

Innlevering 2

Innlevering 2 i Data Science 2021 - Maren Sognefest og Daniel Karstad

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

## v tibble  3.1.4      v dplyr   1.0.7
## v tidyr   1.1.3      v stringr 1.4.0
## v readr   2.0.1      v forcats 0.5.1
## v purrr   0.3.4

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

Er det høyde som bestemmer inntekt?

Denne artikkelen skal ta for seg om det er en, og eventuelt hvilken, sammenheng det er mellom høyde og inntekt. Ved hjelp av flere analyser, skal vi bruke datasettet *heights* fra pakken *modelr* skal vi prøve å finne ut om det er en sammenheng. Kan det stemme at høye personer tjener mest?

I analyse-delen av artikkelen vil vi bruke ulike **plots** for å analysere spørsmålet, og komme frem til en konklusjon.

Vi må ha ca 1 side litteraturgjennomgang her. Jeg har laget et zotero-bibliotek som ligger ved her. Må gjerne ha inn noen flere kilder vi kan bruke til denne delen

income	height	weight	age	marital	sex	education	afqt
Min. : 0.0	Min. :52.0	Min. : 76.0	Min. :47.00	single :1124	male :3402	Min. : 1.00	Min. : 0.00
1st Qu.: 165.5	1st Qu.:64.0	1st Qu.:157.0	1st Qu.:49.00	married :3806	female:3604	1st Qu.:12.00	1st Qu.: 15.12
Median : 29589.5	Median :67.0	Median :184.0	Median :51.00	separated: 366	NA	Median :12.00	Median : 36.76
Mean : 41203.9	Mean :67.1	Mean :188.3	Mean :51.33	divorced :1549	NA	Mean :13.22	Mean : 41.21
3rd Qu.: 55000.0	3rd Qu.:70.0	3rd Qu.:212.0	3rd Qu.:53.00	widowed : 161	NA	3rd Qu.:15.00	3rd Qu.: 65.24
Max. :343830.0	Max. :84.0	Max. :524.0	Max. :56.00	NA	NA	Max. :20.00	Max. :100.00
NA	NA	NA's :95	NA	NA	NA	NA's :10	NA's :262

Over er sammendraget av statistikken i datasettet “heights.” Man har kolonner med inntekt i dollar, høyde i tommer, vekt i pound, alder, sivilstatus, kjønn, utdannelse og score på Armed Forces Qualitication Test.

Under har vi kopiert datasettet “heights” og kalt det “hoyde.” Her har vi regnet dataen om til europeiske standarderer; vi bruker høyde i cm og vekt i kg, i tillegg til at inntekten er i norske kroner, at vi har laget en egen kolonne for BMI og at vi har laget en forenklet utgave av marital (married - not married). Det er også et sammendrag av dette under.

```
hoyde <- heights %>%
mutate(heights, hoyde_cm = height*2.54, #høyde i cm = hoyde i tommer * 2,54 fordi 1 tommer = 2,54 cm
vekt_kg = weight/2.2, #vekt i kg = vekt i pound / 2.2 fordi 2,2 kg = 1 pound
inntekt_nok = income*8.5, #inntekt i nok = inntekt i dollar * 8,5 fordi 1 dollar = 8,5 nok
married = factor(
```

```

      case_when(
        marital == 'married' ~ TRUE,
        TRUE ~ FALSE)
      )
    )
  hoyde$bmi <- hoyde$vekt_kg/(hoyde$hoyde_cm/100)/(hoyde$hoyde_cm/100) #bmi = vekt i kg /
  hoyde$weight <- NULL #fjerner vekt i pounds fra datasett
  hoyde$height <- NULL #fjerner høyde i tommer fra datasett
  hoyde$income <- NULL #fjerner inntekt i dollar fra datasett

knitr::kable(summary(hoyde[1:5]))

