

Er det høyde som bestemmer inntekt?

Assignment 2 i kurset Data Science 2021

Karoline Midtbø

Morten Knutsen

Er det høyde som bestemmer inntekt?

I denne artikkelen skal vi skrive om det er høyden som påvirker inntekten. Vi har sett på to ulike kilder som viser til ulike studier om hvordan høyde påvirker inntekt. Der det utføres ulike undersøkelser for å finne ut om sannheten om at høyden spiller en rolle når det kommer til høyde. Vi skal også utføre ulike datasett der vi selv ser på hvordan høyde påvirker inntekt, ved bruk av ulike dataer. Til slutt skal vi komme frem til en konklusjon og presentere vårt svar på spørsmålet.

Her tar vi å gjør om til målenhetene til metriske. Vi har også lagt til tre nye variabler.

Litteraturgjennomgang

I 2004 skrev Judge og Cable en artikkel der de inkluderte en analyse på om høyde påvirker inntekt. De utførte en analyse med flere kontrollvariabler. De kontrollvariablene var kjønn, alder, vekt og metode. I følge artikkelen kan kjønn ha en påvirkning på både høyde og inntekt. Innen høyde vil de si at gjennomsnittet på høyde på menn og kvinner har en forskjell. I Amerika er gjennomsnittet på menn på 175,5 og på kvinner er 161,7 som tilvarer en forskjell på 12,7(Judge og Cable, 2004). Når det kommer til alder, vil mennesker i gjennomsnitt minske 5 cm i løpet av levetiden. Høyde og vekt er tydeligvis korrelerte, men kan likevel ha effekter i motsatte retninger, de bør derfor isoleres og behandles hver for seg i en analyse. Den siste kontrollvariablene er metode der det vises til at det skal undersøkes på tvers av fire komplementære prøver. De begrenset analysene til personer som hadde et gjennomsnitt på 20 timer eller mer arbeid per uke, bortsett fra studie 1, der vi begrenset analysene til enkeltpersoner som var de primære lønnstakerne i familien.

Studie 1 gikk ut på deltakere og prosedyre der de samlet inn alder, kjønn, høyde og vekt. I studiet 2 samlet de inn lønn, alder, kjønn, høyde og vekt fra en kilde. I studiet 3 samlet de samme data som i studiet 2 men bare i en ny kilde. I siste studiet fra en siste kilde. Resultatene de fikk av dette var at i alle studiene var høyden signifikant positivt korrelert med inntjeningen. I studie 1 viser regresjonsresultatene at alder forutsier positivt inntjening. For studie 2 forutsier kjønn negativ inntjening slik at kvinner tjener mindre enn menn. Alder

forutsier positivt inntjening og vekt forutsier negativt inntjening. For studie 3 forutsier høyden betydelig inntjening. Til slutt, i studie 4, forutsier hver inntekt betydelig. Multippelkorrelasjonen er $R = .29$, og de uavhengige variablene forklarer 8% av variansen i inntjening. Ved å beregne gjennomsnittet på tvers av disse resultatene, finner vi ut at en person som er 72 tommer høy kan forventes å tjene 5.525 dollar mer per år enn noen som er 65 tommer høy, selv etter å ha kontrollert kjønn, vekt og alder.

