Problem Set 1: Getting Started

Claire Duquennois

NAME: Benjamin Houck

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.

Submission instructions:

1) Knit your assignment in PDF.
2) **Assignment shell extra credit:** You will receive a little extra credit if your answers line up correctly with the answer positions of the template on gradescope. For this to work

- **Work by putting your answers directly into this R markdown file that is pre-formatted for you!**
- Make sure you have ONE question and answer per page (this is how the template is structured and allows gradescope to easily find your answers), unless the question indicated that the answer will require two pages because of large tables.
- Your final PDF should be 19 pages.
- You can insert needed page breaks as illustrated below
- Make sure you do not "print" the data. If you change the data, make sure you store with a structure like: newname<-modification(olddata). If you only type modification(olddata), R will display the data rather than saving your modifications

3) Upload your assignment PDF to gradescope.

Finding the data

I have downloaded Washington's `basic.dta` file and made it available in the RCloud assignment workspace. I downloaded this data from the AER's website which links you to the ICPSR's data repository. Anyone can sign in to get access to the replication data files. These include the typical files in a replication folder: 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.

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`, `ggplot2`, `lfe` and `stargazer`.

If you are working on a desktop version of R (i.e not in the cloud workspace) and have not used a package before you will need to install the packages by un-commenting (removing the #) the following code. If you are working in R Studio Cloud these should load automatically or you will be prompted to load them.

```
#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("ggplot2",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 (it is not necessary to reinstall a package every single time).

Question 1.1:

In the following chunk, call all the packages you will be using with the `library` function.

```
#importing necessary packages via library() function
library(ggplot2)
library(haven)
library(dplyr)
```

```
##
## 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(stargazer)
```

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

```
##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.
```

```
##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

Question 1.2:

Below, create a code chunk in which you load your data. Remember, since `basic.dta` is a .dta file, you will use the `read.dta()` function to load it. Hint: code chunks start and end with "‘ and, as above, need to be given a name in {}. When needed, you can also specify setting for the chunk in these brackets.

```
#assigning data to variable basic_data
basic_data <- read_dta('basic.dta')
```

Question 1.3:

How many observations are in the original dataset?

Hint: use the `nrow()` function.

Code and Answer:

```
cat("There are",nrow(basic_data),"observations in the original dataset.")
```

```
## There are 1740 observations in the original dataset.
```

Cleaning the data

Question 2.1:

The original dataset contains data from the 105th to 108th U.S. Congress reported in the variable `congress`. We only want to keep the observations from the 105th congress.

Hint: Use the `filter` function in the `dplyr` package.

Code:

```
#filtering variable congress_105 to be data only on the 105th congress
congress_105 = basic_data%>% filter(congress=="105")
```

Question 2.2:

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
nowtot	NOW 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
age	Age

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:

```
#loading dplyr

library(dplyr)

#cutting down data set

final_dataset<-congress_105 %>% select(aauw,nowtot,totchi,ngirls,party,female,age)
```

Question 2.3:

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:

```
is(final_dataset)
```

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

```
data_df = data.frame(as.matrix(final_dataset))

is(data_df)
```

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

Summary Statistics

Question 3.1:

Report summary statistics for all the remaining variables in the dataset. 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 `type="latex"` for PDF output. If you have trouble knitting to PDF, try installing MikTex (https://miktex.org/download)

Code:

```
library(stargazer)

library(stargazer)

# Generate the summary statistics table in LaTeX format
stargazer(data_df,
          type = "latex",
          summary.stat = c("mean", "sd", "min", "max", "n"),
          title = "Summary Statistics")
```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Wed, Dec 11, 2024 - 12:13:06

Table 2: Summary Statistics

Statistic	Mean	St. Dev.	Min	Max	N
aauw	47.308	42.021	0	100	435
nowtot	41.311	36.534	0	100	431
totchi	2.493	1.648	0	10	434
ngirls	1.274	1.125	0	7	434
party	1.529	0.504	1	3	435
female	0.110	0.314	0	1	435
age	51.671	9.618	26	87	435

Generate Variables

Question 4.1:

Construct a variable called *repub*, a binary set to 1 if the observation is for a republican, 0 otherwise.

