Problem Set 2: Omitted Variable Bias and Fixed Effects

Rohan Krishnan

2024-02-08

I am completing this homework assignment using my local RStudio application. Apologies for any slight discrepancies in formatting!

Empirical Analysis using Data from Washington (2008, AER)

This exercise, like PS1, also 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.

Set up and opening the data

Question 1.1:

Load the basic.dta file like you did for PS1 and call all the packages you will be using with the library function. The packages you will need are haven, lfe, dplyr, and stargazer.

Code:

```
#Load libraries
library(haven)
library(lfe)
library(dplyr)
library(stargazer)
#Load data
basic <- read_dta("~/Downloads/Basic.dta")
```

Cleaning the data

Question 2.1:

Like in PS1, restrict your data to observations from the 105th congress and keep only the variables listed in the table below. Here too, make sure your final dataset is a data frame.

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.
female	Female dummy variable
white	White dummy variable
srvlng	Years of service
age	Age
demvote	State democratic vote share in most recent presidential election
rgroup	religious group
region	region
name	representative's name

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:

```
#Filter for 105th congress and select variables
basic1 <- basic %>%
  filter(congress == 105) %>%
  select(aauw, totchi, ngirls, party, female,
         white, srvlng, age, demvote, rgroup, region, name)

#Check if basic1 is a data frame
is(basic1)
```

```
## [1] "tbl_df"     "tbl"        "data.frame" "list"       "oldClass"
## [6] "vector"
```

```
#Convert basic1 to data frame
basic1 <- as.data.frame(basic1)

#Check if basic1 is a data frame
is(basic1)
```

```
## [1] "data.frame" "list"       "oldClass"   "vector"
```

Analysis

Question 3.1:

Estimate the following linear regression models using the felm command (part of the lfe package). Report your regression results in a formatted table using stargazer. Report robust standard errors in your table.

$$\text{Model 1}: aauw_i = \beta_0 + \beta_1 ngirls_i + \epsilon_i$$

$$\text{Model 2}: aauwi = \beta_0 + \beta_1 ngirls_i + \beta_2 totchi_i + \epsilon_i$$

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 type="latex" for a pdf output. If you have trouble knitting to PDF, try installing MikTex (https://miktex.org/download).

Code:

```
#Estimate regressions
model1 <- felm(aauw ~ ngirls, data = basic1)
model2 <- felm(aauw ~ ngirls + totchi, data = basic1)
#Generate table
stargazer(model1, model2,
          type = "latex",
          title = "Question 3.1 Model Comparison",
          se = list(model1$rse, model2$rse),
          header = FALSE,
          no.space = TRUE,
          table.placement = "H"
          )
```

Table 2: Question 3.1 Model Comparison

	Dependent variable:	
	aauw	
	(1)	(2)
ngirls	−2.784	5.776**
	(1.750)	(2.714)
totchi		−7.992***
		(1.784)
Constant	50.964***	59.982***
	(3.036)	(3.520)
Observations	434	434
R^2	0.006	0.051
Adjusted R^2	0.003	0.047
Residual Std. Error	41.939 (df = 432)	41.010 (df = 431)
Note:		*p<0.1; **p<0.05; ***p<0.01

Question 3.2:

Compare the estimates of β_1 across the two specifications. Why does our estimate of β_1 changes so much? Which control variable is particularly important and why?

Answer:

The estimate of β_1 differs greatly between models one and two. Specifically, it goes from not statistically significant and negative (-2.784) to statistically significant and positive (5.776). The control variable *totchi* is a particularly important factor in this change as it allows β_1 to represent the effect of an additional daughter controlling for total children that way senators with many children are not weighing down the estimate.

Question 3.3:

Consider the second specification which controls for $totchi_i$. Conditional on the number of children, 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:

Given the *totchi* control, I believe *ngirls* is plausibly exogenous. The main assumption for $beta_1$ to be interpreted as causal is that, controlling for one's total number of children, the "assignment" of a daughter to a congressperson is effectively random. This assumption seems reasonable as it relies on biology and the odds of conceiving a girl or boy rather than any social or cultural factors that may be harder to argue as random.