```

age	marital	sex	education	afqt
Min. :47.00	single :1124	male :3402	Min. : 1.00	Min. : 0.00
1st Qu.:49.00	married :3806	female:3604	1st Qu.:12.00	1st Qu.: 15.12
Median :51.00	separated: 366	NA	Median :12.00	Median : 36.76
Mean :51.33	divorced :1549	NA	Mean :13.22	Mean : 41.21
3rd Qu.:53.00	widowed : 161	NA	3rd Qu.:15.00	3rd Qu.: 65.24
Max. :56.00	NA	NA	Max. :20.00	Max. :100.00
NA	NA	NA	NA's :10	NA's :262

```
knitr::kable(summary(hoyde[6:10]), "pipe")
```

hoyde_cm	vekt_kg	inntekt_nok	married	bmi
Min. :132.1	Min. : 34.55	Min. : 0	FALSE:3200	Min. :12.90
1st Qu.:162.6	1st Qu.: 71.36	1st Qu.: 1407	TRUE :3806	1st Qu.:25.14
Median :170.2	Median : 83.64	Median : 251511	NA	Median :28.38
Mean :170.4	Mean : 85.59	Mean : 350234	NA	Mean :29.37

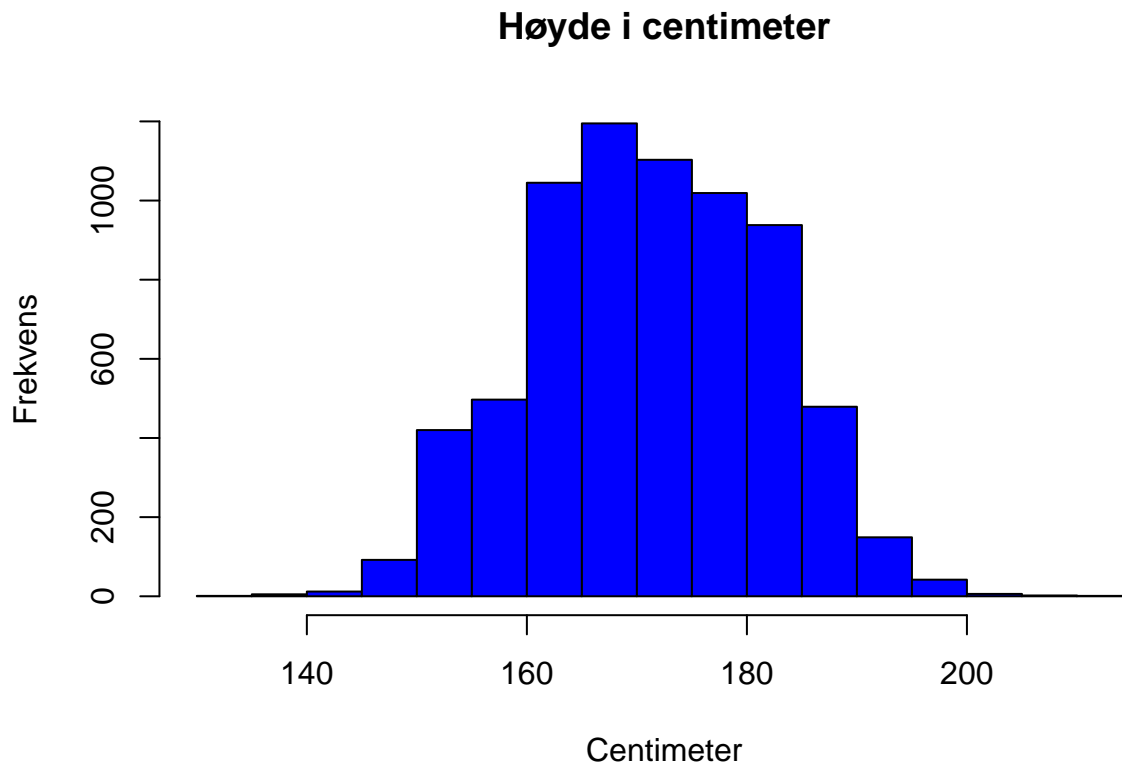
hoyde_cm	vekt_kg	inntekt_nok	married	bmi
3rd Qu.:177.8	3rd Qu.: 96.36	3rd Qu.: 467500	NA	3rd Qu.:32.35
Max. :213.4	Max. :238.18	Max. :2922555	NA	Max. :75.15
NA	NA's :95	NA	NA	NA's :95

EDA (vha. ggplot) av datasettet.

