Problem Set 2: Omitted Variable Bias Key

Claire Duquennois

Group Member 1:

Group Member 2:

Group Member 3:

1 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.

2 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 in a 'data' folder located in the same folder as your project repository. Since some datasets in this course will be big, we want to avoid keeping the data on github so I would recommend not placing the data in the project repository itself.

3 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.

Other packages you will need: `dplyr`, `stargazer` and `lfe`.

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("dplyr",repos = "http://cran.us.r-project.org")
#install.packages("stargazer",repos = "http://cran.us.r-project.org")
#install.packages("lfe",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.

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. Because of this, I would recommend you avoid using any computer specific filepaths in your code. Instead, make sure you and your groupmates structure your project folders in the same way and only specify filepaths within your project folder. R uses the folder where you are saving your code as it's default "working directory" (where things will be saved or be searched for unless specified otherwise). You can move up to the parent folder by using `..` in the file path. Thus, if your data is not saved in the forked github repository but is saved in a folder called `data` next to it you can call your data with the following file path: `"../data/basic.dta"`.

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

Code:

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

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

```r
library(stargazer)
library(lfe)

mydata<-read_dta("../../../PSdata/PS2_OVB_FEdata/data/basic.dta")
nrow(mydata)
```

```
## [1] 1740
```

Answer: The data contains 1740 observations.

4 Cleaning the data

4.1 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:

```
#selecting only the observations from the 105th congress
mydata<-mydata%>%filter(congress==105)
#checking selection was correctly done
nrow(mydata)
```

```
## [1] 435
```

```
summary(mydata$congress)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     105     105     105     105     105     105
```

4.2 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:

```r
#selecting variables we will use
mydata<-mydata %>% select(aauw, totchi, ngirls, party, female, white, srvlng,
                          age, demvote, medinc, perf, perw, perhs, percol, perur,
                          moredef, statabb, district, rgroup, region)
```

4.3 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:

```
#converting to a data frame
mydata<-as.data.frame(mydata)
```

5 Summary Statistics

5.1 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:

```r
mydata2<-mydata[, c("party", "age", "white", "female", "aauw", "totchi", "ngirls")]

stargazer(mydata2)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Oct 18, 2023 - 11:22:43 AM

Table 2:

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
party	435	1.529	0.504	1	1	2	3
age	435	51.671	9.618	26	45	58	87
white	435	0.869	0.338	0	1	1	1
female	435	0.110	0.314	0	0	0	1
aauw	435	47.308	42.021	0	0	100	100
totchi	434	2.493	1.648	0.000	2.000	3.000	10.000
ngirls	434	1.274	1.125	0.000	0.000	2.000	7.000

6 Generate Variables

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

Code:

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

head(mydata)
```

```
##   aauw totchi ngirls party female white srvlng age demvote medinc      perf
## 1  100      0      0     1      0     1      7  58    0.57  46389 0.5000151
## 2   88      3      1     1      0     1     15  54    0.60  57915 0.5283446
## 3    0      0      0     2      0     1      1  31    0.43  25401 0.5291693
## 4  100      2      2     1      0     1      1  51    0.52  36067 0.5252141
## 5  100      2      2     1      0     1      7  39    0.59  40674 0.5291277
## 6    0      7      3     2      0     1     27  68    0.28  51258 0.5126253
##        perw perhs percol     perur   moredef statabb district rgroup region
## 1 0.2909628 0.814  0.266 0.9979499        NA      HI        1      0      9
## 2 0.8401480 0.839  0.348 0.9934435 0.1709234      NY        5      4      2
## 3 0.9251785 0.577  0.081 0.3419504 0.3140097      AL        4      1      6
## 4 0.9851513 0.817  0.224 0.4894656        NA      ME        1      1      1
## 5 0.7835612 0.742  0.169 0.9607588 0.1647510      NJ        1      1      2
## 6 0.8334154 0.901  0.407 0.9580938 0.2549923      TX        7      2      7
##   repub
## 1     0
## 2     0
## 3     1
## 4     0
## 5     0
## 6     1
```

7 Run Estimations

7.1 Question: (2 pages) 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:

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

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

reg3<-felm(aauw~ngirls+totchi+female+repub,mydata)

stargazer( reg1, reg2, reg3, se = list(reg1$rse, reg2$rse, reg3$rse))
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Oct 18, 2023 - 11:22:44 AM

