R-seminar 6: Multippel regresjon

STV1020 Vår 2021

Uke 17

Introduksjon

Datasettet vi skal bruke i dag er hentet fra European Social Survey (ESS) runde 8, og respondentene er fra Sverige. Vi ønsker å undersøke sammenhengen mellom utdanning og tilfredshet med demokratiet. Er det slik at mennesker med høyere utdanning har høyere tilfredshet med demokratiet? Vi ønsker å kontrollere for alder, kjønn, hvor ofte man leser nyheter og om man stemte ved sist valg eller ikke. Vi skal bruke følgende variabler: "agea", "gndr", "eduyrs", "nwspol", "stfdem" og "vote".

- agea = Age of respondent
- gndr = Gender, Male = 1, Female = 2
- eduyrs = Years of full-time education completed. About how many years of education have you completed, whether full-time or part-time? Please report these in full-time equivalents and include compulsory years of schooling.
- nwspol = On a typical day, how much time do you spend watching, reading or listening to the news about politics or current affairs. Answer in minutes.
- stfdem = How statisfied are you with the way democracy works in your country? 0 = extremely dissatisifed, 10 = extremely satisfied
- vote = Did you vote in the last national election? 1 = Yes, 2 = No, 3 = Not eligble to vote.

1 Laste inn data

Først laster vi inn datasettet som vi har lagret i vårt working directory. Vi bruker funksjonen read_dta() fra haven-pakken. Husk å velge den funksjonen som passer til fil-formatet datasettet er i. Vi setter først en working directory, dersom du ikke jobber i prosjekt. Vi laster også ned de pakkene vi skal bruke i seminaret. Til laster vi inn datasettet.

```r
## Setter working directory (hvis du ikke jobber i prosjekt)
setwd("")

## Error in setwd(""):  cannot change working directory

## Laster inn pakker
library(tidyverse)
library(stargazer)
library(haven)

## Laster inn datasettet. Bruker funksjonen read_dta fra haven pakken.
data <- read_dta("ESS8SE.dta")
```

