Problem Set 1: Getting Started

Claire Duquennois

NAME:**_____**

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.

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 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.

Question 1.1:

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

```
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(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
```

```
library(ggplot2)
library(lfe)
```

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

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.

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

Question 1.3:

How many observations are in the original dataset?

Hint: use the `nrow()` function

Code and Answer:

```
nrow(mydata)
```

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

The data contains 1740 observations.

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:

```
#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
```

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:

```
#selecting variables we will use
mydata<-mydata %>% 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(mydata)
```

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

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

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

Code:

```
stargazer(mydata, 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: Wed, Jan 25, 2023 - 11:01:41 AM

Table 2:

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
aauw	435	47.308	42.021	0	0	100	100
nowtot	431	41.311	36.534	0.000	5.000	80.000	100.000
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
party	435	1.529	0.504	1	1	2	3
female	435	0.110	0.314	0	0	0	1
age	435	51.671	9.618	26	45	58	87

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:

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

Question 4.2:

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

Code:

```
mydata$age2<-mydata$age*mydata$age
```

Analysis

Question 5.1:

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.

Hints:

- in stargazer 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).

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:

```
reg1<-felm(aauw~ngirls+totchi,mydata)
reg2<-felm(aauw~ngirls+totchi+female+repub,mydata)
reg3<-felm(aauw~ngirls+totchi+female+repub+ age+ age2,mydata)
stargazer( reg1, reg2, reg3, type = "latex", 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, Jan 25, 2023 - 11:01:43 AM

Table 3:

	Dependent variable:		
	aauw		
	(1)	(2)	(3)
ngirls	5.776**	2.825**	2.899**
	(2.714)	(1.306)	(1.289)
totchi	−7.992***	−3.149***	−3.557***
	(1.784)	(0.964)	(0.964)
female		12.577***	12.064***
		(3.258)	(3.205)
repub		−71.783***	−71.286***
		(2.100)	(2.176)
age			0.814
			(0.971)
age2			−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
Residual Std. Error	41.010 (df = 431)	19.055 (df = 429)	19.023 (df = 427)
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: Controlling for the total number of children, an additional daughter is predicted to increase a congresspersons AAUW score by 5.78 points. This relationship is statistically significant at the 5% level. I can reject the null hypothesis of no effect with a 95% level of confidence.

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:

$\frac{daauw}{dage} = \beta_5 + 2 * \beta_6 * age = 0.8 - 0.012age$ thus age is associated with a higher aauw score until the age of 66.6 at which point it is associated with lower AAUW scores. This relationship is not statistically significant.

Question 5.4:

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 using the felm function:

- 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 these four regressions in a stargazer table. Is there evidence that the effect of a daughter differs for male and female legislators?