Fixed Effects

Question 4.1:

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.

$$\text{Model 1} : aauw_i = \beta_0 + \beta_1 ngirls_i + \beta_2 totchi_i + \epsilon_i$$

$$\text{Model 2} : aauw_i = \beta_0 + \beta_1 ngirls_i + \beta_2 chi1 + ... + \beta_{10} chi10 + \epsilon_i$$

$$\text{Model 3} : aauw_i = \beta_0 + \beta_1 ngirls_i + \gamma_i + \epsilon_i$$

Hints:

- you will need to generate the dummy variables for the second equation or code it as factor(totchi).
- For the third equation, the felm function allows you to specify fixed effects as we saw in class.

Code:

```
#Estimate regressions
model1 <- felm(aauw ~ ngirls + totchi, data = basic1)
model2 <- felm(aauw ~ ngirls + factor(totchi), data = basic1)
model3 <- felm(aauw ~ ngirls |totchi, data = basic1)
#Generate table
stargazer(model1, model2, model3,
          type = "latex",
          title = "Question 4.1 Model Comparison",
          se = list(model1$rse, model2$rse, model3$rse),
          header = FALSE,
          no.space = TRUE,
          omit.stat = "all",
          table.placement = "H"
          )
```

Table 3: Question 4.1 Model Comparison

	Dependent variable:		
	aauw		
	(1)	(2)	(3)
ngirls	5.776**	5.748**	5.748**
	(2.714)	(2.667)	(2.667)
totchi	−7.992***		
	(1.784)		
factor(totchi)1		7.616	
		(8.816)	
factor(totchi)2		−6.182	
		(7.074)	
factor(totchi)3		−17.186**	
		(7.770)	
factor(totchi)4		−25.833***	
		(9.090)	
factor(totchi)5		−28.128**	
		(11.601)	
factor(totchi)6		−34.712	
		(24.334)	
factor(totchi)7		−65.986***	
		(11.828)	
factor(totchi)8		−74.859***	
		(15.283)	
factor(totchi)9		−81.108***	
		(14.386)	
factor(totchi)10		−75.360***	
		(11.957)	
Constant	59.982***	52.367***	
	(3.520)	(5.400)	
Note:			*p<0.1; **p<0.05; ***p<0.01

Question 4.2:

Explain the difference between the three models.

Answer:

The first model controls for the total number of children a congress person has by coding it as an additional variable. The second model creates 10 dummy variables representing the discrete set of values that $totchi_i$ can take on. The third model uses a fixed effect for total children which essentially demeans the data. $beta_1$ is statistically significant for all three models. However, it is the exact same for models two and three. This result indicates that controlling for *totchi* via the addition of dummy variables and by creating a fixed effect for *totchi* have the same effect on the OLS calculation of $beta_1$.

Question 4.3:

Reproduce the EXACT results presented in column 2 of table 2 from Washington's paper. To do this you will need to first build three variables: age^2 and $srvlng^2$ and $repub_i$, an indicator set to 1 if the representative is republican and 0 otherwise. Then estimate the following specification, where γ_i is a fixed effect for total children, ϕ_i is a fixed effect for religious group, and λ_i is a fixed effect for region:

Model A : $aauw_i = \beta_0 + \beta_1 ngirls_i + female_i + white_i + repub_i + age_i + age2_i + srvlng_i + srvlng2_i + demvote_i + \gamma_i + \phi_i + \lambda_i + \epsilon_i$

Code:

```
#Create age2 and srvlng2 column in new df
basic2 <- basic1
basic2$age2 <- (basic2$age)^2
basic2$srvlng2 <- (basic2$srvlng)^2
#Create repub variable
basic2$repub <- NA
basic2$repub[basic2$party == 2] <- 1
basic2$repub[basic2$party != 2] <- 0
#Estimate regression
modelA <- felm(aauw ~ ngirls + female + white +
                 repub + age + age2 +
                 srvlng + srvlng2 + demvote | totchi + rgroup + region, data = basic2
               )
#Generate table -- Can't display religion because it is a fixed effect in the problem
stargazer(modelA,
          type = "latex",
          title = "Question 4.3 Table Replication",
          header = FALSE,
          no.space = TRUE,
          digits = 2,
          omit.stat = "all",
          covariate.labels = c("Number of female children","Female", "White",
                               "Republican", "Age", "Age Squared",
                               "Service length", "Service length squared",
                               "Democratic vote share in district
                               (most recent presidential election)"),
          dep.var.labels = c("AAUW"),
          column.sep.width = "1pt",
          font.size = "small",
          table.placement = "H"
          )
```

Table 4: Question 4.3 Table Replication

	Dependent variable:
	AAUW
Number of female children	2.38^{**}
	(1.12)
Female	9.19^{***}
	(2.91)
White	0.14
	(3.68)
Republican	-60.47^{***}
	(2.28)
Age	0.85
	(0.86)
Age Squared	-0.01
	(0.01)
Service length	-0.21
	(0.32)
Service length squared	0.004
	(0.01)
Democratic vote share in district (most recent presidential election)	62.15^{***}
	(11.57)
Note:	*p<0.1; **p<0.05; ***p<0.01

Question 4.4:

Explain what the region fixed effects are controlling for.

