Problem Set 2: Omitted Variable Bias and Fixed Effects

Claire Duquennois

NAME:**_____**

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:

```
library(haven)
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.1.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(lfe)
```

```
## Loading required package: Matrix
```

```
library(stargazer)
```

```
##
## Please cite as:
```

```
##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
```

```
##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```
mydata<-read_dta("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

```
#selecting only the observations from the 105th congress
mydata<-mydata%>%filter(congress==105)

#selecting variables we will use
mydata<-mydata %>% select(aauw, totchi, ngirls, party,
                          female, white, srvlng, age,
                          demvote, rgroup, region, name)
```

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.

$$\text{Model 1: } aauw_i = \beta_0 + \beta_1 ngirls_i + \epsilon_i$$
$$\text{Model 2: } aauw_i = \beta_0 + \beta_1 ngirls_i + \beta_2 totchi + \epsilon_i$$

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:

```
reg1<-felm(aauw~ngirls,mydata)

reg2<-felm(aauw~ngirls+totchi,mydata)

stargazer( reg1, reg2,  type = "latex", se = list(reg1$rse, reg2$rse))
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sat, Dec 31, 2022 - 12:56:20 PM

Table 2:

	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:

When we compare β_1 across specifications 1 and 2 we see that controlling for the total number of children is particularly important. This makes sense because the number of girls a congress person has will be a function of the number of children they have, the choice of which could be correlated to their political views and how they vote as measured by their aauw score. Using the Omitted Variable Bias formula, we know that the bias is given by $\tilde{\beta}_1 - \beta_1 = \rho\beta_2$. The correlation between the number of girls and the total number of children is positive. The correlation between the total number of children and the aauw score is negative, as seen in specification 2. Thus excluding the control for the total number of children leads our estimate of β_1 to be downward biased.

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:

$ngirls_i$ will be plausibly exogenous if the Conditional Independence Assumption holds. This will be the case if once we control for $totchi_i$ the number of girls is as good as randomly assigned. As discussed in the article, this assumption could be violated if couples follow fertility stopping rules (ie. keep having kids until they get both a girl and boy for example). This assumption could also be violated if voters select their representatives based on the gender composition of their children. In the article, Washington presents evidence that these concerns are not driving her results. She looks at the gender of the first born and finds that it is predictive of the gender mix, but not the total number of children in the sample. She also looks at numerous district characteristics and does not find that there is a concerning relationship between these and the number of daughters the congress person they elect has.

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.

$$\text{Model 1: } aauw_i = \beta_0 + \beta_1 ngirls_i + \beta_2 totchi + \epsilon_i$$
$$\text{Model 2: } aauw_i = \beta_0 + \beta_1 ngirls_i + \beta_2 chi1 + ... + \beta_{10} chi10 + \epsilon_i$$
$$\text{Model 3: } aauw_i = \beta_0 + \beta_1 ngirls_i + \gamma_i + \epsilon_i$$

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:

```
regfe1<-felm(aauw~ngirls+totchi,mydata)
regfe2<-felm(aauw~ngirls+factor(totchi),mydata)
regfe3<-felm(aauw~ngirls|totchi,mydata)

stargazer(regfe1, regfe2, regfe3, type = "latex", se = list(regfe1$rse, regfe2$rse, regfe3$rse))
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sat, Dec 31, 2022 - 12:56:20 PM

Table 3:

	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)	
Observations	434	434	434
R^2	0.051	0.065	0.065
Adjusted R^2	0.047	0.040	0.040
Residual Std. Error	41.010 (df = 431)	41.154 (df = 422)	41.154 (df = 422)
Note:			*p<0.1; **p<0.05; ***p<0.01

Question 4.2:

Explain the difference between the three models.

Answer: Models 2 and 3 are equivalent and generate the exact same estimate for β_1. This is because adding the fixed effects is equivalent to estimating a dummy variable for all of the categories covered by the fixed effect. Model 1 generates an estimate that is similar to the other. However model 1 is a bit different conceptually. Model one imposes a constant linear effect for each additional child, effectively assuming that the first child affects voting patterns in the same way as the 6th child. The other two models are much more flexible. As you can see from the differences in the coefficients on the dummy variables in model 2, this assumption is not entirely valid.

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: $aauw_i = \beta_0 + \beta_1 ngirls_i + female_i + white_i + repub_i + age_i + age_i^2 + srvlng_i + srvlng_i^2 + demvote_i + \gamma_i + \phi_i + \lambda_i + \epsilon_i$

Code:

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

mydata$repub<-0
mydata$repub[mydata$party==2]<-1

regrep<-felm(aauw~ngirls+female+white+repub+age+agesq+srvlng+srvlngsq+demvote
             |totchi+region+rgroup,
             mydata)

stargazer(regrep, type = "latex")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sat, Dec 31, 2022 - 12:56:20 PM

Table 4:

	Dependent variable:
	aauw
ngirls	2.385**
	(1.124)
female	9.194***
	(2.910)
white	0.144
	(3.676)
repub	−60.468***
	(2.280)
age	0.854
	(0.860)
agesq	−0.006
	(0.008)
srvlng	−0.208
	(0.324)
srvlngsq	0.004
	(0.011)
demvote	62.148***
	(11.568)
Observations	434
R^2	0.840
Adjusted R^2	0.828
Residual Std. Error	17.441 (df = 402)
Note:	*p<0.1; **p<0.05; ***p<0.01

Question 4.4:

Explain what the region fixed effects are controlling for.

Answer: he region fixed effects are controlling for the common effect on the aauw score of being from a particular region. Regional patterns could bias our estimates of the effect of girls if being in a particular region correlates with the aauw score and the number of girls, holding the other variables constant.

Question 4.5:

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

Code:

```
mydataall<-read_dta("basic.dta")

mydataall$agesq<-mydataall$age*mydataall$age
mydataall$srvlngsq<-mydataall$srvlng*mydataall$srvlng

mydataall$repub<-0
mydataall$repub[mydataall$party==2]<-1
```

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:

```
regrepall<-felm(aauw~ngirls+female+white+repub+age+agesq+srvlng+srvlngsq+demvote
             |totchi+region+rgroup+congress+name,
             mydataall)
```

```
## Warning in chol.default(mat, pivot = TRUE, tol = tol): the matrix is either
## rank-deficient or indefinite
```

```
stargazer(regrepall, type = "latex")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sat, Dec 31, 2022 - 12:56:20 PM

Table 5:

	Dependent variable:
	aauw
ngirls	2.010
	(3.144)
female	
white	
repub	−3.034
	(6.056)
age	10.393
	(7.686)
agesq	−0.003
	(0.007)
srvlng	−0.990
	(5.376)
srvlngsq	0.0004
	(0.009)
demvote	0.454
	(8.045)
Observations	1,735
R^2	0.973
Adjusted R^2	0.958
Residual Std. Error	8.732 (df = 1117)
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:

Though we do see a positive effect of similar magnitude on the $AAUW$ score, it is not statistically significant at conventional levels. Because we are using individual fixed effects, the coefficient of interest is only identified on individuals who experience a change in the number of daughters within the 4 years of data. Since very few congresspersons experience a change in the number of daughters over this 4 year time period, the standard errors are quite large making any effect using this specification difficult to identify. Furthermore, this approach would be be picking up the effect of an infant/toddler daughter. The effect of an adolescent or adult daughter might be quite different.

Question 4.8:

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

Answer: The individual fixed effect is perfectly colinear with time invariant characteristics about these congresspersons. For these individuals, there is no variation across time in their race or gender. Thus the individual fixed effect will already capture the effect of these characteristics: ie once you control for the individual, there is no variation left with which to estimate the impact of race or gender.

Question 4.9:

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

Answer: We are able to estimate a coefficient for *Repub* (though with large standard errors). This implies that this is not a time invariant variable: some congressperson(s) switch from being republican to not (or vice versa), generating the variation needed to estimate this coefficient. Note though that the large standard errors suggest that there are not many representatives that switch parties.

Submission instructions:

1) Knit your assignment in PDF.
2) Make sure you have ONE question and answer per page, or with tables this might involve two pages, (this allows gradescope to easily find your answers).
3) Upload your assignment PDF to gradescope.