I en annen artikkel skrevet av Deaton og Arora ble det brukt data fra Gallup-Healthways Well-Being Index, der de hadde daglige meningsmåling om å undersøke forholdet mellom høyde og en rekke emosjonelle og evaluerende utfall. De ser på 454 065 voksne i alderen 18 år eller eldre som ble intervjuet fra 2. januar 2008 til 16. april 2009. De ble da spurt om høyde og ble bedt om å plasere livssituasjon i en rank fra 0 til 10, der 0 fra “det verste mulige liv for deg” og 10 var “det best mulige livet for deg”. De bes også om å svare ja eller nei på spørsmål om følelsene de har hatt i løpet av dagen. Menn som er over gjennomsnittlig høyde rapporterer at de er litt mer enn en sjuendedel av trinnet på ranken over menn som er under gjennomsnittlig høyde, gjennomsnittlig poengsum på 6,55 mot 6,41. For kvinner er forskjellen mindre, med kvinner under gjennomsnittlig høyde litt mindre enn en tiendedel av ranken under kvinner over gjennomsnittlig høyde, gjennomsnittlig rank score 6,55 mot 6,64. En av de mest konsekvent kraftige prediktorene for livsevaluering er inntekt. Regresjonskoeffisienten for ranken på logaritmen for familieinntekt er 0,54 for kvinner og 0,60 for menn. Ifølge denne sammenligningen har hver ekstra tomme høyde samme effekt på rapportert livsevaluering som en 3,8% økning i familieinntekt for kvinner og 4,4% økning for menn (Deaton og Arora, 2009). I følge studiet viser det et resultat på at høyere menn og kvinner er mer sannsynlig å rapportere glede og lykke, og mindre sannsynlig å rapportere smerte og tristhet. Stress og sinne er imidlertid mer sannsynlig å oppleve av mennesker over gjennomsnittlig høyde. Med ca. 2,54 cm høyere høyde sier studiet at en vil ha en 4,5–8,5 prosent økning i familieinntekt.

Table 2
Means (*M*), Standard Deviations (*SD*), and Intercorrelations Among Study 1 and Study 2 Variables

Variable	<i>M</i>	<i>SD</i>	1	2	3	4	5	<i>M</i>	<i>SD</i>
1. Gender	1.19	0.39	1.00	.04**	-.53**	-.72**	-.31**	1.46	0.50
2. Age	39.76	12.63	-.03	1.00	.07**	-.01	.11**	38.95	2.25
3. Weight	3.13	0.60	.02	.06	1.00	.63**	.18**	155.62	32.67
4. Height (in inches)	69.09	3.27	-.65**	.04	-.10	1.00	.31**	67.34	4.07
5. Earnings (U.S. dollars)	48,247.52	33,850.18	-.20**	.17**	.09	.24**	1.00	58,776.12	38,272.26

Note. Correlations for Study 1 (listwise $n = 261$) are below the diagonal. Correlations for Study 2 (listwise $n = 4,314$) are above the diagonal. Gender is coded as 1 = male, 2 = female. For Study 1, weight was coded on a 1 = *skinny* to 5 = *obese* scale. For Study 2, weight is in pounds. Earnings were adjusted to reflect 2002 dollars.

** $p < .01$, two-tailed.

Figur 1: Study 1 and 2

Table 3
Means (*M*), Standard Deviations (*SD*), and Intercorrelations Among Study 3 and Study 4 Variables

Variable	<i>M</i>	<i>SD</i>	1	2	3	4	5	<i>M</i>	<i>SD</i>
1. Gender	1.43	0.50	1.00	—	—	—	—	1.46	0.50
2. Age	32.79	2.63	.05	1.00	—	—	—	—	—
3. Weight (in lbs.)	154.42	28.27	-.74**	.15	1.00	.68**	.15**	135.95	22.18
4. Height (in inches)	68.35	4.00	-.76**	.03	.81**	1.00	.26**	61.56	10.06
5. Earnings (in U.S. dollars)	0.00	1.00	-.22**	-.01	.23**	.35**	1.00	36,626.15	14,474.21

Note. Correlations for Study 3 (listwise $n = 118$) are below the diagonal. Correlations for Study 4 (listwise $n = 3,872$) are listed above the diagonal. Age is not included in Study 4 because all individuals are the same age (within 3 weeks). Gender is coded as 1 = male, 2 = female.

** $p < .01$, two-tailed.

Analyse

```
hoyde <- heights
attach(hoyde)
```

Her tar vi å gjør om til målenhetene til metriske. Vi har også lagt til tre nye variabler.

```
hoyde <- hoyde %>%
  mutate(inntekt = income * 8.42,
         hoyde_cm = height * 2.54,
         vekt_kg = weight * 0.454,
         BMI = vekt_kg/(hoyde_cm/100)^2)
```

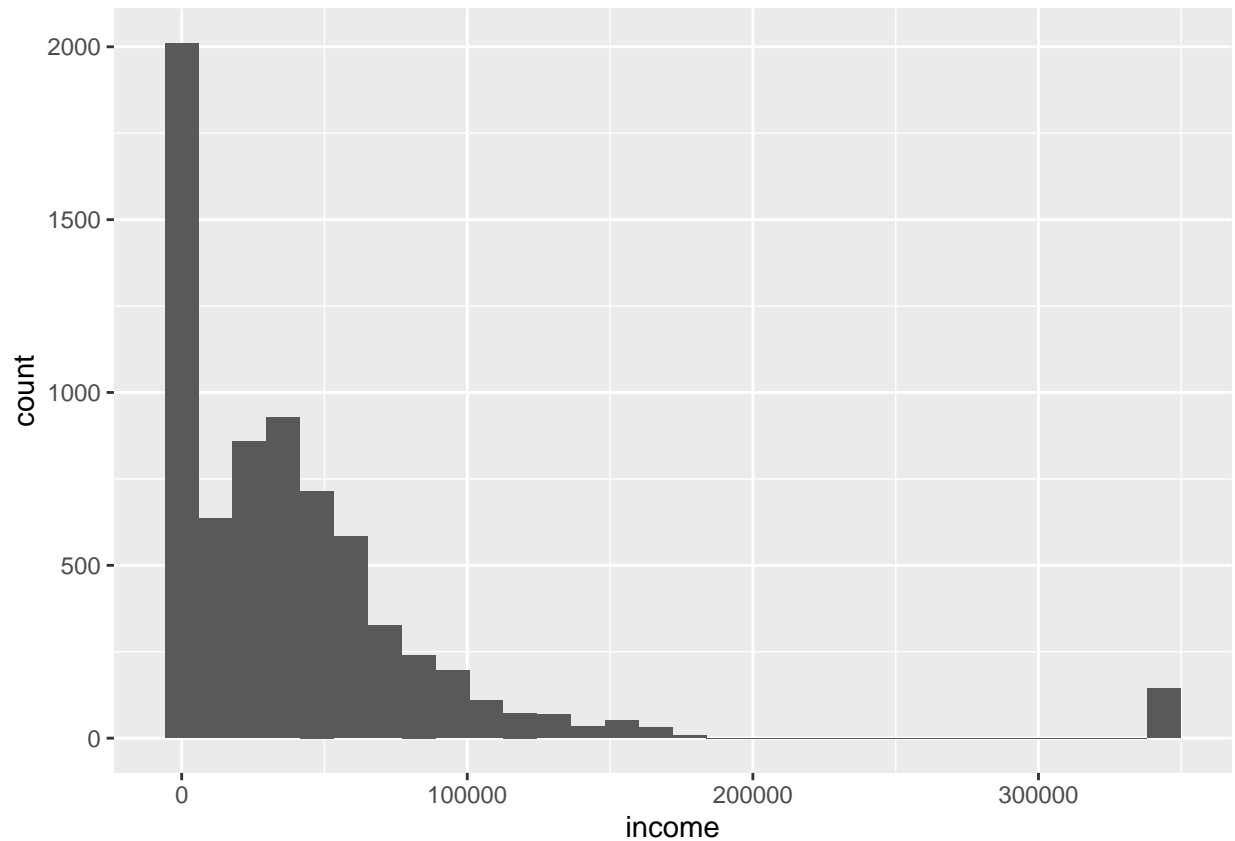
Beskrivende statistikk

- income: Dette er den årlige inntekten. Her ble topp to prosent av inntektene ble gjort om til gjennomsnittet mellom disse og erstatter inntektene som er i toppen.
- height: Høyde i tommer
- weight: Vekt i pund
- age: Alder mellom 47 og 56 år.
- marital: Sivilstatus
- sex: Kjønn:
- education: Antall år med utdanning
- afqt: Prosent i hvor mye du egner deg i militæret