Code and Answer:

```
reg1<-felm(aauw~ngirls+totchi,mydata)
regfem<-felm(aauw~ngirls+totchi,mydata[mydata$female==1,])
regmale<-felm(aauw~ngirls+totchi,mydata[mydata$female==0,])
reginter<-felm(aauw~ngirls+totchi+female+female*ngirls+totchi*female, mydata)

stargazer(reg1, regfem, regmale, reginter, type = "latex", se = list(reg1$rse, regfem$rse, regmale$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, Jan 25, 2023 - 11:01:44 AM

It looks like the effect of daughters is larger for men but the difference is not statistically significant: I cannot reject the null that the effect is the same for men and women since the estimate on the interaction term between female and ngirls is not statistically significant.

Table 4:

	Dependent variable:			
	aauw			
	(1)	(2)	(3)	(4)
ngirls	5.776**	3.043	5.071*	5.071*
	(2.714)	(10.070)	(2.829)	(2.838)
totchi	−7.992***	−5.428	−7.525***	−7.525***
	(1.784)	(6.360)	(1.845)	(1.850)
female				28.176***
				(9.561)
ngirls:female				−2.029
				(10.220)
totchi:female				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
R^2	0.051	0.018	0.052	0.103
Adjusted R^2	0.047	−0.026	0.047	0.092
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: As specified in the code above:

$\beta_1^B = \beta_1^D + \beta_4^D$

$\beta_2^B = \beta_2^D + \beta_5^D$

$\beta_1^C = \beta_1^D$

$\beta_2^C = \beta_2^D$

Question 5.5:

Lets reproduce the first set of columns in the top chart of figure 1:

- Filter your data so that it only includes representatives with two children
- use ggplot, with geom_bar to generate this plot with the NOW score on the vertical axis and the number of daughters on the horizontal axis.
- Hint: geom_bar(position = "dodge", stat = "summary", fun = "mean")

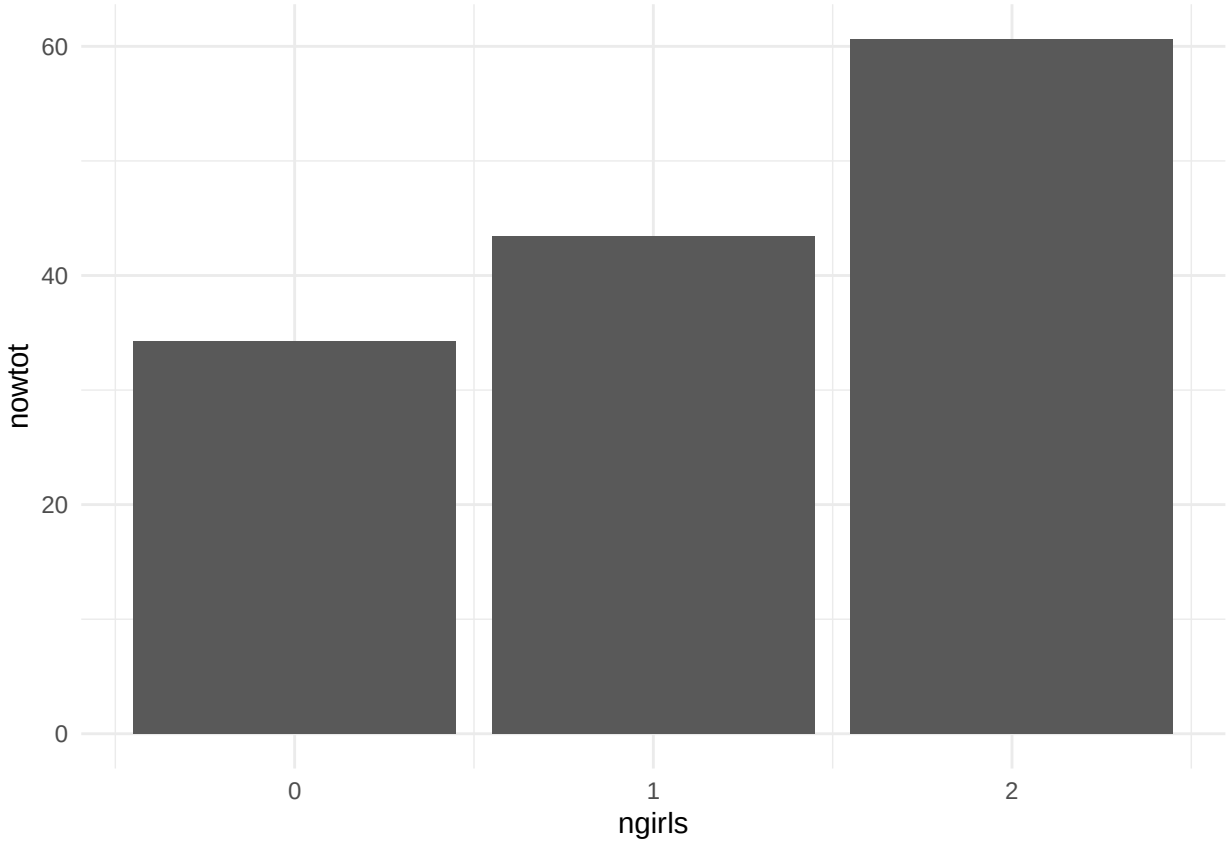
Make you graph as nice as possible!

Code:

```
hist<-ggplot(mydata[mydata$totchi==2,], aes(ngirls, nowtot ))+
  geom_bar(position = "dodge",
           stat = "summary",
           fun = "mean")+
    theme_minimal()

hist
```

```
## Warning: Removed 1 rows containing non-finite values (stat_summary).
```



Submission instructions:

1) Knit your assignment in PDF.
2) Make sure you have ONE question and answer per page (this allows gradescope to easily find your answers).
3) Upload your assignment PDF to gradescope.