Er det høyde som bestemmer inntekt?

Kort innledning

Vi vil i denne oppgaven bruke datasettet ***heights***[1] 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 and Cable (2004a) hevder at høyde kan skape fordeler i en rekke viktige aspekter i både liv og i karrière. Gjennom deres evalueringer med faktorene kjønn, vekt og alder legger de frem en teoretisk model av forholdet mellom høyde of inntekt. Modellen forfatterne fremlegger viser blant annet at høyde er relatert til både utførelse og selvtillit. De fremlegger videre at høyde i relasjon med inntekt eller «karriéresuksess» viste en større relasjon for menn enn for kvinner. 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. Eli Kvisvik (2008a) hevder også i sin statistiske analyse fra 2008 at det eksisterer en sammenheng mellom nettopp høyde og inntekt. Hun uttrykker 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 (2004b) også referer til, nemlig selvtillit (Eli Kvisvik 2008b).

Ved å bruke både datasettet *heights* og gjennom argumentene Eli Kvisvik (2008a) og Judge and Cable (2004b) legger til grunn, vil vi teste påstandene gjennom statistiske tester 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 resultatene fra Judge and Cable (2004b) og Eli Kvisvik (2008a).

```
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:

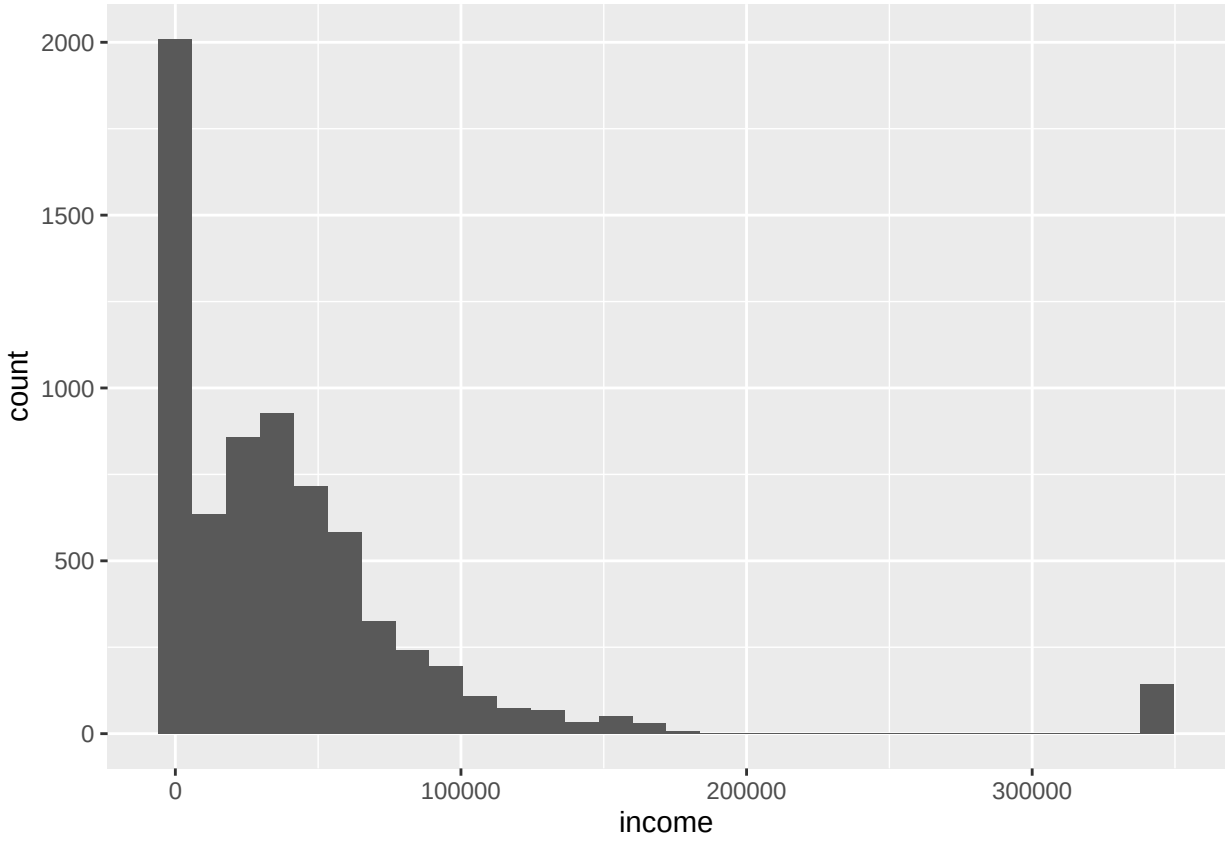
- Income - Basert på årlig inntekt, der topp to prosent er representert av gjennomsnittelig verdi av de to prosentene.
- Height - Høyden til subjektene som har deltatt målt i inch.
- Weight - Vekten til subjektene som har deltatt målt i pounds.
- Age - Alderen til subjektene som har deltatt er mellom 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'

[1] Fra R-pakken *modelr* (Wickham 2020) kjørt under statistikk systemet R (R Core Team 2021).

EDA

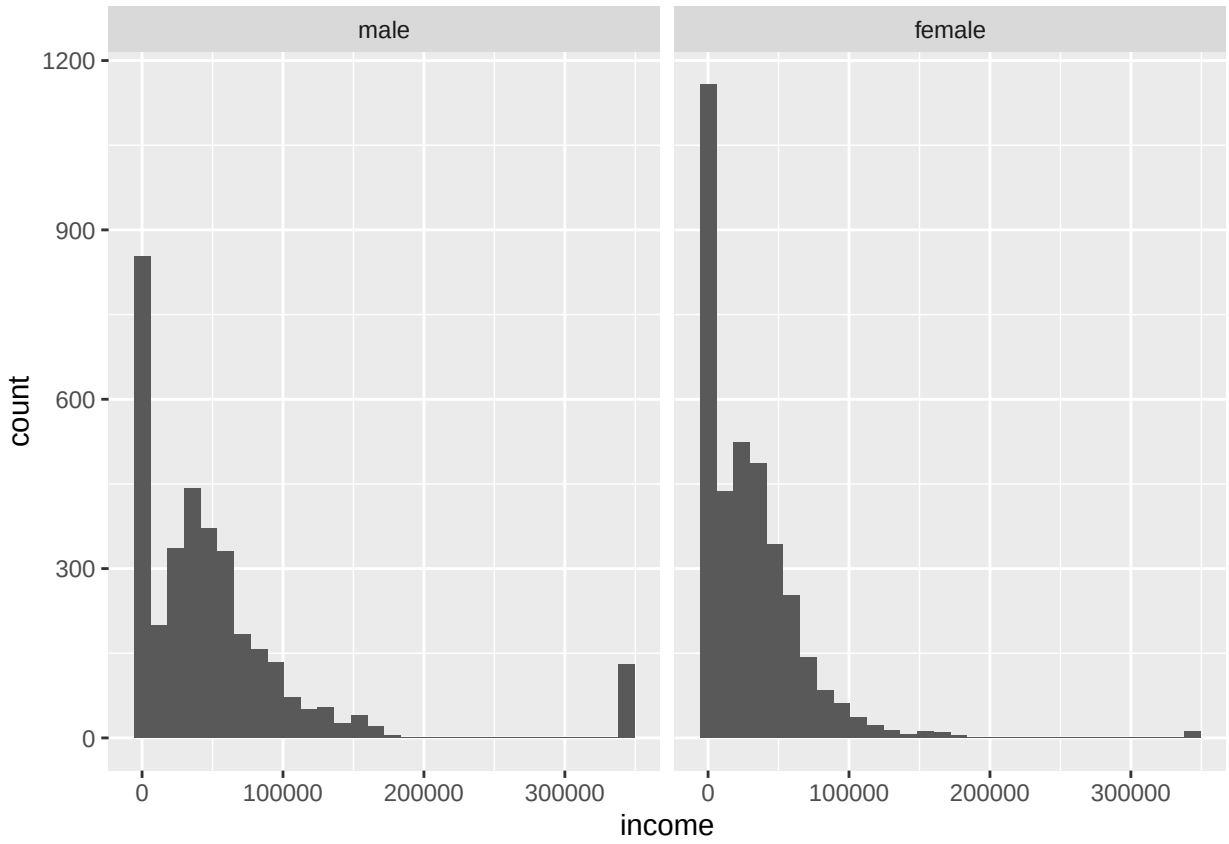
Histogram av variabelen *income*

