

**EC395 Applied Econometrics**  
**Wilfrid Laurier University**  
**Fall 2020**  
**Stata assignment #1**

Due date: **11 pm on Thursday, October 8<sup>th</sup>**

In this assignment, you are going to explore how foreign direct investment (FDI) has influenced labour market outcomes for workers in Vietnam.

I have posted ten different datasets on MyLS of data extracted from the 2009 Vietnam population census. These are the zip files “Assignment 1 dataset version X.zip” where X ranges from 0 to 9. You are to download and use the version of the dataset that corresponds to the second-last digit in your student number. You can open a zipped Stata dataset directly in Stata using the `zipfile` command, but it is probably better to unzip it first and then load the Stata dataset.

Additionally, I have uploaded a dataset called “dn2009.dta”. All students use the same version of this dataset.

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

Submission instructions:

- By the assignment deadline, you are to upload a do file, a log file, and a report that provides answers to the questions below
- 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 report to Gradescope by **11 pm on Thursday, October 8<sup>th</sup>**
  - Upload the do file and log file to MyLS by **11 pm on Thursday, October 8<sup>th</sup>**
- 

[54 marks]

1. [7 marks] Calculate total employment in foreign invested firms in a province:

Begin by loading the dataset “dn2009.dta” into Stata.

a) Restrict the dataset to firms that are (i) 100% foreign, (ii) a joint venture between state and foreign, or (iii) a joint venture between others and foreign. The variable ownership stores the ownership type. You should end up with 6,547 observations. [1 mark]

b) Calculate total employment in foreign invested firms by province. You will likely find it helpful to learn how to use the collapse command if you have not used it before and the following video tutorial could be helpful:

<https://www.youtube.com/watch?v=zOkJmJINXNA>. [2 marks]

c) Save your dataset for later combining with the population census data. [1 mark]

d) Summarize the distribution of employment in foreign invested firms across provinces. There are numerous ways you could do so. I want you to think about and explore different methods. [3 marks]

2. [6 marks] Combine the provincial FDI employment values with the population census data

Open the population census data in Stata. Please sure to use the version that corresponds to the second last digit of your student number.

a) Merge the census data with the provincial FDI employment values. If you are not familiar with the merge command, you will likely find this video tutorial useful:

<https://www.youtube.com/watch?v=niGZBRyyDuY&t=73s>. You may also find it useful to refer to the Stata manual on how to choose the “Type of merge”. [2 marks]

b) What do you notice about the output that Stata provides after merging? Look at the help file to learn what the codes mean. Is this result what you expected? [2 marks]

c) For provinces in the census data that did not have an observation in the FDI employment data from the firm dataset, replace the FDI employment value with a 0. [2 marks]

### 3. Working [22 marks]

a) Create an indicator variable that takes the value 1 if the individual is working and 0 otherwise. Call this variable `working`. See the variable `empstat` for various values describing the individual's economic activity. [2 marks]

b) How does the likelihood of working change with age? Create a graph or table that nicely demonstrates the pattern. [2 marks]

c) Estimate the following regression:  $\text{working}_{ip} = \beta_0 + \beta_1 \text{employment}_p + \epsilon_{ip}$  where  $\text{working}_{ip}$  is an indicator variable for whether individual  $i$  in province  $p$  is working and  $\text{employment}_p$  is the number of workers in foreign-invested firms in province  $p$ . [2 marks]

d) How do you think you should calculate consistent standard errors? Do you see any problems? [2 marks]

e) How do you interpret the coefficient for the regression estimated in (c)? [2 marks]

f) Do you think the relationship you estimated in (c) represents a causal relationship? Why or why not? Be specific and give examples if you find that helpful in your explanation. [3 marks]

g) Merge the dataset `wap.dta` into the current dataset in memory. [2 marks]

h) Create a new variable called `fdiwap` that is equal to FDI employment divided by the working age population. [1 mark]

i) What do you notice about the ratio of FDI employment to working age population across provinces? [2 marks]

j) Estimate the following regression:  $\text{working}_{ip} = \beta_0 + \beta_1 \text{fdiwap}_p + \epsilon_{ip}$  where  $\text{working}_{ip}$  is an indicator variable for whether individual  $i$  in province  $p$  is working and  $\text{fdiwap}_p$  is the ratio of the number of workers in foreign-invested firms in province  $p$  to the working age population in province  $p$ . [2 marks]

k) How do you interpret the coefficient for the regression estimated in (j)? [2 marks]

### 4. Further regression analysis [19 marks]

- a) Suppose provinces with greater FDI employment also have better educated individuals on average and better educated individuals are more likely to work. How is that likely to effect whether your estimator is unbiased? Be specific and give a full explanation. [2 marks]
- b) Using the variable `edattand`, create the following three indicator variables [4 marks]:
- i) `primary` – the individual has completed primary education, but not a higher level [1 mark]
  - ii) `lowersec` – the individual has completed lower secondary education, but not a higher level [1 mark]
  - iii) `uppersec` – the individual has completed secondary general track, or some college, or secondary technical track, or university [2 marks]
- c) Add the newly created education indicator variables as control variables in the regression estimated in 3j). How do the results change? [2 marks]
- d) How do you interpret the coefficient on `primary`? [2 marks]
- e) Create a new variable, `agegroup`, that is a categorical variable that stores the age of the individual by 5-year age bins: the value 1 will correspond to individuals age 15 to 19, the value 2 will correspond to individuals age 20 to 24, and so on, to the value 10 will correspond to individuals age 60 to 64. [2 marks]
- f) Using a loop, estimate the regression in 4c) separately for each age group. If you are not familiar with looping, try watching this video first <https://www.youtube.com/watch?v=eNFqzNBSBF8&t=511s> [4 marks]
- g) What do you notice about how the results vary by age group? Why do you think that might be? [3 marks]

Bonus: [5 marks]

Create a plot that displays the coefficient and 95% confidence intervals for the coefficient on `fdiwap` for each age group. The graph is to be automatically created within Stata. You are not to copy the coefficient and confidence interval bound estimates into Excel or Stata and then make the graph from the manually created dataset. Your specific graph will have different values for the coefficients and confidence intervals due to different versions of the dataset, but the graph should look something like:

