ECON 2123 Introduction of Econometrics
Stata Instruction Session - Practice Questions

Instructor: Jeffrey Je-uei Kuo **Email:** jeffkuo@gwu.edu Lecture Date: March 29, April 1-2, 2019

Section I. Setting-up

1. In Windows OS, create a new folder called "StataCB123" on the desktop.
2. Right-click the folder, click "properties," copy and paste the location path.
3. GW public PC default path for the desktop is "C:\Users\GWNetID\Desktop\"
4. `clear all` and `set more off` in Stata.

Question 1 (convert .xlsx to .dta)
Download the excel file, "caschool.xslx." Using `import` function to create a dta file.

Question 2 (`cd`, `use`)
Without clicking on the Stata user interface, type in the code to load the data.

Question 3 (`log using`)
Open and save the work log, name it as "today.log" .Type `sum` or `br` in Stata command section. Find and open "today.log" file in your PC.

Question 4 (do.File and footnote)
Open up a new DO file, save it as "mydofile" and comment out your name and date.

Question 5 (`clear all`+DO file)
Copy and paste all the successful commands to your DO file, type "exit" in the Stata command. Run all previous commend in 1 second and resume to where we were. (Goal: Use Do file to run multiple lines of commands in one time.)

Section II: Browsing Dataset

Question 6 (Data Structure)
Is this dataset a cross-sectional, time-series, or panel dataset? How many "string variables" do we have? How many "float variables" do we have? How many "double variables" do we have?

Question 7 (`gsort + gsort −, sort`)
Which 3 observations have the top-three in the number of enrollment in caschool.dta ?
Which 3 observations have the low-three in the student-teacher ratio in caschool.dta ?

Question 8 (`gsort, sort`)
List top-three observations of the math score in Orange County in caschool.dta

Question 9 (destring method: `encode`) syntax formula: `encode` string variables, `gen` (new variable)
Convert the string format "county" into long format "county1."
Is it more useful for us to control the data? (i.e. index the data)

Question 10 (`rename, (ren), label var`)
Could you rename the "county1" by "countynew" ?
Could you label it as "long format of the county data" ?

Section III Summary Statistics

Question 11 (tabulate, (tab))

How many observations are there in the County of San Diego?

Among all schools observed in San Diego, how many percentage are K-6 and how many percentage are K-8?

Question 12 (tabstat, sum, sum;d)

Find the mean, standard deviation, min, median(p50), and max in testscore.

Find the mean, standard deviation, min, median(p50), and max in testscore in Orange County

Find the mean, standard deviation, min, median(p50), and max in testscore for income below average.

Question 13 (summarize, (sum) if)

Comparing the counties, does Sacramento or Los Angeles has the higher mean of math score?

Comparing the counties, does Sacramento or Los Angeles has the higher variance of math score?

Question 14(corr)

Find and interpret the correlation coefficient between income and Calworks. (vs your expectation)

Find and interpret the correlation coefficient between teachers and Calworks. (vs your expectation)

Find the variance-covariance (VC) matrix for all variables. What is the dimension of the VC matrix?

Section IV Generate New Variables and Regression

Question 15 (gen)

Generate a Z-value for test score. Z could be treated as the relative performance on the test. In Japan, the colleges use this score to evaluate the performance of the applicants. Do you think this kind of transformation will change the analytic result? Now, you should be able to find out mean and sd for the testscore.

Run the regression,

$$Z_i = \beta_0 + \beta_1 \times \text{AverageIncome}_\text{i}$$

compare the result with

$$\text{testscore}_\text{i} = \beta_0 + \beta_1 \times \text{AverageIncome}_\text{i}$$

Do you expect the hypothesis-testing result of the coefficients estimated will be different?

Question 16 (gen, replace)

Run the (level-level) OLS model, estimate and interpret $\hat{\beta}_2$

$$\text{TestScore}_\text{i} = \beta_0 + \beta_1 \text{StudentTeacherRatio}_\text{i} + \beta_2 \text{AverageIncome}_\text{i} + \mu_\text{i}$$

Run the (log-log) OLS model, estimate and interpret $\hat{\beta}_2$

$$\log(\text{TestScore})_\text{i} = \beta_0 + \beta_1 \text{StudentTeacherRatio}_\text{i} + \beta_2 \log(\text{AverageIncome})_\text{i} + \mu_\text{i}$$

Run the (higher order) OLS model, estimate and interpret $\hat{\beta}_2$ and $\hat{\beta}_3$

$$\text{TestScore}_\text{i} = \beta_0 + \beta_1 \text{StudentTeacherRatio}_\text{i} + \beta_2 \text{AverageIncome}_\text{i} + \beta_3 \text{AverageIncome}_\text{i}^2 + \mu_\text{i}$$

Question 17 (Fixed Effect Model) If we consider the fixed effects in different counties, some other variables that might not be able to see in the dataset which might also affect the test score. How do we control them? We need to add the fixed-effect variables in each counties.

```
xi: reg math_scr i.county str claw_pct meal_pct avginc, robust cluster(county)
```

Compare its results with the OLS-robustness (GLS) model

```
reg math_scr str claw_pct meal_pct avginc, robust cluster(county)
```