```
ggplot(data = hoyde, aes(income)) +
  # satt inn bins=30 for å slippe warning
  geom_histogram(bins=30)
```



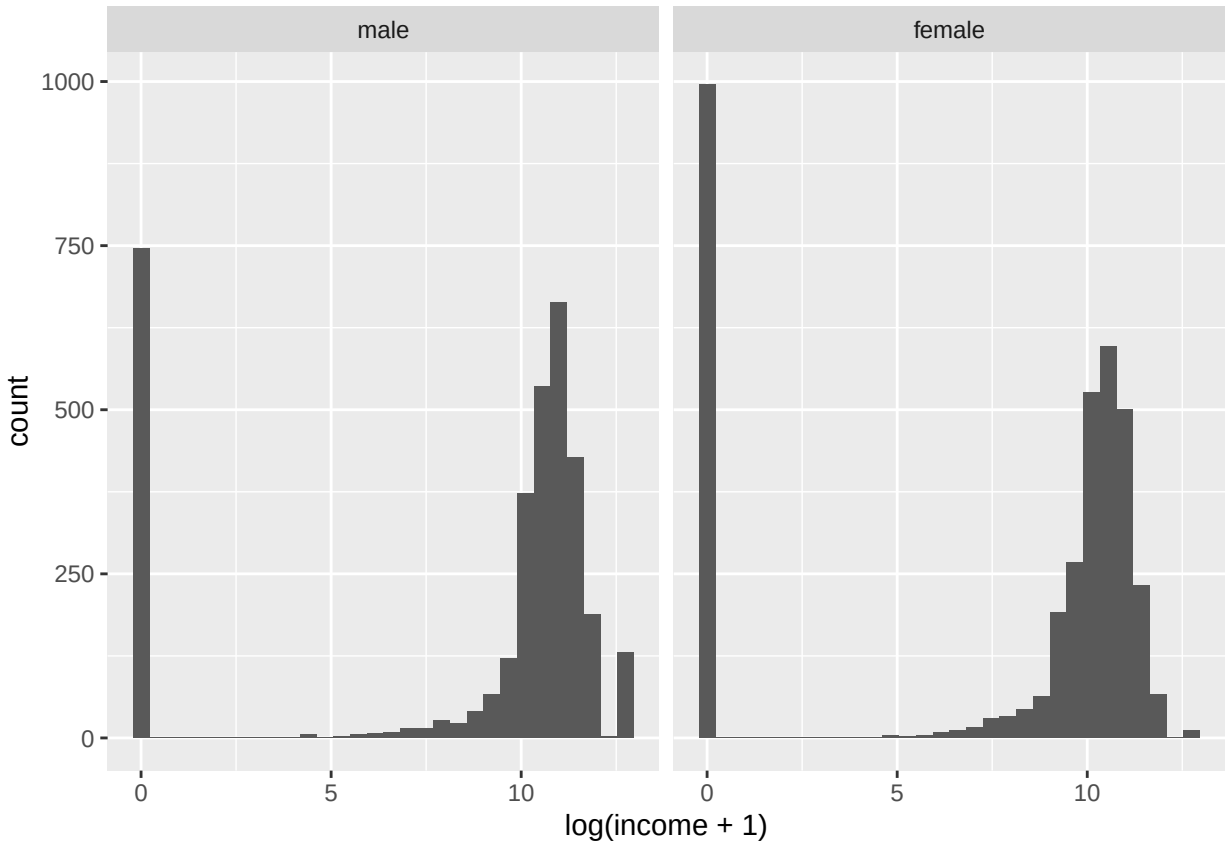
Histogram av *income* fordelt på kjønn

```
# For gjort å lage facet på kjønn
ggplot(data = hoyde, aes(income)) +
  facet_wrap(~sex) +
  geom_histogram(bins=30)
```



Histogram av *ln(income+1)* fordelt på kjønn

```
# For gjort å lage ln tranformasjon og facet på kjønn
ggplot(data = hoyde, aes(log(income+1))) +
  facet_wrap(~sex) +
  geom_histogram(bins=30)
```



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

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.

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

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

Lag to nye variabler + lag ny variabel *bmi*

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

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

```r
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*

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

Seks ulike regresjonsmodeller

Modell 1

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

```r
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   <2e-16 ***
```

```
## hoyde_cm       10047.9      552.8   18.18   <2e-16 ***
## ---
## 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: < 2.2e-16
```

```
# 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 variablene *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    < 2e-16 ***
## hoyde_cm       11430.3      624.3  18.308    < 2e-16 ***
## 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: < 2.2e-16
```

```
# 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 resultat av funksjonen gir lønnsøkning og personen som er høyere enn den andre tjener mer.

Modell 3

Modell 3 viser sammenhengen mellom variablene *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     132300  -9.722  < 2e-16 ***
## hoyde_cm        9413       1184   7.951 2.16e-15 ***
## 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: < 2.2e-16
```

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 variablene *"hoyde_cm"* og *"sex"*.

```
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
```

	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.

```
##
## 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: < 2.2e-16
```

Modell 5

Modell 5 inneholder interaksjon mellom flere variabler.