Table 3:

	Dependent variable:		
	aauw		
	(1)	(2)	(3)
ngirls	−2.784	5.776**	2.825**
	(1.750)	(2.714)	(1.306)
totchi		−7.992***	−3.149***
		(1.784)	(0.964)
female			12.577***
			(3.258)
repub			−71.783***
			(2.100)
Constant	50.964***	59.982***	87.822***
	(3.036)	(3.520)	(1.809)
Observations	434	434	434
R^2	0.006	0.051	0.796
Adjusted R^2	0.003	0.047	0.794
Residual Std. Error	41.939 (df = 432)	41.010 (df = 431)	19.055 (df = 429)
Note:			*p<0.1; **p<0.05; ***p<0.01

7.2 Question: (2 pages) 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:

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.

Specification 3 adds additional controls that further help reduce bias on β_1, though these controls are not as important since once we control for the total number children, the number of girls a congress person has is only weakly correlated with being female or republican, as you can see in the table below.

```r
regcont1<-felm(ngirls~totchi+female,mydata)

regcont2<-felm(ngirls~totchi+repub,mydata)


stargazer( regcont1, regcont2, type = "latex", se = list(regcont1$rse, regcont2$rse))
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Oct 18, 2023 - 11:22:44 AM

Table 4:

	Dependent variable:	
	ngirls	
	(1)	(2)
totchi	0.500^{***}	0.503^{***}
	(0.027)	(0.028)
female	0.168	
	(0.108)	
repub		-0.087
		(0.076)
Constant	0.009	0.065
	(0.057)	(0.063)
Observations	434	434
R^2	0.537	0.536
Adjusted R^2	0.535	0.534
Residual Std. Error (df = 431)	0.768	0.768
Note:	*p<0.1; **p<0.05; ***p<0.01	

7.3 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:

$ngirls_i$ will be plausibly exogenous if the Conditional Independence Assumption holds. This will be the case if once we control for $totchi_i$, $female_i$ and $repub_i$, 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.

7.4 Question: (2 pages) 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?

```r
regfem<-felm(aauw~ngirls+totchi+repub,mydata[mydata$female==1,])
regmale<-felm(aauw~ngirls+totchi+repub,mydata[mydata$female==0,])
reg4<-felm(aauw~ngirls+totchi+female+repub+female*ngirls,mydata)
reg5<-felm(aauw~ngirls+totchi+female+repub+female*ngirls+totchi*female+repub*female,mydata)


stargazer(regfem, regmale, reg4, reg5,  se = list(regfem$rse, regmale$rse, reg4$rse, reg5$rse))
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Oct 18, 2023 - 11:22:44 AM

Answer:

Looking at columns 1 and 2, we see that the effect of $ngirls_i$ on the aauw score is driven by a positive effect on male legislators. There does not seem to be any effect on the aauw score of female legislators though this coefficient is very imprecisely estimated as there are only 48 female observations.

The estimates in columns 1 and 2 are equivalent to the single specification estimates reported in column 4. The omitted category in column 4 is men and we can see that the non-interacted coefficients in that specification are equivalent to those obtained using the male subset. When we add the interaction coefficient we get coefficients equivalent to those on the female subset (ex: 3.021+(-3.267)=-0.246).

The difference between columns 3 and 4 is that in column 4 we are allowing for $totch_i$ and $repub_i$ to have a differential effect on males vs. females, effectively making our controls more refined. This switched the sign of the $ngirls_i * female_i$ coefficient (though in both cases it is imprecisely estimated).

Table 5:

	Dependent variable:			
	aauw			
	(1)	(2)	(3)	(4)
ngirls	−0.246	3.021**	2.706**	3.021**
	(6.281)	(1.302)	(1.281)	(1.308)
totchi	1.222	−3.417***	−3.133***	−3.417***
	(3.971)	(0.986)	(0.958)	(0.990)
female			11.103**	5.669
			(4.659)	(3.789)
repub	−70.881***	−71.943***	−71.791***	−71.943***
	(9.462)	(2.102)	(2.098)	(2.111)
ngirls:female			1.097	−3.267
			(3.295)	(6.209)
totchi:female				4.639
				(3.963)
female:repub				1.062
				(9.384)
Constant	94.006***	88.337***	87.936***	88.337***
	(3.441)	(1.810)	(1.831)	(1.818)
Observations	48	386	434	434
R^2	0.729	0.791	0.796	0.797
Adjusted R^2	0.711	0.789	0.794	0.794
Residual Std. Error	20.365 (df = 44)	18.920 (df = 382)	19.074 (df = 428)	19.074 (df = 426)
Note:		*p<0.1; **p<0.05; ***p<0.01		

8 Fixed Effects:

8.1 Question: (2 pages) 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.

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

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.

