

Er det høyde som bestemmer inntekt?

Assignment 2 i MSB105 Data Science

Kevin Ha - 571821

Ola Andre Olofsson - 170745

Innledning

Dette er oppgave 2 i kurset MSB105 Data Science. I den følgende artikkelen anvendes datasettet **heights** fra pakken **modelr** for å besvare følgende problemstilling; **Er det høyde som bestemmer inntekt?**

En kort litteraturgjennomgang på ca. 1 side

I Judge and Cable (2004), kommer de med utsagnet at “høyde påvirker inntekten” er ved første øyekast en gammel myte, men at det kanskje er mer til det enn man først skulle trodd. For å støtte dette utsagnet referer de til Robert & Herman sin forskning som viser til at høyde er et trekk som er ettertraktet i en sosial sammenheng.

Denne forskningen mener også at mennesker som er høyere er mer overbevisende. De viser også til Higham and Carment (1992) som påstår at høye folk er mer sannsynlig til å komme i en lederstilling. Judge & Cable teoriserer at dette muligens har røtter i biologi, ettersom at høyde i naturen er en måling får styrke.

En studie utført av Kurtz Burns (1993) viste at 78% av ansettelse innen salg, var mennesker med over gjennomsnittlig høyde. Dette ble argumentert av rekrutterene å være fordi mennesker over gjennomsnittlig høyde ville være mer utmerket ovenfor kundene, i forhold til små mennesker.

Judge and Cable (2004), ville svare på dette med å utføre en studie med tre hovedpunkter. Først fremstille en modell som viser forholdet mellom høyde og karrieresuksess. De begrunnet dette med at det ikke var blitt utført tidligere. Steg to var å utføre en metaanalyse på tidligere analyser og litteratur for å se etter generelle implikasjoner. Siste steget var å utføre fire nye undersøkelser på forholdet mellom en persons høyde og inntekt.

Argumentet for å se på sammenhengen/forholdet mellom en persons høyde og inntekt var fordi de anså inntekt som den primære faktoren for karrieresuksess. Men i følge “AN EMPIRICAL INVESTIGATION OF THE PREDICTORS OF EXECUTIVE CAREER SUCCESS - JUDGE - 1995 - Personnel Psychology - Wiley Online Library” (n.d.) og Whitely, Dougherty, and Dreher (1991), så er det nesten ingen støttende forskning på dette.

Judge and Cable (2004), tok i sin undersøkelse utgangspunkt i flere menneskelige faktorer for å produsere sin modell for forholdet mellom en persons høyde og inntekt.

Noen av punktene de så på var selvtillit og sosial rang. Judge & Cable Judge and Cable (2004) mente dette var to viktige punkter i analysen fordi disse to faktorene påvirker en persons arbeidsinnsats og hvordan en person blir behandlet i arbeidslivet av arbeidsgiveren. Dette mente faktorene ville påvirke en persons suksess i arbeidslivet, og medføre at en høyere person ville ha høyere inntekt enn en lavere person.

Modellen til Judge og Cable Judge and Cable (2004), kom frem til at det var en form for “rekke” av påvirkninger på en høy person i arbeidslivet. Høyden til en person medfører bedre selvtillit. Denne selvtilliten vil føre til høyere sosial rang, som igjen ville til bedre arbeidsinnsats og muligheter, som videre fører til karrieresuksess.

For å støtte opp under dette grunnlaget for modellen, henviste de til flere tidlegere studier gjort rundt høyde, karrieremuligheter og suksess, samt til flere andre studier om hvordan selvtillit blir påvirket av ulike personlige faktorer.

Studien som Judge og Cable Judge and Cable (2004) gjennomførte konkluderte med at det var en direkte sammenheng mellom en persons høyde og inntekt. I senere tid har dette resultatet blitt sett på og folk mener at sunn fornuft tilsier at dette ikke kan stemme eller at det må være andre eller flere faktorer som spiller inn.

Analyse med egen versjon av datasettet

I henhold til oppgaveteksten, angir vi datasettet for *hoyde*.

```
# Vi selekterer ut dataene for heights fra pakken modelr, og angir deretter benevnelse  
  
data('heights', package = 'modelr')  
hoyde <- heights  
  
# Vi rydder videre opp i benevnelsene ved å slik at de blir enklere å jobbe med. Vi ov  
  
hoyde$inntekt <- hoyde$income*8.5  
hoyde$height_cm <- hoyde$height*2.54  
kable(summary(hoyde[,9:10]))
```

inntekt	height_cm
Min. : 0	Min. :132.1
1st Qu.: 1407	1st Qu.:162.6
Median : 251511	Median :170.2
Mean : 350234	Mean :170.4
3rd Qu.: 467500	3rd Qu.:177.8
Max. :2922555	Max. :213.4

```
# Til slutt kan vi oppsummere de interessante variablene i metrisk form, samt oversatt
```

Beskrivende statistikk (beskrivelse av dataer)

Datasettet vi bruker, *modelr* er hentet fra National Longitudinal Study, som er sponset av U.S. Bureau of Labor Statistics. Dataene stammer fra 2012. Følgende er forklaringene på variablene:

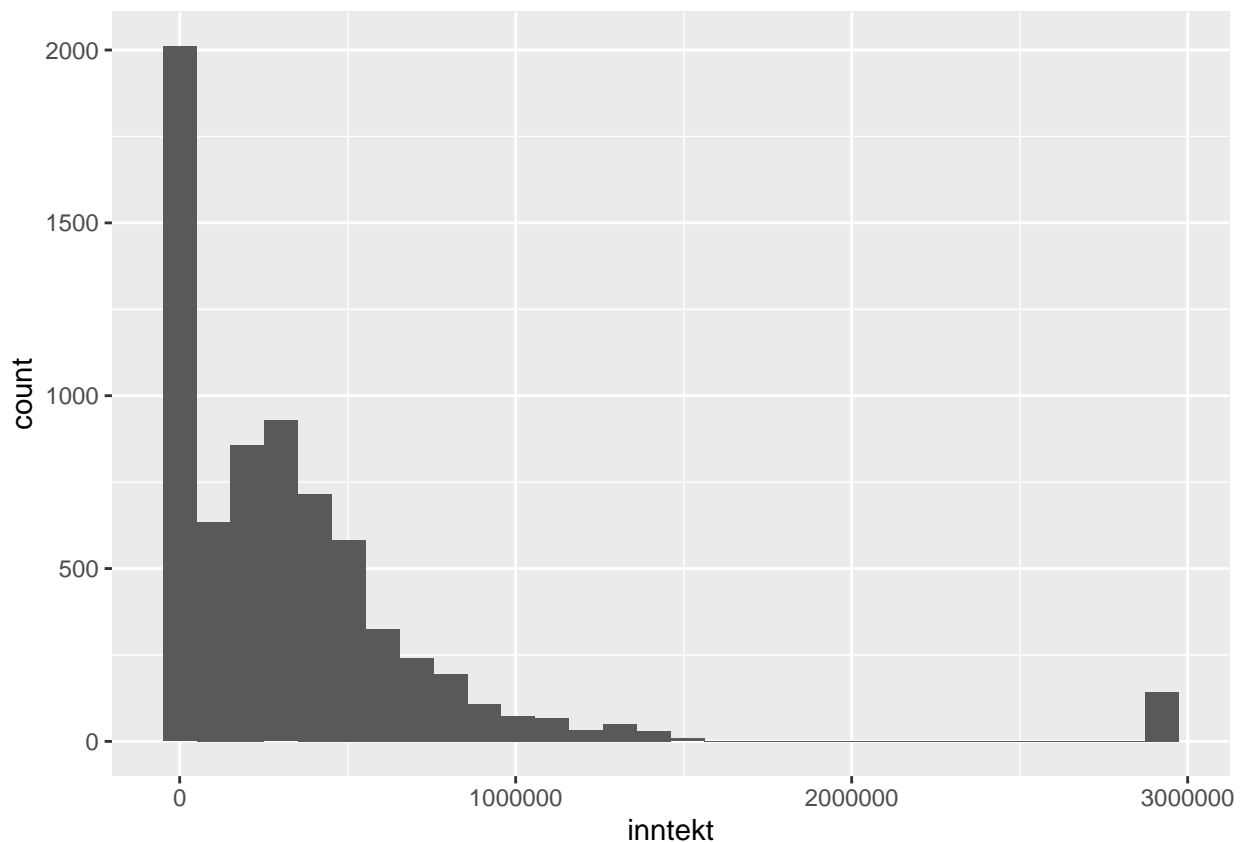
- *height* = høyde i tommer
- *weight* = vekt i pund

- *age* = alder mellom 47 og 56
- *marital* = sivilstatus
- *sex* = kjønn
- *education* = år med utdanning
- *afqt* = prosentskår på test for militær egnethet

Exploratory Data Analysis (EDA) vha. ggplot

```
# Her har vi laget et histogram av variablene income (også kalt inntekt)
ggplot(data = hoyde,
       aes(x = inntekt)) +
  geom_histogram()
```

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



Her ser vi noen utliggere på høyresiden. Dette er 143 observasjoner av personer som tjener rett under 3MNOK. De skiller seg fra resten av observasjonene i histogrammet grunnet at både median- og snittlønn er langt lavere.

Vi har også personer *uten* inntekt i datasettet.

Regresjonsanalyser

```
(lm(inntekt ~ height_cm, data = hoyde)) %>%
  summary()

##
## Call:
## lm(formula = inntekt ~ height_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 <0.0000000000000002 ***
## height_cm     9978.5       534.3    18.68 <0.0000000000000002 ***
## ---
## 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: < 0.00000000000000022
```

Her ser vi at en økning i høyden på 1 cm, gir 9978.5 kr mer i årlig inntekt. La oss prøve med datasett uten de 2% med toppinntekt, og uten de med inntekt = 0.

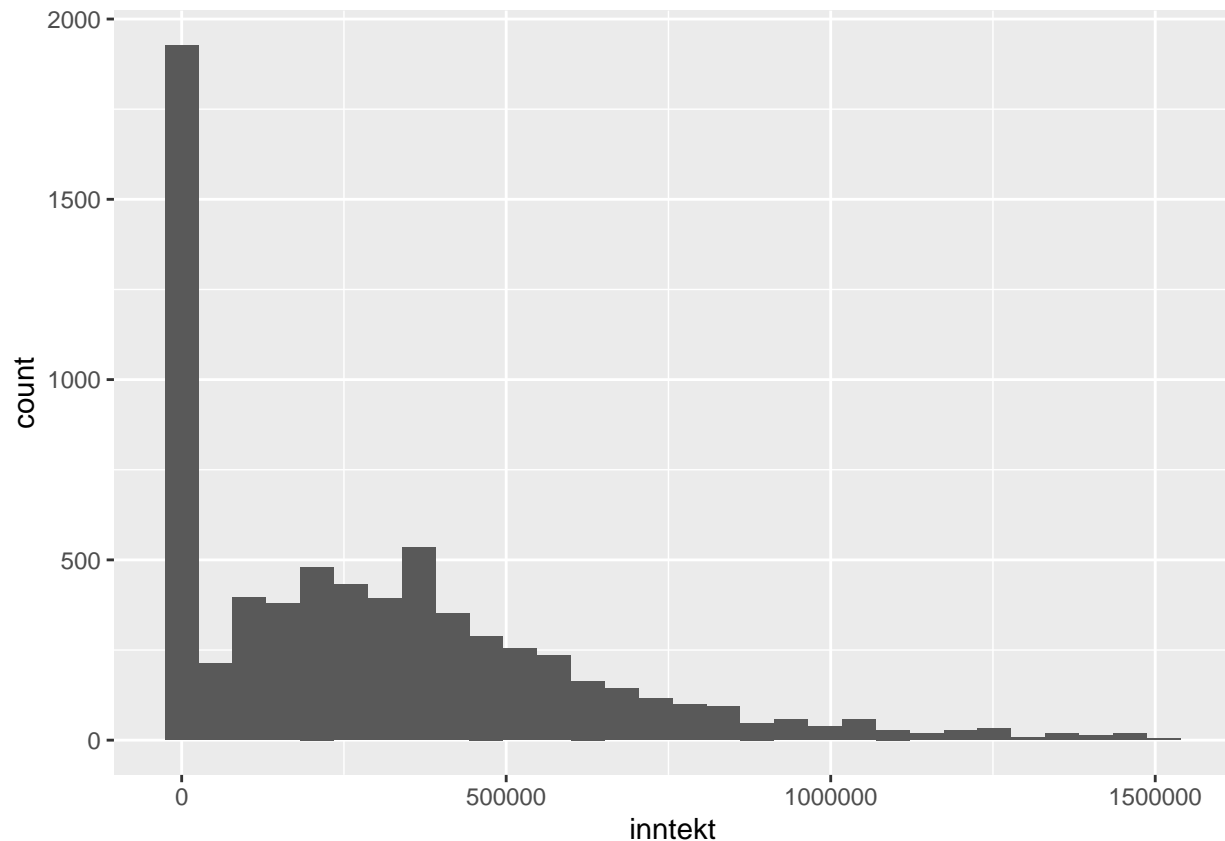
Nå filtrerer vi ut de observasjonene med inntekt lavere enn 1,6MNOK. Vi ser da under

```
hoyde_max_inntekt <- hoyde %>%
  filter(inntekt < 1600000)
```

