Introduction to Matlab & Stata

Stata Exam

Universitat de Barcelona – MSc in Economics

Instructor: Salvatore Viola

Academic Year: 2025-2026

The aim of this examination is to gauge students' understanding of some core concepts of the Stata statistical software as well as their ability to produce readable tables and figures. The code and written responses to the questions should be recorded in a **single .do file**. Students will have **one hour** to complete the exam and must submit their .do file to sviola@ub.edu before the end of the exam period at **11:00h**. Collaboration between students is not allowed and the use of the internet to search for solutions is not permitted. Students may, however, refer to their own notes and the materials provided for each class.

Question 1 (1 point)

Open Stata and create a new .do file to record all of your answers for the exam, including a comment to denote the beginning of each response. Add a comment at the beginning indicating your **name, ID number (DNI, NIE, Passport or NIUB), and the purpose of the .do file**. Then, complete the following tasks:

1. Clear everything in Stata memory.
2. Change the current working directory to a location on your device.
3. Create a path to a folder called "Output" within your working directory using the `global` command.

(Note: Task 3 will be used throughout the exam to store results. If you are unable to successfully complete this task, just save the output to the working directory.)

Question 2 (1 point)

Load the National Longitudinal Survey of Mature and Young Women (nlsw88.dta) dataset from Stata's sample datasets which contains individual demographic and occupational information. Next, respond to the following questions:

1. In which format – or type – are the variables stored/represented in the dataset?
2. What common type of variable is not present in the dataset?
3. How (i.e., using which command) could you convert variables like *race*, *married*, or *collgrad* to this other, common variable type?

Question 3 (1 point)

Produce a table containing summary statistics for the variables *age*, *race*, *married*, *grade* and *wage*. Export this table to the output folder you set up in Question 1 using one of the export methods covered in class.

(Hint: you may need to install a command in order to export the table and using quotes around the file path to save to table may not be necessary depending on which command you use)

Question 4 (2 points)

Produce the following figures in Stata:

- A histogram of hourly wages (*wage*)
- A histogram of total work experience (*ttl_exp*)
- A scatter plot of hourly wages (Y) and total work experience (X) which includes a line of best fit

Then, complete the following tasks:

1. Combine the three graphs into one figure and give the combined figure a title.
2. Save (export) the combined figure to the output folder you set up in Question 1.
3. In your .do file, briefly describe the distribution of the variables shown in the histograms as well as the observed relationship in the scatter plot.

(Hint: the `lfit` and `graph combine` commands as well as the `name()` option of the `graph` command are useful here)

Question 5 (1 point)

Create the following three variables:

- *full_time*: which takes the value 1 if a given individual works 35 or more hours a week and 0 otherwise
- *one_job*: which takes the value 1 if total experience and tenure are equal and 0 otherwise
- *ln_wage*: which is equal to the log of wages

Using the *full_time* variable you just created, drop the observations from the dataset corresponding to individuals who do **NOT** work full-time. Report the number of observations removed.

Question 6 (2 points)

Taking the following model as an example:

$$ln\_wage_i = \beta_0 + \beta_1 age_i + \beta_2 collgrad_i + \beta_3 race_i + \beta_4 married_i + \beta_5 tenure_i + u_i \quad (1)$$

Complete the following tasks:

1. Run an OLS regression in Stata based on the example model
2. Export the resulting table using one of the methods covered in class, saving it to the output folder you set up in Question 1

3. Interpret the coefficients of the *tenure* and *married* variables

4. Repeat 1-2, but only include individuals who have had more than one job in the regression (i.e. *one_job* is equal to 0)

(Hint: a special prefix for variable names is used to include categorical variables in regressions)

Question 7 (1 point)

Using the `collapse` command in Stata, collapse the current dataset in memory by the categorical *industry* variable to obtain the mean values of *age*, *wage* and *total experience*. It is expected that some variables will be dropped as a result of this action.

Question 8 (1 point)

Save the current dataset in memory as **BOTH** a Stata data file (.dta) and as an Excel spreadsheet (.xlsx) to the output folder you set up in Question 1. Make sure that both files are *overwriteable* (i.e. the code corresponding to this action can be run more than once to replace the files if they already exist and Stata does not return any errors).

SUBMIT ONLY A .DO FILE WITH YOUR WRITTEN AND CODE RESPONSES TO sviola@ub.edu BEFORE 11:00h