Seminar 3

Martin Søyland

Disposisjon

1. Laste inn data #advanced
2. Omkoding #intermediate
3. Subsetting av data
4. OLS Samspill
5. Plotte predikerte effekter

Laste inn data

Så langt i seminarene har vi gjort det veldig enkelt ved å laste inn data fra helt standard .csv-filer. Det finnes, dessverre, en haug med andre filtyper. La oss laste ned et datasett i mange forskjellige format fra http://folk.uio.no/martigso/stv4020/aiddata.zip

Her ligger data fra Burnside & Dollar (2000) i STATA- (.dta), SPSS- (.sav) og to versjoner av R-format (.rda / .RData). Pakk ut disse i mappen du vil jobbe fra, og la oss starte med det siste:

R-data

```
load("./data/aidgrowth/aidgrowth.rda")

head(aid, 3)
```

```
##    country period gdp_growth       aid     policy gdp_pr_capita
## 34     ARG      2   1.700300 0.0182389  0.6565560          5637
## 35     ARG      3   1.077615 0.0171555 -0.5792648          6168
## 36     ARG      4  -1.115285 0.0239942 -0.1356454          5849
##    ethnic_frac assasinations sub_saharan_africa fast_growing_east_asia
## 34        0.31          2.75                  0                      0
## 35        0.31          9.75                  0                      0
## 36        0.31          1.00                  0                      0
##    institutional_quality m2_gdp_lagged
## 34               4.28125      24.82476
## 35               4.28125      28.79304
## 36               4.28125      30.23452
```

```
rm(aid) # rm() fjerner objektet fra Environment
```

SPSS-data

```
#install.packages("haven")

library(haven)

aid <- read_sav("./data/aidgrowth/aidgrowth.sav")
head(aid, 3)
```

```
## # A tibble: 3 x 12
##   country period gdp_growth       aid     policy gdp_pr_capita ethnic_frac
##     <chr>  <dbl>      <dbl>     <dbl>      <dbl>         <dbl>       <dbl>
## 1     ARG      2   1.700300 0.0182389  0.6565560          5637        0.31
## 2     ARG      3   1.077615 0.0171555 -0.5792648          6168        0.31
## 3     ARG      4  -1.115285 0.0239942 -0.1356454          5849        0.31
## # ... with 5 more variables: assasinations <dbl>,
## #   sub_saharan_africa <dbl>, fast_growing_east_asia <dbl>,
## #   institutional_quality <dbl>, m2_gdp_lagged <dbl>
```

```
rm(aid)
```

STATA-data

```
#install.packages("haven")
# library(haven)

aid <- read_dta("./data/aidgrowth/aidgrowth.dta")
head(aid, 3)
```

```
## # A tibble: 3 x 12
##   country period gdp_growth       aid     policy gdp_pr_capita ethnic_frac
##     <chr>  <dbl>      <dbl>     <dbl>      <dbl>         <dbl>       <dbl>
## 1     ARG      2   1.700300 0.0182389  0.6565560          5637        0.31
## 2     ARG      3   1.077615 0.0171555 -0.5792648          6168        0.31
## 3     ARG      4  -1.115285 0.0239942 -0.1356454          5849        0.31
## # ... with 5 more variables: assasinations <dbl>,
## #   sub_saharan_africa <dbl>, fast_growing_east_asia <dbl>,
## #   institutional_quality <dbl>, m2_gdp_lagged <dbl>
```

Kort introduksjon til subsetting av data

Noen ganger vil vi fjerne enheter fra datasettene våre. Det kan være vi kun vil se på trender innad i et land, sammenligne spesifikke grupper, fjerne enheter som skaper systematisk missing, og så videre. Under viser jeg to måter dette kan gjøres på – det finnes mange flere.

```
# aid$country == "ARG"
# which(aid$country == "ARG")
# aid[which(aid$country == "ARG"), ]
argentina <- aid[which(aid$country == "ARG"), ]
argentina <- subset(aid, country == "ARG")

nomiss_policy <- aid[which(is.na(aid$policy) == FALSE), ]
nomiss_policy <- subset(aid, is.na(policy) == FALSE)

# "&" Betyr AND, "|" betyr OR
nomiss_policy_growth <- aid[which(is.na(aid$gdp_growth) == FALSE & is.na(aid$policy) == FALSE), ]
nomiss_policy_growth <- subset(aid, is.na(gdp_growth) == FALSE | is.na(policy) == FALSE)

rm(argentina, nomiss_policy, nomiss_policy_growth)
```