Code:

```
repub = ifelse(data_df$party == 2, 1, 0)
```

Question 4.2:

Construct a variable called $age2$, where age2=age^2.

Code:

```
data_df$age2 = (data_df$age)^2
```

Analysis

Question 5.1 (2 pages):

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

Hints:

- in stargazer, to get robust standard errors, specify `se = list(model1$rse, model2$rse, model3$rse)` and `type = "latex"`.
- your estimates of β_1 should be similar, but not exactly the same, as the estimate in the first row, second column of table 2 in Washington(2008).
- make sure you specify results = "asis" in the chunk header to print the table

Model 1: $aauw_i = \beta_0 + \beta_1 ngirls_i + \beta_2 totchi + \epsilon_i$

Model 2: $aauw_i = \beta_0 + \beta_1 ngirls_i + \beta_2 totchi + \beta_3 female_i + \beta_4 repub_i + \epsilon_i$

Model 3: $aauw_i = \beta_0 + \beta_1 ngirls_i + \beta_2 totchi + \beta_3 female_i + \beta_4 repub_i + \beta_5 age_i + \beta_6 age_i^2 + \epsilon_i$

Code:

```
library(lfe)
```

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

```
library(stargazer)

# Model 1
model1 = felm(aauw ~ ngirls + totchi, data = data_df)

# Model 2
model2 = felm(aauw ~ ngirls + totchi + female + repub, data = data_df)

# Model 3
model3 = felm(aauw ~ ngirls + totchi + female + repub + age + age2, data = data_df)

# Report the results using stargazer with robust standard errors
stargazer(model1, model2, model3,
          se = list(model1$rse, model2$rse, model3$rse),
          type = "latex",
          title = "Regression Results",
          dep.var.labels = "AAUW",
          covariate.labels = c("Number of Girls", "Total Children", "Female", "Republican", "Age", "Age
          omit.stat = c("f", "ser"),
          header = FALSE)
```

Table 3: Regression Results

	Dependent variable:		
	AAUW		
	(1)	(2)	(3)
Number of Girls	5.776**	2.825**	2.899**
	(2.714)	(1.306)	(1.289)
Total Children	−7.992***	−3.149***	−3.557***
	(1.784)	(0.964)	(0.964)
Female		12.577***	12.064***
		(3.258)	(3.205)
Republican		−71.783***	−71.286***
		(2.100)	(2.176)
Age			0.814
			(0.971)
Age Squared			−0.006
			(0.010)
Constant	59.982***	87.822***	63.184***
	(3.520)	(1.809)	(23.987)
Observations	434	434	434
R^2	0.051	0.796	0.798
Adjusted R^2	0.047	0.794	0.795
Note:		*p<0.1; **p<0.05; ***p<0.01	

Question 5.2:

Interpret your estimate of β_1 from the first regression. Be sure to touch upon Sign, Size and Significance

Answer:

The coefficient β_1 in the first regression represents the increase in women's rights voting for each additional daughter a politician has. This increase is reflected by a rise in the AAUW scoring on women's rights issues. The value of β_1 is 5.776 with a standard deviation of 2.714, indicating a much larger impact on voting than in other regressions. This likely occurs because the independent variable serves as a proxy for gender, political affiliation, and age. Although fewer variables lead to a larger standard error, the results remain significant with $p < 0.05$.

Question 5.3:

How does age relate to the aauw score? At what age does the relationship between the aauw score and age "flip''? Is this relationship statistically significant?

Answer:

Age shows an increase in the AAUW score until a politician turns 67.83, when the function

$$0.814a - 0.006a^2$$

becomes negative. However, this relationship is not statistically significant, as it does not meet the 90% confidence level.

Question 5.4 (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:

- Model A: Model 1
- Model B: Model 1 on women only
- Model C: Model 1 on men only
- Model D: Model 1 with the addition of $female$, $female \times ngirls$ and $female \times totchi$

Present all four regressions in one stargazer table with robust standard errors. Is there evidence that the effect of a daughter differs for male and female legislators?

Code and Answer:

```
data_women = data_df%>% filter(female==1)

data_men = data_df%>% filter(female==0)

# Model A
modelA = felm(aauw ~ ngirls + totchi, data = data_df)

#Model B
modelB = felm(aauw ~ ngirls + totchi, data = data_women)

#Model C
modelC = felm(aauw ~ ngirls + totchi, data = data_men)

# Model D
modelD = felm(aauw ~ ngirls + totchi + female + female*ngirls + female*totchi, data = data_df)


stargazer(modelA, modelB, modelC, modelD,
          se = list(modelA$rse, modelB$rse, modelC$rse, modelD$rse),
          type = "latex",
          title = "Regression Results",
          dep.var.labels = "AAUW",
          covariate.labels = c("Number of Girls", "Total Children", "Female", "Female $\\times$ Daughte
          omit.stat = c("f", "rsq", "adj.rsq"),

          header = FALSE)
```

Table 4: Regression Results

	Dependent variable:			
	AAUW			
	(1)	(2)	(3)	(4)
Number of Girls	5.776**	3.043	5.071*	5.071*
	(2.714)	(10.070)	(2.829)	(2.838)
Total Children	−7.992***	−5.428	−7.525***	−7.525***
	(1.784)	(6.360)	(1.845)	(1.850)
Female				28.176***
				(9.561)
Female × Daughters				−2.029
				(10.220)
Female × Total Children				2.097
				(6.471)
Constant	59.982***	84.532***	56.356***	56.356***
	(3.520)	(9.058)	(3.650)	(3.661)
Observations	434	48	386	434
Residual Std. Error	41.010 (df = 431)	38.347 (df = 45)	40.213 (df = 383)	40.021 (df = 428)
Note:				*p<0.1; **p<0.05; ***p<0.01

Question 5.4:

How do the coefficients in models B and C relate to those in model D? Specifically, how can I calculate β_1 and β_2 from models B and C using the results in model D?

Answer:

Using Model D, we can express Model B when female = 1 as follows:

$$56.356 + 5.071 \times \text{ngirls} - 7.525 \times \text{totchi} + 28.176 \times 1 - 2.029 \times 1 \times \text{ngirls} + 2.097 \times 1 \times \text{totchi}$$

$$= 84.532 + 3.042 \times \text{ngirls} - 5.428 \times \text{totchi}$$

This is the same as Model B which shows the predicted AAUW score for women. They are the same because when female = 1 it only includes the women's data, therefore they have the same coefficients.

Similarly, for Model C when female = 0:

$$56.356 + 5.071 \times \text{ngirls} - 7.525 \times \text{totchi} + 28.176 \times 0 - 2.029 \times 0 \times \text{ngirls} + 2.097 \times 0 \times \text{totchi}$$

$$= 56.356 + 5.071 \times \text{ngirls} - 7.525 \times \text{totchi}$$

Through the same reasoning as Model B, this is the same as Model C because they use solely the data on how the number of daughters affect male voting.

Question 5.5 (LAST QUESTION!):

Code:

```
library(ggplot2)

data_twokids = data_df%>%filter(totchi == 2)

ggplot(data_twokids, aes(x = factor(ngirls), y = nowtot, fill = factor(ngirls),fig.width=6, fig.height=4
  geom_bar(position = "dodge", stat = "summary", fun = "mean") +
  geom_text(stat = "summary", aes(label = round(after_stat(y), 2)), fun = "mean", vjust = -0.5) +
  labs(
    title = "NOW Score by Number of Daughters (2 Children)",
    x = "Number of Daughters",
    y = "Mean NOW Score",
    fill = "Number of Daughters"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 14, face = "bold"),
    axis.text = element_text(size = 12),
    axis.title = element_text(size = 13),
    legend.position = "none"
  )
```