Analyse med 1-variabel

Denne delen inneholder noen grafer med 1 variabel. Dette for å se om variablene er normalfordelt.

```
## Warning: Unknown or uninitialised column: 'height_cm'.
```



Utliggerne til høyre er der fordi det er beregnet gjennomsnittsinntekt av de to prosentene med høyest inntekt. Dette gjennomsnittet er brukt, og har erstattet alle de øverste verdiene. Vi merker oss også at histogrammet er tilnærmet normalfordelt, der de fleste observasjonene ligger mellom 160-180 cm.

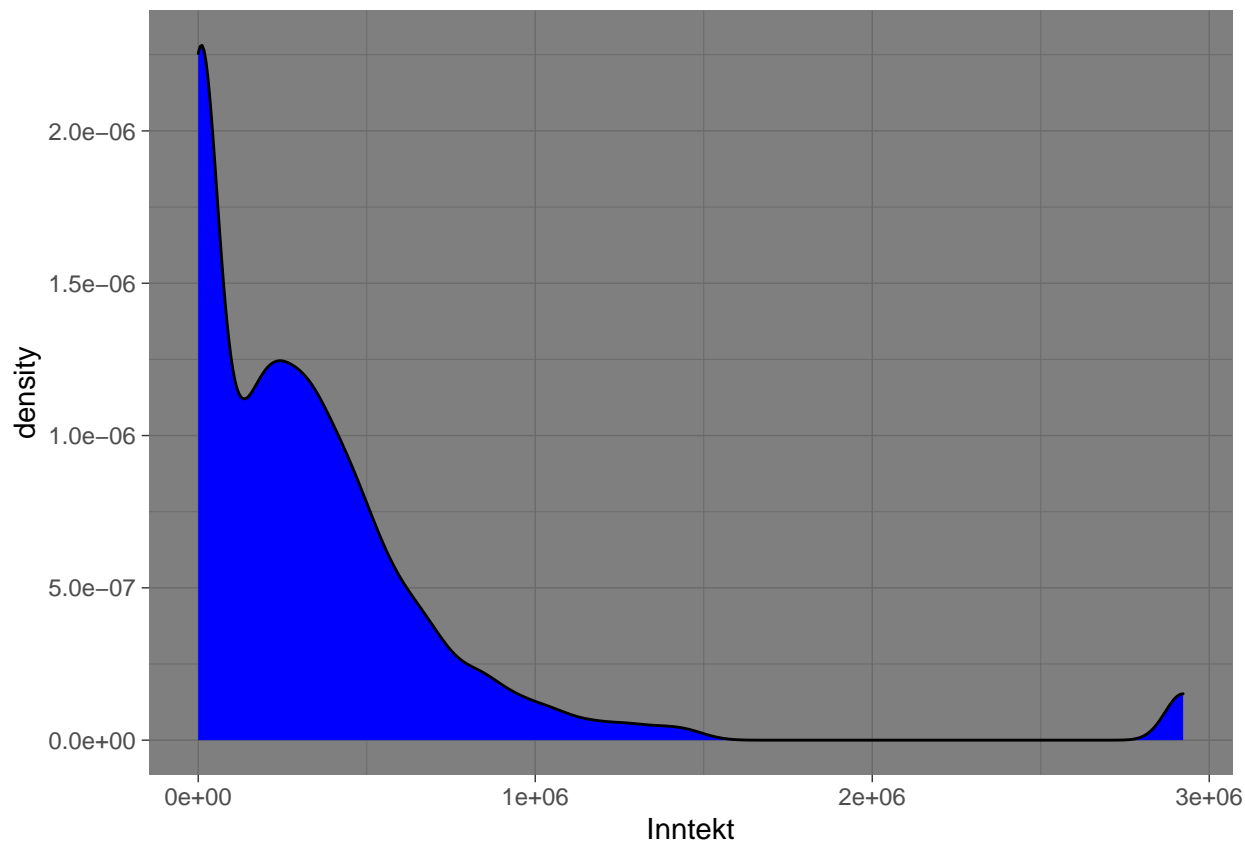
Som man kan se utfra histogrammet er det med flere uten inntekt i datasettet. Det kan man også se slik:

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

Det er altså 1740 personer uten inntekt med i datasettet.