Det er alltid lurt å ta en titt på datasettet for å få en oversikt over hvordan datamaterialet er organisert.

```r
View(data)
head(data)

## # A tibble: 6 x 535
##   name  essround edition proddate  idno cntry nwspol netusoft     netustm
##   <chr>    <dbl> <chr>   <chr>    <dbl> <chr> <dbl+> <dbl+lb>   <dbl+lbl>
## 1 ESS8~        8 2.2     10.12.2~     1 SE       240 1 [Neve~ NA(a) [Not~
## 2 ESS8~        8 2.2     10.12.2~     3 SE        60 1 [Neve~ NA(a) [Not~
## 3 ESS8~        8 2.2     10.12.2~    10 SE       120 1 [Neve~ NA(a) [Not~
## 4 ESS8~        8 2.2     10.12.2~    14 SE        20 1 [Neve~ NA(a) [Not~
## 5 ESS8~        8 2.2     10.12.2~    15 SE       180 2 [Only~ NA(a) [Not~
## 6 ESS8~        8 2.2     10.12.2~    16 SE       210 1 [Neve~ NA(a) [Not~
## # ... with 526 more variables: ppltrst <dbl+lbl>, pplfair <dbl+lbl>,
## #   pplhlp <dbl+lbl>, polintr <dbl+lbl>, psppsgva <dbl+lbl>,
## #   actrolga <dbl+lbl>, psppipla <dbl+lbl>, cptppola <dbl+lbl>,
## #   trstprl <dbl+lbl>, trstlgl <dbl+lbl>, trstplc <dbl+lbl>, trstplt <dbl+lbl>,
## #   trstprt <dbl+lbl>, trstep <dbl+lbl>, trstun <dbl+lbl>, vote <dbl+lbl>,
## #   prtvtbat <dbl+lbl>, prtvtcbe <dbl+lbl>, prtvtfch <dbl+lbl>,
## #   prtvtdcz <dbl+lbl>, prtvede1 <dbl+lbl>, prtvede2 <dbl+lbl>,
## #   prtvtfee <dbl+lbl>, prtvtdes <dbl+lbl>, prtvtdfi <dbl+lbl>,
## #   prtvtcfr <dbl+lbl>, prtvtbgb <dbl+lbl>, prtvtehu <dbl+lbl>,
## #   prtvtbie <dbl+lbl>, prtvtcil <dbl+lbl>, prtvtbis <dbl+lbl>,
## #   prtvtbit <dbl+lbl>, prtvblt1 <dbl+lbl>, prtvblt2 <dbl+lbl>,
## #   prtvblt3 <dbl+lbl>, prtvtfnl <dbl+lbl>, prtvtbno <dbl+lbl>,
## #   prtvtdpl <dbl+lbl>, prtvtcpt <dbl+lbl>, prtvtdru <dbl+lbl>,
## #   prtvtbse <dbl+lbl>, prtvtesi <dbl+lbl>, contplt <dbl+lbl>,
## #   wrkprty <dbl+lbl>, wrkorg <dbl+lbl>, badge <dbl+lbl>, sgnptit <dbl+lbl>,
## #   pbldmn <dbl+lbl>, bctprd <dbl+lbl>, pstplonl <dbl+lbl>, clsprty <dbl+lbl>,
## #   prtclcat <dbl+lbl>, prtclcbe <dbl+lbl>, prtclfch <dbl+lbl>,
## #   prtcldcz <dbl+lbl>, prtclede <dbl+lbl>, prtclfee <dbl+lbl>,
## #   prtclees <dbl+lbl>, prtcldfi <dbl+lbl>, prtclefr <dbl+lbl>,
## #   prtclbgb <dbl+lbl>, prtclfhu <dbl+lbl>, prtcldie <dbl+lbl>,
## #   prtcldil <dbl+lbl>, prtclbis <dbl+lbl>, prtclcit <dbl+lbl>,
## #   prtclblt <dbl+lbl>, prtclenl <dbl+lbl>, prtclbno <dbl+lbl>,
## #   prtclgpl <dbl+lbl>, prtclept <dbl+lbl>, prtcldru <dbl+lbl>,
## #   prtclbse <dbl+lbl>, prtclesi <dbl+lbl>, prtdgcl <dbl+lbl>,
## #   lrscale <dbl+lbl>, stflife <dbl+lbl>, stfeco <dbl+lbl>, stfgov <dbl+lbl>,
## #   stfdem <dbl+lbl>, stfedu <dbl+lbl>, stfhlth <dbl+lbl>, gincdif <dbl+lbl>,
```

```
## #   mnrgtjb <dbl+lbl>, freehms <dbl+lbl>, hmsfmlsh <dbl+lbl>,
## #   hmsacld <dbl+lbl>, euftf <dbl+lbl>, imsmetn <dbl+lbl>, imdfetn <dbl+lbl>,
## #   impcntr <dbl+lbl>, imbgeco <dbl+lbl>, imueclt <dbl+lbl>, imwbcnt <dbl+lbl>,
## #   happy <dbl+lbl>, sclmeet <dbl+lbl>, inprdsc <dbl+lbl>, sclact <dbl+lbl>,
## #   crmvct <dbl+lbl>, aesfdrk <dbl+lbl>, ...
```

2 Subsetting

Vi trenger ikke alle variablene, og da er det nyttig å lage et subset av det opprinnelige datasettet. Her tar vi kun med de variablene vi vil ha til resten av analysen, og er nyttig for å få bedre oversikt over datasettet vårt. Vi bruker select()-funksjonen fra dplyr-pakken til å velge ut variablene vi vil ha med i datasettet. Det er lurt å lagre det nye datasettet som et nytt slik at vi ikke overskrider det originale datasettet. Det er nyttig å ha det opprinnelige datasettet dersom vi gjør noe feil senere.

```
## Tar kun med variablene vi trenger til regresjonen
data2 <- data %>%
  select(gndr, agea, eduyrs, nwspol, stfdem, vote)
```

Det er lurt å sjekke hvor mange NAs/missing det er i datasettet. Vi kan bruke table() og complete.cases(). FALSE viser hvilke enheter (rader) som har NAs på minst en variabel, mens TRUE viser enhetene (radene) som har complete verdier på alle variablene. Se også dokumentet til seminar 5 for informasjon.

```
## Bruker table() og complete.cases() for å sjekke omfanget av missing verdier.
table(complete.cases(data2))

##
## FALSE  TRUE
##    46  1505
```

