Module 4 - Instructions

Oliver Engist

3/25/2020

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)
basic <- import("basic.dta")
genold108 <- import("genold108.dta")
con108 <- subset(basic, congress == 108)
joined <- left_join(con108,genold108)
```

```
## Warning: Column `district` has different attributes on LHS and RHS of join
```

```
## Warning: Column `statenam` has different attributes on LHS and RHS of join
```

```
## Warning: Column `name` has different attributes on LHS and RHS of join
```

- import the data sets *basic.dta* and *genold108.dta*
- create a subset of the 108th congress from the *basic* dataset
- join this subset with the *genold* dataset

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)
- variables needed: white, female, party, age, age squared, srvlng, srvlng squared, religion/rgroup, region, ngirls
- drop all other variables.
- Recode *genold* such that gender is a factor variable and missing values are coded as NAs.
- Recode *party* as a factor with 3 levels (D, R, I).
- Recode *rgroup* and *region* as factors.
- generate variables for age squared and service length squared
- create an additional variable of the number of children as factor variable

```
tbl1_data <-select(joined, white, female, party, age, srvlng, rgroup, region, genold, ngirls, totchi)
tbl1_data$genold <- factor(tbl1_data$genold)
tbl1_data$genold <- tbl1_data$genold %>% na_if("")
tbl1_data$party <- factor(tbl1_data$party)
tbl1_data <- tbl1_data %>%
   mutate(party = fct_recode(party, "D" = "1", "R" = "2", "I" = "3"))
tbl1_data$rgroup <- factor(tbl1_data$rgroup)
tbl1_data$region<- factor(tbl1_data$region)
tbl1_data$agesq <- tbl1_data$age^2
tbl1_data$srvlngsq <- tbl1_data$srvlng^2
tbl1_data$totchi_factor<- factor(tbl1_data$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. 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] Regression with total children as depedent variable across all parties

```
reg1 <- lm(totchi ~ genold + white + female + party + age + srvlng + rgroup + region + srvlngsq + agesq
totchi_congress_beta <- summary(reg1)$coefficients["genoldG","Estimate"]
totchi_congress_se <- summary(reg1)$coefficients["genoldG","Std. Error"]
con_obs <- nobs(reg1)
```

Regression with number of daughters as dependent variable across all parties

```
ngreg <- lm(ngirls ~ genold + totchi + white + female + party + age + srvlng + rgroup + region + srvlngs
ngirls_congress_beta <- summary(ngreg)$coefficients["genoldG","Estimate"]
ngirls_congress_se <- summary(ngreg)$coefficients["genoldG","Std. Error"]
```

Regression of total children as dependent variable across democrats

```
ddata <- filter(tbl1_data, party == "D")
dreg <- lm(totchi ~ genold + white + female + age + srvlng + rgroup + region + srvlngsq + agesq, data =
summary(dreg)
```

```
##
## Call:
## lm(formula = totchi ~ genold + white + female + age + srvlng +
##     rgroup + region + srvlngsq + agesq, data = ddata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4310 -0.5141 -0.1786  0.5735  2.8399
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.8435835  2.7338828   0.309   0.7584
## genoldG      0.0921043  0.1797167   0.512   0.6096
## white        0.2847018  0.2127998   1.338   0.1845
```

[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.

```
## female      -0.3083708  0.2692256  -1.145   0.2553
## age          0.0008919  0.0982974   0.009   0.9928
## srvlng      -0.0369403  0.0352454  -1.048   0.2976
## rgroup1      0.3666101  0.6620016   0.554   0.5812
## rgroup2      0.0940324  0.6621437   0.142   0.8874
## rgroup3     -0.9358030  0.9178807  -1.020   0.3109
## rgroup4      0.5046389  0.6843133   0.737   0.4629
## region2      0.7101795  0.4227530   1.680   0.0967 .
## region3      0.3352997  0.4326423   0.775   0.4405
## region4      0.8315216  0.5251712   1.583   0.1171
## region5      0.6535005  0.4230134   1.545   0.1261
## region6      0.2005181  0.5463099   0.367   0.7145
## region7      0.8237174  0.4199574   1.961   0.0531 .
## region8      0.9622232  0.5219550   1.843   0.0688 .
## region9      0.5950433  0.4040783   1.473   0.1446
## srvlngsq     0.0009718  0.0010699   0.908   0.3663
## agesq        0.0000995  0.0008942   0.111   0.9117
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8392 on 84 degrees of freedom
##   (101 observations deleted due to missingness)
## Multiple R-squared:  0.187,  Adjusted R-squared:  0.003133
## F-statistic: 1.017 on 19 and 84 DF,  p-value: 0.4512
```

```
totchi_d_beta <- summary(dreg)$coefficients["genoldG","Estimate"]
totchi_d_se <- summary(dreg)$coefficients["genoldG","Std. Error"]
d_obs <- nobs(dreg)
```

Regression with number of daughters as dependent variable across democrats

```
ngdreg <- lm(ngirls ~ genold + totchi + white + female + age + srvlng + rgroup + region + srvlngsq + ag
ngirls_d_beta <- summary(ngdreg)$coefficients["genoldG","Estimate"]
ngirls_d_se <- summary(ngdreg)$coefficients["genoldG","Std. Error"]
```

Regression of total children as dependent variable across republicans

```
rdata <- filter(tbl1_data, party == "R")
rreg <- lm(totchi~ genold + white + female + age + srvlng + rgroup + region + srvlngsq + agesq, data = r
totchi_r_beta <- summary(rreg)$coefficients["genoldG","Estimate"]
totchi_r_se <- summary(rreg)$coefficients["genoldG","Std. Error"]
r_obs <- nobs(rreg)
```

Regression with number of daughters as dependent variable across rebublicans

```
ngrreg <- lm(ngirls ~ genold + totchi + white + female + age + srvlng + rgroup + region + srvlngsq + ag
ngirls_r_beta <- summary(ngrreg)$coefficients["genoldG","Estimate"]
ngirls_r_se <- summary(ngrreg)$coefficients["genoldG","Std. Error"]
```

#Create table with coefficients and std. errors

```
final_data <- matrix(c(ngirls_congress_beta, totchi_congress_beta, ngirls_d_beta, totchi_d_beta, ngirls
colnames(final_data) <- c("C Number of daugthers", "C Number of children", "D Number of daugthers", "D
rownames(final_data) <- c("First Child Female", "Std Error", "Observations")
header <- data_frame(c("Congress", "Democrates", "Republicans"))
```

```
## Warning: `data_frame()` is deprecated as of tibble 1.1.0.
## Please use `tibble()` instead.
```

```
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
```

```
final_data <- round(final_data, digits = 2)
print(final_data)
```

```
##                     C Number of daugthers C Number of children
## First Child Female                   1.41                -0.08
## Std Error                            0.08                 0.15
## Observations                       227.00               227.00
##                     D Number of daugthers D Number of children
## First Child Female                   1.40                 0.09
## Std Error                            0.11                 0.18
## Observations                       104.00               104.00
##                     R Number of daugthers R Number of children
## First Child Female                   1.36                -0.28
## Std Error                            0.12                 0.23
## Observations                       122.00               122.00
```

```
library(knitr)
```

- Save the main coefficient of interest (β_1)
- Run the same regression separately for Democrats (including Bernie) and Republicans. Save the coefficient and standard error of *genold*
- Collect all the *genold* coefficients from the six regressions, including their standard errors and arrange them in a table as in the paper.
- print the table.