Assignment 1

EC295 – Fall 2020
Due: Friday, October 9, 9:00pm

Assignment Description

In this assignment you are asked to manipulate data, estimate statistical relationships, and interpret the findings. The main goal behind the assignment is to help you get more comfortable applying statistical methods and using software, but also to think about a policy-relevant topic that economists actively research today.

The questions below guide you through the process of statistical estimation. You are provided the relevant Stata commands you will need, some of which you will not have seen before. It will therefore be useful for you to use the "help" function in Stata, and/or to look up the command in the Stata reference manuals (which are available within Stata as PDFs), or Google. You are also, as always, welcome to ask me for help. Finally, some of the relevant Stata commands are discussed in tutorials, so you may want to review those Zoom videos, which are posted in mylearningspace.

I strongly suggest that you start this assignment early because it will not be possible (in my opinion) to do well if you start close to the due date. There are parts that you may find difficult; you will want to identify them and leave enough time to ask questions if necessary.

Assignment Instructions

Data analysis

In mylearningspace, you will find a datafile called "assign1.dta" that contains the data for this assignment. Download it on to your computer and make note of the folder where you save it.

I have also provided a template dofile that all students must use to write their assignment dofile. Store it in the same folder where you put your data. You will need to manipulate that template in the following way:

- Rename the file from "assign1 template.do" to your family name followed by your student number (no spaces)
- After *cd*, replace INSERT THE PATH TO THE FOLDER WHERE YOU STORE THE DATA with the path to the folder where you stored the assign1.dta dataset. Do not remove the quotation marks.
- After *log using*, replace INSERT YOUR LAST NAME AND STUDENT NUMBER HERE with your last name and student number, with no space between the two. Do not remove the quotation marks
- After *set seed*, replace INSERT YOUR STUDENT NUMBER HERE with your full student number.

Leave all other commands and comments untouched. You should type in your Stata commands below the line that says "Insert your stata commands below here", but above "Insert your stata commands above here".

Note that the *set seed* and *sample* commands will take a random 90% subsample of the data that is different for every student. For this reason, the numbers that you get with your output will not be the same for any two students. Be mindful of this if you are comparing your work with your peers.

Submission

You are required to submit three documents to the dropbox in mylearningspace:

a) **A report** containing your answers to all the questions. I outline below how I would like your report to look. The overall goal is that the answers to each question must be easily identifiable in a readable, professional-looking document;
b) **Stata dofile**;
c) **Stata log file**

In the report described in (a) above, please answer all questions in the same order as they are stated on the question sheet. For each question and sub-question, include the relevant Stata code (if any) that you used, the output generated by that command if there was any, and an interpretation if you are asked to provide it. For example, if you were answering the following hypothetical question, it might look like this:

****

1) Locate the variable y

    a. Using the *tab* command, provide a frequency distribution for y

        Stata commands:

```
tab y;
```

        Output:

```
          y |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |     23,844       10.05       10.05
          2 |    138,568       58.40       68.45
          3 |      9,049        3.81       72.26
          4 |     63,162       26.62       98.88
          5 |      2,651        1.12      100.00
------------+-----------------------------------
      Total |    237,274      100.00
```

****

You could also format your own output tables rather than copying and pasting Stata output from your log file if you find it easier. The key is that as long as the questions are answered in order, and the Stata commands used for each subquestion and associated output are clear, it will be fine.

A note on plagiarism: this is an independent assignment, which I expect you to complete on your own. It is plagiarism to copy someone else's work verbatim, which includes Stata dofiles. Any work you submit should be yours only.

In mylearningspace you will find a dataset called "assign1.dta". Please use this datafile to answer the following questions. **Each subquestion is worth 5 points, for a total of 65.**

1) Imagine you are interested in learning about the education level of teachers in schools that serve kids from different backgrounds.

    a. Using the *tabulate* command, generate a table with the joint probability distribution between "teacher has a masters degree" and "student lives in suburb". Describe the results [hint: you will need to add an option to this command to produce the joint probability distribution].

    b. Use the *tabulate* command to produce the probability distribution for "teacher has a masters degree". Then use the *tabulate* command again to produce the probability distribution for "teacher has a masters degree" conditional on the student living in a suburb. How does the probability that a teacher has a masters degree change when you use only students who live in a suburb? Based on this, is teacher education independent of students living in the suburbs?

2) Suppose you are interested in learning about the math scores of kindergarteners, and testing some hypotheses about the population average test score.

    a. Using the *summarize* command, compute the mean, standard deviation, and median spring math score. Interpret each value.

    b. Manually compute the t-statistic for testing the null hypothesis that the spring math score equals 51.8 against the alternative that it does not equal 51.8. Interpret the value of this statistic.

    c. Using a significance level of 5%, use the *invttail* function combined with the *display* fucntion to compute the critical value for the hypothesis test. Do you accept or reject the null hypothesis? Explain why. Recall that the degrees of freedom are n-1.

    d. Plot the t probability density function for the test in (b) using the *twoway function* command. Plot over the range -5 to 5, and include a vertical line at the critical value(s).

    e. Using the *ci* command, compute a 95% confidence interval for the mean spring math score. What is the set of null hypotheses would we accept at the 5% level?

    f. Using the *ci* command, compute a 90% confidence interval for the mean spring math score. Explain why the interval is narrower than the one in (e).

    g. Using the *ttest* command, perform a one-sided test where the null hypothesis is that the mean spring math score is less than or equal to 51.8, and the alternative is that it is greater than 51.8. Based on the p-value for this test, what is the range of significance levels that would lead you to accept the null hypothesis?

3) question asks you to relate the spring math score to class size

a. Using the *correlate* command, compute the <u>covariance</u> between spring math scores and class size. Next, use the *generate* command to create a new variable called *xmathtc2* that equals spring math scores divided by 10. Compute the covariance between this new variable and class size. Compare the two covariances and explain any differences.

b. Use the *correlate* command to compute the <u>correlation</u> between spring math scores and class size. Next use the same command to compute the correlation between *xmathtc2* and class size. Compare the two correlations and explain any differences.

c. Find the average spring math score at each class size by typing *egen meanmath = mean(mathtc2), by(classsize).* Using the *twoway scatter* command, draw a scatterplot between average spring math scores and class size. Based on this graph, do math scores and class size appear independent?

d. Suppose you were to model the relationship between math scores and class size using the following linear regression model:

$$math = \beta_0 + \beta_1 class\_size + u$$

Precisely interpret β_0, β_1, and u in this context.