llaszewski - Replication Project Diary - 2024-01-19

Scripts Verified By: Asatko - Aaron Satko

01/8/2024- Article and Variables

My article: **Hopcroft Martin 2014.pdf**
Data Set: GSS data set

Next, I looked at the Descriptive Statistics to see which variables I would Need:

Using the Codebook and also this site https://gssdataexplorer.norc.org/variables/vfilter

Statistics	Variable Name	Description
Years of education	educ	highest year of education completed
Sex(Male-1)	sex	respondents sex
Fathers SEI	pasei10	r's fathers socioeconomic index (2010)
Age	age	age of respondent
Race(White-1)	race	race of respondent
Number of Siblings	sibs	number of brothers and sisters
<HS	degree	r's highest degree
HS	degree	r's highest degree
Associate degree	degree	r's highest degree
Bachelor's degree	degree	r's highest degree
Graduate degree	degree	r's highest degree

Info for variables:
Degree: <HS= 0 , HS= 1, Junior College = 2, Bachelors Degree = 3 , Graduate/Professional Degree = 4
Educ: Years of education attained , ranging from 0 to 20.
Sex: 1 or 2. 1= Male 2 = Female (article does 1=Male 0=Female)
Race: Race 1 = White 2 = Black 3 = Other
Pasei10: On a 1-100 scale
NOTE: * PASEI10 was chosen since it's based on 2010. It could possibly be PASEI needed instead *

01/15/2024 - Data Mangement

Renamed Sex variable to sex_updated and Race variable to race_updated

After reading my article thoroughly and observing the descriptive statistics table I've decided to work through, I was ready to use R. Table 1 is the descriptive statistics table, and Table 3 is the regression model Ill eventually try and replicate.

I worked on my data management scripts throughout the day. I was beginning with just reading in the RDS for the GSS dataset.

The scripts I ended up with are what they contain are as follows:

llaszewski-ReplProj-LoadingDataV1

- I loaded in the GSS Data set and saved it as an RDS.

llaszewski-ReplProj-subsetyearsV1

- Created a subset that only includes observations from 2000-2010.

llaszewski-ReplProj-VariablesV1

- Removed all the unneeded variables, have 7 remaining. Also restructured any variable in the dataset that needed it. Race and Sex needed to be adjusted to only have observations of 0 or 1.

llaszewski-ReplProj-RemovePaseiV1

- Removed missing values from Pasei10 variable. Were in the form ".i"

llaszewski-ReplProj-RemoveAge_EducV1

- Removed missing values in the variables Age in the form ".n", and Educ which were ".a"

llaszewski-ReplProj-RemoveSibs_DegreeV1

- Removed missing values from values in the variables Sibs in the form ".n" and ".d". Then Degree, ".a".

llaszewski-ReplProj-MetaDataV1

- Making all variables numeric. So descriptive statistics can be done. Added metadata for each of the variables. Also for the subset (did this for each subset).

I also included code in these scripts that double-checks my work and that the code does what I want it to do.

The process in these scripts was different techniques to clean the data and get it down to the article's observation amount of N=11,857. After the days of work, I was able to get the observations to N=12,210.

I did a quick check of the means to see how far off I am. Most are very close, other than Race where I need to do more work to find what is making it so far off.

01/16/2024 - Data Management (Continued) / Descriptive Statistics

Figured out that Race variable should be changed so White = 1 and Other = 0.
This made the mean and standard deviation almost exactly what they should be.

The only remaining obvious difference is looking at the PASEI10 variable. The actual mean is 47.10, where I got 45.41. The standard deviation is also off by a little more than 1. There's a slight chance this has to do with the variable choice I made when I mentioned I could have used PASEI instead. But more likely it is just me being 300 observations off so I see a slight difference in the statistics.

I then created the table of means and standard deviations using a loop. I was able to use my loop from assignment 6 and just switch around the names. This uses mean(), sd(), data.frame(), and then format() to make the rounding look nicer.

Then I had to make a second table that shows the percentages of respondents who graduated/completed the following: Less than High School, High School, Associate Degree, Bachelor's Degree, and Graduate Degree. This uses data.frame() as well, along with prop.table() and then sprintf() to format the percentages.

In the end, a rbind() is used to combine these two tables and make it match up with my article (table 1).

01/17/2024 Data Analysis

I went back through my data management to check through my work. I'm now also adding metadata for each subset at the end of all my scripts.

The new scripts I created today for my descriptive statistics were:

llaszewski-ReplProj-DescStatsFrequenciesV1

- Made a table for degree variable frequencies. Using the prop.table() function.

llaszewski-ReplProj-DescStats_MeanSDV1

- Used a for loop to create a table of means and standard deviations for the rest of the variables. Then, rbind() to combine this table with the frequency table. Creating the full descriptive statistics table. (Table 1)

01/18/2024 Analysis , Read Me

After wrapping up my work with the Descriptive statistics scripts, I created my regression models for Table 3 and Table 4 in the article.

llaszewski-ReplProj-RegressionTable3V1.R

- This script is where I got my numbers for Table 3. Using the lm() function, I made Model 1 and 2, for all cases.

llaszewski-ReplProj-RegressionTable4V1.R

- This script is where I got my numbers for Table 4. I made Model 1 and 2, which was for cases where age > 24

I then made these tables neat and look like they do in my article. Now, I have all the tables completed. Table 1 (Descriptive Statistics), Table 3 (all cases), Table 4 (cases > 24).

I then started making my Readme file. The Readme file will include where all my files will be stored for the project, and what needs to be set at your working directory to replicate my work. Then I made my document that will hold all of the tables. This includes the descriptive statistics table created by the authors, along with mine. Then the regression models created by the authors and mine. This way they are all next to each other and easily comparable.

I then read through some example papers and looked at how I should organize my paper into the correct format and started doing that.

01/19/2024 Writing Paper

My tables/graphs and Read Me files are officially completed. I now have begun writing my paper. Completing the abstract and introduction

01/20/2024 Finish Writing Paper

Finishing up the paper, saving hidden/shown versions of both paper and table documents.

Scripts Verified

Completed project, and got scripts checked and verified By Asatko - Aaron Satko