Vi bruker funksjonen drop_na() til å fjerne enheter med NA-verdier. Dette er også en måte å subsette datasettet på. Å fjerne NA-verdier er noe man alltid bør tenke gjennom om er lurt eller ikke, særlig dersom enkelte variabler har svært mye missing. Er det systmatisk eller tilfeldig at det er NA?

```
## Fjerner NA observasjoner
data2 <- data2 %>%
  drop_na()
```

3 Omkoding

Før en kjører en regresjonsanalyse er det lurt å gjøre seg kjent med variablene. AV her er "stfdem".

```r
summary(data2$stfdem)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   5.000   7.000   6.429   8.000  10.000
```

Vi gjør oss videre kjent med resten av variablene. Hoveduavhengig variabel er "eduyrs".

```r
summary(data2$eduyrs)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   11.00   13.00   13.31   16.00   30.00
```

Hva med "gndr"?

```r
table(data2$gndr)
```

```
##
##   1   2
## 758 747
```

Vi vil omkode "gndr" slik at mann = 0 og kvinne = 1. Det kan på forhånd være lurt å tenke over hvilke forventninger vi har til variabelen. Her forventer vi kanskje at kvinner er mer fornøyde med demokratiet enn menn. Da kvinne kodes som 1 vil vi i så fall få en positiv koeffisient dersom forventningen innfris. Vi bruker mutate() til å opprette en ny variabel for kjønn. Vi bruker ifelse() for å omkode variabelen. Koden sier: dersom gdnr har verdier 1 skal den nye variabelen gndr_new får verdien 0, dersom gndr har andre verdier får disse verdien 1 i den nye variabelen. Deretter sjekker vi at omkodingen har gått i orden.

```r
data2 <- data2 %>%
  mutate(gndr_new = ifelse(gndr == 1, 0, 1))

# Sjekker at omkodingen ser ok ut
table(data2$gndr, data2$gndr_new)
```

```
##
##       0   1
##   1 758   0
##   2   0 747
```

Hva med "vote"?

```r
table(data2$vote)
```

```
##
##    1    2    3
## 1336   85   84
```

Denne variabelen har informasjon om hvordan respondentene stemte forrige valg, og er kodet "ja", "nei" eller "ikke stemmeberettiget". Vi vil heller ha en

