Module 4 - Instructions

Oliver Engist

04/04/2021

In the next assignment we want to replicate some plots from the paper "Female Socialization: How Daughters Affect Their Legislator Fathers' Voting on Women's Issues" (Washington, 2008). The paper explores whether having a daughter makes politicians more sensitive to women's rights issues and how this is reflected in their voting behavior. The main identifying assumption is that after controlling for the number of children, the gender composition is random. This might be violated if families that have a preference for girls keep having children until they have a girl. In this assignment we will prepare a dataset that allows us to test whether families engage in such a "female child stopping rule".

Setup

- Load the libraries "Rio" and "tidyverse"
- Change the path of the working directory to your working directory.

```
library(rio)
```

```
## Warning: package 'rio' was built under R version 4.0.5
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.5
```

```
## Warning: package 'stringr' was built under R version 4.0.4
```

```
setwd("C:/Users/Adam Json Lindh/Desktop/Module_4")
```

- import the data sets *basic.dta* and *genold108.dta*

```
basic <- import("basic.dta")
genold108 <- import("genold108.dta")
```

- create a subset of the 108th congress from the *basic* dataset

```
basic <- basic %>% filter(congress == "108")
```

- join this subset with the *genold* dataset

```
combined_dataset <- left_join(genold108,basic, by = c("district", "statenam", "name"))
```

Data preparation

- check table 1 in the appendix of the paper and decide which variables are necessary for the analysis (check the footnote for control variables)

To recreate appendix Table 1, the following variables are included: total children (totchi), number of daughters (ngirls), Political party (party), First child gander (genold). For the set of control variables, we need: race (white), gender (female), age(age), experience proxied by age squarred (age^2), service length (srvlng), service length squarred (srvlng^2), religion(rgroup) and region (region).

- drop all other variables.

```
combined_dataset <- select(combined_dataset,totchi,ngirls,party,genold,
        female,age,srvlng,rgroup,region, white)
```

- Recode *genold* such that gender is a factor variable and missing values are coded as NAs.

```
combined_dataset$genold <- combined_dataset$genold  %>% na_if("") %>% as.factor()
combined_dataset$genold <- combined_dataset$genold
```

```
#control
combined_dataset$genold %>% glimpse() %>% is.na()
```

- Recode *party* as a factor with 3 levels (D, R, I)

```
combined_dataset$party <- combined_dataset$party %>% as.factor() %>% recode("1"="D",
                                                                             "2"="R","3"="I")
```

- Recode *rgroup* and *region* as factors.

```
combined_dataset$rgroup <- combined_dataset$rgroup %>% as.factor()
combined_dataset$region <- combined_dataset$region %>% as.factor()
```

- generate variables for age squared and service length squared

```
combined_dataset <- combined_dataset %>% mutate(age_sq = age^2 , srvlng_sq = srvlng^2)
```

- create an additional variable of the number of children as factor variable

```
combined_dataset <- combined_dataset %>% mutate(n_children = as.factor(totchi))
```

Replicationg Table 1 from the Appendix

We haven't covered regressions in R yet. Use the function *lm()*. The function takes the regression model (formula) and the data as an input. The model is written as $y \sim x$, where x stands for any linear combination of regressors (e.g. $y \sim x_1 + x_2 + female$). Use the help file to understand the function.

- Run the regression $total.children = \beta_0 + \beta_1 gender.oldest + \gamma' X$ where γ stands for a vector of coefficients and X is a matrix that contains all columns that are control variables.[1]

```
reg_totchi_congress <- lm(totchi~ genold + age + age_sq + party + srvlng + srvlng_sq +
                            rgroup + region + white, combined_dataset)
```

- Save the main coefficient of interest (β_1)

```
beta1_totchi_congress <- reg_totchi_congress[["coefficients"]][["genoldG"]]
```

- Run the same regression separately for Democrats and Republicans (assign the independent to one of the parties). Save the coefficient and standard error of *genold*

```
#I perform the regression on total child for two subsets of data: if democrat,
# if republican.

reg_totchi_D <- lm(totchi~ genold + age + age_sq +
        srvlng + srvlng_sq + rgroup + region + white, combined_dataset, party == "D")
reg_totchi_R <- lm(totchi~ genold + age + age_sq +
                     srvlng + srvlng_sq + rgroup + region + white,
                   combined_dataset, party == "R")

# I extract the beta coefficients for gender of
# oldest children from the two regressions.

beta1_totchi_D <- reg_totchi_D[["coefficients"]][["genoldG"]]
beta1_totchi_R <- reg_totchi_R[["coefficients"]][["genoldG"]]

# Analysis of Variance, I create a summary table where I can view the
# standard errors of the beta1 coefficients.

anova_totchi_congress <- summary(reg_totchi_congress)
anova_totchi_D <- summary(reg_totchi_D)
anova_totchi_R <- summary(reg_totchi_R)

# I collect the standard errors for the estimate parameters.

beta1_SE_totchi_congress <- anova_totchi_congress[["coefficients"]][["genoldG",2]]

beta1_SE_totchi_congress <- anova_totchi_congress[["coefficients"]][["genoldG",2]]

beta1_SE_totchi_congress <- anova_totchi_congress[["coefficients"]][["genoldG",2]]
```