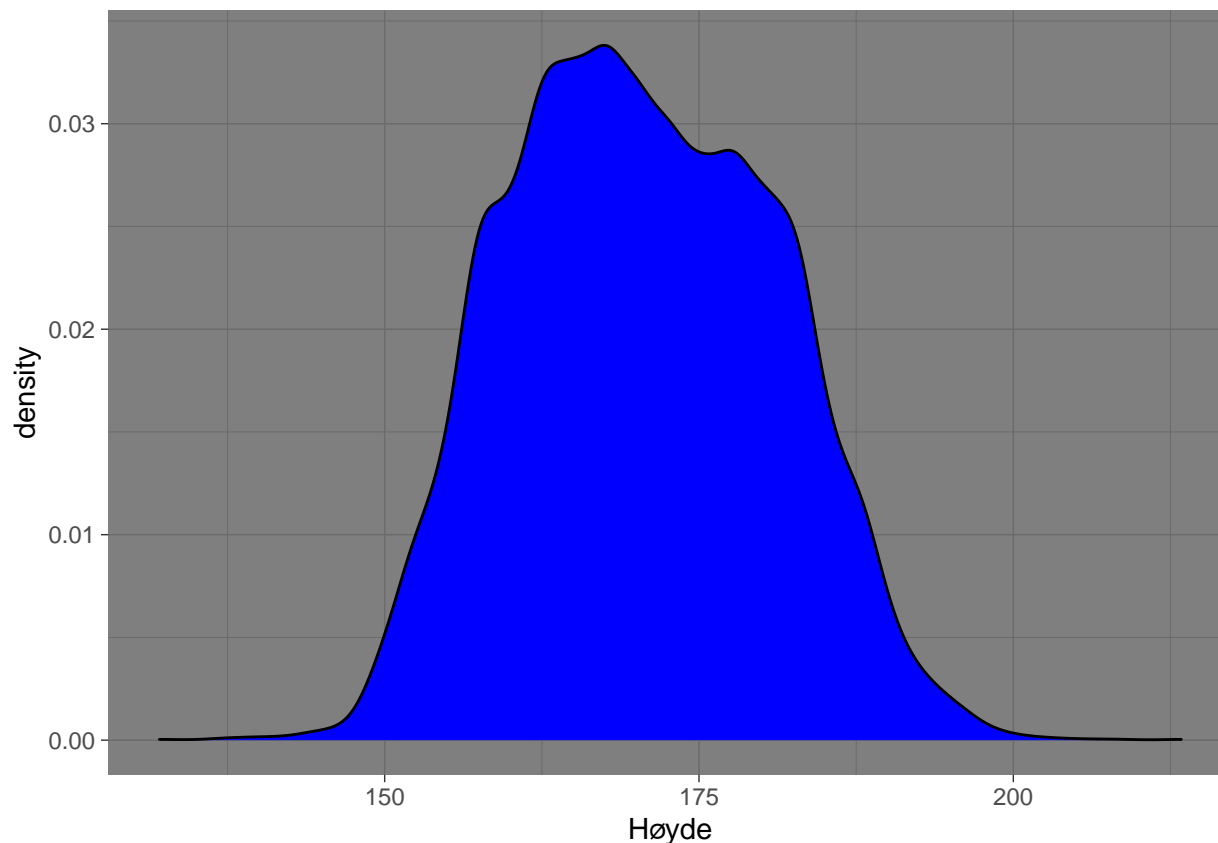
Vi må ha regresjonsanalyse (dokumentet Liten introduksjon til å kjøre regresjonsanalyser i R kan være til hjelp)

```
Inntekt <- hoyde$inntekt_nok  
ggplot(hoyde = Inntekt) + theme_dark() +  
  geom_density(aes(x = Inntekt), fill = "blue")
```



Denne grafen illustrerer inntekt, og her ser man ved første øyekast at det ikke er normalfordelt. Dette fordi man har mange observasjoner tilnærmet 0, og hovedvekten er <500,000.00 NOK. I andre enden av skalaen, de med 2% høyest inntekt er illustrert ved et gjennomsnitt av lønnen. Derfor foreligger det ingen observasjoner mellom 1,600,000.00 NOK og 3,000,000.00 NOK.

```
Høyde <- hoyde$hoyde_cm
Inntekt <- hoyde$inntekt_nok
ggplot(hoyde = Høyde) + theme_dark() +
  geom_density(aes(x = Høyde), fill = "blue")
```