Replikasjon uten å gjøre noe

La oss kjapt kjøre alle variablene fra modell 5 i artikkelen inn i en regresjon og se om vi får samme resultat. Da bare putter vi inn variablene fra artikkelen i en `lm()`.

```
model5 <- lm(gdp_growth ~ gdp_pr_capita + ethnic_frac * assasinations +
               institutional_quality + m2_gdp_lagged +
               sub_saharan_africa + fast_growing_east_asia + policy * aid,
             data = aid)
summary(model5)
```

```
##
## Call:
## lm(formula = gdp_growth ~ gdp_pr_capita + ethnic_frac * assasinations +
##     institutional_quality + m2_gdp_lagged + sub_saharan_africa +
##     fast_growing_east_asia + policy * aid, data = aid)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.1538  -1.5874   0.0386   1.5829  13.6089
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)               -1.0410219  0.9747372  -1.068  0.28652
## gdp_pr_capita             -0.0003812  0.0001368  -2.786  0.00574 **
## ethnic_frac               -0.4433386  0.8564135  -0.518  0.60513
## assasinations             -0.4570900  0.3181723  -1.437  0.15204
## institutional_quality      0.7537407  0.1858934   4.055 6.65e-05 ***
## m2_gdp_lagged             -0.0137827  0.0162300  -0.849  0.39655
## sub_saharan_africa        -1.9458082  0.7089091  -2.745  0.00648 **
## fast_growing_east_asia     1.1109101  0.7742178   1.435  0.15253
## policy                     0.7251472  0.2465391   2.941  0.00357 **
## aid                       -0.1492503  0.1808464  -0.825  0.40997
## ethnic_frac:assasinations  0.6781451  0.6604810   1.027  0.30550
## policy:aid                 0.1563717  0.1064776   1.469  0.14316
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.07 on 258 degrees of freedom
##   (61 observations deleted due to missingness)
## Multiple R-squared:  0.2947, Adjusted R-squared:  0.2647
## F-statistic: 9.802 on 11 and 258 DF,  p-value: 8.051e-15
```

Hvis vi sammenligner disse resultatene med resultatene i artikkelen ser vi kjapt at vi ikke har klart å replisere resultatene. Da må vi tilbake og lese teksten i artikkelen. Jeg har juksa og lest den på forhånd. Så jeg kan avsløre at det er noe som mangler her:

1. Gdp pr capita skal log-transformeres
2. De har kontrollert for periode, men ikke rapportert det
3. Område-dummiene er ikke satt til riktig målenivå
4. De viser ikke konstantledd; alltid vis konstantledd
5. (De bruker også "heteroskedasticity-consistent standard errors")

