

Er det høyde som bestemmer inntekt?

```
library(ggplot2)
library(tinytex)
library(tidyverse)
```

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

```
## v tibble  3.1.3    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()
```

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

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

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

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

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

```
library(car)
```

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

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

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

```
## The following object is masked from 'package:purrr':  
##  
##      some
```

```
library(carData)  
options(scipen = 999)
```

Kort innledning

Vi vil i denne oppgaven bruke datasettet *heights* for å undersøke riktigheten av påstanden om at høyde bestemmer inntekt. Undersøkelsen vil bli gjort gjennom analyser av ulike variabler og ved regresjonsanalyser. Resultatene vil bli fremstilt i ulike modeller, før vi vil avslutte med en konklusjon av spørsmålet “er det høyde som bestemmer inntekt?”.

Litteraturgjennomgang

Analyse

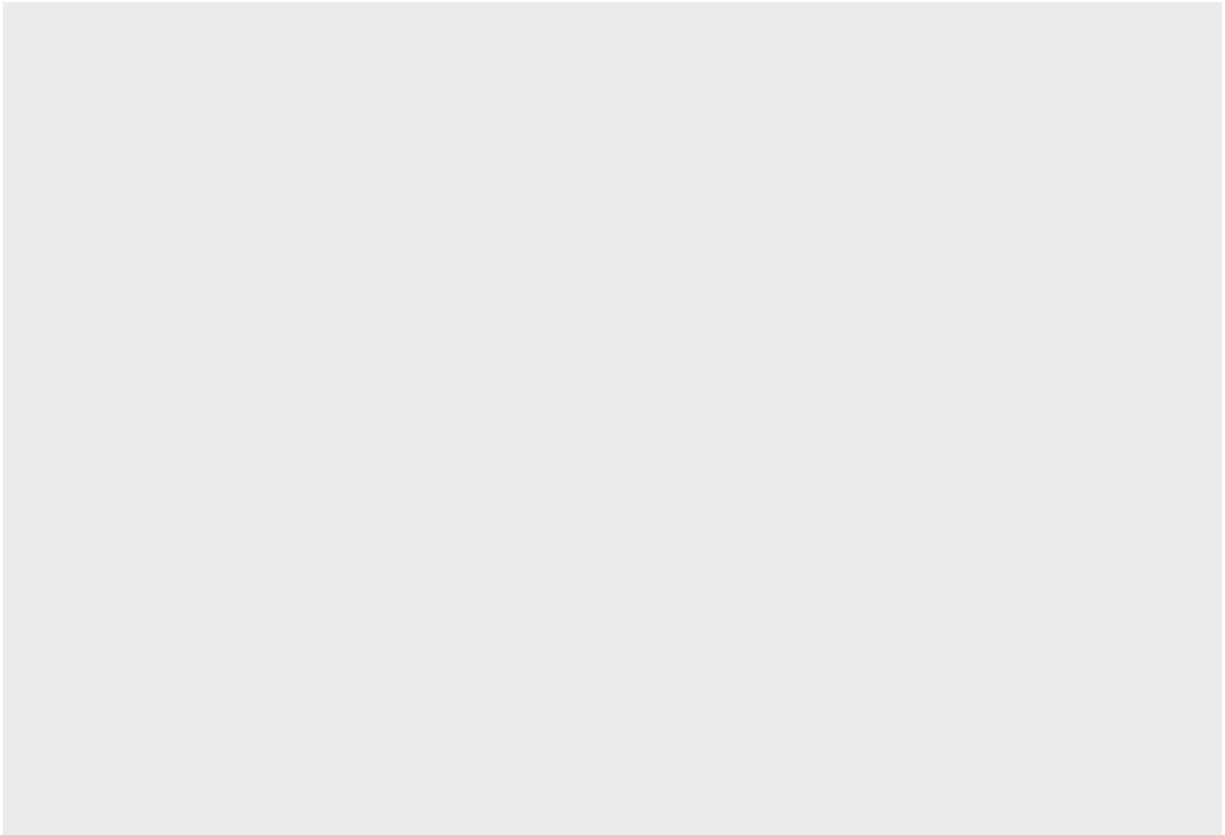
```
hoyde <- heights
```

Beskrivende statestikk

Datasettet ‘Heights’ som har fått navnet ‘hoyde’ består opprinnelig av åtte variabler og 7006 observasjoner. Basert på disse variablene og observasjonene skal det være mulig å konkludere om ens inntekt bestemmes av høyde. Variablene som er inkludert i datasettet er: * income - Basert på årlig inntekt, der topp to prosent er representert av gjennomsnittelig verdi av de to prosentene. * height - Høyden til subjektene som har deltatt målt i inch. * weight - Vekten til subjektene som har deltatt målt i pounds. * age - Alderen til subjektene som har deltatt er mellom 47 og 56. * martial - Subjektenes sivilstatus, om der gift eller skilt. * sex - Subjektenes kjønn, mann eller kvinne. * education - Subjektenes år med utdanning. * afqt - Subjektenes poengsum i prosent på ‘Armed Forces Qualification Test’

EDA

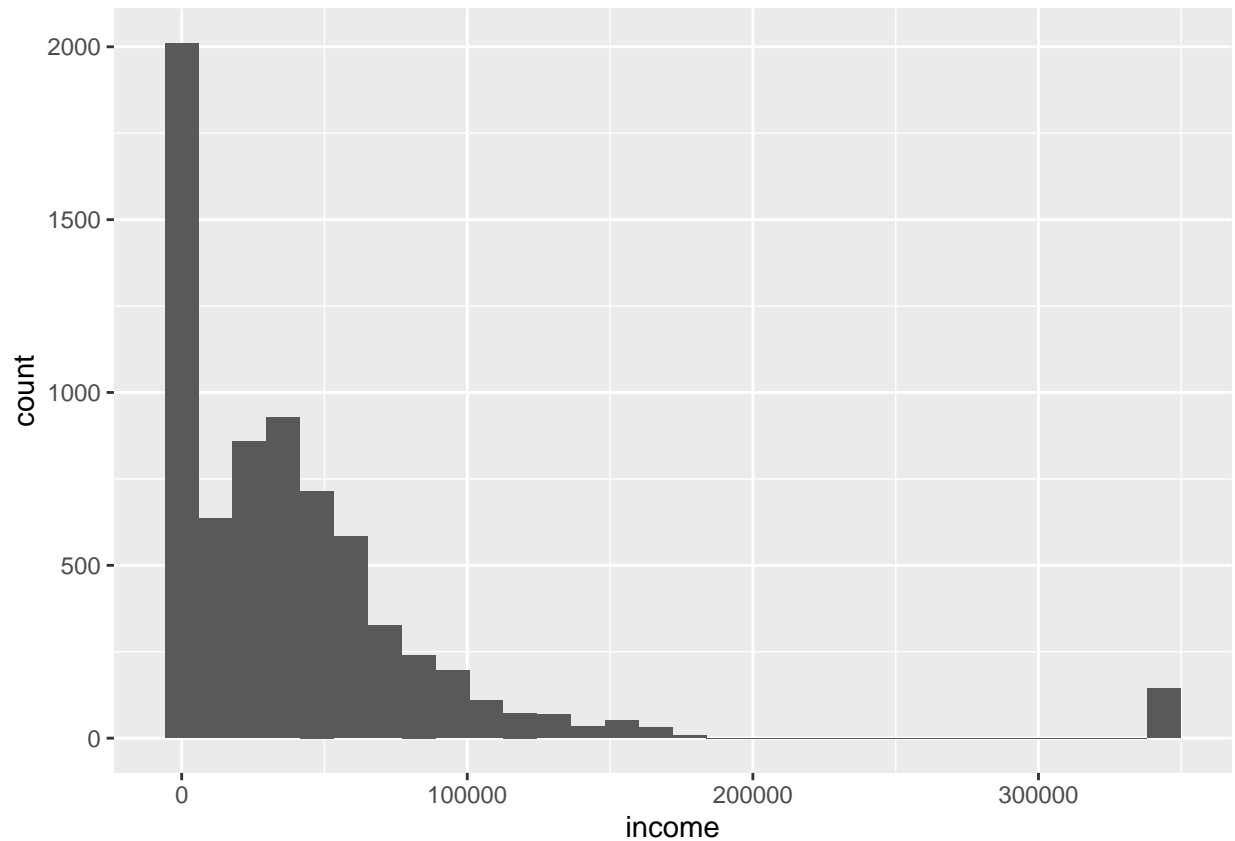
```
ggplot(data = hoyde)
```



Lag et histogram av variablen *income*

```
ggplot(data = hoyde, aes(income)) +  
  geom_histogram()
```