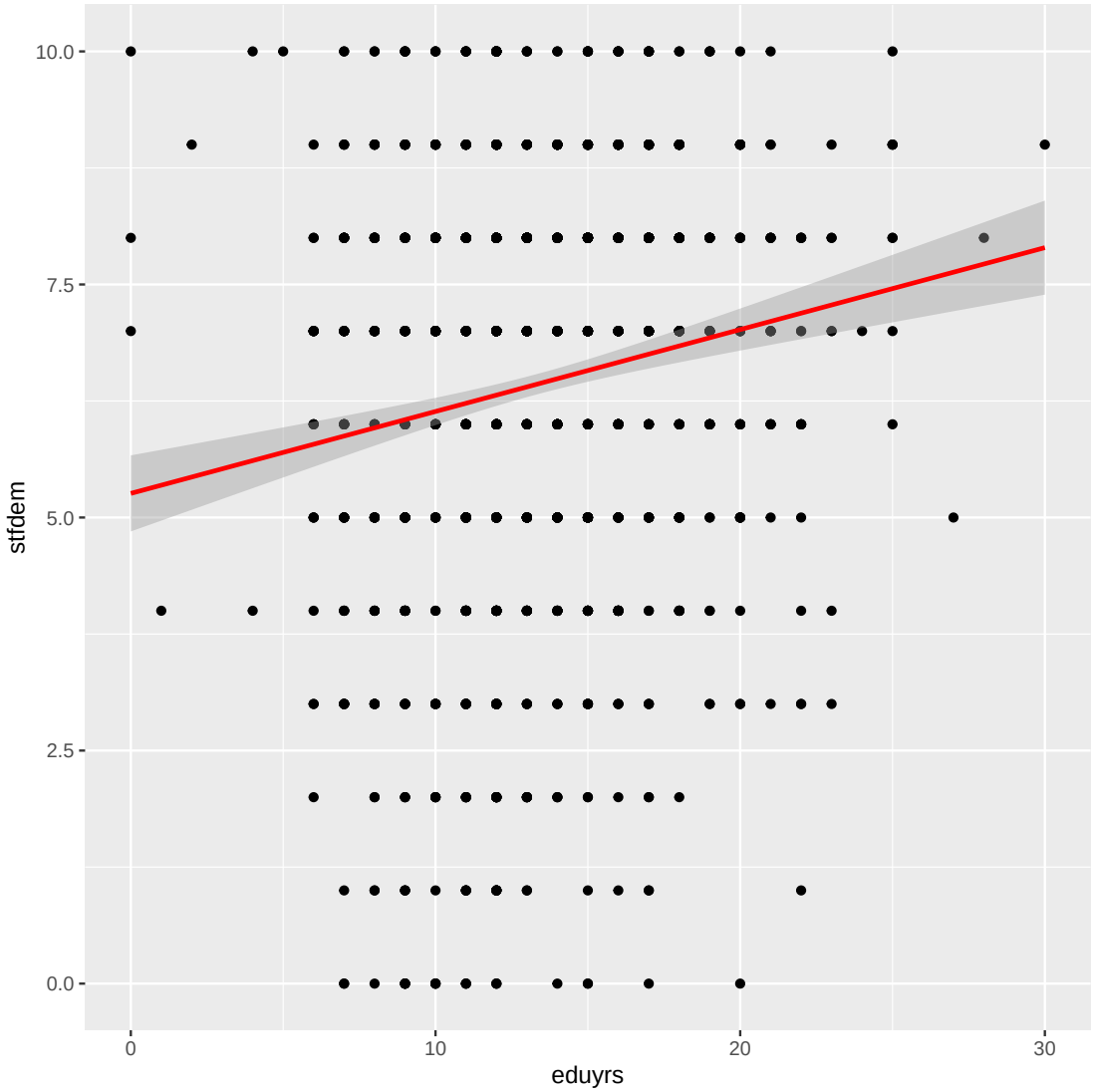
dikotom variabel som fanger opp hvorvidt personen stemte ved sist valg eller ikke. Det betyr at vi vil ha kategori 2 og 3 i samme kategori i den nye variabelen. Dette gjør vi også ved hjelp av mutate(). Respondentene som har verdien 2 eller høyere får verdien 0, mens resten får 1. >= betyr større enn eller er lik. Vi sjekker at omkodingen gikk i orden.

```
## Dikotom variabel for hvorvidt respondentene stemte ved siste valg eller ikke
data2 <- data2 %>%
  mutate(vote_new = ifelse(vote >= 2, 0, 1))

# Sjekker at omkodingen ser ok ut
table(data2$vote_new)

##
##    0    1
##  169 1336
```

4 Plotting

Før vi kjører i gang med regresjonsanalyse kan det være greit å lage et spredningsplott over AV (stfdem) og UV (eduyrs). Vi kan også legge til en regresjonslinje ved hjelp av geom_smooth()-argumentet

```
ggplot(data2, aes(x = eduyrs, y = stfdem)) +
  geom_point() +
  geom_smooth(method = "lm", col = "red") # line color is red
```



5 Multippel regresjon

Når vi kjører lineære regresjonsmodeller (OLS), bruker vi lm()-funksjonen. Se seminar 5 for mer informasjon. Når vi vil legge til flere variabler, bruker vi +. Slik:

```
mod <- lm(AV ~ UV + UV1 + UV2 + UV3,
          data = data)

## Error in eval(predvars, data, env):  object 'AV' not found
```