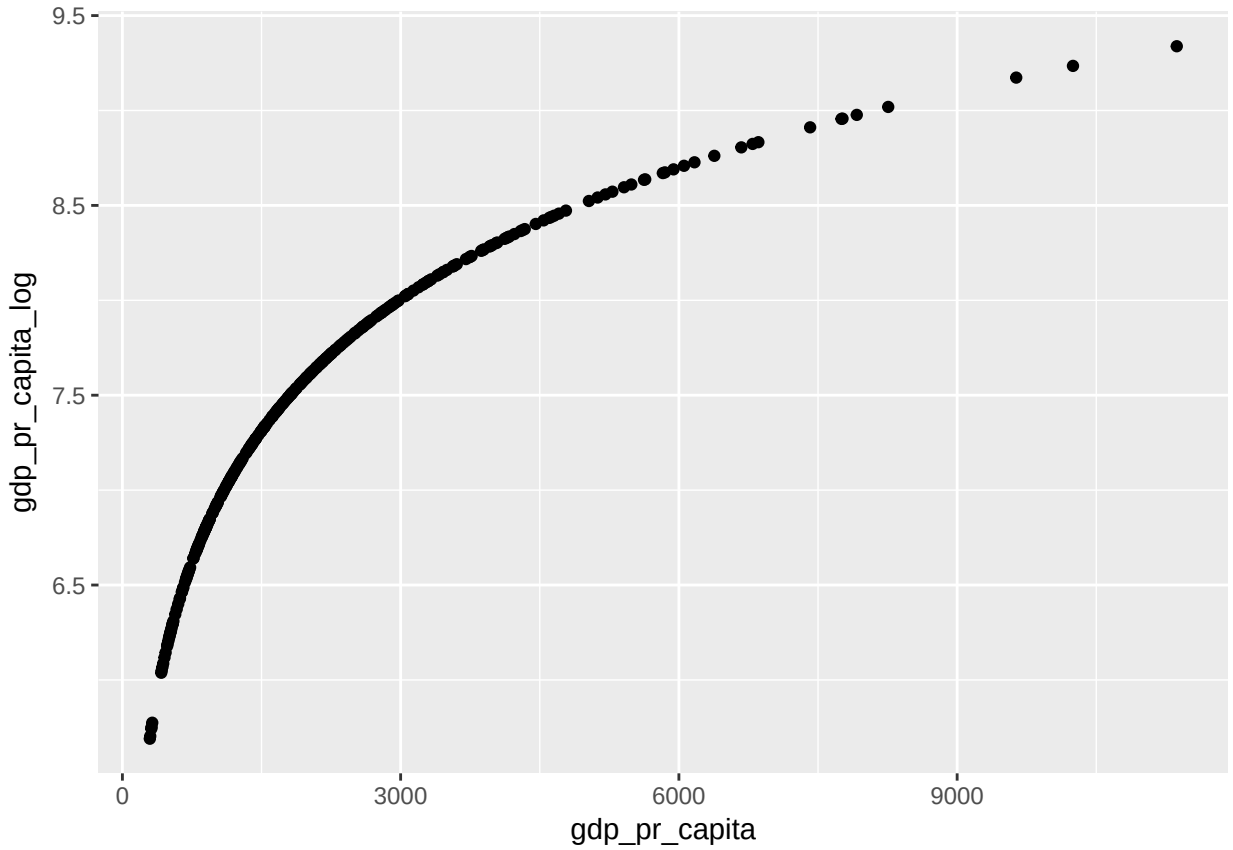
Omkoding

Logging av variabler

I artikkelen blir gdp per capit log-transformert. Derfor må vi også gjøre det for å klare å replisere resultatene. Det er også lurt å ta vare på den gamle variabelen ved å gi den nye et nytt navn. Her `gdp_pr_capita_log`.

```
aid$gdp_pr_capita_log <- log(aid$gdp_pr_capita)

library(ggplot2)

ggplot(aid, aes(x = gdp_pr_capita, y = gdp_pr_capita_log)) + geom_point()
```

```
## Warning: Removed 7 rows containing missing values (geom_point).
```



Her ser vi veldig tydelig hva som skjer når vi logger en variabel; en økning på 1 er viktigere for lave tall enn for høye tall. Kort sagt betyr dette at vi mener det er viktigere å gå fra å være veldig fattig til litt mindre fattig, enn å gå fra å være veldig rik til enda rikere.

1. ~~Gdp pr capita skal log-transformeres~~
2. De har kontrollert for periode, men ikke rapportert det
3. Område-dummiene er ikke satt til riktig målenivå
4. ~~De viser ikke konstantledd; alltid vis konstantledd~~
5. (De bruker også "heteroskedasticity-consistent standard errors")

