Replicating Research (2): Washington 2008

Using 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

```
library(haven)
library(lfe)
library(dplyr)
library(stargazer)
basic<- read_dta('C:\\Users\\roryq\\Downloads\\Data.dta')
```

Cleaning the data

Restrict your data to observations from the 105th congress and keep only the variables listed in the table below.

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.

```
basic<- basic %>% filter(congress==105) %>% select(aauw,totchi,ngirls,party,female,white,
                                                   srvlng,age,demvote,rgroup,region)
```

Analysis

Estimate the following linear regression models. Report your regression results in a formatted table using. Report robust standard errors in your table.

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

Code:

```r
reg1<- felm(basic$aauw~basic$ngirls,data=basic)
reg2<- felm(basic$aauw~basic$ngirls+basic$totchi, data=basic)
stargazer(reg1,reg2,se = list(reg1$rse, reg2$rse), type="latex")
```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Thu, Oct 31, 2024 - 9:34:32 AM

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

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:

It changes so much because the number of daughters is correlated with the number of children, so the more children you have the more likely you will have daughter, possible a lot by chance. controlling for total number of children allows for more accurate comparisons of having a daughter vs a son with the underlying variable of number of children confounding the results.

In the regression we can see without controlling for number of children number of girls actually has a negative effect, but this could be because people who have a lot of daughters, have a lot of children overall, and having more children overall is correlated with a lower aauw score. By controlling for this we eliminate this underlying negative correlation related to total children to see only the effect of having a daughter. After this control we can see that this effect has now flipped meaning having a daughter holding number of children constant does increase aauw score by over 5 points, and this is significant at the 95% confidence level.

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, i.e. the assumption that must hold for β_1 to be interpreted as a causal estimate? What evidence does Washington give to support this assumption?

Answer:

To hold for model assumptions, number of daughters must be a random variable that is not biases in favor or against having a daughter, ie having a daughter is random and 50% true and 50% of the time false. Therefore we are assuming there is no sex selection, ie no stopping rule, having kids until you get x number of sons, or adoption, where you pick the sex of the child. In those cases where sex is selected, they are determined endogenously by factors that are not random, and violate the assumption of sex being random, and invalidate the claim of causality.

Fixed Effects:

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. Use robust standard errors.

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

Code:

```
model1<-reg1
model2<- felm(basic$aauw~basic$ngirls+factor(basic$totchi), data=basic)
model3<- felm(basic$aauw~basic$ngirls | basic$totchi)
stargazer(model1,model2,model3,se = list(reg1$rse, model2$rse, model3$rse), type="latex")
```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Thu, Oct 31, 2024 - 9:34:32 AM

Table 3:

	Dependent variable:		
	aauw		
	(1)	(2)	(3)
ngirls	−2.784	5.748**	5.748**
	(1.750)	(2.667)	(2.667)
totchi)1		7.616	
		(8.816)	
totchi)2		−6.182	
		(7.074)	
totchi)3		−17.186**	
		(7.770)	
totchi)4		−25.833***	
		(9.090)	
totchi)5		−28.128**	
		(11.601)	
totchi)6		−34.712	
		(24.334)	
totchi)7		−65.986***	
		(11.828)	
totchi)8		−74.859***	
		(15.283)	
totchi)9		−81.108***	
		(14.386)	
totchi)10		−75.360***	
		(11.957)	
Constant	50.964***	52.367***	
	(3.036)	(5.400)	
Observations	434	434	434
R^2	0.006	0.065	0.065
Adjusted R^2	0.003	0.040	0.040
Residual Std. Error	41.939 (df = 432)	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:

The first model does not control for the total number of children, therefore the coefficient is affected by the underlying negative correlation of number of children to aauw score. The second model controls for that negative correlation, therefore the coefficient for number of daughters is now positive. The fixed effect for having each specified number of children compared to the baseline of 0 children is displayed as β_2 through β_{12}. Model 3 is the same as 2 except we did not display the fixed effects compared to the baseline. Meaning that β_1 is the same in models two and three.