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



Forklaringen på utliggerne langt til høyre

På modellen ser man at utliggeren som i dette tilfellet er ekstremalpunktet ligger langt til høyre. Dette er på grunn av at de har blitt regnet ut et gjennomsnitt av topp to prosentene av inntektene.

Personer uten inntekt inkludert i datasettet?

```
sum(hoyde$income == 0)
```

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

Det er altså 1740 av 7006 personer som har null i inntekt.

Regresjonsanalyse

```
hoyde <- heights
```

Dollar til norske kroner

```
hoyde <- hoyde %>%  
  mutate(inntekt = income * 8.42)
```

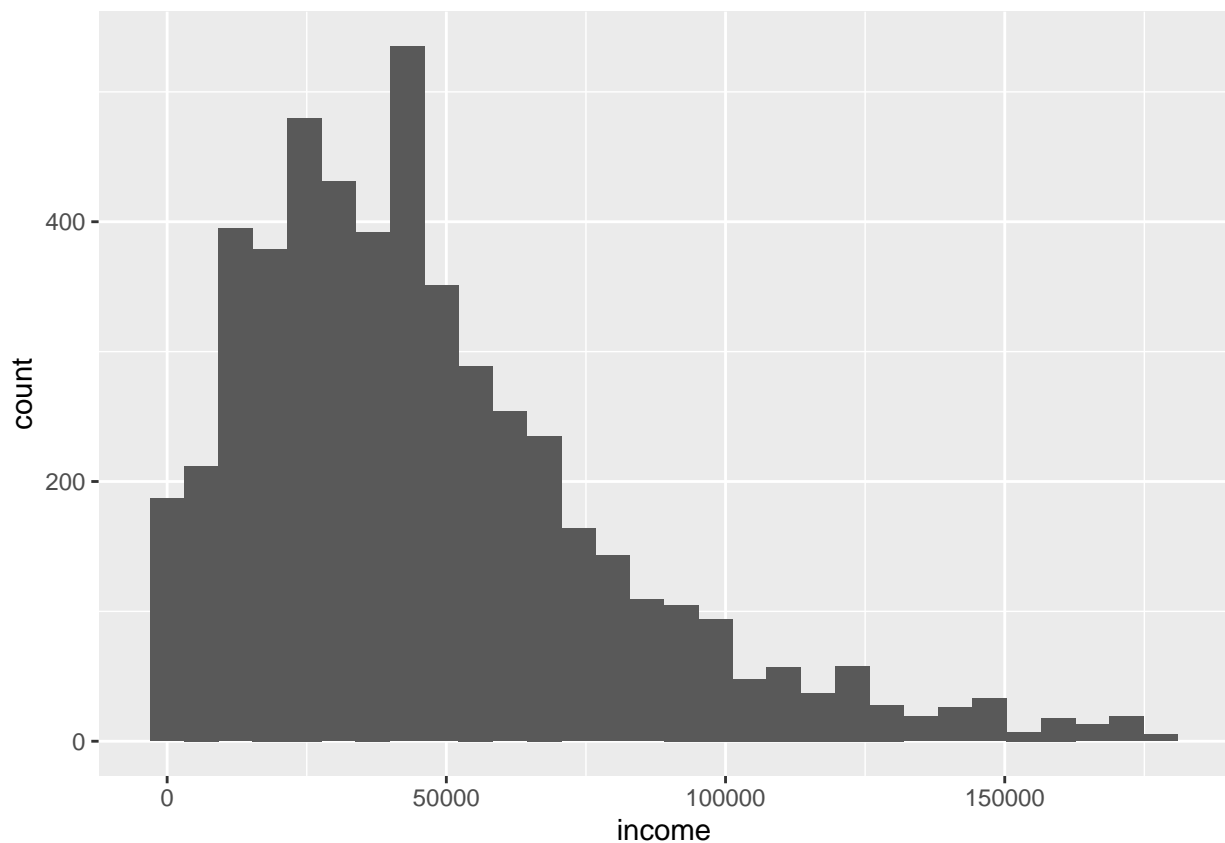
Vi gjør Amerikanske dollar om til norske kroner. Valutakursen ligger på 8.42.

