Problem Set 2: Omitted Variable Bias and Fixed Effects

Claire Duquennois

6/5/2020

Group Member 1:

Group Member 2:

Group Member 3:

Empirical Analysis using Data from Washington (2008, AER)

This exercise uses data from Ebonya Washington's paper, "Female Socialization: How Daughters Affect their Legislator Father's voting on Women's Issues," published in the *American Economic Review* in 2008. This paper studies whether having a daughter affects legislator's voting on women's issues.

Finding the data

The data can be found by following the link on the AER's website which will take you to the ICPSR's data repository. You will need to sign in to get access to the data files. Once logged in, you will find the set of files that are typically included in a replication file. These include several datasets, several .do files (which is a STATA command file), and text files with the data descriptions which tell you about the different variables included in the dataset. For this assignment we will be using the `basic.dta` file.

Download it and save it on the project repository. The next time you `$git pull` you will download a copy to your local file.

Set up and opening the data

Because this is a `.dta` file, you will need to open it with the `read.dta` function that is included in the `haven` packages.

Additionally, this .Rmd file will be opened on different computers. But you don't want to have to change the filepaths each time you pull a new version off of GitHub. The `here` package will help us with that. Some information on the `here` package: https://github.com/jennybc/here_here.

Other packages you will need: `dplyr`.

Remember, if you have not used a package before you will need to install the package as follows

```
#install.packages('haven',repos = "http://cran.us.r-project.org")
#install.packages("here",repos = "http://cran.us.r-project.org")
#install.packages("dplyr",repos = "http://cran.us.r-project.org")
```

Hint: Once you have run these once, on your machine, you may want to comment them out with a # so that your code runs faster.

Question: Now that the packages are installed, call all your packages and load your data. How many observations are in the original dataset?

Code and Answer:

Cleaning the data

Question: The original dataset contains data from the 105th to 108th U.S. Congress. We only use the observations from the 105th congress. Refer to the data documentation to find the relevant variable and then use the `filter` function in the `dplyr` package to extract observations from the 105th congress.

Code:

Question:The dataset contains many variables, some of which are not used in this exercise. Keep the following variables in the final dataset (Hint: use the `select` function in `dplyr`).

Name	Description
aauw	AAUW score
totchi	Total number of children
ngirls	Number of daughters
party	Political party. Democrats if 1, Republicans if 2, and Independent if 3.
famale	Female dummy variable
white	White dummy variable
srvlng	Years of service
age	Age
demvote	State democratic vote share in most recent presidential election
medinc	District median income
perf	Female proportion of district voting age population
perw	White proportion of total district population
perhs	High school graduate proportion of district population age 25
percol	College graduate proportion of district population age 25
perur	Urban proportion of total district population
moredef	State proportion who favor more defense spending
statabb	State abbreviation
district	id for electoral district
rgroup	religious group
region	region

You can find the detailed description of each variable in the original paper. The main variable in this analysis is `AAUW`, a score created by the American Association of University Women (AAUW). For each congress, AAUW selects pieces of legislation in the areas of education, equality, and reproductive rights. The AAUW keeps track of how each legislator voted on these pieces of legislation and whether their vote aligned with the AAUW's position. The legislator's score is equal to the proportion of these votes made in agreement with the AAUW.

Code:

Question: Make sure your final dataset is a data frame. You can check your data's format with the command `is`. If the first element of the returned vector is not "data.frame", convert your dataset with the function `as.data.frame`.

Code:

Summary Statistics

Question: Report summary statistics of the following variables in the dataset: political party, age, race, gender, AAUW score, the number of children, and the number of daughters. Present these summary statistics in a formatted table, you can use `stargazer` or other packages. Make this table as communicative as possible.

Hints: If you want RMarkdown to display your outputted table, include the code `results = "asis"` in the chunk header. This is true for all chunks that output a formatted table. In the stargazer command, you will want to specify the format of the table by including the code `results="html"` for html output or `results="latex"` for a pdf output.

Code:

Generate Variables

Question:Construct a variable called $repub_i$, a binary set to 1 if the observation is for a republican.

Code:

Run Estimations

Question: Estimate the following linear regression models using the `felm` command (part of the lfe package). Report your regression results in a formatted table using a package such as `stargazer`. Report robust standard errors in your table (Hint: in stargazer specify `se = list(model1$rse, model2$rse, model3$rse)`). Make this table as informative as possible by adding needed information and removing superfluous information.

$$aauw_i = \beta_0 + \beta_1 ngirls_i + \epsilon_i$$
$$aauw_i = \beta_0 + \beta_1 ngirls_i + \beta_2 totchi + \epsilon_i$$
$$aauw_i = \beta_0 + \beta_1 ngirls_i + \beta_2 totchi + \beta_3 female_i + \beta_4 repub_i + \epsilon_i$$

Code:

Question: Compare the OLS estimates of β_1 across the above three specifications. Discuss what explains the difference (if any) of the estimate across three specifications? Which control variable is particularly important and why?

Answer and Code:

Question: Consider the third specification (with 3 controls in addition to $ngirls_i$. Conditional on the number of children and other variables, do you think $ngirls_i$ is plausibly exogenous? What is the identifying assumption necessary for β_1 to be interpreted as a causal estimate? What evidence does Washington give to support this assumption?

Answer:

Question: It is possible that the effects of having daughters might be different for female and male legislators. Estimate four different models to think about this question: the equivalent of model 3 separately on men and women, model 3 with a single interaction term added, and model 3 with three interaction terms added. Present your results in a table. Is there evidence that the effect of a daughter differs for male and female legislators? Of the four models you estimated, which are equivalent, which are different, and why?

Code and Answer:

Fixed Effects:

Question: Equation 1 from Washington's paper is a little bit different from the equations you have estimated so far. Estimate the three models specified below (where γ_i is a fixed effect for the number of children). Present your results in a table and explain the difference between the three models.

$$aauw_i = \beta_0 + \beta_1 ngirls_i + \beta_2 totchi + \epsilon_i$$
$$aauw_i = \beta_0 + \beta_1 ngirls_i + \beta_2 chi1 + ... + \beta_{10} chi10 + \epsilon_i$$
$$aauw_i = \beta_0 + \beta_1 ngirls_i + \gamma_i + \epsilon_i$$

Hint: you will need to generate the dummy variables for the second equation or code it as `factor()`. For the third equation, the `felm` function allows you to specify fixed effects.

Code and Answer:

Question: Reproduce the results in column 2 of table 2 from Washington's paper.

Code:

Question: Explain what the region fixed effects are controlling for?

Answer:

Question: Reload the data and this time we will keep observations from all of the congresses. Generate a variable that creates a unique identifier for region by year. Estimate the following models and present your results in a table.

$$aauw_i = \beta_0 + \beta_1 ngirls_i + \gamma_i + \phi_i + \epsilon_i$$
$$aauw_i = \beta_0 + \beta_1 ngirls_i + \gamma_i + \phi_i + \eta_i + \epsilon_i$$
$$aauw_i = \beta_0 + \beta_1 ngirls_i + \gamma_i + \theta_i + \epsilon_i$$
$$aauw_i = \beta_0 + \beta_1 ngirls_i + \rho_i + \epsilon_i$$

γ_i is a fixed effect for the total number of children, ϕ_i is a region fixed effect, η_i is a year (congress session) fixed effect and θ_i is a region by year fixed effect and ρ_i is a total children by region by year fixed effect. Explain what the differences between these four different estimation. Is there a downside to a specification like the fourth specification?

Code and Answer:

Question: In her paper, Washington chooses not to pool the data for all four congresses and instead estimates her main specification on each year separately. Why do you think she makes this choice?

Answer:

Question: Check to see that names uniquely identify each congress person. If you are not sure if they do, make a unique identifier for each congress person.

Answer and Code:

Question:Because we have data for four congress sessions, we may be able to see how an individual congress person's voting patterns change as the number of daughters they have changes. Propose an estimating equation that would allow you to estimate this, run your estimation and present your results. Be sure to define all new variables. What do your results tell you? Why?

Answer and Code:

Question: Can you think of any identification concerns with this approach?

Answer:

Question: Using data from all four congresses, estimate the same specification as that used in column 2 of table 2 with the addition of year and individual fixed effects and report your results. Why aren't you able to estimate a coefficient for certain covariates?

Code:

Answer:

Question: Which fixed effects from the original specification are now redundant?

Answer:

Question: Can you estimate a coefficient for *Repub*? What does this imply?

Answer:

Submission instructions:

1) Make sure the final version of your assignment is uploaded on GitHub in both html and Rmarkdown format.
2) Knit your final version as a Word or Pdf document and submit this to Gradescope by the due date.