For illustrasjonshensikter, kan vi også se hvordan dette histogrammet har endret seg

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

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



Her ser vi at utliggerne forsvinner, ettersom den vannrette akse kun viser observasjoner hvor inntekt er lavere enn 1.600.000.

```
(lm(inntekt ~ height_cm, data = hoyde_max_inntekt)) %>%
  summary()
```

```
##
## Call:
## lm(formula = inntekt ~ height_cm, data = hoyde_max_inntekt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -547811 -236923  -54031  158327 1265382
##
## Coefficients:
##              Estimate Std. Error t value      Pr(>|t|)
## (Intercept) -695742.7    58424.7  -11.91 <0.0000000000000002 ***
## height_cm     5828.4      342.5    17.02 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 293300 on 6861 degrees of freedom
## Multiple R-squared:  0.0405, Adjusted R-squared:  0.04036
## F-statistic: 289.6 on 1 and 6861 DF,  p-value: < 0.000000000000000022
```

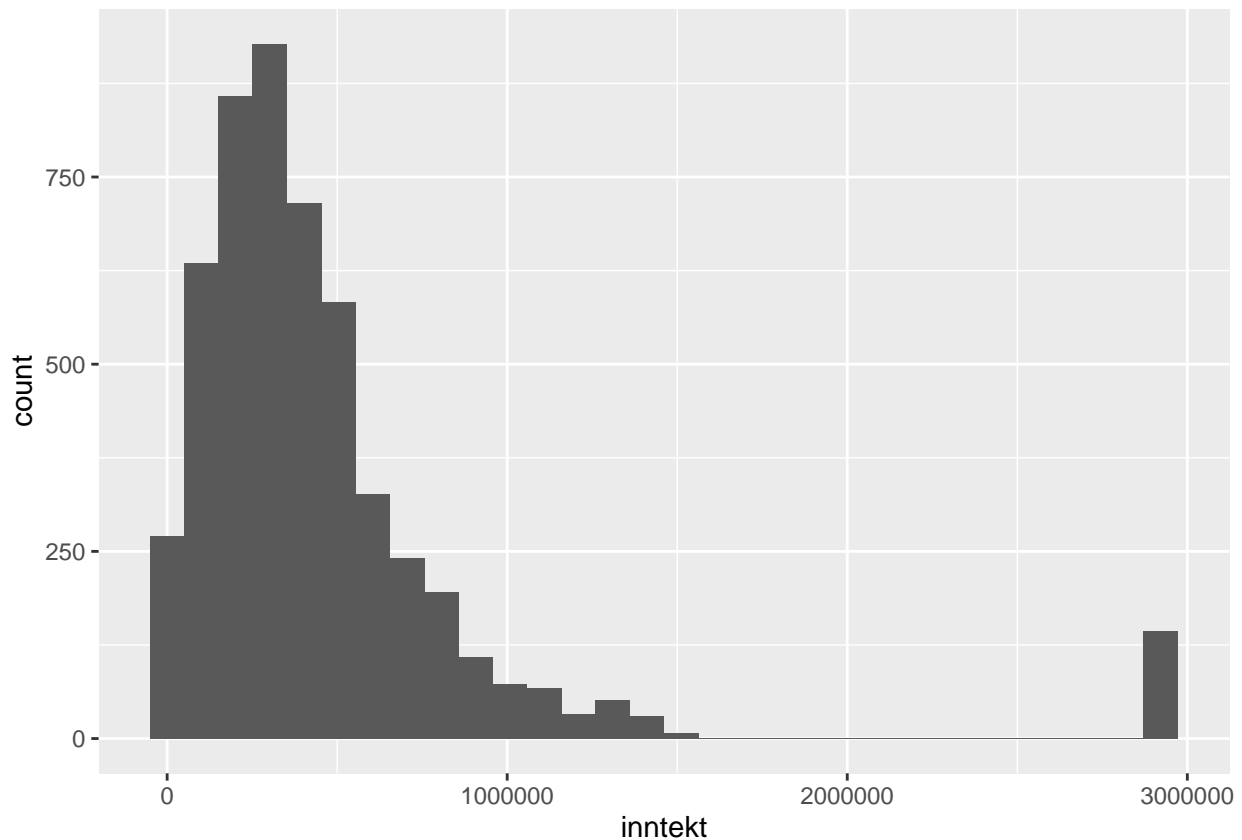
Her ser vi at en økning i høyden på 1 cm, gir 5828.4 kr mer i årlig inntekt.

```
# Nå filtrerer vi ut de observasjonene med inntekt høyere enn 0. Vi ser da under "Envi
hoyde_min_inntekt <- hoyde %>%
  filter(inntekt > 0)

# For illustrasjonshensikter, kan vi også se hvordan dette histogrammet har endret seg

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

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