Først prøver vi med en bivarat regresjonsmodell, som var tema for seminar 5. Her undersøker vi sammenhengen mellom utdanning ("eduyrs") (UV) og tilfredshet med demokratiet ("stfdem") (AV). Regresjonen lagres i objektet mod1. Vi bruker summary() for å se resultatene av modellen vår.

```r
## Kjører bivariat regresjon først, der AV: stfdem og UV: eduyrs
mod1 <- lm(stfdem ~ eduyrs,
           data = data2)

summary(mod1)

##
## Call:
## lm(formula = stfdem ~ eduyrs, data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.0162 -1.3131  0.5111  1.5990  4.7414
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.25860    0.20792   25.29  < 2e-16 ***
## eduyrs       0.08788    0.01505    5.84  6.4e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.157 on 1503 degrees of freedom
## Multiple R-squared:  0.02219,Adjusted R-squared:  0.02153
## F-statistic:  34.1 on 1 and 1503 DF,  p-value: 6.401e-09
```

Vi prøver med de seks variablene våre for å undersøke sammenhengen mellom utdanning (”eduyrs”) og tilfredshet med demokratiet (”stfdem”), kontrollert for hvor ofte man leser nyheter, om man stemte ved sist valg eller ikke, alder og kjønn. Vi bruker dermed følgende kontrollvariabler: ”nwspol”, ”vote_new”, ”agea” og ”gndr_new”.

```r
## Multivariat regresjon der vi legger til 4 kontrollvariabler
mod2 <- lm(stfdem ~ eduyrs +
             nwspol +
             vote_new +
             agea +
             gndr_new,
           data = data2)

summary(mod2)

##
## Call:
## lm(formula = stfdem ~ eduyrs + nwspol + vote_new + agea + gndr_new,
##     data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.1939 -1.3181  0.4166  1.5766  4.3671
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.9953181  0.2927606  20.479  < 2e-16 ***
```

```
## eduyrs       0.0857999  0.0155097   5.532 3.73e-08 ***
## nwspol       0.0003356  0.0006568   0.511  0.60944
## vote_new    -0.5278922  0.1909703  -2.764  0.00577 **
## agea        -0.0054872  0.0033306  -1.647  0.09967 .
## gndr_new     0.0382974  0.1112279   0.344  0.73066
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.148 on 1499 degrees of freedom
## Multiple R-squared:  0.03272,Adjusted R-squared:  0.0295
## F-statistic: 10.14 on 5 and 1499 DF,  p-value: 1.438e-09
```

Det går også an å legge til samspillseffekter i modellen vår ved å bruke * mellom de to variablene vi vi ha et samspill mellom. eduyrs*nwspol vil si at vi legger til et samspill mellom utdanning og konsum av politiske nyheter. Samspill kan også kalles for statistisk interaksjon, og innebærer at effekten av en uavhengig variabel (for eksempel sigarettrøyking) på en avhengig variabel (for eksempel helsetilstand) varierer med verdiene på en annen uavhengig variabel (for eksempel å være utsatt for asbest).

```
## Legger til et samspillsledd mellom eduyrs og nwspol med *
mod3 <- lm(stfdem ~ eduyrs * nwspol +
             vote_new +
             agea +
             gndr_new,
           data = data2)

summary(mod3)

##
## Call:
## lm(formula = stfdem ~ eduyrs * nwspol + vote_new + agea + gndr_new,
##     data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.2467 -1.3207  0.4177  1.5787  4.4767
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.8222123  0.3363187  17.312  < 2e-16 ***
## eduyrs         0.0994616  0.0202791   4.905 1.04e-06 ***
## nwspol         0.0026828  0.0023388   1.147  0.25154
## vote_new      -0.5357177  0.1911109  -2.803  0.00513 **
## agea          -0.0054747  0.0033305  -1.644  0.10043
## gndr_new       0.0367189  0.1112347   0.330  0.74137
## eduyrs:nwspol -0.0001791  0.0001712  -1.046  0.29590
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.148 on 1498 degrees of freedom
## Multiple R-squared:  0.03343,Adjusted R-squared:  0.02956
## F-statistic: 8.635 on 6 and 1498 DF,  p-value: 2.962e-09
```

Legg merke til at vi får koeffisienter for både samspillsleddet (eduyrs:nwspol) og enkeltvariablene (eduyrs) (nwspol) selv. Dette er viktig informajson når vi skal tolke resultatene fra en modell med samspill.