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

```
##
```

```
## Call:
## lm(formula = Modell5, data = hoyde)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -817997 -247845  -91322  125631 2682274
##
## Coefficients:
##                            Estimate  Std. Error t value    Pr(>|t|)
## (Intercept)            -1735256.55   352726.42  -4.920 0.000000888 ***
## sexfemale               1235228.84   286424.66   4.313 0.000016361 ***
## hoyde_cm                   9211.45     2510.05   3.670    0.000245 ***
## vekt_kg                   14077.13     5211.47   2.701    0.006926 **
## I(vekt_kg^2)                -74.61       20.58  -3.625    0.000291 ***
## BMI                      -541074.91   859723.71  -0.629    0.529135
## I(BMI^2)                  651407.12   340295.70   1.914    0.055631 .
## sexfemale:hoyde_cm         -5387.86     1635.05  -3.295    0.000988 ***
## sexfemale:vekt_kg          -8222.31     3379.45  -2.433    0.014998 *
## sexfemale:I(vekt_kg^2)        35.98       17.81   2.021    0.043359 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 458200 on 6901 degrees of freedom
##   (95 observations deleted due to missingness)
## Multiple R-squared:  0.06222,    Adjusted R-squared:  0.06099
## F-statistic: 50.87 on 9 and 6901 DF,  p-value: < 2.2e-16
```

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

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: < 2.2e-16
```

Legge til residualene til datasettet

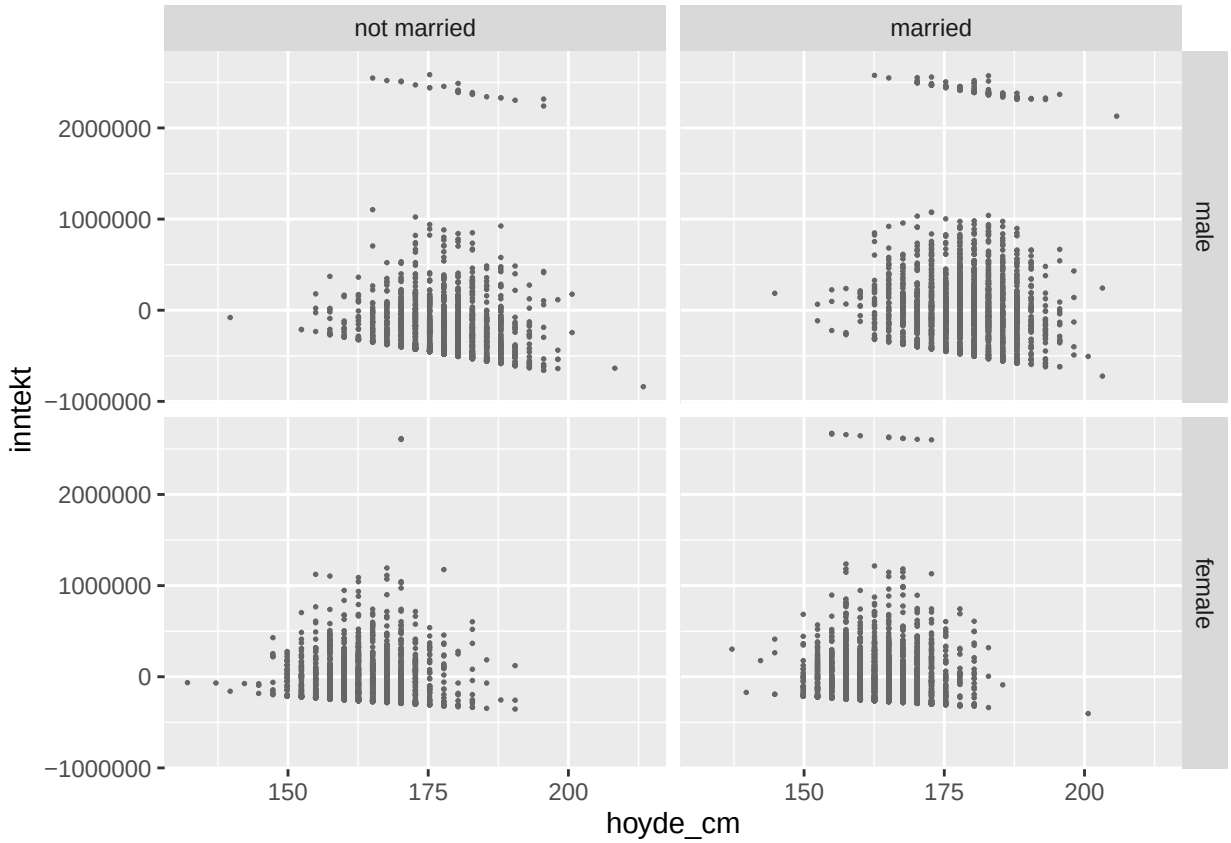
```
# Bruk verdiene fra begrenset datasett
hoyde_begr <- hoyde %>%
  add_residuals(lm6)
hoyde_begr %>%
  head(n=10)
```

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

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

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



Modell 7 på redusert datasett