[1] This is just a short notation instead of writing the full model with all control variables $totchi = \beta_0 + \beta_1 genold + \gamma_1 age + \gamma_2 age^2 + \gamma_3 Democrat + ... + \epsilon$ which quickly gets out of hand for large models.

- Collect all the *genold* coefficients from the six regressions, including their standard errors and arrange them in a table as in the paper.

```
# I redo all of the regressions above, but with number of daughters as dependent
# variable instead, and extract all of the variables we are interested in.

reg_ngirls_congress <- lm(ngirls~ genold + age + age_sq + party + srvlng +
                            srvlng_sq + rgroup + region + white + n_children,
                          combined_dataset)
reg_ngirls_D <- lm(ngirls~ genold + age + age_sq + srvlng + srvlng_sq + rgroup +
                     region + white + n_children, combined_dataset, party == "D")
reg_ngirls_R <- lm(ngirls~ genold + age + age_sq + srvlng + srvlng_sq + rgroup +
                     region + white + n_children, combined_dataset, party == "R")

beta1_ngirls_congress <- reg_ngirls_congress[["coefficients"]][["genoldG"]]
beta1_ngirls_D <- reg_ngirls_D[["coefficients"]][["genoldG"]]
beta1_ngirls_R <- reg_ngirls_R[["coefficients"]][["genoldG"]]

anova_ngirls_congress <- summary(reg_ngirls_congress)
anova_ngirls_D <- summary(reg_ngirls_D)
anova_ngirls_R <- summary(reg_ngirls_R)

beta1_SE_ngirls_congress <- anova_ngirls_congress[["coefficients"]][["genoldG",2]]

beta1_SE_ngirls_congress <- anova_ngirls_congress[["coefficients"]][["genoldG",2]]

beta1_SE_ngirls_congress <- anova_ngirls_congress[["coefficients"]][["genoldG",2]]
```

- print the table

```
# I use the stargazer package to print the necessary code for a LaTeX table.
# install.packages("knitr")
library(knitr)
# install.packages("stargazer")
library(stargazer)
```

```
##
## Please cite as:

##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.

##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```
stargazer(reg_totchi_congress,reg_ngirls_congress,reg_totchi_D,reg_ngirls_D,
          reg_totchi_R,reg_ngirls_R,type='text', title=
"APPENDIX TABLE 1: Evidence on Legislator Child Gender Mix Selection, 108th Congress",
          align=TRUE,column.labels = c("Congress","Democrats", "Republicans"),
          column.separate = c(2,2,2), digits=2,keep.stat="n", keep = "genoldG",
notes = c(
"The sample includes the 227 of the 381 parent members of the 108th Congress,",
"for whom gender of the first born could be established Number of children",
"regressions include controls for legislator race, gender, party, age, ",
```

```
"age squared, service length and its square, religion, and region Number of",
"daughters regressions include the preceding covariates, as well as fixed",
"effects for total number of children Standard errors in parentheses"),
notes.align = "l")
```

```
##
## APPENDIX TABLE 1: Evidence on Legislator Child Gender Mix Selection, 108th Congress
## ================================================================================================
##                                              Dependent variable:
##                -------------------------------------------------------------------------------
##                   totchi        ngirls        totchi        ngirls        totchi       ngirls
##                         Congress                    Democrats                  Republicans
##                     (1)           (2)           (3)           (4)           (5)           (6)
## ------------------------------------------------------------------------------------------------
## genoldG           -0.08        1.36***        0.12         1.41***       -0.26        1.24***
##                  (0.15)        (0.08)        (0.18)        (0.11)        (0.23)        (0.11)
##
## ------------------------------------------------------------------------------------------------
## Observations       227           227           104           104           122           122
## ================================================================================================
## Note:          *p<0.1; **p<0.05; ***p<0.01
##                The sample includes the 227 of the 381 parent members of the 108th Congress,
##                for whom gender of the first born could be established Number of children
##                regressions include controls for legislator race, gender, party, age,
##                age squared, service length and its square, religion, and region Number of
##                daughters regressions include the preceding covariates, as well as fixed
##                effects for total number of children Standard errors in parentheses
```