```
(lm(inntekt ~ height_cm, data = hoyde_min_inntekt)) %>%
  summary()
```

```
##
## Call:
## lm(formula = inntekt ~ height_cm, data = hoyde_min_inntekt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -714128 -253106 -103101   95637 2634963
##
## Coefficients:
##              Estimate Std. Error t value      Pr(>|t|)
## (Intercept) -1435793.6   110687.8  -12.97 <0.0000000000000002 ***
## height_cm    11122.9     646.2    17.21 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 483000 on 5264 degrees of freedom
## Multiple R-squared:  0.05328,    Adjusted R-squared:  0.0531
## F-statistic: 296.3 on 1 and 5264 DF,  p-value: < 0.00000000000000022
```

Her ser vi at en økning i høyden på 1 cm, gir 11122.9 kr mer i årlig inntekt.

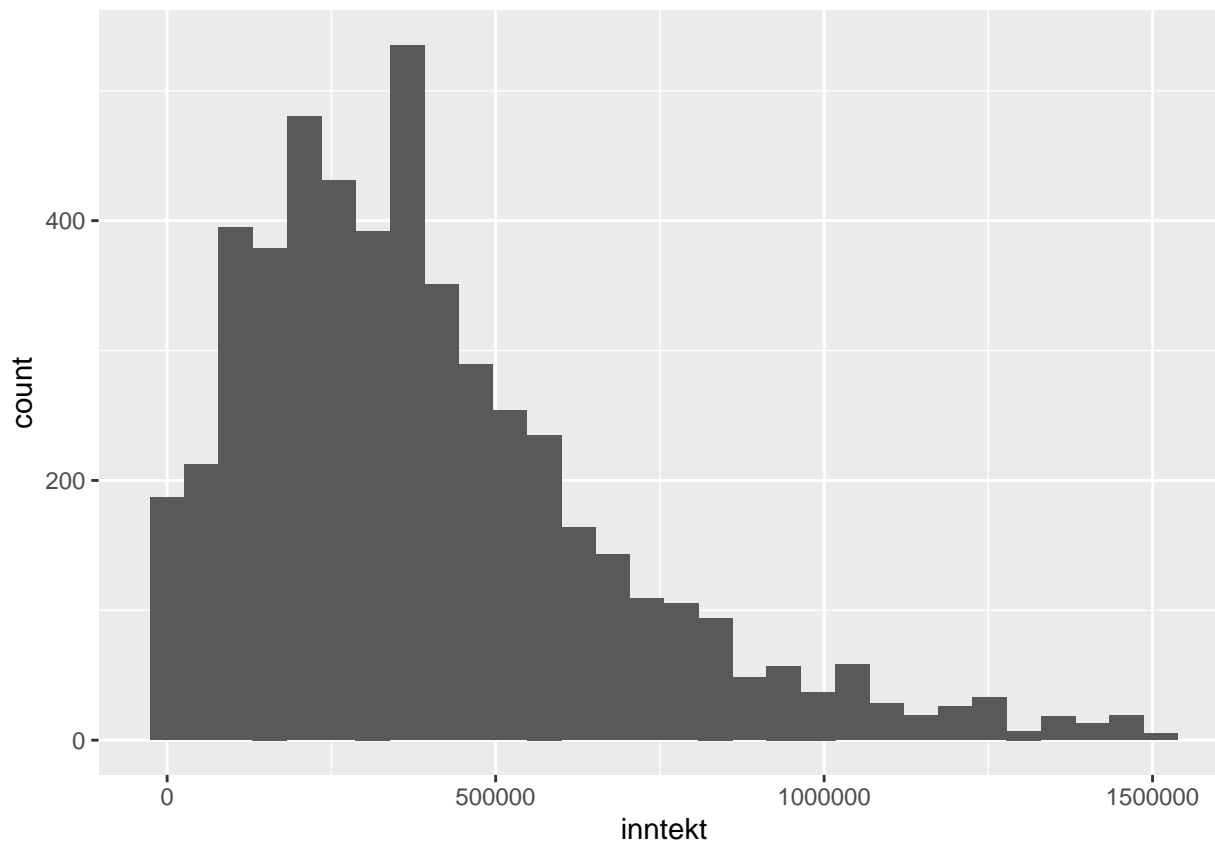
Tar her å ser på når vi tar vekk både 0 inntekt og topp 2% i samme modell i stedenfor

```
hoyde_min_og_max_inntekt <- hoyde %>%
  filter(inntekt < 1600000) %>%
  filter(inntekt > 0)
```

Fremstiller dette i ggplott for å illustrere forskjellen mellom modellene ovenfor gr

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

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

```
(lm(inntekt ~ height_cm, data = hoyde_min_og_max_inntekt)) %>%
summary()
```

```
##
## Call:
## lm(formula = inntekt ~ height_cm, data = hoyde_min_og_max_inntekt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -532259 -190685  -57109   135445 1170911
##
## Coefficients:
##              Estimate Std. Error t value      Pr(>|t|)
## (Intercept) -642281.3    64244.0  -9.998 <0.0000000000000002 ***
## height_cm     6088.8     375.6   16.212 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 276000 on 5121 degrees of freedom
## Multiple R-squared:  0.04882,    Adjusted R-squared:  0.04863
## F-statistic: 262.8 on 1 and 5121 DF,  p-value: < 0.00000000000000022
```

Ser her at om vi tar vekk både 0 inntekt og topp 2% inntekt, så vil 1 cm tilsvare en lønnsøkning på 6088.8 kr.

Forklaring til Ytterligere i *plots* (Prøver å unngå Merge Conflict) Som vi ser ut fra grafen er det en stor ujevnheter. I datasettet er den største andelen av observasjonene fra ca 700 000 kroner og ned, med en mindre andel over dette. 143 observasjoner har rett i underkant av 3 millioner kroner. Dette er den høyre ytterliggigheten i datasettet.

Vi har også med observasjoner *uten* lønn. Dette er den venstre ytterliggigheten. Det er 2000 observasjoner der vedkommende ikke har lønn.

Disse ytterliggighetene påvirker resultatet av analysen. Disse to ekstreme observasjonene resulterer at sammenhengen mellom høyde og snitt- og medianlønn blir feilaktig fremstilt. Vi får dermed feil informasjon ut av dataene vi analyserer. Vi får tilfeller der en lav person er arbeidsledig eller at en høy person har langt høyere inntekt, slik som analysen fra National Longitudinal Study kom frem til.

For å finne et mer reelt resultat må vi se vekk ifra de ekstreme ytterliggighetene. I dette tilfellet vil resultat blir reelt om vi ser vekk ifra både 0 inntekt og 3 millioner i inntekt. Dette gjennomførte vi under kode-chunken “regresjonsanalyse4” ovenfor.

Mutate: Nye Variabler

Vi lager to nye datasett med nye variabler ved å bruke *mutate()* funksjonen. Et datasett der vi tar med hele tidligere datasettet, dvs. med 0 inntekt og topp 2%. Vi lager så enda et datasett der vi tar uten 0 inntekt og topp 2%.

tommer er 2.54cm

pund er 450g, eller 0.45kg

Helt datasett, nye variabler, BMI og gift - ikke gift:

```
hoyde <- hoyde %>%
  mutate(
    height_cm = 2.54 * height,
    weight_kg = weight * 0.45,
    bmi = (weight / (height_cm / 100)^2),
    married = factor(
      case_when(
        # note, summary showed no NA for marital
        marital == 'married' ~ TRUE,
        # all other categories FALSE
        TRUE ~ FALSE
      )
    )
  )
```

Oppsummerer resultatet via `summary()`