Grunnen til utliggerne ut til høyre er på grunn av personvern under deltakelsen. Der de høyeste lønningene ble lagt sammen og regnet ut gjennomsnittet på disse lønningene.

```
ggplot(hoyde, mapping = aes(x = income)) +
  geom_histogram()
```

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



```
geom_histogram(bins = 30)
```

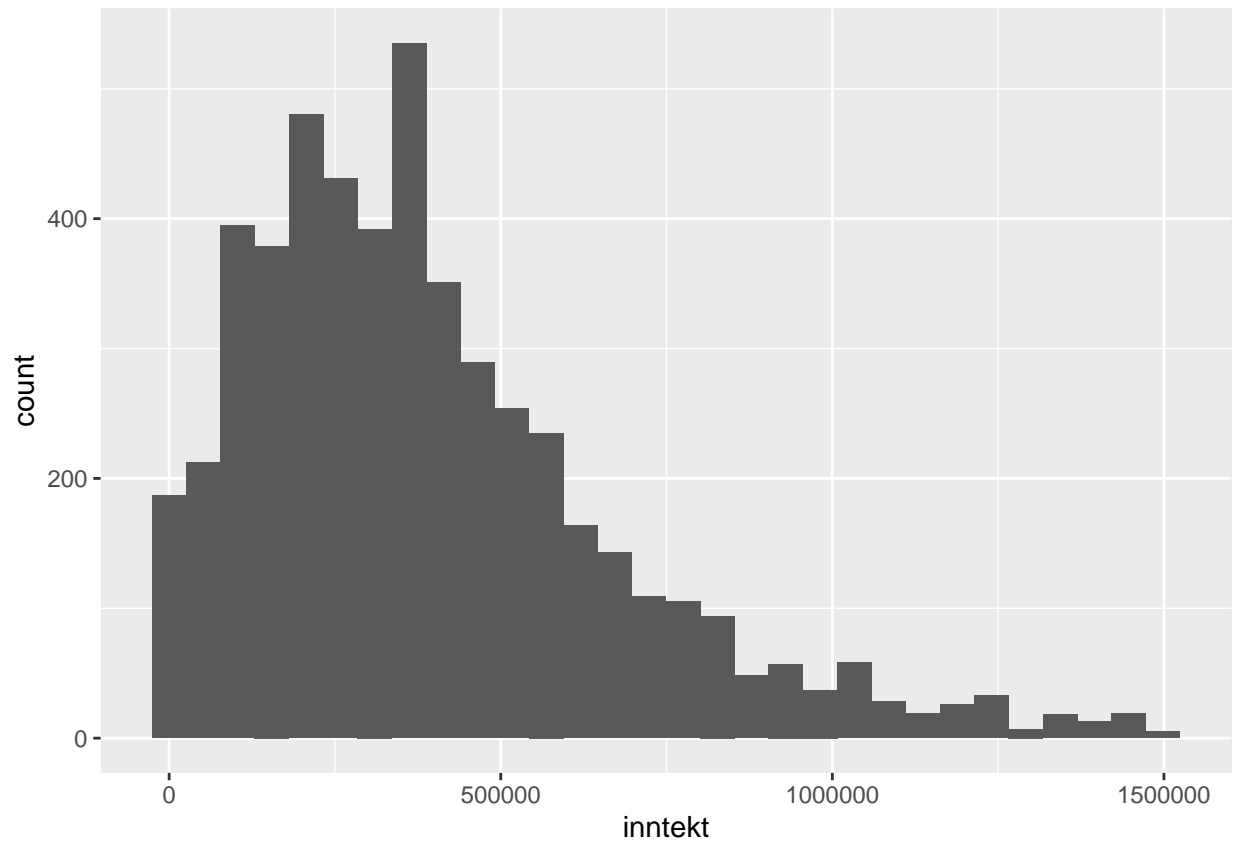
```
## geom_bar: na.rm = FALSE, orientation = NA
## stat_bin: binwidth = NULL, bins = 30, na.rm = FALSE, orientation = NA, pad = FALSE
## position_stack
```