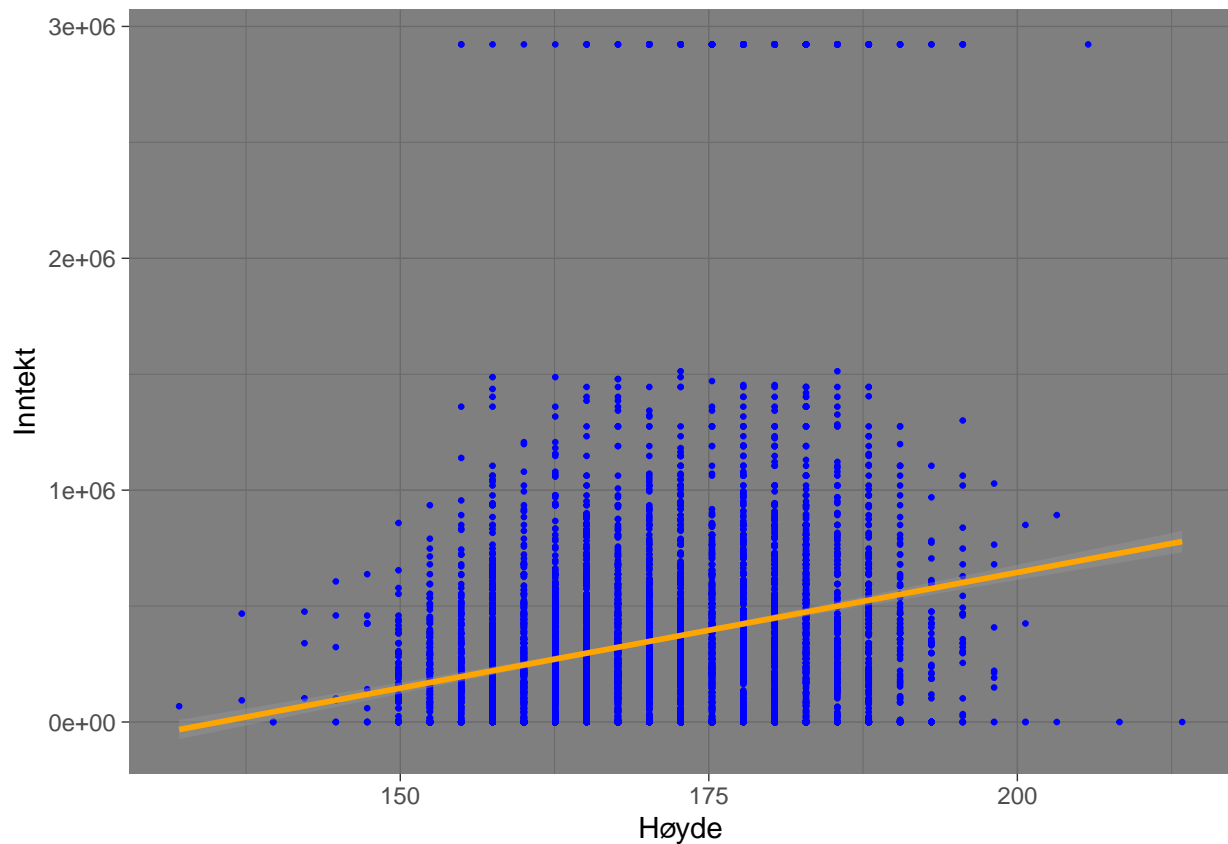
Ovenfor kan sees et density plot. Det viser hovedsaklig det samme som histogrammet, men man ser enklere tettheten av observasjonene her ved en jevn og fin kurve. Her ser man igjen, en tilnærmet normalfordeling.

Analyse med 2-variabler

I grafene og analysene over har vi sett på variablene inntekt og høyde hver for seg. For å undersøke om det finnes noe sammenheng mellom disse, må vi putte de 2 variablene inn i samme plot.

```
#options(scipen = 999) #disabling scientific notation in R. vet ikke om vi skal ha med  
Høyde <- hoyde$hoyde_cm  
Inntekt <- hoyde$inntekt_nok  
ggplot(hoyde, aes(Høyde, Inntekt)) + theme_dark() +  
  geom_point(color= "blue", size = 0.5) +
```

```
geom_smooth(formula = y ~ x, method = "lm", color= "orange")
```