```
summary(hoyde)
```

```
##      income      height      weight      age
## Min.   :    0.0   Min.   :52.0   Min.   : 76.0   Min.   :47.00
## 1st Qu.:  165.5   1st Qu.:64.0   1st Qu.:157.0   1st Qu.:49.00
## Median : 29589.5   Median :67.0   Median :184.0   Median :51.00
## Mean   : 41203.9   Mean   :67.1   Mean   :188.3   Mean   :51.33
## 3rd Qu.: 55000.0   3rd Qu.:70.0   3rd Qu.:212.0   3rd Qu.:53.00
## Max.   :343830.0   Max.   :84.0   Max.   :524.0   Max.   :56.00
##
##                NA's      :95
##      marital      sex      education      afqt
## single  :1124   male  :3402   Min.   : 1.00   Min.   : 0.00
## married :3806   female:3604   1st Qu.:12.00   1st Qu.: 15.12
## separated: 366                Median :12.00   Median : 36.76
## divorced :1549                Mean   :13.22   Mean   : 41.21
## widowed  : 161                3rd Qu.:15.00   3rd Qu.: 65.24
##
##                Max.   :20.00   Max.   :100.00
##                NA's   :10     NA's   :262
##      inntekt      height_cm      weight_kg      bmi
## Min.   :    0   Min.   :132.1   Min.   : 34.20   Min.   : 28.38
## 1st Qu.:  1407   1st Qu.:162.6   1st Qu.: 70.65   1st Qu.: 55.31
## Median : 251511   Median :170.2   Median : 82.80   Median : 62.44
## Mean   : 350234   Mean   :170.4   Mean   : 84.74   Mean   : 64.61
## 3rd Qu.: 467500   3rd Qu.:177.8   3rd Qu.: 95.40   3rd Qu.: 71.17
## Max.   :2922555   Max.   :213.4   Max.   :235.80   Max.   :165.32
##
##                NA's      :95     NA's      :95
##      married
## FALSE:3200
## TRUE  :3806
##
##
##
##
##
```

Filtrert datasett, nye variabler, BMI og gift - ikke gift:

```
hoyde_filtrert <- hoyde_min_og_max_inntekt %>%
  mutate(
    height_cm = 2.54 * height,
    weight_kg = weight * 0.45,
    bmi = (weight / (height_cm / 100)^2),
```

```

married = factor(
  case_when(
    # note, summary showed no NA for marital
    marital == 'married' ~ TRUE,
    # all other categories FALSE
    TRUE ~ FALSE)
)
)

```

Oppsummerer resultatet via `summary()`

```
summary(hoyde_filtret)
```

```

##      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      height_cm      weight_kg      bmi
##  Min.   :   382.5  Min.   :132.1  Min.   : 35.10  Min.   : 28.38
##  1st Qu.: 195500.0  1st Qu.:162.6  1st Qu.: 71.55  1st Qu.: 55.36
##  Median : 340000.0  Median :170.2  Median : 83.25  Median : 62.39
##  Mean   : 397386.4  Mean   :170.7  Mean   : 84.78  Mean   : 64.37
##  3rd Qu.: 527000.0  3rd Qu.:177.8  3rd Qu.: 95.40  3rd Qu.: 70.76
##  Max.   :1513000.0  Max.   :203.2  Max.   :216.00  Max.   :147.59
##
##                NA's :69      NA's :69
##      married
##  FALSE:2140
##  TRUE :2983
##
##
##
##
##
##

```

HuxReg

Setter opp for å sette opp en HuxTable på datasettene med nye variabler.

lm_hoyde er fulle datasettet, men med nye variabler.

lm_hoyde_filtrert er datasettet uten 0 inntekt og 2% topp, men med nye variabler.

```
lm_hoyde <- (lm(
  inntekt ~ height_cm + weight_kg + marital + bmi,
  data = hoyde))
lm_hoyde_filtrert <- (lm(
  inntekt ~ height_cm + weight_kg + marital + bmi,
  data = hoyde_filtrert))
```

Setter opp til liste med avvik innenfor statistikk. Gir navn til tabellene våres for bedre oversikt.

hoyde er med alle observasjoner **hoyde_filtrert** er uten 0 inntekt og topp 2%

```
huxreg(
  list("hoyde"=lm_hoyde, "hoyde_filtrert"=lm_hoyde_filtrert),
  error_format = "[{statistic}]")
```

Ser at det er en betraktelig forskjell mellom *hoyde* i det fulle datasettet, og *hoyde_filtrert* i datasettet som er uten 0 inntekt og 2% topp.

De mest akutelle faktorene for studien denne innleveringen baserer seg på er:

height_cm, *weight_kg*, *N*, R^2

Vi ser ut fra tabellene at disse fire faktorene er alle blitt påvirket i stor grad. Inntekt fra *høyde* har falt nesten 10 000kr pr cm, i tillegg til å gå ifra $p < 0.001$ signifikans nivå til $p < 0.01$ signifikans nivå. Vekt har gått ifra $p < 0.05$ til ingen signifikans. N har gått ned ifra **6911 kr pr cm**, til **5054 kr pr cm**. R^2 har gått ifra **0.88** ned til **0.82**.

Det vi kan tolke ut fra dette er at 0 inntekt og 2% topp inntekt har hatt en betydelig påvirkning på studien til Judge & Cable.

Test av robusthet

Robushets refereres til styrken av den anvendte statistiske modellen, og kan eksempelvis være å utføre en t-test. Dette type testen er en hypotesetest og brukes for å teste hvorvidt gjennomsnittssverdien i et normalfordelt datasatt er signifikant forskjellig fra en nullhypotese. Det finnes ut varianter av t-tester; *paret t-test* og *uavhengig t-test* og vi vil her demonstrere _____.

Vi må først definere H_0 og H_1 , og deretter tar vi en titt på t-verdiene til de ulike variablene og ser om de er signifikante.

H_0 : *Liten endring i t-verdi og signifikansnivå for høyde* H_1 : *Høyde har mindre betydning enn antatt, støttes av verdier*

Modellene

Modeller uten 0 inntekt:

```
modell_1 <- "inntekt ~ height_cm"
lm1 <- lm(modell_1, data = hoyde_min_inntekt)
summary(lm1)

##
## Call:
## lm(formula = modell_1, data = hoyde_min_inntekt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -714128 -253106 -103101   95637 2634963
##
## Coefficients:
##              Estimate Std. Error t value      Pr(>|t|)
## (Intercept) -1435793.6   110687.8  -12.97 <0.0000000000000002 ***
## height_cm    11122.9     646.2    17.21 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 483000 on 5264 degrees of freedom
## Multiple R-squared:  0.05328,    Adjusted R-squared:  0.0531
## F-statistic: 296.3 on 1 and 5264 DF,  p-value: < 0.00000000000000022
```

•

```
modell_2 <- "inntekt ~ height_cm + weight + marital"
lm2 <- lm(modell_2, data = hoyde_min_inntekt)
summary(lm2)
```

```
##
## Call:
## lm(formula = modell_2, data = hoyde_min_inntekt)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -746084 -252292  -98264   100382  2585620
##
## Coefficients:
##              Estimate Std. Error t value      Pr(>|t|)
## (Intercept)   -1572607.8    116340.3  -13.517 < 0.0000000000000002 ***
## height_cm      12121.0       748.7   16.189 < 0.0000000000000002 ***
## weight        -741.9        179.9   -4.125     0.0000377 ***
## maritalmarried  166514.8     20054.3    8.303 < 0.0000000000000002 ***
## maritalseparated -66105.1     36509.9   -1.811     0.0703 .
## maritaldivorced  52220.2     23190.2    2.252     0.0244 *
## maritalwidowed   7776.9      50210.9    0.155     0.8769
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 478700 on 5190 degrees of freedom
## (69 observations deleted due to missingness)
## Multiple R-squared:  0.07994,    Adjusted R-squared:  0.07887
## F-statistic: 75.15 on 6 and 5190 DF,  p-value: < 0.00000000000000022
```

•