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

Exploratory Data Analysis (EDA) vha. ggplot

```
ggplot(data = hoyde_begr, aes(x = inntekt)) +
  geom_histogram()
```

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



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

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

Det er 1740 personer som står uten inntekt i dette datasettet

Regresjon

```
mod1 <- "inntekt ~ hoyde_cm"  
lm1 <- lm(mod1, data = hoyde, subset = complete.cases(hoyde))
```

```
summary(lm1)
```

```
##  
## Call:  
## lm(formula = mod1, 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
```

```
-1361001 + (10047.9*162)
```

```
## [1] 266758.8
```

```
-1361001 + (10047.9*161)
```

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

Vi ser at inntekten vil øke med 10047,9 kr per cm en øker i høyden.

```
mod2 <- "inntekt ~ hoyde_cm + vekt_kg"
lm2 <- lm(mod2, data = hoyde, subset = complete.cases(hoyde))
```

```
summary(lm2)
```

```
##
## Call:
## lm(formula = mod2, 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
```

```
-1466873.6 + (11430.3*162) + (-1518.4*62)
```

```
## [1] 290694.2
```

```
-1466873.6 + (11430.3*161) + (-1518.4*61)
```

```
## [1] 280782.3
```

Det vi ser her er at når høyden øker så øker lønnen, men når vekten øker så går lønnen ned. En kombinasjon av disse vil gi en økt inntekt.

```
mod3 <- "inntekt ~ hoyde_cm + vekt_kg + BMI"
lm3 <- lm(mod3, data = hoyde, subset = complete.cases(hoyde))
```

```
summary(lm3)
```

```
##
## Call:
## lm(formula = mod3, data = hoyde, subset = complete.cases(hoyde))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -886295 -261634  -93597  124905 2709981
##
## Coefficients:
##              Estimate Std. Error t value    Pr(>|t|)
## (Intercept) -2015890    447005  -4.510 0.0000066012 ***
## hoyde_cm      14669       2649   5.537 0.0000000319 ***
## vekt_kg       -4723       2567  -1.840    0.0658 .
## BMI           9224       7332   1.258    0.2084
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 466600 on 6641 degrees of freedom
## Multiple R-squared:  0.05081,    Adjusted R-squared:  0.05038
## F-statistic: 118.5 on 3 and 6641 DF,  p-value: < 0.00000000000000022
```

```

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

```

Huxreg

```

huxreg(
  list("mod1" = lm1, "mod2" = lm2, "mod3" = lm3),
  error_format = "[{statistic}]",
  note = "Regresjonstabell 3: {stars}. T statistics in brackets."
)

```

	mod1	mod2	mod3
(Intercept)	-1361000.990 *** [-14.413]	-1466873.555 *** [-15.139]	-2015889.845 *** [-4.510]
hoyde_cm	10047.860 *** [18.175]	11430.259 *** [18.308]	14669.413 *** [5.537]
vekt_kg		-1518.381 *** [-4.737]	-4722.577 [-1.840]
BMI			9224.408 [1.258]
N	6645	6645	6645
R2	0.047	0.051	0.051
logLik	-96177.211	-96166.004	-96165.212
AIC	192360.423	192340.008	192340.424