Answer:

The region fixed effects controls for any variation that occurs across observations due to the region they are from. Since there are multiple observations from each region, the time=invariant individual effect of region can be removed when calculating the most unbiased estimate of $beta_1$.

Question 4.5:

Reload the data and this time keep observations from all of the four congresses. Add the the three variables you built for question 4.3 to this data set.

Code:

```
#Reload data
basic <- read_dta("~/Downloads/Basic.dta")
##Create age2 and srvlng2 column in new df
basic4.5 <- basic
basic4.5$age2 <- (basic$age)^2
basic4.5$srvlng2 <- (basic$srvlng)^2
#Create repub variable
basic4.5$repub <- NA
basic4.5$repub[basic$party == 2] <- 1
basic4.5$repub[basic$party != 2] <- 0
```

Question 4.6:

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. Estimate model A with the addition of congress and name fixed effects. Present your results in a table.

Code:

```
#Estimate model A with congress and name fixed effects
modelA <- felm(aauw ~ ngirls + female + white +
                 repub + age + age2 +
                 srvlng + srvlng2 + demvote |
                 totchi + rgroup + region + congress + name,
               data = basic4.5
               )
#Generate table
stargazer(
  modelA,
  type = "latex",
  title = "Question 4.6 Model A with Congress and Name Fixed Effects",
  no.space = TRUE,
  header = FALSE,
  digits = 2,
  omit.stat = "all",
  covariate.labels = c("Number of female children","Female", "White",
                               "Republican", "Age", "Age Squared",
                               "Service length", "Service length squared",
                               "Democratic vote share in district
                               (most recent presidential election)"),
  dep.var.labels = c("AAUW"),
  column.sep.width = "1pt",
  font.size = "small",
  table.placement = "H"
)
```

Table 5: Question 4.6 Model A with Congress and Name Fixed Effects

	Dependent variable:
	AAUW
Number of female children	2.01
	(3.14)
Female	
White	
Republican	−3.03
	(6.06)
Age	10.39
	(7.69)
Age Squared	−0.003
	(0.01)
Service length	−0.99
	(5.38)
Service length squared	0.0004
	(0.01)
Democratic vote share in district (most recent presidential election)	0.45
	(8.04)
Note:	*p<0.1; **p<0.05; ***p<0.01

Question 4.7:

How does this estimate compare to your estimate in question 4.3? Why are the standard errors so much bigger? Why doesn't Washington use this approach in her paper?

Answer:

The $\widehat{beta_1}$ from question 4.6 is slightly smaller (2.01) than the $\widehat{beta_1}$ from question 4.3 (2.38) and, unlike $\widehat{beta_1}$ from question 4.3, is not significant. The standard errors are so much bigger because of the fixed effects for congress and name. By "de-meaning" the regression by name and congress, $\widehat{beta_1}$ represents the effect on AAUW of a specific congress person having an additional daughter in that specific time frame, which is a very small sample size.

Washington does not use this approach in this paper because she is interested in the broader effect on voting behavior caused by having a daughter. It is likely that the name fixed effect is absorbing much of the variation attributed to $ngirls$. Ultimately, this approach highlights the effect on AAUW of having a daughter within a specific time frame as a specific congressperson, which is not what Washington is interested in exploring.

Question 4.8:

Why are you not able to generate a coefficient for $female_i$ or $white_i$?

Answer:

We cannot generate a coefficient for $female_i$ or $white_i$ because the fixed effect for name absorbs all of the variation in female and white attributes. In other words, once you add a fixed effect for the individual congressperson, they will not change race or gender (at least not within this data from 2008 or at a scalable sample size). Thus, there is no variation across an individual to use to explain variation in AAUW score, resulting in no coefficients for either variable.

Question 4.9:

You are able to generate an estimate for $repub_i$. What does this imply?

Answer:

This implies that some congresspeople switched parties over time. As an example, they may have been a republican in one session of congress and an independent in another session. Because there is some variation in party across individual congresspeople, we can calculate a coefficient in the OLS regression.