Øverst i grafen kan man se ulempen ved å inkludere de med 2% høyest lønn, da det blir ingen punkter/observasjoner mellom 1,600,000.00 NOK og 3,000,000.00 NOK. Det som er interessant er jo at de på toppen er representert gjennom hele høydespekteret. Den oransje regresjonslinjen stiger på x- og y-aksen jo høyere man kommer på høydespekteret. Men den stiger ikke nevneverdig, så det vil ikke gi sterke indisier på at høyde påvirker inntekt i særlig grad. Men man observerer jo at inntekt er økende ved stigende høyde.

```
##
```

```
## Call:
```

```
## lm(formula = inntekt_nok ~ hoyde_cm, data = hoyde)
```

```
##
```

```
## Residuals:
```

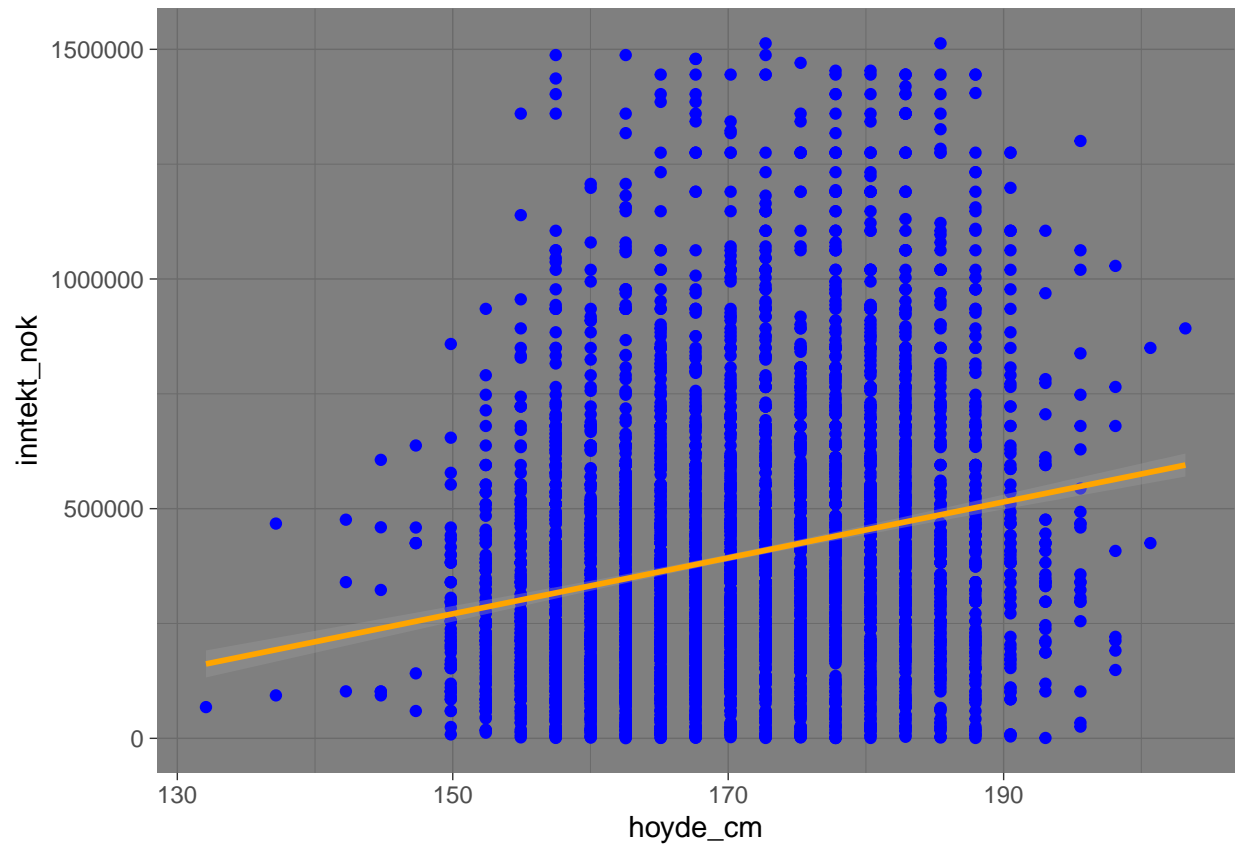
```
##      Min       1Q   Median       3Q      Max
```