Omkode dummies til èn variabel

Mest for å vise litt forskjellige måter å kode på, fikser vi også dummysettet fra artikkelen inn i èn variabel. Dette kan gjøres på en million måter, men jeg skal vise to. Den første utnytter indeksering. Den andre bruker en snarvei, via funksjonen `ifelse()`.

```
# Illustrasjon på which():
aid$country[which(aid$sub_saharan_africa == 0 & aid$fast_growing_east_asia == 1)]
which(aid$sub_saharan_africa == 0 & aid$fast_growing_east_asia == 1)
aid$sub_saharan_africa == 0 & aid$fast_growing_east_asia == 1
```

```
aid$regions <- NA
aid$regions[which(aid$sub_saharan_africa == 0 & aid$fast_growing_east_asia == 0)] <- "Other"
```

```r
aid$regions[which(aid$sub_saharan_africa == 1 & aid$fast_growing_east_asia == 0)] <- "Sub-Saharan Africa"
aid$regions[which(aid$sub_saharan_africa == 0 & aid$fast_growing_east_asia == 1)] <- "East Asia"
```

I den neste funksjonen bruker vi **nesting** med `ifelse()` for å gjøre akkurat det samme. Dette viser at de to variablene er helt identiske. Vi sjekker at det blir riktig, og setter kategorien "Other" til referansekategori med funksjonen `factor()`.

```r
# ?ifelse()

aid$regions2 <- ifelse(aid$sub_saharan_africa == 1, "Sub-Saharan Africa",
                       ifelse(aid$fast_growing_east_asia == 1, "East Asia", "Other"))

table(aid$regions, aid$regions2)
```

```
##
##                       East Asia Other Sub-Saharan Africa
##   East Asia                  30     0                  0
##   Other                       0   177                  0
##   Sub-Saharan Africa          0     0                124
```

```r
aid$regions <- factor(aid$regions, levels = c("Other", "Sub-Saharan Africa", "East Asia"))
```

1. ~~Gdp pr capita skal log-transformeres~~
2. De har kontrollert for periode, men ikke rapportert det
3. ~~Område-dummiene er ikke satt til riktig målenivå~~
4. ~~De viser ikke konstantledd; alltid vis konstantledd~~
5. (De bruker også "heteroskedasticity-consistent standard errors")

Korrekt modell

Nå er vi faktisk klar til å kjøre regresjonen! Legg merke til at vi har satt inn `factor(period)` direkte i regresjonen

```r
model5 <- lm(gdp_growth ~ gdp_pr_capita_log + ethnic_frac * assasinations +
               institutional_quality + m2_gdp_lagged + regions + policy * aid +
               factor(period),
             data = aid, na.action = "na.exclude")
results <- summary(model5)
results
```

```
##
## Call:
## lm(formula = gdp_growth ~ gdp_pr_capita_log + ethnic_frac * assasinations +
##     institutional_quality + m2_gdp_lagged + regions + policy *
##     aid + factor(period), data = aid, na.action = "na.exclude")
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.7213  -1.6078  -0.1369   1.5895  12.0507
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 3.39109    2.94549   1.151 0.250702
## gdp_pr_capita_log          -0.59961    0.39262  -1.527 0.127965
## ethnic_frac                -0.42359    0.81009  -0.523 0.601507
## assasinations              -0.44895    0.30119  -1.491 0.137311
## institutional_quality       0.68684    0.17452   3.936 0.000107 ***
## m2_gdp_lagged               0.01222    0.01627   0.751 0.453130
## regionsSub-Saharan Africa -1.87248    0.68095  -2.750 0.006393 **
## regionsEast Asia            1.30739    0.73063   1.789 0.074747 .
## policy                      0.71245    0.24359   2.925 0.003760 **
## aid                        -0.02078    0.17808  -0.117 0.907182
## factor(period)3            -0.01252    0.61994  -0.020 0.983901
## factor(period)4            -1.41449    0.62920  -2.248 0.025434 *
## factor(period)5            -3.46987    0.64085  -5.415 1.43e-07 ***
## factor(period)6            -2.01030    0.66149  -3.039 0.002622 **
## factor(period)7            -2.25625    0.70848  -3.185 0.001631 **
## ethnic_frac:assasinations   0.79154    0.62031   1.276 0.203111
## policy:aid                  0.18622    0.10113   1.841 0.066752 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.873 on 253 degrees of freedom
##   (61 observations deleted due to missingness)
## Multiple R-squared:  0.3944, Adjusted R-squared:  0.3561
## F-statistic:  10.3 on 16 and 253 DF,  p-value: < 2.2e-16
```

```r
# "Heteroskedasticity-consistent standard errors"
library(sandwich)
results$coefficients[, "Std. Error"] <- sqrt(diag(vcovHC(model5, type = "HC")))
results$cov.unscaled <- vcovHC(model5, type = "HC")
round(results$coefficients[, c("Estimate", "Std. Error")], digits = 2)
```

```
##                            Estimate Std. Error
## (Intercept)                    3.39       4.11
## gdp_pr_capita_log             -0.60       0.57
## ethnic_frac                   -0.42       0.72
## assasinations                 -0.45       0.26
## institutional_quality          0.69       0.17
```

```
## m2_gdp_lagged                 0.01       0.01
## regionsSub-Saharan Africa    -1.87       0.75
## regionsEast Asia              1.31       0.58
## policy                        0.71       0.19
## aid                          -0.02       0.16
## factor(period)3              -0.01       0.57
## factor(period)4              -1.41       0.63
## factor(period)5              -3.47       0.59
## factor(period)6              -2.01       0.53
## factor(period)7              -2.26       0.64
## ethnic_frac:assasinations     0.79       0.44
## policy:aid                    0.19       0.07
```