Vi kan presentere resultatene våre med stargazer(). Her kan vi inkludere flere modeller ved siden av hverandre, slik:

```
stargazer(mod1, mod2,
          type = "text")
```

Sjekk hjelpefilen ?stargazer. Vi bruker type = til å spesifisere hvilken format vi vil ha tabellen i. De følgende spesifikasjonene under er ikke viktig at dere kan, men det er greit å vite at man kan tilpasse tabellen enda mer. covariate.labels bruker vi for å legge til nye navn på de uavhengige variablene i modellen. De må skrives i samme rekkefølge som i lm()-formulaen. dep.var.labels gir navn til den avhengige variabelen på toppen.

```
## Presenterer resultatet av mod1 og mod2 i stargazer
stargazer(mod1, mod2, # Legger til begge modellene
          type = "text", # Spesifiserer tabell-type
          title = "Regresjonstabeller", # Tittel på tabellen
          covariate.labels = c("Utdanning", # Forklarende navn på variabler
                               "Politiske nyheter",
                               "Stemte ved forrige valg",
                               "Alder",
                               "Kjønn"),
          dep.var.labels = "Tilfredshet med demokratiet") # Forklarende navn AV

##
## Regresjonstabeller
## ================================================================================
##                                          Dependent variable:
##                              ---------------------------------------------------
##                                     Tilfredshet med demokratiet
##                                    (1)                          (2)
## --------------------------------------------------------------------------------
## Utdanning                         0.088***                     0.086***
##                                   (0.015)                      (0.016)
##
## Politiske nyheter                                               0.0003
##                                                                 (0.001)
##
## Stemte ved forrige valg                                        -0.528***
##                                                                 (0.191)
##
## Alder                                                           -0.005*
##                                                                 (0.003)
##
## Kjønn                                                            0.038
##                                                                 (0.111)
##
## Constant                          5.259***                     5.995***
##                                   (0.208)                      (0.293)
```

```
##
## ------------------------------------------------------------------------
## Observations                      1,505                    1,505
## R2                                0.022                    0.033
## Adjusted R2                       0.022                    0.029
## Residual Std. Error        2.157 (df = 1503)        2.148 (df = 1499)
## F Statistic             34.101*** (df = 1; 1503) 10.142*** (df = 5; 1499)
## ========================================================================
## Note:                                          *p<0.1; **p<0.05; ***p<0.01
```

Vi kan lagre tabellen vår i html-format også, dersom vi ønsker å bruke den i et word dokument for eksempel. Det er nyttig å vite dersom du skal skrive en kvantitativ bacheloroppgave. Da endrer vi type-argumentet til type = "html", og legger til et out-argument, slik:

```
## Stargazer i html-format
stargazer(mod1, mod2,
          type = "html", #html-format
          out = "regresjonstabell.html", # Spesifisere filnavn
          title = "Regresjonstabeller",
          covariate.labels = c("Utdanning",
                               "Politiske nyheter",
                               "Stemte ved forrige valg",
                               "Alder",
                               "Kjønn"),
          dep.var.labels = "Tilfredshet med demokratiet")
```

Tabellen lagres i working directory, og vi kan åpne den fra mappen vår.