Redusert datasett

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

```
ggplot(data = hoyde_begr, aes(income)) +  
  geom_histogram()
```

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



Lag to nye variabler + lag ny variabel “bmi”

```

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

```

Forenklet utgave av variabelen “Marital”

```

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

```

Minst 6 modeller

```

modell1 <- "inntekt ~ hoyde_cm"
lm1 <- lm(modell1, data = hoyde, subset = complete.cases(hoyde))
summary(lm1)

```

```

##
## Call:
## lm(formula = modell1, 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

```

```

#trenger vi å ta med dette??
-1337837.4 + (9884.6 *173)

```

```
## [1] 372198.4
```

```
-1337837.4 + (9884.6 *161)
```

```
## [1] 253583.2
```

```
Model2 <- "inntekt ~ hoyde_cm + vekt_kg"
lm2 <- lm(Model2, data = hoyde, subset = complete.cases(hoyde))
summary(lm2)
```

```
##
## Call:
## lm(formula = Model2, 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
```

```
#Må vi ta med dette??
-697374.7 + (6735.7*173) + (-691.5*70)
```

```
## [1] 419496.4
```

```
Model3 <- "inntekt ~ hoyde_cm + vekt_kg + BMI"
lm3 <- lm(Model3, data = hoyde, subset = complete.cases(hoyde))
summary(lm3)
```

```
##
## Call:
## lm(formula = Model3, data = hoyde, subset = complete.cases(hoyde))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -810031 -262631  -92854   124005  2705975
##
## Coefficients:
##              Estimate Std. Error t value      Pr(>|t|)
## (Intercept) -1286218    132300  -9.722 < 0.0000000000000002 ***
## hoyde_cm       9413      1184    7.951 0.000000000000000216 ***
## vekt_kg        2039      1803    1.131    0.258
## BMI          -538714    268709  -2.005    0.045 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 466500 on 6641 degrees of freedom
## Multiple R-squared:  0.05115,    Adjusted R-squared:  0.05073
## F-statistic: 119.3 on 3 and 6641 DF,  p-value: < 0.00000000000000022
```

```
# Slett denne
summary(hoyde_begr)
```

```
##      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
##      inntekt
## Min.   :    378.9
## 1st Qu.: 193660.0
## Median : 336800.0
## Mean   : 393646.3
## 3rd Qu.: 522040.0
## Max.   :1498760.0
##
```

Huxreg

```
huxreg (list
  ("model1" = lm1, "model2" = lm2, "model3" = lm3),
  error_format = "[{statistic}]",
  note = "Regresjonstabell 3: {stars}.T statistics in brackets."
)
```

Interaksjon

```
# Riktig, fjern denne linjen.
Model14 <- "inntekt ~ sex*hoyde_cm + vekt_kg + I(vekt_kg^2) + BMI + I(BMI^2)"
```


	model1	model2	model3
(Intercept)	-1361000.990 *** [-14.413]	-1466873.555 *** [-15.139]	-1286217.908 *** [-9.722]
hoyde_cm	10047.860 *** [18.175]	11430.259 *** [18.308]	9413.347 *** [7.951]
vekt_kg		-1518.381 *** [-4.737]	2039.260 [1.131]
BMI			-538714.354 * [-2.005]
N	6645	6645	6645
R2	0.047	0.051	0.051
logLik	-96177.211	-96166.004	-96163.994
AIC	192360.423	192340.008	192337.987