1. ~~Gdp pr capita skal log-transformeres~~
2. ~~De har kontrollert for periode, men ikke rapportert det~~
3. ~~Område-dummiene er ikke satt til riktig målenivå~~
4. ~~De viser ikke konstantledd; alltid vis konstantledd~~
5. ~~(De bruker også "heteroskedasticity-consistent standard errors")~~

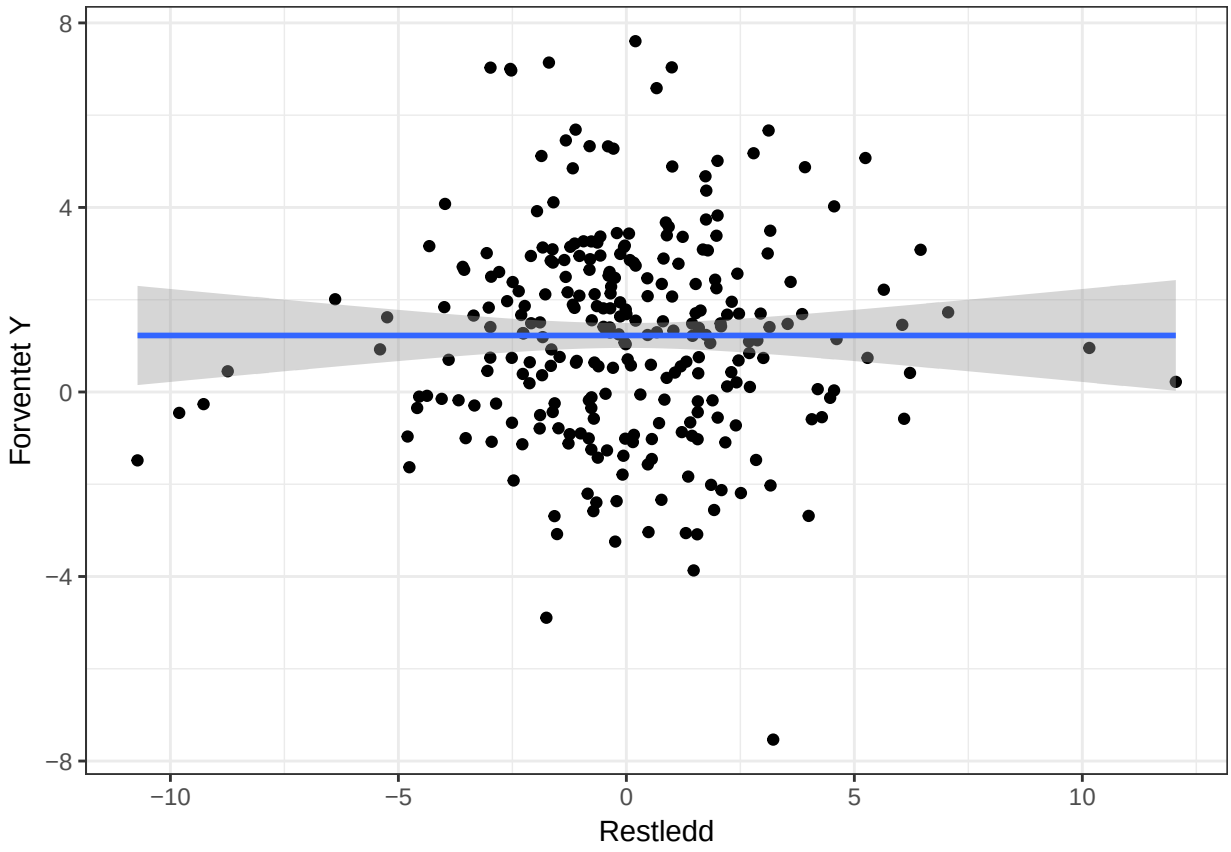
Plot av forventede verdier (og restledd)