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

interaksjon

```
mod4 <- "inntekt ~ sex*(hoyde_cm + vekt_kg + I(vekt_kg^2) + BMI + I(BMI^2))"
lm4 <- lm(mod4, data = hoyde)
summary(lm4)
```

```
##
## Call:
## lm(formula = mod4, data = hoyde)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -849879 -244503  -90876  126516 2680060
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -2310702.51   3080033.03  -0.750    0.453
## sexfemale       367086.85   3901078.89   0.094    0.925
## hoyde_cm       14261.90    17112.40   0.833    0.405
## vekt_kg         218.88    29727.93   0.007    0.994
## I(vekt_kg^2)    -29.66      71.43  -0.415    0.678
## BMI            17943.84    96811.11   0.185    0.853
## I(BMI^2)        -64.41      770.19  -0.084    0.933
## sexfemale:hoyde_cm    -31.53    22594.40  -0.001    0.999
## sexfemale:vekt_kg   -18366.14   40808.50  -0.450    0.653
## sexfemale:I(vekt_kg^2)  67.48     106.08   0.636    0.525
## sexfemale:BMI       25644.64   121508.62   0.211    0.833
## sexfemale:I(BMI^2)   -194.11     941.57  -0.206    0.837
##
## Residual standard error: 458400 on 6899 degrees of freedom
## (95 observations deleted due to missingness)
## Multiple R-squared:  0.06166, Adjusted R-squared:  0.06016
## F-statistic: 41.21 on 11 and 6899 DF, p-value: < 0.000000000000000022
```

LinearHypothesis

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

```
mod5 <- "inntekt ~ hoyde_cm + vekt_kg + BMI"
lm5 <- lm(mod5, data = hoyde_begr)
summary(lm5)
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
6.90e+03	1.46e+15				
6.9e+03	1.45e+15	6	1.51e+13	12	1.81e-13

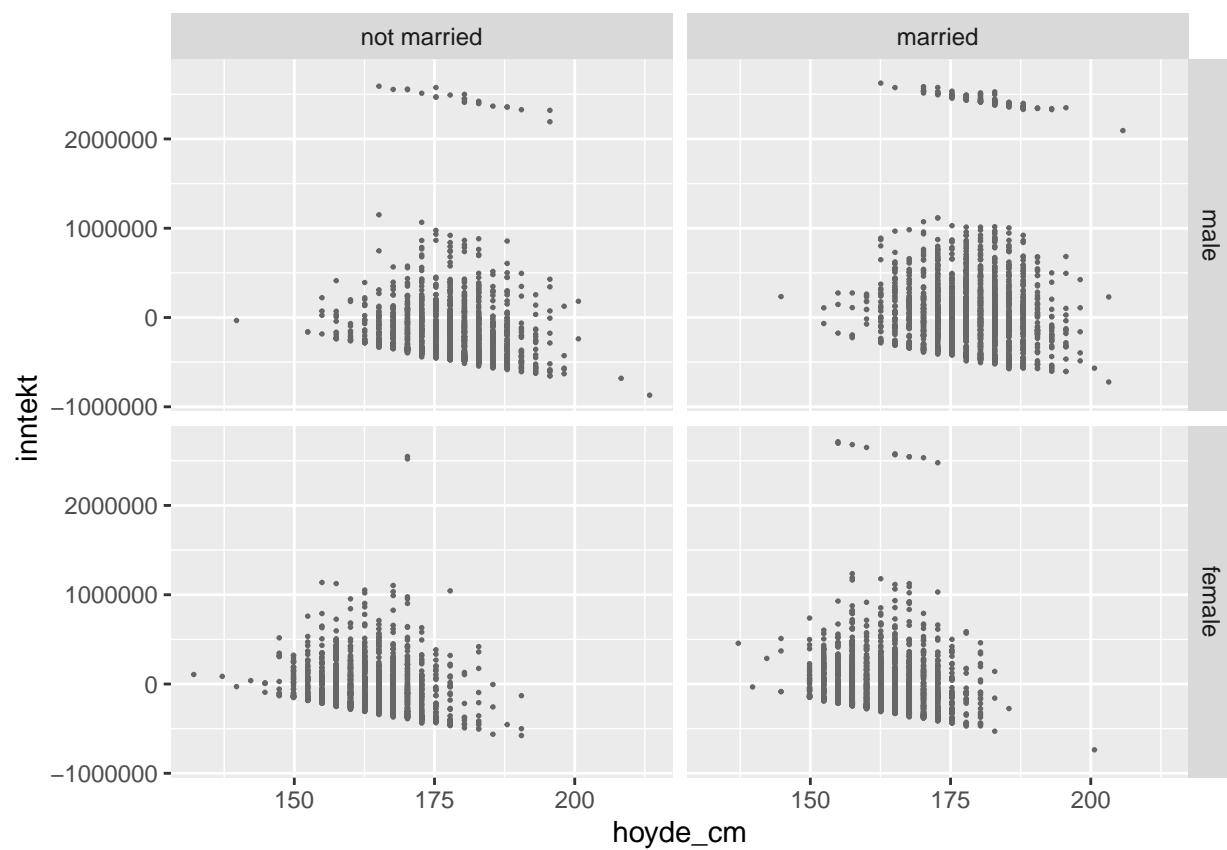
```
##
## Call:
## lm(formula = mod5, data = hoyde_begr)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -869705 -258953  -92227   124770  2712987
##
## Coefficients:
##              Estimate Std. Error t value    Pr(>|t|)
## (Intercept) -1918884     430470  -4.458 0.0000084174 ***
## hoyde_cm      14055         2549   5.513 0.0000000365 ***
## vekt_kg       -4219         2468  -1.709    0.0874 .
## BMI           7873         7061   1.115    0.2649
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 460700 on 6907 degrees of freedom
## (95 observations deleted due to missingness)
## Multiple R-squared:  0.05094,    Adjusted R-squared:  0.05053
## F-statistic: 123.6 on 3 and 6907 DF,  p-value: < 0.00000000000000022
```

