**EC395 Applied Econometrics**

**Wilfrid Laurier University**

**Fall 2021**

**Stata assignment #1**

Due date: **11 pm on Thursday, October 21st**

In this assignment, you are going to explore how labour market outcomes vary depending on the type of firm a worker works for. We’re going to use household survey data from Vietnam, specifically the Vietnam Household Living Standards Surveys. In Vietnam, there are five types of firms: informal firms (i.e., not registered with the government), collectives, state-owned firms, private-owned firms, and foreign-owned firms.

I have posted data on MyLS from the 2010, 2012, 2014, and 2018 household surveys. I have also included the associated questionnaires. You are to use data from only one survey depending on the second-last digit in your student number:

* 0 and 1: Use 2010
* 2 and 3: Use 2012
* 4, 5 and 6: Use 2014
* 7, 8 and 9: Use 2018

The datasets have been uploaded as zip files with multiple datasets in each zip file. Download the file based on your student number and unzip the files.

Learning objectives targeted by the activity include:

KB3 – Empirical analysis of economic relationships;

SB4 – Use computer software common in economic research, including statistical analysis software

Stata skills targeted included:

1. Cleaning data for analysis
2. Data management: merging
3. Creating new variables
4. Looping
5. Regression analysis, including controlling the sample of observations included by using logical operators
6. Graphic creation

Submission instructions:

* By the assignment deadline, you are to upload one file to Gradescope
* The file should contain the answers to the questions, a copy of your entire do file, and a copy of your entire log file
* The report needs to include the tables and figures that you generate
* All tables and figures should be clearly labelled and easy to read. The tables cannot simply be copied and pasted from the Stata results window. You are to create nicely formatted, easy to read tables
* When preparing your report, please organize it by question and include both your code to generate the output and the output for the specific question
* The do file should perfectly replicate all aspects of your analysis, including loading the dataset
* The log file should contain all results displayed in the Stata output window
* Upload the file to Gradescope by **11 pm on Thursday, October 21st**

[35 marks]

1. We’re going to start by examining the raw datasets. Open the dataset from section 1A of your data. [7 marks]

a) Examine the variable **ivid** (note that it will have a 2-digit end, such as 10, indicating the survey year, e.g., ivid10, ivid12, etc.). This variable is supposed to uniquely identify observation in the data. Does it do so? How do you know? [1 mark]

b) Remove observations without a unique individual identifier, if there are such observations. Hint, you might find the command **duplicates** helpful. [1 mark]

c) Find the variable that stores the individual’s gender and rename it **gender**. [1 mark]

d) Are there any unexpected values in the variable **gender**? How do you know? [1 mark]

e) Find the variable that stores the individual’s age and rename it **age**. [1 mark]

f) Are there any unexpected values in the variable **age**? How do you know? [1 mark]

g) Save a version of the data with only the variables **ivid, gender,** and **age**. Name the dataset demographics.dta. [1 mark]

2. Use the dataset from section 2A of the questionnaire to create a dataset that has the individual identifier and the highest grade completed by the individual. Call the variable storing the highest grade completed **grade**. Keep ivid10 and grade and save this dataset with the name grade.dta. Pay attention to the data cleaning steps used in question 1 and whether they apply to this dataset as well. [5 marks]

3. Use the dataset from section 4A of the questionnaire to create a dataset that has the individual identifier, the salary/wages received over the past 12 months for their main job, and the type of organization (farming…., independent …, collective, private, state, or foreign) the individual worked for in their main job. Call the variable storing the salary/wage **salary** and the variable storing the type of organization **ownership**. Keep ivid, ownership, and salary and save this dataset with the name work.dta. Pay attention to the data cleaning steps used in question 1 and whether they apply to this dataset as well. [5 marks]

4. Merge the three datasets together. What do you notice about the merge? Did it work as you expected? [2 marks]

5. [16 marks]

a) How does the tendency to work for a foreign firm vary by gender, education, and age? Focus on working age individuals, 15 to 54 inclusive. Decide on an appropriate table or figure to show these relationships. All tables and figures should be cleanly labelled and easy to understand. [3 marks]

b) Estimate a regression that allows for testing the null hypothesis that individuals that work in a foreign-invested firm earn a higher salary/wage (use the ln of salary/wage). Restrict the sample to individuals ages 15 to 54. Write down an equation to represent this regression and explain all of its terms. Then estimate the regression. How do you interpret the coefficient? [3 marks]

c) A classmate suggests that foreign-invested firms hire better educated workers than other types of firms. Is this a threat to internal validity? If so, which threat is it and why? If it is a threat, how can you address it? Show results addressing it and interpret and changes in the estimated coefficients. [2 marks]

d) A classmate suggests that foreign-invested firms employ a lot of women and women tend to earn less than men. Using a loop, loop over men and women to estimate the regression in 5b) separately for men and women. Interpret the results. [3 marks]

e) Critically discuss whether the regressions estimated in (d) are internally valid. That is, do they represent the causal effect on salary of working in a foreign firm? Discuss all five threats discussed in section 9.2 of the text. [5 marks]

6. Include a copy of your complete do file.

7. Include a copy of your complete log file.

Bonus: [5 marks]

Use a loop to loop over 5-year age groups (15-19, 20-24, …, 50-54) to run separate regressions for each age group. Use just women and control for highest grade completed. Store the coefficients from each regression and generate the following plot in Stata (all work should be done in Stata).

2010:



2012:



2014:



2018:

