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

Judge og Cable hevder i deres akademiske artikkel fra 2004 at høyde kan skape fordeler i en rekke viktige aspekter i både liv og i karriere (Judge and Cable 2004). Gjennom deres evalueringer med faktorene kjønn, vekt og alder legger de frem en teoretisk model av forholdet mellom høyde of inntekt (Judge and Cable 2004). Modellen forfatterne fremlegger viser blant annet at høyde er relatert til både utførelse og selvtillit (Judge and Cable 2004). De fremlegger videre at høyde i relasjon med inntekt eller "karrieresuksess" viste en større relasjon for menn enn for kvinner (Judge and Cable 2004). Siste del av forskningen utført av forfatterne, fremlegger at det foreligger en tydelig sammenheng mellom høyde og inntekt - hvor de tidligere nevnte faktorene kjønn, vekt og alder er hensyntatt (Judge and Cable 2004). Eli Kvisvik (2008a) hevder også i sin statistiske analyse fra 2008 at det eksisterer en sammenheng mellom nettopp høyde og inntekt. Hun utrykker videre at "den høyeste fjerdedelen av den amerikanske befolkningen tjener 9-10% mer enn den laveste fjerdedelen (. . .)," hvor forskere begrunner dette med andre psykologiske årsaker som Judge and Cable (2004) også referer til, nemlig selvtillit (Eli Kvisvik 2008b).

Ved å bruke både dattasettet ***heights*** og gjennom argumentene Eli Kvisvik (2008a) og Judge and Cable (2004) legger til grunn, vil vi teste påstandene gjennom statistikken som utføres videre i oppgaven. Ved bruk av datasettet og dets variabler vil vi kunne se flere aspekter relatert til høyde og inntekt hvor variablene vekt, kjønn, alder og utdannelse også vil være inkludert. Konklusjonen av oppgaven vil være basert på utfallet av de framlagte data, hvor vi naturligvis vil se om våre resultater støtter Judge and Cable (2004) og Eli Kvisvik (2008a) resultater.

```
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'

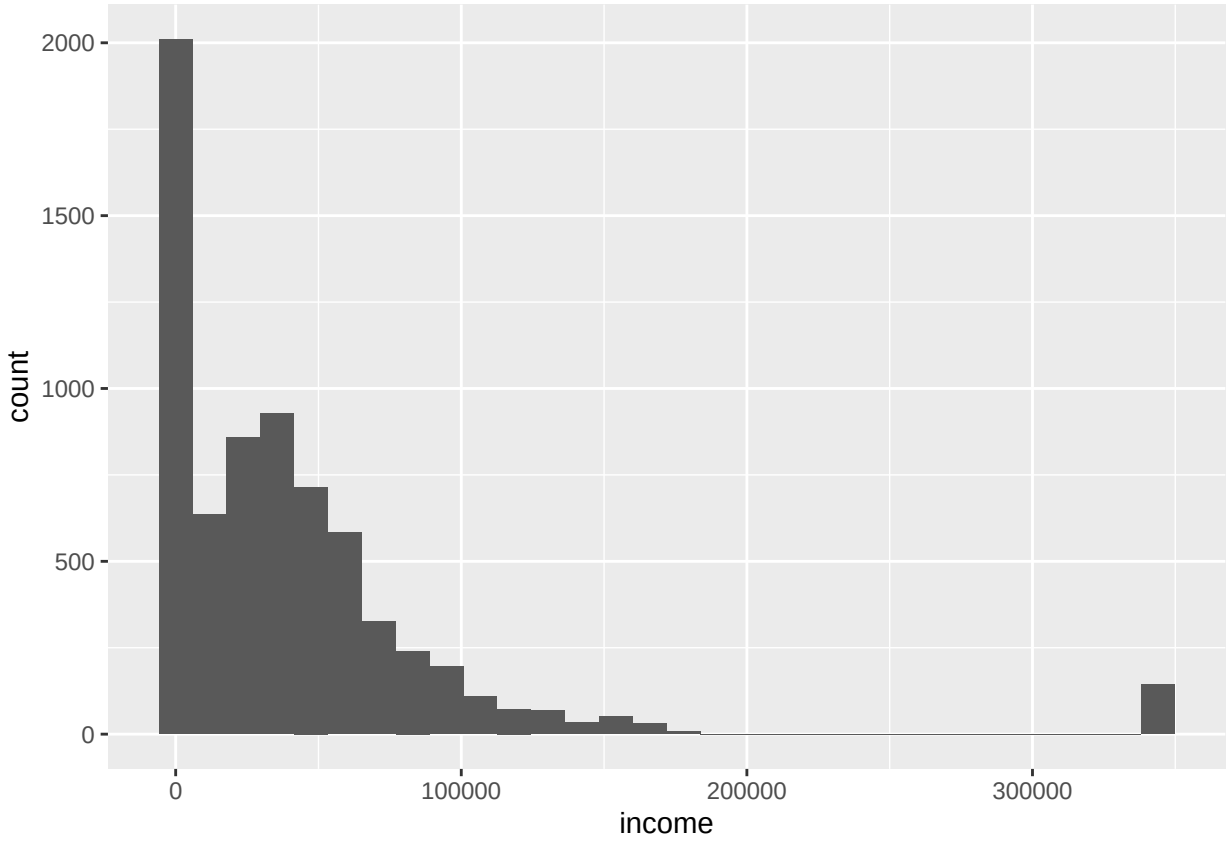
EDA

```
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

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

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

Redusert datasett

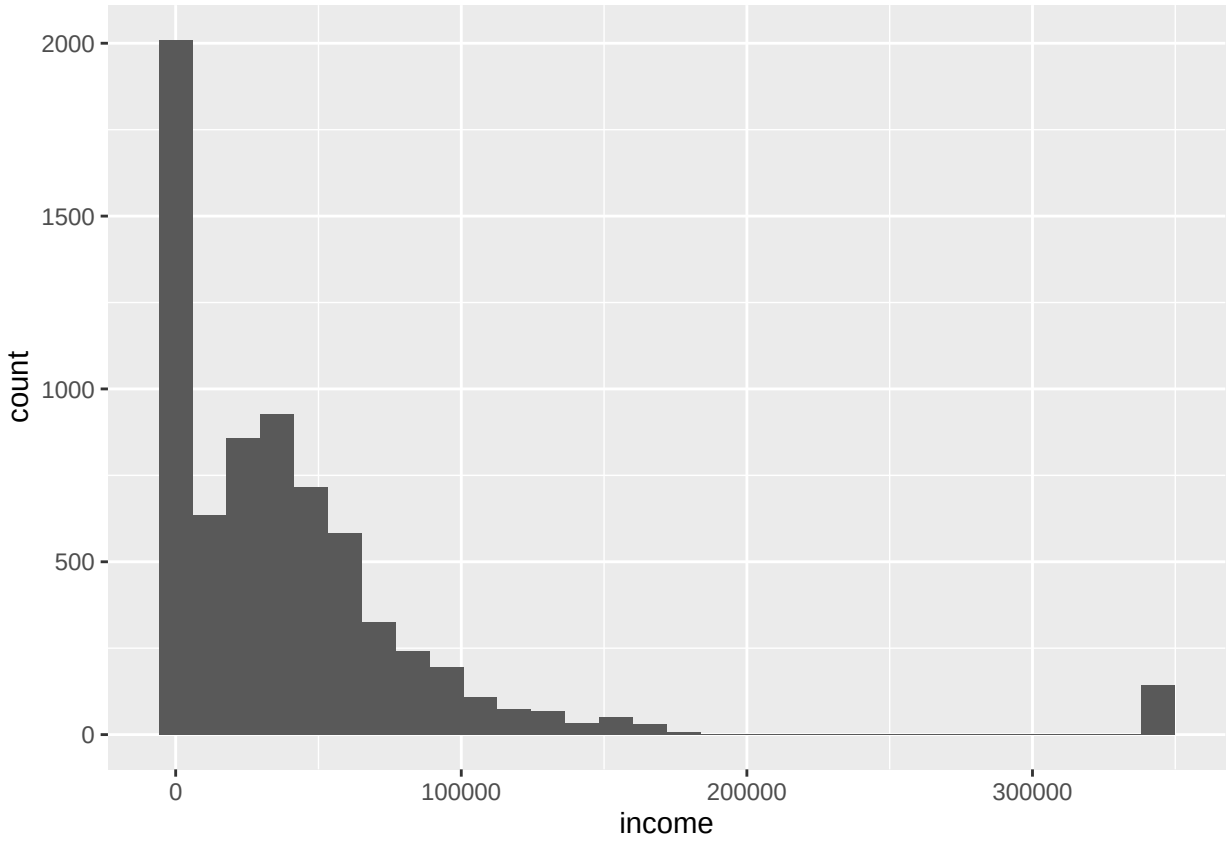
Vi begrenser datasettene *"hoyde"* og *"hoyde_begr"* ved å utelate personer med topp 2 prosent inntekt og uten inntekt.

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

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

```
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"

Datasettene *"hoyde"* og *"hoyde_begr"* får tre nye variabler ved å gjøre det om til meterisk standard.

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

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

Forenklet utgave av variabelen "Martial"

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

Minst 6 modeller

Modell 1

Modell 1 viser sammenhengen mellom variablene *inntekt* og *hoyde_cm*.

```
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
```

```
# Eksempel 1
-1361001.0 + (10047.9 *173)
```

```
## [1] 377285.7
```

```
-1361001.0 + (10047.9 *161)
```

```
## [1] 256710.9
```

En person som er 1,73 meter høy vil tjene 377285.7 NOK og en som er 1,61 meter høy vil tjene 256710.9 NOK. Resultatet viser at den som er høyere tjener mer.

Modell 2

Modell 2 viser sammenhengen mellom variabelene *inntekt*, *hoyde_cm* og *vekt_kg*.

```
Modell2 <- "inntekt ~ hoyde_cm + vekt_kg"
lm2 <- lm(Modell2, data = hoyde, subset = complete.cases(hoyde))
summary(lm2)
```

```
##
## Call:
## lm(formula = Modell2, 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
```

```
# Eksempel 2
-1466873.6 + (11430.3*173) + (-691.5*70)
```

```
## [1] 462163.3
```

```
-1466873.6 + (11430.3*161) + (-691.5*70)
```

```
## [1] 324999.7
```

Eksemplet viser at hoyde gir økning i inntekt og vekt gir reduksjon i inntekt. Et samlet resulat av funksjonen gir lønnsøkning og personen som er høyere enn den andre tjener mer.

Modell 3

Modell 3 viser sammenhengen mellom variabelene *inntekt*, *hoyde_cm*, *vekt_kg* og *BMI*.

```
Modell3 <- "inntekt ~ hoyde_cm + vekt_kg + BMI"
lm3 <- lm(Modell3, data = hoyde, subset = complete.cases(hoyde))
summary(lm3)
```

```
##
## Call:
## lm(formula = Modell3, 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     132301  -9.722 < 0.0000000000000002 ***
## hoyde_cm        9413       1184   7.951  0.00000000000000216 ***
## 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
```

Huxreg

```
huxreg (list
  ("Modell1" = lm1, "Modell2" = lm2, "Modell3" = lm3),
       error_format = "[{statistic}]",
       note = "Regresjonstabell 3: {stars}.T statistics in brackets."
)
```

Interaksjon

Modell 4

Modell 4 fire inneholder en interaksjon mellom variabelene *"hoyde_cm"* og *"sex"*.

	Modell1	Modell2	Modell3
(Intercept)	-1361000.990 ***	-1466873.555 ***	-1286217.908 ***
	[-14.413]	[-15.139]	[-9.722]
hoyde_cm	10047.860 ***	11430.259 ***	9413.347 ***
	[18.175]	[18.308]	[7.951]
vekt_kg		-1518.381 ***	2039.260
		[-4.737]	[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.

```
Modell4 <- "inntekt ~ sex*hoyde_cm + vekt_kg + I(vekt_kg^2) + BMI + I(BMI^2)"
lm4 <- lm(Modell4, data = hoyde)
summary(lm4)
```

```
##
## Call:
## lm(formula = Modell4, 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
```

Modell 5

Modell 5 inneholder interaskjon mellom flere variabler.

```
Modell5 <- "inntekt ~ sex*(hoyde_cm + vekt_kg + I(vekt_kg^2)) + BMI + I(BMI^2)"
lm5 <- lm(Modell5, data = hoyde)
summary(lm4)
```

```
##
## Call:
## lm(formula = Modell4, 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)
6.9e+03	1.46e+15				
6.9e+03	1.45e+15	2	1.37e+13	32.5	8.64e-15

Begrensing

Modell 6

```
Modell6 <- "inntekt ~ sex*hoyde_cm + vekt_kg + I(vekt_kg^2) + BMI + I(BMI^2)"
lm6 <- lm(Modell6, data = hoyde_begr)
summary(lm6)
```

```
##
## Call:
## lm(formula = Modell6, data = hoyde_begr)
##
## 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 resiudalene til datasettet

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

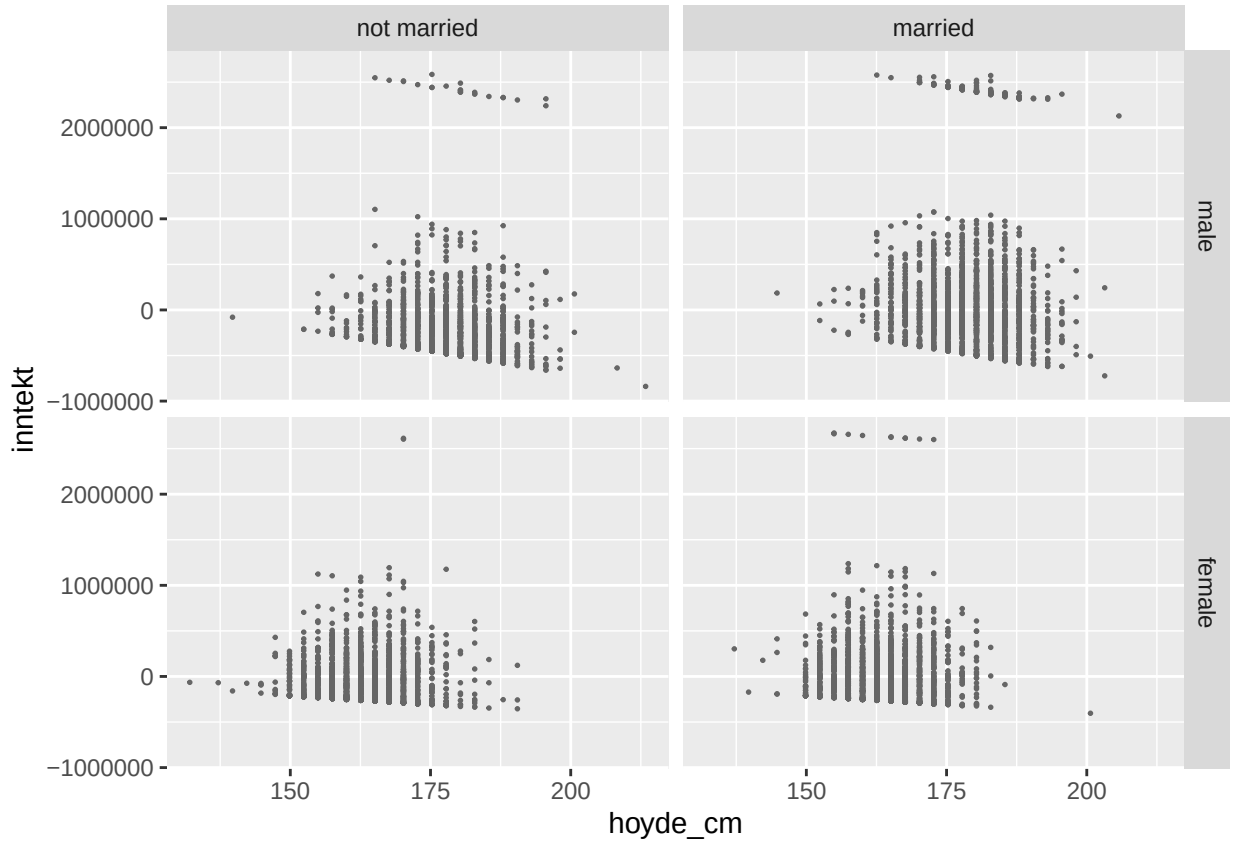
Plot av samtilige observasjoner

```
ggplot(data = hoyde_begr, mapping = aes(x = hoyde_cm, y = inntekt)) +
  geom_point(
```

eight	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

```
    data = hoyde_begr,
    mapping = aes(x = hoyde_cm, y = resid),
    colour = "grey40",
    size = 0.3
    ) +
  facet_grid(sex ~ factor(married, labels = c("not married", "married")))
```

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



Konklusjon

Kilder

Eli Kvisvik. 2008a. "Sammenhengen Mellom høyde Og Inntekt : Kvisvik :consulting." http://kvisvikconsulting.no/sammenhengen-mellom-hoyde-og-inntekt/.

———. 2008b. "Sammenhengen Mellom høyde Og Inntekt : Kvisvik :consulting." http://kvisvikconsulting.no/sammenhengen-mellom-hoyde-og-inntekt/.

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.