```
mod7 <- 'sqrt(income) ~ sex*(education + afqt + age + married)'
lm7 <- hoyde %>%
  filter(income >0, income < 300000) %>%
  filter(complete.cases(.) == TRUE) %>%
  lm(mod7, data = .)
```

```
summary(lm7)
```

```
##
## Call:
## lm(formula = mod7, data = .)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -279.819  -39.601    0.899   41.711  230.105
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            106.68768   32.47402   3.285  0.00103 **
## sexfemale              -80.33900   45.36066  -1.771  0.07660 .
## education                7.97504    0.68254  11.684  < 2e-16 ***
## afqt                     0.66405    0.05793  11.464  < 2e-16 ***
## age                     -0.85418    0.61394  -1.391  0.16419
## marriedTRUE             34.35031    2.85321  12.039  < 2e-16 ***
## sexfemale:education      0.91711    0.91753   1.000  0.31758
## sexfemale:afqt          -0.20054    0.08210  -2.443  0.01462 *
## sexfemale:age            1.20999    0.86016   1.407  0.15958
## sexfemale:marriedTRUE -37.89356    3.97439  -9.534  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 66.72 on 4861 degrees of freedom
## Multiple R-squared:  0.2653, Adjusted R-squared:  0.2639
## F-statistic:   195 on 9 and 4861 DF,  p-value: < 2.2e-16
```

Bør *age* være med i modellen?

```
# robyst standard errors
linearHypothesis(
  lm7,
  c("age = 0", "sexfemale:age"),
  white.adjust = "hc3"
  )
```

Res.Df	Df	F	Pr(>F)
4.86e+03			
4.86e+03	2	1.12	0.326

Ser at vi ikke kan forkaste H0 om at *age* koeffisientene er lik 0. Bør derfor trolig droppe *age* fra modellen.

```
mod8 <- 'sqrt(income) ~ sex*(education + afqt + married)'
lm8 <- hoyde %>%
  filter(income >0, income < 300000) %>%
  filter(complete.cases(.) == TRUE) %>%
  lm(mod8, data = .)
```