For å evaluere hvor bra modellen er, må vi gjøre flere ting. Koden `plot(modell5)` vil vise mange forskjellige model-fit plot. Nedefor lager jeg ett av disse manuelt: restledd mot forventede verdier.

```
theme_set(theme_bw())

aid$pred <- predict(model5)
aid$restledd <- resid(model5)

ggplot(aid, aes(x = restledd, y = pred)) + geom_point() +
  geom_smooth(method = "lm") +
  labs(x = "Restledd", y = "Forventet Y")
```

```
## Warning: Removed 61 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 61 rows containing missing values (geom_point).
```



Ser dette bra ut?

Bonus!

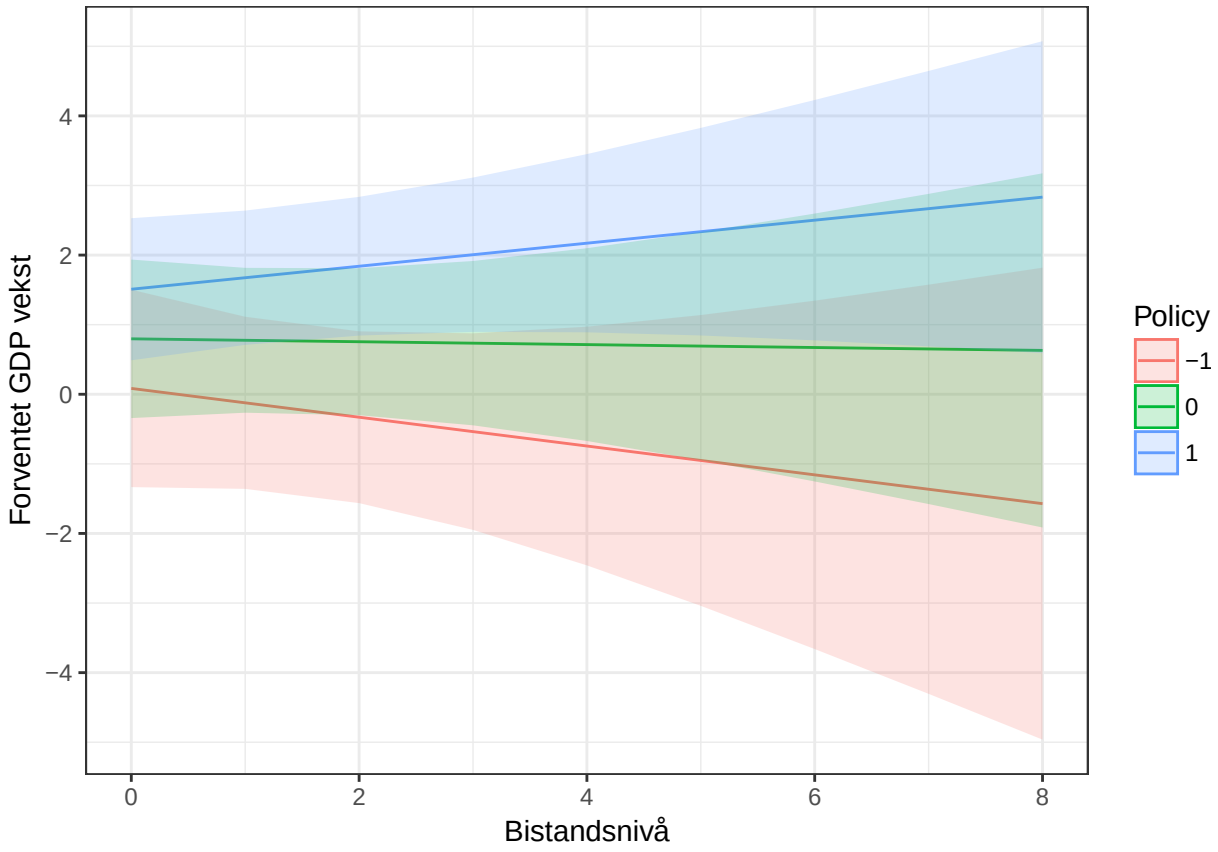
Samspillsledd er vanskelige å tolke fra en regresjonstabell. Jeg foretrekker å plotte effektene. Her er et eksempel på hvordan dette kan gjøres. Det er mye som skjer her, så ha tunga rett i munnen når dere prøver dere på dette.

```
snitt_data <- data.frame(gdp_pr_capita_log = mean(aid$gdp_pr_capita_log, na.rm = TRUE),
                         ethnic_frac = mean(aid$ethnic_frac, na.rm = TRUE),
                         assasinations = mean(aid$assasinations, na.rm = TRUE),
                         institutional_quality = mean(aid$institutional_quality, na.rm = TRUE),
                         m2_gdp_lagged = mean(aid$m2_gdp_lagged, na.rm = TRUE),
                         regions = "Other",
                         policy = c(rep(-1, 9), rep(0, 9), rep(1, 9)),
                         aid = rep(0:8, 3),
                         period = median(aid$period, na.rm = TRUE))

predict(model5, newdata = snitt_data, se = TRUE)
```

```
## $fit
##           1           2           3           4           5           6
```

```