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

```
lm4 <- lm(Model4, data = hoyde)
summary(lm4)
```

```
##
## Call:
## lm(formula = Model4, data = hoyde)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -838791 -246526  -91148  127011 2671450
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1483862.35    298212.30  -4.976 0.000000665 ***
## sexfemale      1013185.90    267052.10   3.794  0.00015 ***
## hoyde_cm        9605.99     2224.53   4.318 0.000015950 ***
## vekt_kg         7666.44     4523.70   1.695  0.09017 .
## I(vekt_kg^2)    -41.62       15.13  -2.752  0.00594 **
## BMI            -574174.60    743591.63  -0.772  0.44004
## I(BMI^2)        508980.64    335754.40   1.516  0.12958
## sexfemale:hoyde_cm -6597.83     1566.60  -4.212 0.000025683 ***
## ---
## 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:  0.06121,    Adjusted R-squared:  0.06026
## F-statistic:  64.3 on 7 and 6903 DF,  p-value: < 0.00000000000000022
```

```
Model5 <- "inntekt ~ sex*(hoyde_cm + vekt_kg + I(vekt_kg^2)) + BMI + I(BMI^2)"
lm5 <- lm(Model5, data = hoyde)
summary(lm4)
```

```
##
## Call:
## lm(formula = Model4, data = hoyde)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -838791 -246526  -91148   127011  2671450
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1483862.35    298212.30  -4.976 0.000000665 ***
## sexfemale      1013185.90    267052.10   3.794   0.00015 ***
## hoyde_cm        9605.99     2224.53   4.318 0.000015950 ***
## vekt_kg         7666.44     4523.70   1.695   0.09017 .
## I(vekt_kg^2)    -41.62       15.13  -2.752   0.00594 **
## BMI            -574174.60    743591.63  -0.772   0.44004
## I(BMI^2)        508980.64    335754.40   1.516   0.12958
## sexfemale:hoyde_cm -6597.83     1566.60  -4.212 0.000025683 ***
## ---
## 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:  0.06121,    Adjusted R-squared:  0.06026
## F-statistic:  64.3 on 7 and 6903 DF,  p-value: < 0.00000000000000022
```

Test av koeffisientene

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

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
6.90e+03	1.46e+15				
6.9e+03	1.45e+15	2	1.37e+13	32.5	8.64e-15

Begrensing

```
Model6 <- "inntekt ~ sex*hoyde_cm + vekt_kg + I(vekt_kg^2) + BMI + I(BMI^2)"
lm6 <- lm(Model6, data = hoyde)
summary(lm6)
```

```
##
## Call:
## lm(formula = Model6, data = hoyde)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -838791 -246526  -91148  127011  2671450
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1483862.35   298212.30  -4.976 0.000000665 ***
## sexfemale     1013185.90   267052.10   3.794   0.00015 ***
## hoyde_cm       9605.99     2224.53   4.318 0.000015950 ***
## vekt_kg        7666.44     4523.70   1.695   0.09017 .
## I(vekt_kg^2)    -41.62       15.13  -2.752   0.00594 **
## BMI           -574174.60   743591.63  -0.772   0.44004
## I(BMI^2)        508980.64   335754.40   1.516   0.12958
## sexfemale:hoyde_cm -6597.83    1566.60  -4.212 0.000025683 ***
## ---
## 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:  0.06121, Adjusted R-squared:  0.06026
## F-statistic: 64.3 on 7 and 6903 DF, p-value: < 0.00000000000000022
```

Legge til residualene til datasettet

```
# Bruk verdiene fra begrenset datasett
hoyde_begr <- hoyde %>%
  add_residuals(lm6)
hoyde_begr %>%
  head(n=10)
```