```
# robust standard errors
lmtest::coeftest(
  lm8,
  vcov = sandwich::vcovHC, type = "HC3"
)
```

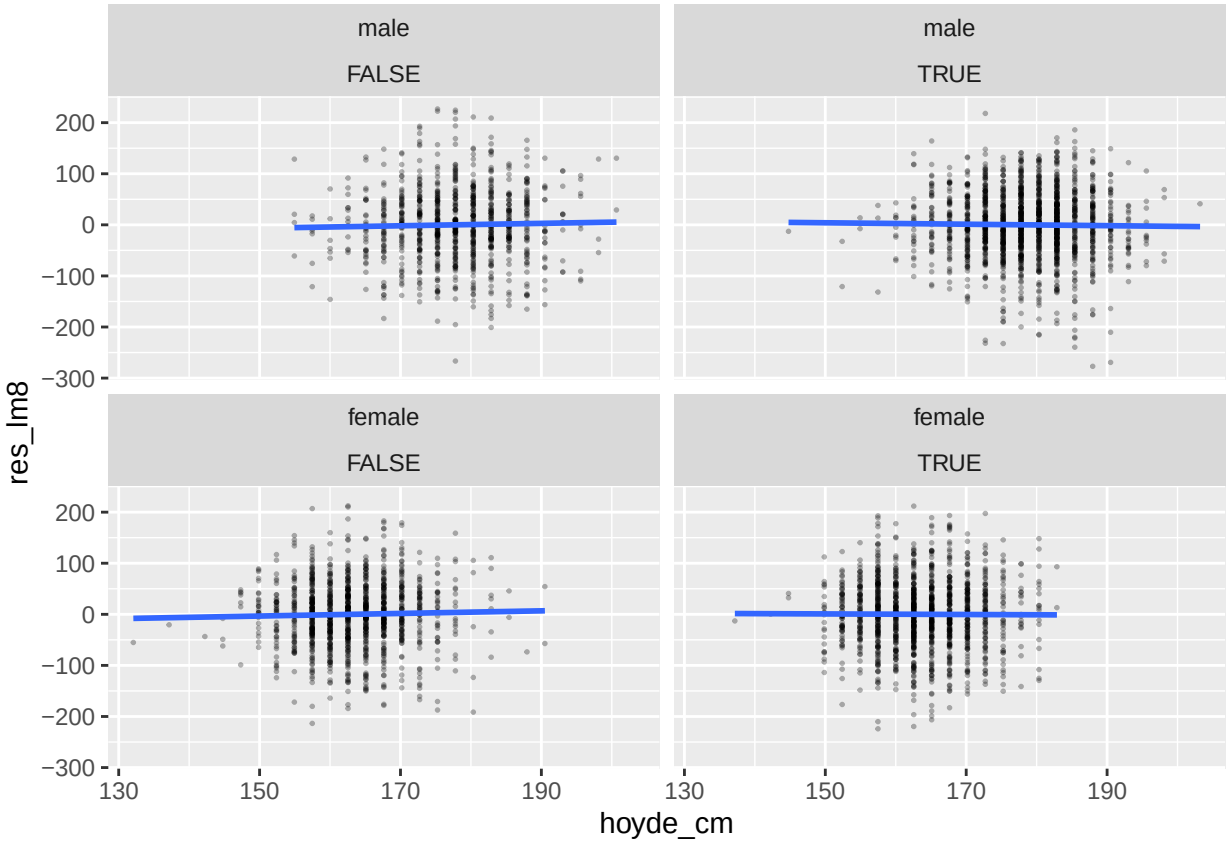
```
##
## t test of coefficients:
##
##                          Estimate Std. Error t value  Pr(>|t|)
```

```
## (Intercept)            62.887874   8.389245  7.4962 7.759e-14 ***
## sexfemale             -18.386491  11.408622 -1.6116   0.10711
## education               7.988504   0.718425 11.1195 < 2.2e-16 ***
## afqt                    0.663643   0.058626 11.3199 < 2.2e-16 ***
## marriedTRUE            34.135888   2.929092 11.6541 < 2.2e-16 ***
## sexfemale:education     0.912114   0.955949  0.9541   0.34006
## sexfemale:afqt         -0.200926   0.082860 -2.4249   0.01535 *
## sexfemale:marriedTRUE -37.667435   3.975194 -9.4756 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Sjekker om det er noen sammenheng mellom residualene (uforklarte forskjeller i inntekt) og høyde i centimeter.

```
res_lm8 <- residuals(lm8)

hoyde %>%
  filter(income >0, income < 300000) %>%
  filter(complete.cases(.) == TRUE) %>%
  ggplot(mapping = aes(x = hoyde_cm, y = res_lm8)) +
  facet_wrap(sex~married) +
  geom_point(size = 0.3, alpha = 0.3) +
  geom_smooth(formula = "y ~ x", method = "lm", se = FALSE)
```



Justerer vi for kjente faktorer som kjønn, utdanning, evner og sivilstatus ser det ikke ut til å være noen som helst sammenheng mellom høyde og inntekt. Dette gjelder for grupene gifte/ugifte menn og gifte/ugifte kvinner. Det kan selvsagt tenkes at det finnes undergrupper hvor høyde har betydning for inntekt, men å avgjøre om dette er tilfelle vil kreve nærmere undersøkelser.

Konklusjon

Selv om både Judge and Cable (2004b) og Eli Kvisvik (2008a) har gode argumenter for sammenhengen mellom høyde og inntekt, viser statistikken utført at vi burde inkludere flere variabler for å kunne konkludere med en definitiv sammenheng for om det er høyde som bestemmer inntekt. Dette kan forstås bedre ved å lese av den siste grafen som viser at gifte kvinner som er 150 cm og ca. 180 cm høye tjener omtrent det samme.

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. 2004b. "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.

———. 2004a. "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.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2020. *Modelr: Modelling Functions That Work with the Pipe*. https://CRAN.R-project.org/package=modelr.