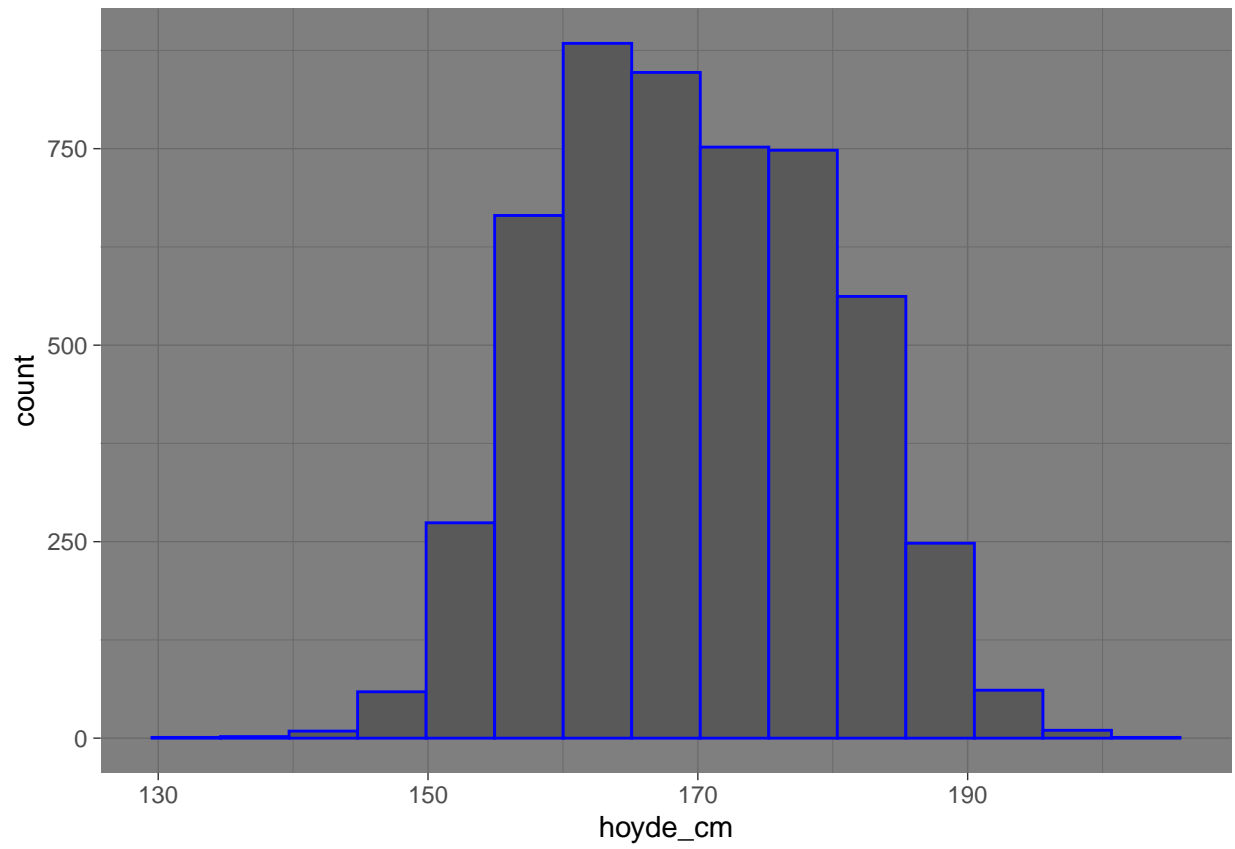
```
## -778460 -267842 -92589 126498 2727038
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1350548.5    91236.9  -14.80  <2e-16 ***
## hoyde_cm      9978.5      534.3   18.68  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 463700 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
```