```
hoyde_begr <- hoyde_begr %>%
  add_residuals(lm5)
hoyde_begr %>%
  head(n = 10)
```

```
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).
```

weight	age	marital	sex	education	afqt	inntekt	hoyde_cm	vekt_kg	BMI	ma
155	53	married	female	13	6.84	1.6e+05	152	70.4	30.3	TR
156	51	married	female	10	49.4	2.95e+05	178	70.8	22.4	TR
195	52	married	male	16	99.4	8.84e+05	165	88.5	32.5	TR
197	54	married	female	14	44	3.37e+05	160	89.4	34.9	TR
190	49	married	male	14	59.7	6.32e+05	168	86.3	30.7	TR
200	49	divorced	female	18	98.8	8.59e+05	173	90.8	30.4	FA
225	48	married	male	16	82.3	0	188	102	28.9	TR
160	54	divorced	female	12	50.3	5.89e+05	163	72.6	27.5	FA
162	55	divorced	male	12	89.7	5.05e+05	175	73.5	23.9	FA
194	54	divorced	male	13	96	1.26e+06	175	88.1	28.7	FA



Det vi ser i dette plottet er høyde ikke bestemmer inntekt. Vi ser at høye menn kan tjene det samme som lave menn, det samme gjelder hos kvinner. En annen ting vi kan observere

her er at menn tjener mer en kvinner. Vi ser også at det er litt høyere lønninger hos de som er gift, både hos kvinner og menn

Referanser

- Deaton, A., og Arora, R. (2009). *Life at the Top: The Benefits of Height*. <https://www.sciencedirect.com/science/article/pii/S1570677X0900046X>
- Judge, T. A., og Cable, D. M. (2004). *The Effect of Physical Height on Workplace Success and Income: Preliminary Test of a Theoretical Model*. <http://www.w.timothy-judge.com/Height%20paper--JAP%20published.pdf>