Module 4

Filippo Allegri - 41872

23/04/2021

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"

```
library("tidyverse")
library("rio")
```

- Change the path of the working directory to your working directory.

```
setwd("~/Documents/SSE/A4/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

```
Basic108 <- filter(Basic, Basic$congress=="108")
```

- join this subset with the *genold* dataset

```
Data108 <- merge(x = Genold108, y = Basic108, by = c("name","district","statenam"), all=FALSE)
```

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.

The following variables are necessary for the analysis. All the others are automatically dropped.

```
variables <- c("white", "female", "party", "age", "srvlng", "rgroup",
               "region", "totchi", "genold", "ngirls")
Data108 <- Data108[variables]
```

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

```
Data108$genold <- as.factor(Data108$genold)
Data108$genold[Data108$genold == ""] <- NA
```

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

```
Data108$party <- factor(Data108$party, levels=c(1,2,3), labels=c("D", "R", "I"))
```

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

```
Data108$rgroup <- as.factor(Data108$rgroup)
Data108$region <- as.factor(Data108$region)
```

- generate variables for age squared and service length squared

```
Data108$agesq <- Data108$age*Data108$age
Data108$srvlngsq <- Data108$srvlng*Data108$srvlng
```

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

```
Data108$totchiFactor <- as.factor(Data108$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<- lm(Data108$totchi ~ genold + white + female + party + age + agesq
         + srvlng + srvlngsq + rgroup + region, data=Data108)
summary(reg)
```

```
##
## Call:
## lm(formula = Data108$totchi ~ genold + white + female + party +
##     age + agesq + srvlng + srvlngsq + rgroup + region, data = Data108)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0485 -0.6572 -0.1433  0.5604  5.7808
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.9274759  2.2432982   1.305  0.19336
## genoldG     -0.0838833  0.1491133  -0.563  0.57436
## white        0.2470309  0.2360813   1.046  0.29662
## female       0.0769063  0.2296767   0.335  0.73808
## partyR       0.2841199  0.1672608   1.699  0.09090 .
## partyI      -1.0256429  1.0981026  -0.934  0.35140
## age         -0.0770507  0.0791304  -0.974  0.33134
## agesq        0.0007331  0.0007433   0.986  0.32516
## srvlng       0.0147286  0.0292026   0.504  0.61455
## srvlngsq    -0.0005872  0.0009700  -0.605  0.54562
## rgroup1      0.2886397  0.8016202   0.360  0.71917
## rgroup2      0.2989158  0.8070999   0.370  0.71150
## rgroup3      0.8818639  0.8906008   0.990  0.32325
## rgroup4      0.4976885  0.8363469   0.595  0.55245
## region2      0.6056196  0.3928366   1.542  0.12470
## region3      0.3083420  0.3957894   0.779  0.43685
## region4      1.4289800  0.4334300   3.297  0.00115 **
## region5      0.5879700  0.3841897   1.530  0.12746
## region6      0.4984320  0.4545767   1.096  0.27416
## region7      0.6306864  0.4020012   1.569  0.11822
## region8      0.8961855  0.4250930   2.108  0.03623 *
## region9      0.4769235  0.3969444   1.201  0.23095
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.072 on 205 degrees of freedom
##   (208 observations deleted due to missingness)
## Multiple R-squared:  0.1451, Adjusted R-squared:  0.05757
## F-statistic: 1.657 on 21 and 205 DF,  p-value: 0.03992
```

[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.

- Save the main coefficient of interest (β_1)

```
Beta1 <- summary(reg)$coefficients[2,1]
Beta1
```

```
## [1] -0.08388331
```

- 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 assign the independent to the Republicans

```
regDEM <- lm(Data108$totchi ~ genold + white + female + age + agesq
         + srvlng + srvlngsq + rgroup + region, data=Data108,
         Data108$party=="D")
summary(regDEM)
```

```
##
## Call:
## lm(formula = Data108$totchi ~ genold + white + female + age +
##     agesq + srvlng + srvlngsq + rgroup + region, data = Data108,
##     subset = Data108$party == "D")
##
## 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
## female      -0.3083708  0.2692256  -1.145   0.2553
## age          0.0008919  0.0982974   0.009   0.9928
## agesq        0.0000995  0.0008942   0.111   0.9117
## srvlng      -0.0369403  0.0352454  -1.048   0.2976
## srvlngsq     0.0009718  0.0010699   0.908   0.3663
## 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
## ---
## 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
```

```
regREP<- lm(Data108$totchi ~ genold + white + female + age + agesq
         + srvlng + srvlngsq + rgroup + region, data=Data108,
         Data108$party=="R"| Data108$party=="I")

summary(regREP)
```

```
##
## Call:
## lm(formula = Data108$totchi ~ genold + white + female + age +
##     agesq + srvlng + srvlngsq + rgroup + region, data = Data108,
##     subset = Data108$party == "R" | Data108$party == "I")
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9573 -0.6811 -0.1033  0.5509  5.3673
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.036019   3.462793   1.166   0.2465
## genoldG     -0.306415   0.231073  -1.326   0.1877
## white        0.542704   0.665789   0.815   0.4169
## female       0.387909   0.373832   1.038   0.3018
## age         -0.115353   0.126987  -0.908   0.3658
## agesq        0.001116   0.001227   0.910   0.3651
## srvlng       0.044823   0.048548   0.923   0.3580
## srvlngsq    -0.001994   0.001788  -1.116   0.2671
## rgroup2      0.098713   0.286941   0.344   0.7315
## rgroup3      1.447772   0.559834   2.586   0.0111 *
## rgroup4     -0.120283   1.265021  -0.095   0.9244
## region2      0.645931   0.734018   0.880   0.3809
## region3      0.352638   0.709324   0.497   0.6201
## region4      1.736424   0.744059   2.334   0.0215 *
## region5      0.396216   0.683978   0.579   0.5637
## region6      0.648018   0.768887   0.843   0.4013
## region7      0.600272   0.750450   0.800   0.4256
## region8      1.114642   0.722929   1.542   0.1261
## region9      0.417692   0.747539   0.559   0.5775
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.209 on 104 degrees of freedom
##   (107 observations deleted due to missingness)
## Multiple R-squared:  0.217,  Adjusted R-squared:  0.08151
## F-statistic: 1.601 on 18 and 104 DF,  p-value: 0.07298
```

- 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