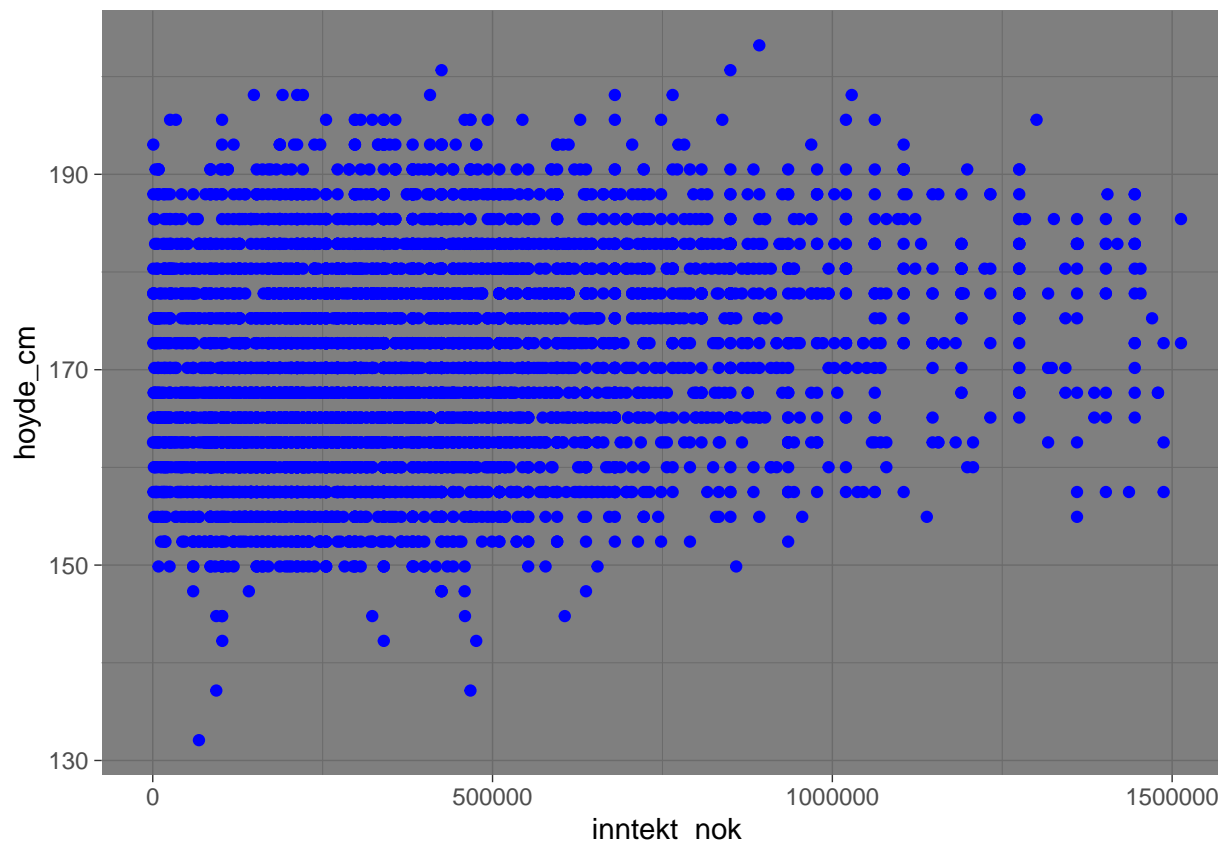
Denne regresjonsanalysen viser at 1cm ekstra høyde, vi gi 9,978.5 NOK ekstra i årsinntekt. Ved å se på R-squared, ser man også at høyde ikke har særlig innvirkning på inntekt, med en forklaringsgrad på bare 4.74%.

```
library(dplyr)
options(scipen = 999) #disabling scientific notation in R. vet ikke om vi skal ha med
hoyde = filter(hoyde, inntekt_nok != 0, inntekt_nok != "2922555") #brukt filter til å f
ggplot(hoyde, aes(hoyde_cm, inntekt_nok)) +
  geom_point(color = "blue") + theme_dark() +
  geom_smooth(formula = y ~ x, method = "lm", color = "orange")
```



```
ggplot(hoyde,aes(hoyde_cm))+geom_histogram(bins=15, color = "blue") + theme_dark()
```





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 (huskfilterfunksjonen fraTidyverse)

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 somførste argument til `huxreg()` kan du styre hva modellene skal hete, f.eks (gir også t-verdier istedenforstandard error)

*- Den endelige modellen skal testes for robusthet på et datasett uten de 2% høyeste inntektene og på etdatasett som i tillegg ikke inneholder observasjoner der inntekten er 0.

Disse modellene på redusert datasett teller med blant de 6.Minst en av model-

lene 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 `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

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

Referanser

::: {#refs} #jeg får ikke denne til å funke, kan du se Daniel? :::