6 OLS forutsetninger

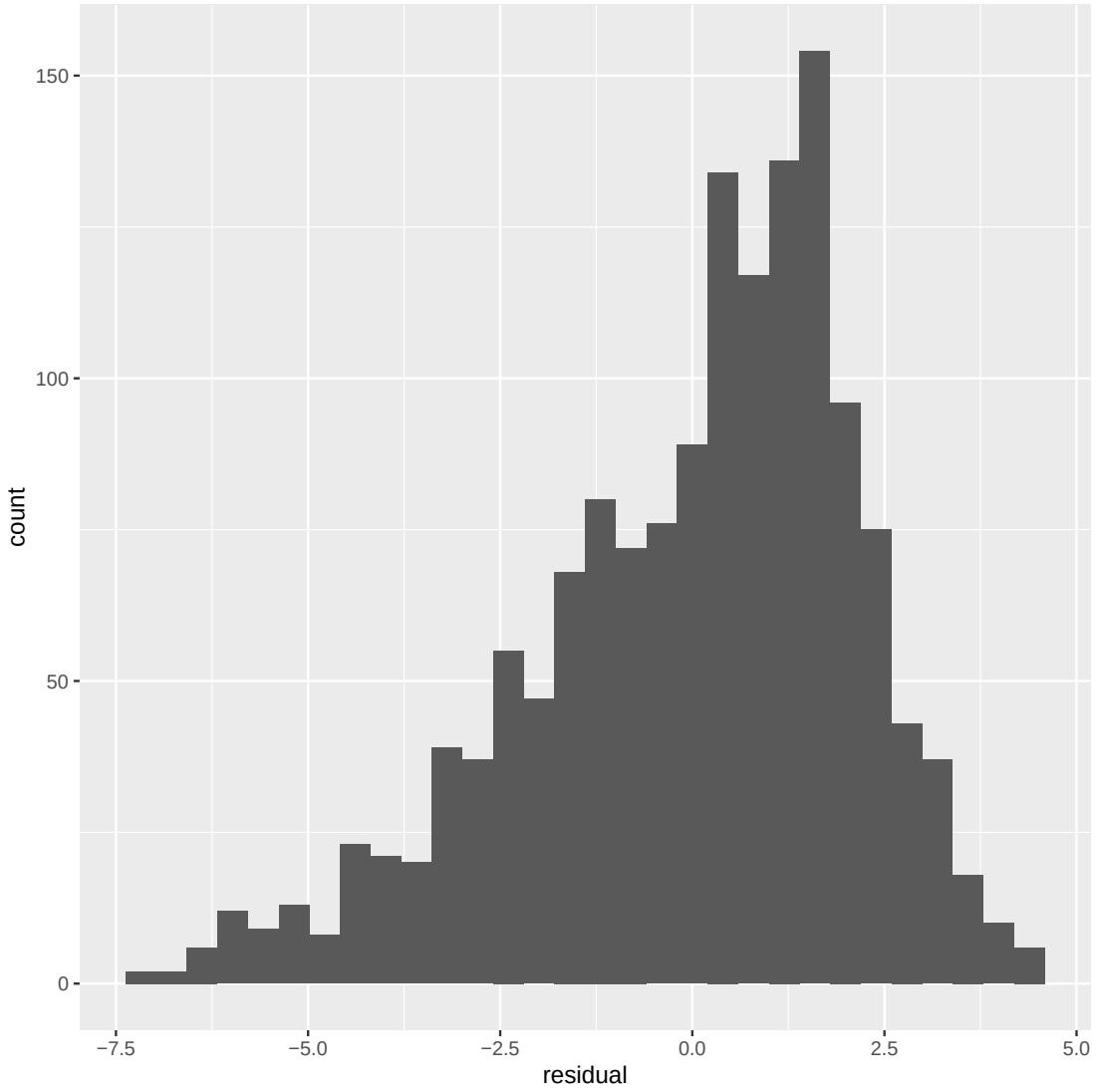
Til slutt skal vi se på noen grafiske verktøy for å vurdere om enkelte forutsetninger for OLS er oppfylt. Dette vil vil ikke gå gjennom på seminar og er mindre relevant for prøve. Les gjerne likevel dette dersom du lurer på hva noen av regresjonsforutsetningene er, som normalfordelte og heteroskedastiske restledd.

Vi må først lagre restleddene og verdiene fra modellen vår i datasettet vårt. Dette kan gjøres med mutate. Vi lager variablene "residual" og "fit". fitted og resid ligger allerede innbakt i modellen vår.

```
data2 <- data2 %>%
  mutate(residual = resid(mod2),
         fit = fitted(mod2))
```