```
modell_3 <- "inntekt ~ sex*height_cm + weight + marital"
lm3 <- lm(modell_3, data = hoyde_min_inntekt)
summary(lm3)
```

```
##
## Call:
## lm(formula = modell_3, data = hoyde_min_inntekt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -748598 -252959  -97185   103409  2657787
##
## Coefficients:
##              Estimate Std. Error t value      Pr(>|t|)
## (Intercept)  -1087464.6    226208.6   -4.807 0.000001572603235 ***
## sexfemale      750581.1    319561.7    2.349     0.01887 *
## height_cm       9685.6     1312.6    7.379 0.0000000000000185 ***
## weight        -829.8       179.3   -4.628 0.000003783773448 ***
## maritalmarried  165425.3     19946.9    8.293 < 0.0000000000000002 ***
## maritalseparated -58590.2     36319.5   -1.613     0.10676
## maritaldivorced  54144.2     23073.3    2.347     0.01898 *
```

```
## maritalwidowed      31950.8    50034.9    0.639          0.52313
## sexfemale:height_cm -5248.1     1875.1   -2.799          0.00515 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 476000 on 5188 degrees of freedom
## (69 observations deleted due to missingness)
## Multiple R-squared:  0.09087,    Adjusted R-squared:  0.08946
## F-statistic: 64.82 on 8 and 5188 DF,  p-value: < 0.000000000000000022
```

Test av koeffisienter:

```
linearHypothesis(lm3, c("sexfemale = 0", "sexfemale:height_cm = 0"))
```

Modeller uten topp 2% inntekt:

```
modell_4 <- "inntekt ~ height_cm"
lm4 <- lm(modell_4, data = hoyde_max_inntekt)
summary(lm4)
```

```
##
## Call:
## lm(formula = modell_4, data = hoyde_max_inntekt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -547811 -236923  -54031  158327 1265382
##
## Coefficients:
##              Estimate Std. Error t value      Pr(>|t|)
## (Intercept) -695742.7    58424.7  -11.91 <0.0000000000000002 ***
## height_cm    5828.4      342.5    17.02 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 293300 on 6861 degrees of freedom
## Multiple R-squared:  0.0405, Adjusted R-squared:  0.04036
## F-statistic: 289.6 on 1 and 6861 DF,  p-value: < 0.000000000000000022
```

•


```
modell_5 <- "inntekt ~ height_cm + weight + marital"
lm5 <- lm(modell_5, data = hoyde_max_inntekt)
summary(lm5)
```

```
##
## Call:
## lm(formula = modell_5, data = hoyde_max_inntekt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -552630 -213609  -54102   149212 1282160
##
## Coefficients:
##              Estimate Std. Error t value      Pr(>|t|)
## (Intercept)   -805106.56    59964.37  -13.426 < 0.0000000000000002 ***
## height_cm       6188.47     382.24   16.190 < 0.0000000000000002 ***
## weight        -268.85      88.48   -3.039    0.00239 **
## maritalmarried  155353.49    9876.94   15.729 < 0.0000000000000002 ***
## maritalseparated -12818.20   17385.95   -0.737    0.46098
## maritaldivorced  68638.64   11379.36    6.032    0.00000000171 ***
## maritalwidowed  19719.18   24496.96    0.805    0.42087
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 286200 on 6761 degrees of freedom
## (95 observations deleted due to missingness)
## Multiple R-squared:  0.09079,    Adjusted R-squared:  0.08999
## F-statistic: 112.5 on 6 and 6761 DF,  p-value: < 0.00000000000000022
```

•

```
modell_6 <- "inntekt ~ sex*height_cm + weight + marital"
lm6 <- lm(modell_6, data = hoyde_max_inntekt)
summary(lm6)
```

```
##
## Call:
## lm(formula = modell_6, data = hoyde_max_inntekt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -534187 -212964  -53532   149687 1276649
##
```

```
## Coefficients:
##              Estimate Std. Error t value      Pr(>|t|)
## (Intercept)   -542258.82  118177.26  -4.589  0.0000045445601 ***
## sexfemale      192793.51  163358.92   1.180    0.237969
## height_cm      4812.42    686.42   7.011  0.00000000000026 ***
## weight       -291.14     88.43  -3.292    0.000998 ***
## maritalmarried 158245.57   9870.39  16.032 < 0.0000000000000002 ***
## maritalseparated -8214.16  17360.89  -0.473    0.636127
## maritaldivorced 71219.50  11360.29   6.269  0.00000000003854 ***
## maritalwidowed 31331.27  24494.07   1.279    0.200893
## sexfemale:height_cm -1498.89   958.11  -1.564    0.117763
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 285400 on 6759 degrees of freedom
## (95 observations deleted due to missingness)
## Multiple R-squared:  0.0964, Adjusted R-squared:  0.09533
## F-statistic: 90.13 on 8 and 6759 DF,  p-value: < 0.00000000000000022
```

Test av koeffisienter:

```
linearHypothesis(lm6, c("sexfemale = 0", "sexfemale:height_cm = 0"))
```

Kommentarer til modellene ovenfor.

Ser at når vi kun tar hensyn til høyde og inntekt er t-verdien 17.27 og er signifikant helt opp til et 0.001 nivå. Ved første øyenkast kan det da se ut til at høyde faktisk bestemmer inntekten. Men visst vi studerer resultatet og da spesielt R^2 så ser vi at vi den verdien er bare 0.05328. Det betyr at høyde kun forklarer 5.3% av resultatet vårt. Som en da ser i modell 2 og spesielt i modell 3, at desto flere variabler vi legger inn og må ta hensyn til, desto mindre betydning får høyde.

Modell 3 tar med flere variabler og vi ser da at t-verdien til høyde faller til 7.379, mens den fortsatt er signifikant på 0.001 nivå, som virker lovendes. Men vi ser også nå at kjønn har en t-verdi på 2.349 og signifikansnivå på 0.05, og om vedkommende er gift har t-verdi på 8.293 med 0.001 signifikans nivå. Vi ser her at høyde har fått en betraktelig mindre betydning når vi har lagt til flere variabler, der flere av de variablene har en stor betydning i iht. t-verdiene og signifikansnivåene.

For modellene uten topp 2% inntekt så ser vi akkurat samme tendens. Flere desto flere variabler, desto mindre betydning.

Vi ser også enn annen tendens, som er at det øyeblikket vi legger til kjønn som en variabel, så stuper t-verdien til høyde med 10. Vi kan dermed si med ganske stor sannsynlighet at kjønn har en enorm stor påvirkning på inntekten til personer.

Modell uten både 0 og topp 2%

Men va skjer om vi lager en modell der vi tar bort både arbeidsledige og topp 2% inntekts-gruppa?

Bruker her da datasettet *hoyde_filtrert* fra tidligere som er uten både 0 inntekt og topp 2% inntekt:

```
modell_7 <- "inntekt ~ height_cm"
lm7 <- lm(modell_7, data = hoyde_filtrert)
summary(lm7)
```

```
##
## Call:
## lm(formula = modell_7, data = hoyde_filtrert)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -532259 -190685  -57109   135445 1170911
##
## Coefficients:
##              Estimate Std. Error t value      Pr(>|t|)
## (Intercept) -642281.3    64244.0  -9.998 <0.0000000000000002 ***
## height_cm    6088.8      375.6   16.212 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 276000 on 5121 degrees of freedom
## Multiple R-squared:  0.04882,    Adjusted R-squared:  0.04863
## F-statistic: 262.8 on 1 and 5121 DF,  p-value: < 0.00000000000000022
```

•

```
modell_8 <- "inntekt ~ sex*height_cm + weight_kg + marital"
lm8 <- lm(modell_8, data = hoyde_filtrert)
summary(lm8)
```

```
##
## Call:
## lm(formula = modell_8, data = hoyde_filtrert)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -543316 -186680  -53277   129072 1180287
```

```
##
## Coefficients:
##               Estimate Std. Error t value      Pr(>|t|)
## (Intercept)    -352138.9   131673.9   -2.674      0.00751 **
## sexfemale       108210.9   184020.3    0.588      0.55653
## height_cm       4441.2     764.6     5.808     0.00000000669 ***
## weight_kg       -671.5     229.7    -2.924      0.00347 **
## maritalmarried  118350.8   11461.8   10.326 < 0.0000000000000002 ***
## maritalseparated -26036.7   20702.3   -1.258      0.20857
## maritaldivorced  58440.0   13221.6    4.420     0.00001007603 ***
## maritalwidowed  10754.6   28617.5    0.376      0.70708
## sexfemale:height_cm -1055.7   1079.1   -0.978      0.32800
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 270800 on 5045 degrees of freedom
## (69 observations deleted due to missingness)
## Multiple R-squared:  0.08998,    Adjusted R-squared:  0.08854
## F-statistic: 62.36 on 8 and 5045 DF,  p-value: < 0.00000000000000022
```

•

```
modell_9 <- "inntekt ~ sex*height_cm + weight_kg + marital + bmi"
lm9 <- lm(modell_9, data = hoyde_filttrert)
summary(lm9)
```

```
##
## Call:
## lm(formula = modell_9, data = hoyde_filttrert)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -540548 -186297  -53638  128611 1179361
##
## Coefficients:
##               Estimate Std. Error t value      Pr(>|t|)
## (Intercept)    -178419.1   347742.5   -0.513      0.6079
## sexfemale       98602.6   184892.1    0.533      0.5939
## height_cm       3423.8    2034.0    1.683      0.0924 .
## weight_kg        312.1    1836.8    0.170      0.8651
## maritalmarried  118033.3   11477.7   10.284 < 0.0000000000000002 ***
## maritalseparated -26161.3   20705.0   -1.264      0.2065
## maritaldivorced  58219.1   13228.9    4.401     0.000011 ***
## maritalwidowed  10577.2   28621.4    0.370      0.7117
```

```
## bmi -1286.4 2383.3 -0.540 0.5894
## sexfemale:height_cm -1000.1 1084.1 -0.922 0.3563
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 270800 on 5044 degrees of freedom
## (69 observations deleted due to missingness)
## Multiple R-squared: 0.09004, Adjusted R-squared: 0.08841
## F-statistic: 55.45 on 9 and 5044 DF, p-value: < 0.000000000000000022
```

Test av koeffisienter:

```
linearHypothesis(lm9, c("sexfemale = 0", "sexfemale:height_cm = 0"))
```

Kommentar til modell 7, 8 og 9

I modell 7 ser vi samme tendens som i modell 1 og 4, der høyde har en stor betydning. Men så ser vi her på modell 8 og 9 uten både 0 inntekt og topp 2% inntekt. t-verdien til høyde har kollapset. Den er nå i modell 9 under 1.92 som er den gylne standarden for t-verdier. Signifikans nivået har også kollapset fra 0.001 nivå helt ned til et 0.1 nivå. Med andre ord, høyde er faktisk ikke lengre signifikant, da de fleste modeller opererer på et 5% nivå for å være signifikant.

Vi ser dermed at når vi ser vekk fra de arbeidsledige samt topp 2% inntektsgruppa, i tillegg til å legge til flere variabler, så er ikke høyde lengre de-facto for inntekt.

Resultat Vi forkaster H_0 da vi tydelig ser at både t-veriden og $p < \alpha$ verdien kolliderer når vi legger til flere faktorer fremfor høyde. H_1 Er dermed gjeldende og resultatet vårt er:

Høyde alene er ikke de-facto grunnlag for høyere lønn, det er andre faktorer med som påvirker.

Residualer til datasettet “hoyde”

Vi velger å bruke modell 9, da denne er mest realistisk og inneholder mest informasjon.