Answer:

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


stargazer(regfe1, regfe2, regfe3,  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: Wed, Oct 18, 2023 - 11:22:44 AM

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.

Table 6:

	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

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

Answer:

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


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

stargazer(regrep)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Oct 18, 2023 - 11:22:44 AM

Table 7:

	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)
factor(rgroup)1	−5.671
	(7.606)
factor(rgroup)2	−10.175
	(7.596)
factor(rgroup)3	−2.466
	(8.860)
factor(rgroup)4	4.012
	(8.223)
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

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

Answer: The 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.

8.4 Question: (2 pages) 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.

$$
\begin{aligned}
aauw_i &= \beta_0 + \beta_1 ngirls_i + \gamma_i + \phi_i + \epsilon_i \\
aauw_i &= \beta_0 + \beta_1 ngirls_i + \gamma_i + \phi_i + \eta_i + \epsilon_i \\
aauw_i &= \beta_0 + \beta_1 ngirls_i + \gamma_i + \theta_i + \epsilon_i \\
aauw_i &= \beta_0 + \beta_1 ngirls_i + \rho_i + \epsilon_i
\end{aligned}
$$

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?

Answer:

```r
mydataall<-read_dta("../../../PSdata/PS2_OVB_FEdata/data/basic.dta")

mydataall$regyear<-paste(mydataall$region, mydataall$congress,sep="_")
mydataall$regyearchi<-paste(mydataall$regyear, mydataall$totchi,sep="_")


regfe1<-felm(aauw~ngirls|totchi+region,mydataall)
regfe2<-felm(aauw~ngirls|totchi+region+congress,mydataall)
regfe3<-felm(aauw~ngirls|totchi+regyear,mydataall)
regfe4<-felm(aauw~ngirls|regyearchi,mydataall)

stargazer(regfe1,regfe2,regfe3, regfe4)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Oct 18, 2023 - 11:22:44 AM

Table 8:

	Dependent variable:			
	aauw			
	(1)	(2)	(3)	(4)
ngirls	5.058***	5.043***	5.125***	4.987***
	(1.257)	(1.257)	(1.265)	(1.405)
Observations	1,735	1,735	1,735	1,735
R^2	0.148	0.151	0.155	0.238
Adjusted R^2	0.138	0.139	0.131	0.110
Residual Std. Error	39.354 (df = 1714)	39.330 (df = 1711)	39.511 (df = 1687)	39.982 (df = 1486)
Note:				*p<0.1; **p<0.05; ***p<0.01

The first estimation pools all the data. It controls for regional voting patterns, and voting patterns that are common across representatives with the same number of children but it does not account for the fact that there may be common voting patterns in each year. The second adds year fixed effects thus it controls for any common voting patterns that are shared by all representatives in that year. The third specification is much more restrictive. It controls for voting pattern specific to a particular region in a particular year. Finally the fourth is the most restrictive. It controls for voting patterns that are common to representatives in a particular region in a particular year that have the same number of children. It is worth noting that for specification 4, the fixed effects are estimated off of a small number of observations since there are not always that many representatives that share all these characteristics and that the fixed effect will not be estimated, thus effectively dropping the observation, for observations that are a unique combination of these characteristics. So even though for this specification the R-squared increased, the adjusted r squared decreases suggesting that the ρ_i fixed effects over fit the model.

8.5 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: First, showing that the results hold in each congress separately is a more compelling result. It is a more transparent way to show the effect since it helps alleviate concerns that the result might be driven by voting on an outlier issue that was only on the docket in a particular year. Second, the panel is not balanced. Depending on the election cycle and reelection, some congress persons appear in the data multiple time whereas others only appear once. There are potential concerns about selection here, which are discussed in the paper and appendix 2. Furthermore, it is not entirely clear conceptually whether the unit of observation should be a congressperson (since for the most part the number of girls each has is unchanged over the 4 year time frame) or a congressperson per congress session. By presenting results for each session, Washington avoids having to take a position on this.

8.6 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:

```
table(mydataall$name)

#it looks like no name appears more then 4 times.