##  0.08409742 -0.12290217 -0.32990176 -0.53690135 -0.74390094 -0.95090054
##           7           8           9          10          11          12
## -1.15790013 -1.36489972 -1.57189931  0.79654815  0.77576455  0.75498094
##          13          14          15          16          17          18
##  0.73419734  0.71341373  0.69263013  0.67184652  0.65106292  0.63027931
##          19          20          21          22          23          24
##  1.50899889  1.67443127  1.83986365  2.00529603  2.17072841  2.33616079
##          25          26          27
##  2.50159317  2.66702555  2.83245794
##
## $se.fit
##         1         2         3         4         5         6         7
## 0.7202650 0.6279388 0.6266786 0.7169646 0.8707720 1.0608217 1.2709582
##         8         9        10        11        12        13        14
## 1.4927225 1.7216269 0.5782802 0.5284831 0.5362879 0.5994488 0.7032047
##        15        16        17        18        19        20        21
## 0.8325135 0.9772843 1.1315980 1.2920401 0.5183695 0.4898167 0.5059652
##        22        23        24        25        26        27
## 0.5629815 0.6502018 0.7572607 0.8769218 1.0046925 1.1378441
##
## $df
## [1] 253
##
## $residual.scale
## [1] 2.872583
```

```
snitt_data <- cbind(snitt_data, predict(model5, newdata = snitt_data,
                                        se = TRUE, interval = "confidence"))

ggplot(snitt_data, aes(x = aid, y = fit.fit,
                       group = factor(policy),
                       color = factor(policy),
                       fill = factor(policy))) +
  geom_line() +
  scale_y_continuous(breaks = seq(-12, 12, 2)) +
  geom_ribbon(aes(ymin = fit.lwr, ymax = fit.upr, color = NULL), alpha = .2) +
  labs(x = "Bistandsnivå", y = "Forventet GDP vekst", color = "Policy", fill = "Policy")
```


Forventet GDP vekst
4
2
0
−2
−4
0
2
4
6
8
Bistandsnivå
Policy
−1
0
1