Plot av 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
```

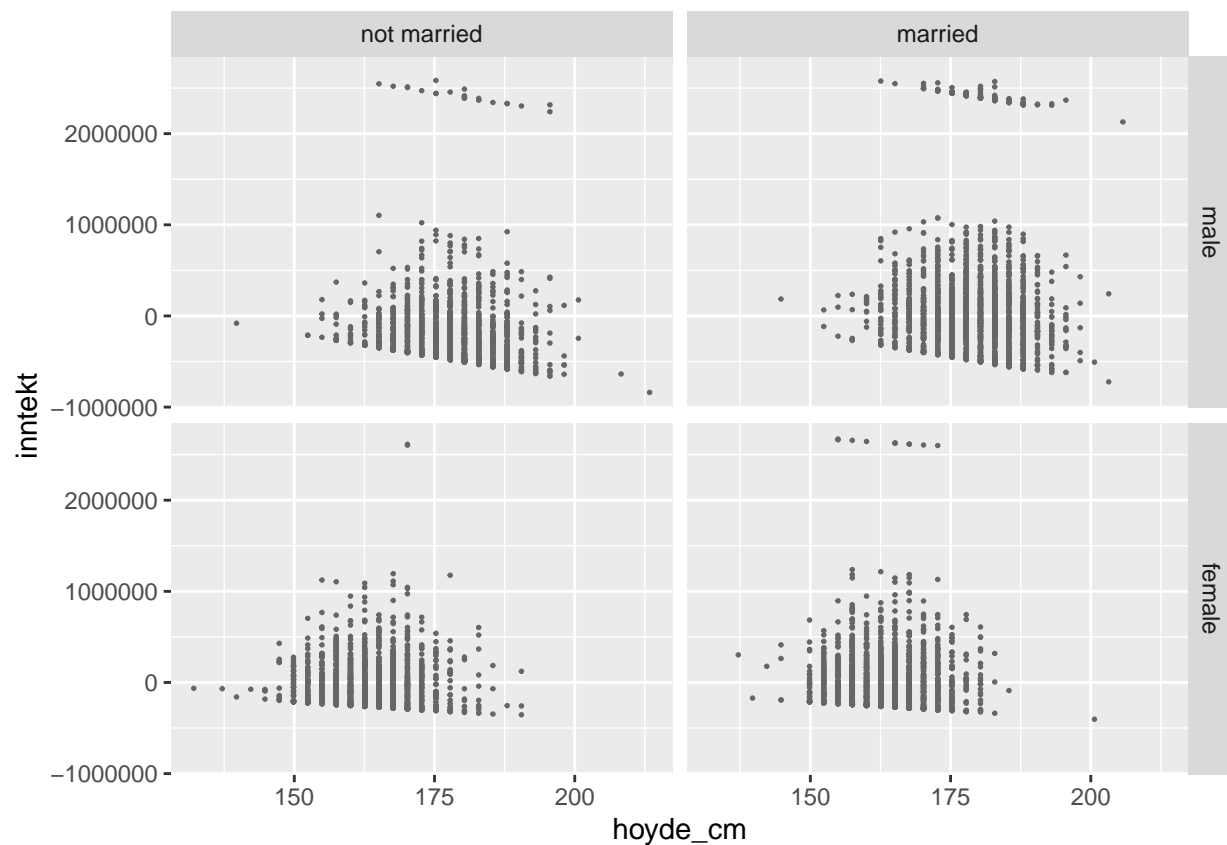
height	weight	age	marital	sex	education	afqt	inntekt	hoyde_cm	vekt_kg	BMI	married
60	155	53	married	female	13	6.84	1.6e+05	152	70.4	0.213	TRUE
70	156	51	married	female	10	49.4	2.95e+05	178	70.8	0.159	TRUE
65	195	52	married	male	16	99.4	8.84e+05	165	88.5	0.288	TRUE
63	197	54	married	female	14	44	3.37e+05	160	89.4	0.312	TRUE
66	190	49	married	male	14	59.7	6.32e+05	168	86.3	0.265	TRUE
68	200	49	divorced	female	18	98.8	8.59e+05	173	90.8	0.276	FALSE
74	225	48	married	male	16	82.3	0	188	102	0.295	TRUE
64	160	54	divorced	female	12	50.3	5.89e+05	163	72.6	0.2	FALSE
69	162	55	divorced	male	12	89.7	5.05e+05	175	73.5	0.176	FALSE
69	194	54	divorced	male	13	96	1.26e+06	175	88.1	0.253	FALSE

```

) +
facet_grid(sex ~ factor(married, labels = c("not married", "married")))

```

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



Konklusjon