```
hoyde_filtrert <- hoyde %>%
  add_residuals(lm9)
```

•

```
hoyde_filtrert %>%  
  head(n=10)
```

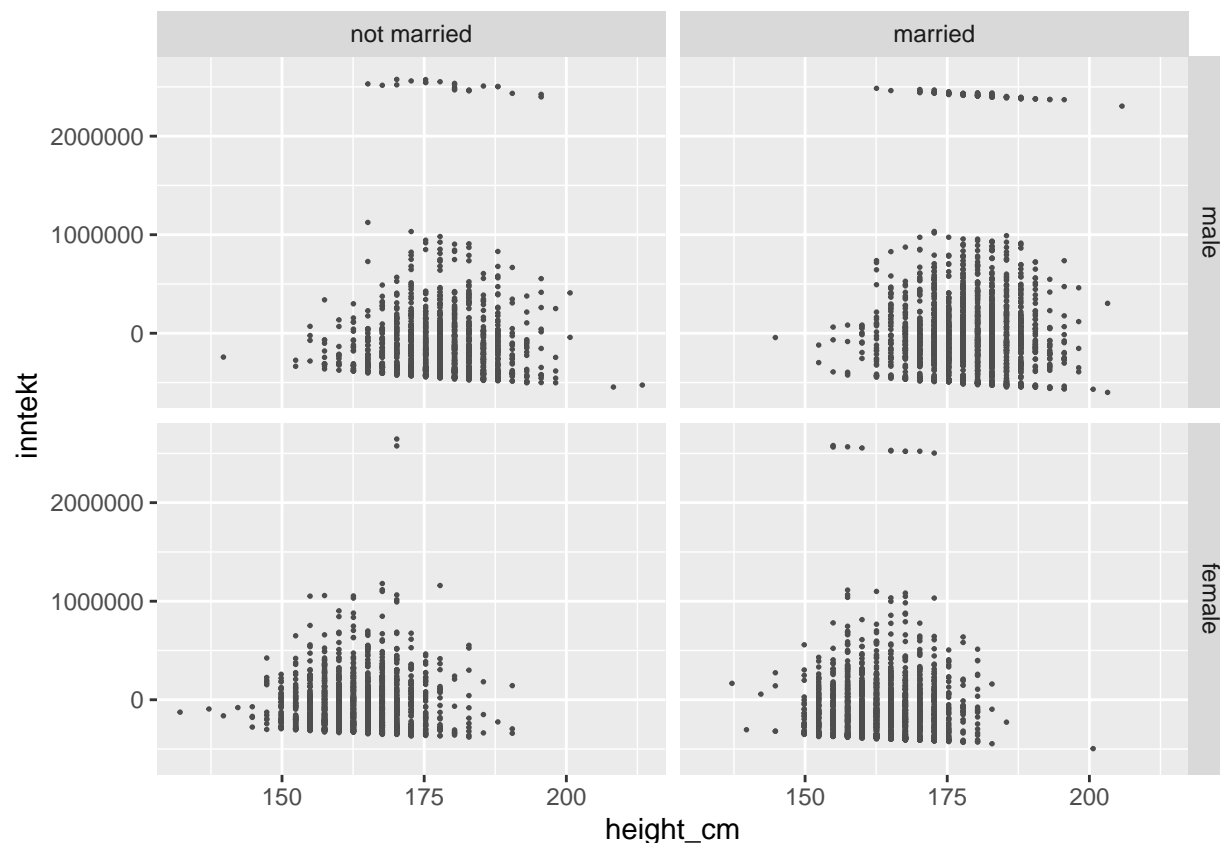
Residualene fra endelig modell skal legges til datasettet hoyde. height_cm skal plottes mot residualene for

- Høyra med Arnstein om detta her e rett?
-

GGplot av observasjonene, med svak bakgrunn.

```
ggplot(data = hoyde_filtrert,  
  mapping = aes(  
    x = height_cm,  
    y = inntekt)) +  
geom_point(  
  data = hoyde_filtrert,  
  mapping = aes(  
    x = height_cm,  
    y = resid),  
  colour = "grey30",  
  size = 0.3  
) +  
facet_grid(sex ~ factor(married, labels = c("not married", "married")))
```

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



Konklusjon

Vi kan se ut ifra modellene våres, spessielt tydelig i modell 9, at høyde ikke er den avgjørende faktoren for inntekt. Det er mange flere faktorer som bestemmer dette slik som utdanning, BMI, kjønn, alder, antall år i en jobb, osv.

Studien gjennomført av Judge & Cable ser ut til å ha oversett viktige data i analysen deres, eller ikke gått nok i dybden på hvordan alle faktorer påvirker. En kan selvsagt ikke se helt bort fra at det er tilfeller hvor en høy person får en jobb som betaler bedre fremfor en lav person, men ut fra dataene og resultatene våres kan vi si en ting ganske sikkert:

Høyde alene er ikke de-facto grunnlag for høy lønn.

Referanser

“AN EMPIRICAL INVESTIGATION OF THE PREDICTORS OF EXECUTIVE CAREER SUCCESS - JUDGE - 1995 - Personnel Psychology - Wiley Online Library.” n.d.
<https://onlinelibrary.wiley.com/doi/10.1111/j.1744-6570.1995.tb01767.x>.

- Burns, David J. 1993. "Retail Salespersons: An Inquiry into Need Recognition." *Journal of Marketing Theory and Practice* 1 (3): 11–28.
- Higham, Philip, and D. Carment. 1992. "The Rise and Fall of Politicians: The Judged Heights of Broadbent, Mulroney and Turner Before and After the 1988 Canadian Federal Election." *Canadian Journal of Behavioural Science/Revue Canadienne Des Sciences Du Comportement* 24 (July): 404–9. <https://doi.org/10.1037/h0078723>.
- 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>.
- Whitely, William, Thomas W. Dougherty, and George F. Dreher. 1991. "Relationship of Career Mentoring and Socioeconomic Origin to Managers' and Professionals' Early Career Progress." *The Academy of Management Journal* 34 (2): 331–51. <https://doi.org/10.2307/256445>.

	hoyde	hoyde_filtretr
(Intercept)	-2321309.587 ***	-573873.741
	[-5.427]	[-1.801]
height_cm	15611.913 ***	5514.450 **
	[6.176]	[2.939]
weight_kg	-6129.694 *	270.450
	[-2.483]	[0.147]
maritalmarried	207941.582 ***	117904.614 ***
	[13.291]	[10.234]
maritalseparated	-32637.320	-30084.379
	[-1.179]	[-1.448]
maritaldivorced	70126.781 ***	56707.469 ***
	[3.882]	[4.272]
maritalwidowed	33568.334	-1930.571
	[0.863]	[-0.067]
bmi	6226.873	-1128.840
	[1.958]	[-0.474]
N	6911	5054
R2	0.088	0.082
logLik	-99857.326	-70408.867
AIC	199732.653	140835.734

*** p < 0.001; ** p < 0.01; * p < 0.05.

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
5.19e+03	1.19e+15				
5.19e+03	1.18e+15	2	1.41e+13	31.2	3.46e-14

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
6.76e+03	5.54e+14				
6.76e+03	5.5e+14	2	3.41e+12	21	8.46e-10

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
5.05e+03	3.73e+14				
5.04e+03	3.7e+14	2	3.14e+12	21.4	5.69e-10

weight	age	marital	sex	education	afqt	inntekt	height_cm	weight_kg	bmi
155	53	married	female	13	6.84	1.62e+05	152	69.8	66.7
156	51	married	female	10	49.4	2.98e+05	178	70.2	49.3
195	52	married	male	16	99.4	8.92e+05	165	87.8	71.5
197	54	married	female	14	44	3.4e+05	160	88.7	76.9
190	49	married	male	14	59.7	6.38e+05	168	85.5	67.6
200	49	divorced	female	18	98.8	8.67e+05	173	90	67
225	48	married	male	16	82.3	0	188	101	63.7
160	54	divorced	female	12	50.3	5.95e+05	163	72	60.5
162	55	divorced	male	12	89.7	5.1e+05	175	72.9	52.7
194	54	divorced	male	13	96	1.28e+06	175	87.3	63.2