten inntekt 0) for å teste modellens robusthet (husk `filter` funksjonen fra `Tidyverse`)
  - Lag to nye variabler height_cm og weight_kg (vha. mutate()) der du konverterer variablene height (inch) og weight (pound) til metrisk standard.
  - Lag også en ny variabel bmi (der bmi=vekt i kg/(høyde i cm)^2).
  - Lag en forenklet utgave av variabelen marital, dvs. married not-married. Kan enkelt gjøres vha.

```
# inne i en pipe med hoyde
mutate(
  married = factor(
      case_when(
      # note, summary showed no NA for marital
        marital == 'married' ~ TRUE,
      # all other categories FALSE
        TRUE ~ FALSE
      )
  )
)
```

- Totalt skal minst 6 modeller estimeres.

- Resultatet fra estimeringen skal rapporteres vha. huxreg(). Se dokumentet ex_reg_tables.pdf under Filer > Assignment 2 på Canvas, hvis du har glemt hvordan det gjøres. Tips: angir du en liste som første argument til huxreg() kan du styre hva modellene skal hete, f.eks (gir også t-verdier istedenfor standard error)

```
huxreg(
   list("Modell 4" = lm3, "Modell 5" = lm3_nhi),
   error_format = "[{statistic}]",
   note = "Regresjonstabell 3: {stars}. T statistics in brackets."
   )
```

- Den endelige modellen skal testes for robusthet på et datasett uten de 2% høyeste inntektene og på et datasett som i tillegg ikke inneholder observasjoner der inntekten er 0.
- Disse modellene på redusert datasett teller med blant de 6.

  Minst en av modellene skal inneholde interaksjon mht. variabelen sex. (Se eksempel 7.10 i dokumentet Liten intro)](https://elastic-turing-41462a.netlify.app/presentations_ag/intro_econometrics/w_4c1_and_4c3)))
- Det skal gjøres test av koeffisientene vha. linearHypothesis() fra car pakken
- Residualene fra endelig modell skal legges til datasettet hoyde. height_cm skal plottes mot residualene for 'facet_grid(sex ~ factor(married, labels = c("not married," "married")))'
- Plot av samtlige observasjoner svakt i bakgrunnen kan en få til med

```
  geom_point(
    # MÅ velge kun de to variabelen for at dette skal virke
    data= select(hoyde, height_cm, resid_lm3),
    colour = "grey80",
    size = 0.3
    )
```

- Konklusjon Svar på spørsmålet: Er det høyde som bestemmer inntekt?
- Referanser

Referanser

Det forutsettes at git/Github brukes under arbeidet med oppgaven. Det er lov å spørre om alt ;-)

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.