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:

```r
# Create Dummy
basic$repub<- ifelse(basic$party==2,1,0)
basic$age_sq<- basic$age^2
basic$serv_sq<- basic$srvlng^2

# omit na value
basic<-na.omit(as.data.frame(basic))

#model
model_A <- felm( data = basic, aauw ~
                   female+white+repub+age+age_sq+srvlng+serv_sq+demvote+ ngirls|totchi+rgroup+region)

# Create table

# Create names to match table
new_coef_names <- c("Number of Female Children", "Female", "White", "Republican", "Age",
"Age Squared", "Service Length", "Service Length Squared",
"Democratic Vote Share in District")

# reorder columns and add title

stargazer(model_A,
         covariate.labels = new_coef_names,
          order = c("ngirls", "female", "white", "repub", "age", "age_sq", "srvlng", "serv_sq",
"demvote"), digits=2,
         title="Impact of Female Children on Legislator Voting on Women's Issues", type = "latex")
```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Thu, Oct 31, 2024 - 9:34:33 AM

Table 4: Impact of Female Children on Legislator Voting on Women's Issues

	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	62.15***
	(11.57)
Observations	434
R^2	0.84
Adjusted R^2	0.83
Residual Std. Error	17.44 (df = 402)
Note:	*p<0.1; **p<0.05; ***p<0.01

Explain what the region fixed effects are controlling for.

Answer:

Region is controlling for any underlying heterogeneity in number of daughters across regions. For example maybe those that live in a region with high pollution are more likely to have a daughter because of genetic mutations due to the high pollution or radiation rates.

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:

```
# Load data
df<- read_dta('C:\\Users\\roryq\\Downloads\\Data.dta')

# Create Dummy
df$repub<- ifelse(df$party==2,1,0)
df$age_sq<- df$age^2
df$serv_sq<- df$srvlng^2
```

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:

```
#model
model_A <- felm( data = df, aauw ~
                   female+white+repub+age+age_sq+srvlng+serv_sq+demvote+
  ngirls|totchi+rgroup+region+congress+name)
```

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

```
# Create table

# Create names to match table
new_coef_names <- c("Number of Female Children", "Female", "White", "Republican", "Age",
"Age Squared", "Service Length", "Service Length Squared",
"Democratic Vote Share in District")

# reorder columns and add title
stargazer(model_A,
         covariate.labels = new_coef_names,
          order = c("ngirls", "female", "white", "repub", "age", "age_sq", "srvlng", "serv_sq",
"demvote"), digits=2,
         title="Impact of Female Children on Legislator Voting on Women's Issues", type = "latex")
```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Thu, Oct 31, 2024 - 9:34:33 AM

Table 5: Impact of Female Children on Legislator Voting on Women's Issues

	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	0.45
	(8.04)
Observations	1,735
R^2	0.97
Adjusted R^2	0.96
Residual Std. Error	8.73 (df = 1117)
Note:	*p<0.1; **p<0.05; ***p<0.01

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:

This actually decreases the estimated effect of having daughters has decreased compared to the previous model suggesting this more complicated adds some explanatory power to voting score not seen in previous model. Additionally since the errors are so much larger the results are no longer significant at traditional confidence levels.

Because the errors are so much larger this indicates overfitting in the model. Partly because the additional degrees of freedom the model uses by adding more variables that dont contribute much explanatory power. Additionally There could be colinearity between these variables, increases the standard errors.

She doesnt use this approach to reduce overfitting in the model, and to make sure that she doesnt violate the assumption of multicolinearity in her model.

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

Answer:

These variables dont change over time, so through the congresses these do not change. Therefore there is not sufficient variation to estimate these coefficients.

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

Answer:

This implies that legislators can and have changed parties, because there was enough variation to estimate this coefficient.