Module 4 - Solutions

Jacob Stevens (41652)

21/04/2020

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".

I encourage you to take a look at the paper, as we will come back to it later in the course.

Setup

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

```
library(rio)
library(tidyverse)
library(kableExtra)
library(knitr)
```

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

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

# covert to tibble
basic <- as_tibble(basic)
genold <- as_tibble(genold)
```

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

```
# take subset, order alphabetcally by congressman
I08 <- basic[basic$congress == 108,]
genold <- genold[order(genold$name),]
```

- join this subset with the *genold* dataset

```
# drop identical columns and merge the two
genold <- subset(genold, select = -c(name, statenam, district))
comb <- bind_cols(genold, I08)
```

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)
- drop all other variables.

```
# keep only relevant variables
comb <- comb[, c("name", "genold", "party", "ngirls", "nboys", "female",
                 "age", "srvlng", "white", "region", "totchi", "rgroup")]
```

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

```
# gender -> factor, missing -> NA
comb$genold <- as.factor(comb$genold)
comb$genold[comb$genold == ''] <- NA
```

- Recode *party* as a factor with 3 levels (D, R, I)
- Recode *rgroup* and *region* as factors.

```
# party, rgroup, region -> factor
comb$party <- factor(comb$party, levels = c(1,2,3), labels=c("D", "R", "I"))
comb$rgroup <- factor(comb$rgroup)
comb$region <- factor(comb$region)
```

- generate variables for age squared and service length squared

```
# square variables
comb$sqage = comb$age ^ 2
comb$sqsrv <- comb$srvlng ^ 2
```

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

```
# drop if children unknown
comb <- subset(comb, !is.na(comb$totchi))
comb <- subset(comb, !is.na(comb$genold))

# total children
comb$totchi <- as.numeric(comb$totchi)
```

Replicating 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]
- Save the main coefficient of interest (β_1)

```
# regression
lfit <- lm(formula = totchi ~ 1 + genold + party + female + age + sqage + srvlng
          + sqsrv + white + region + rgroup, data = comb)

beta1all <- summary(lfit)[["coefficients"]][2,1]
sbeta1all <- summary(lfit)[["coefficients"]][2,2]
```

[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.

- 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*

```
# by party
dcomb <- subset(comb, comb$party == "D")
rcomb <- subset(comb, comb$party == "R" | comb$party == "I")

dfit <- lm(formula = totchi ~ 1 + genold + female + age + sqage + srvlng
           + sqsrv + white + region + rgroup, data = dcomb)

rfit <- lm(formula = totchi ~ 1 + genold + female + age + sqage + srvlng
           + sqsrv + white + region + rgroup, data = rcomb)

beta1d <- summary(dfit)[["coefficients"]][2,1]
sbeta1d <- summary(dfit)[["coefficients"]][2,2]

beta1r <- summary(rfit)[["coefficients"]][2,1]
sbeta1r <- summary(rfit)[["coefficients"]][2,2]
```

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

```
# form coefficients into a table
nm <- c("All", "Dems", "Reps + Ind")
b <- c(beta1all, beta1d, beta1r)
names(b) <- nm
s <- c(sbeta1all, sbeta1d, sbeta1r)
names(s) <- nm

tab <- bind_rows(b,s)
names(tab) <- nm
```

- print the table

```
tab <- as.data.frame(tab)
rownames(tab) <- c("Estimate", "SE")

# nice formatting for the table
kable(tab) %>%
  kable_styling() %>%
  add_header_above(c(" " = 1, "Party Affiliation" = 3), bold = TRUE)
```

	Party Affiliation		
	All	Dems	Reps + Ind
Estimate	-0.0729718	0.0032585	-0.2009949
SE	0.1558773	0.1875687	0.2379813