#Just to be sure I create the following variable since it is very unlikely that two congress persons of
mydataall$statename<-paste(mydataall$state, mydataall$name,sep="_")
```

```
## Warning: Unknown or uninitialised column: ‘state‘.
```

8.7 Question:(2 pages) 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:

We can estimate the following specifications:

$$\begin{aligned}
aauw_i &= \beta_0 + \beta_1 ngirls_i + \gamma_i + \epsilon_i \\
aauw_i &= \beta_0 + \beta_1 ngirls_i + \gamma_i + \eta_i + \epsilon_i \\
aauw_i &= \beta_0 + \beta_1 ngirls_i + \gamma_i + \lambda_i + \epsilon_i \\
aauw_i &= \beta_0 + \beta_1 ngirls_i + \gamma_i + \lambda_i + \eta_i + \epsilon_i
\end{aligned}$$

```
regindivfe1<-felm(aauw~ngirls|totchi,mydataall)
regindivfe2<-felm(aauw~ngirls|totchi+congress,mydataall)
regindivfe3<-felm(aauw~ngirls|totchi+statename,mydataall)
regindivfe4<-felm(aauw~ngirls|totchi+congress+statename,mydataall)


stargazer(regindivfe1, regindivfe2, regindivfe3, regindivfe4)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Oct 18, 2023 - 11:22:44 AM

Table 9:

	Dependent variable:			
	aauw			
	(1)	(2)	(3)	(4)
ngirls	5.864***	5.847***	1.167	1.964
	(1.309)	(1.309)	(3.225)	(3.118)
Observations	1,735	1,735	1,735	1,735
R^2	0.062	0.064	0.971	0.973
Adjusted R^2	0.055	0.056	0.955	0.958
Residual Std. Error	41.206 (df = 1722)	41.184 (df = 1719)	8.973 (df = 1141)	8.671 (df = 1135)
Note:				*p<0.1; **p<0.05; ***p<0.01

Though we do see a positive effect 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.

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

Answer: In addition to the problem of not having sufficient statistical power, there is also the issue that the effect identified on the congresspersons when using individual fixed effects is going to be measuring the effect of having an additional female infant/toddler. This effect could be different from that of having a teenage or adult daughter. Furthermore, the appropriate counter factual would be congresspersons who have an additional male infant/toddler. The specification above, as written does not control for the addition of a young child to the household.

8.9 Question: (2 pages) 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?

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


regcol<-felm(aauw~ngirls+female+white+repub+age+agesq+srvlng+srvlngsq+factor(rgroup)+demvote|totchi+regi
```

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

```
stargazer(regcol)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Oct 18, 2023 - 11:22:45 AM

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, gender, and in this data, their religious groups. Thus the individual fixed effect will already capture the effect of these characteristics.

Table 10:

	Dependent variable:
	aauw
ngirls	2.010 (3.139)
female	
white	
repub	−3.034 (6.045)
age	10.393 (7.672)
agesq	−0.003 (0.007)
srvlng	−0.990 (5.366)
srvlngsq	0.0004 (0.009)
factor(rgroup)1	
factor(rgroup)2	
factor(rgroup)3	
factor(rgroup)4	
demvote	0.454 (8.031)
Observations	1,735
R^2	0.973
Adjusted R^2	0.958
Residual Std. Error	8.716 (df = 1121)
Note:	*p<0.1; **p<0.05; ***p<0.01

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

Answer: The region fixed effect is now redundant (assuming that this is a time invariant variable).

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

Answer: We are able to estimate a coefficient for *Repub* (though with large standard errors). This implies that this 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.

9 Submission instructions:

- Since this is a group assignment only one member of the group will upload it to gradescope.
- Make sure the final version of your assignment is knit in pdf format and uploaded to gradescope. Make sure you have one question response per page (unless otherwise indicated) so that question positions align with the template in gradescope.The final PDF should be 29 pages long.