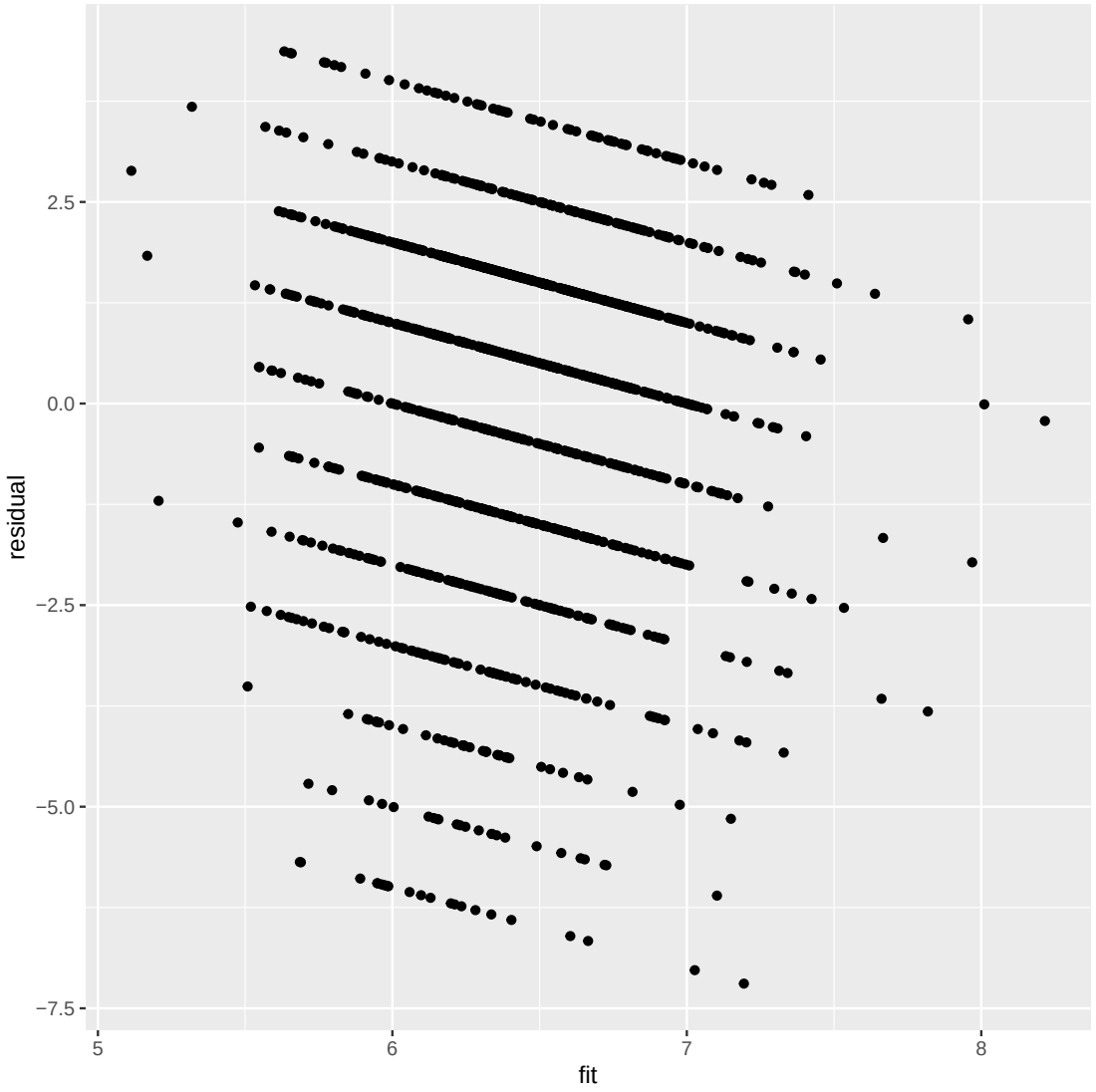
Vi vil så vurdere restleddenes fordeling med et histogram. Er restleddene våre normalfordelte?

```
ggplot(data2, aes(residual)) +
  geom_histogram()
```



Nå vil vi lage en figur som plotter restleddene mot modelles verdier. Dette gjør vi for å vurdere eventuell heteroskedastisitet. Vi bruker de to nye fit-verdiene som er lagret i datasettet på x-aksen, og residual på y-aksen. Vi lagrer et spredningsplot. Slike plot kan være vanskelige å tolke, og hvor enkel eller vanskelig tolkningen blir avheniger ofte av hvordan variablene våre er kodet.

```
ggplot(data2, aes(x= fit, y = residual)) +
  geom_point()
```


residual
2.5
0.0
−2.5
−5.0
−7.5
5
6
7
8
fit