

# Stata Class 2

Alexander C. Lembcke

## Exercise 1

Start a new do-file and add the commands to reproduce the following exercises.

- a) Start a log-file at the beginning of your do-file, make sure you close it at the end of your do-file.
- b) Import the file "additional.csv". Save the file in Stata format and reopen the original data set ("replicate.dta").
- c) Label the variables. The variable "net" indicates whether the "Establishment has a network" and "totserver" tells you the "total servers using site file info".
- d) What are the mean and standard deviation of "net"?
- e) Create a histogram of the total servers using site file info in 2003.
- f) Use the "merge" command to add the additional variables. Save the new data set.
- g) Check whether you obtained the desired result.
- h) How many firms are in your sample (the "duplicate" command is one way to figure this out).

## Exercise 2

Keep using the same do-file and add the commands to reproduce the following exercises.

- a) Rename the management score variables to start with the same prefix, e.g. "m\_peeps".
- b) Use the "order" command to bring all management variables together.
- c) Generate a variable that is equal to the total number of hours worked in a firm.
- d) List the first 10 observations to see whether your new variable is correctly defined.
- e) Drop that variable from the sample.
- f) Create a dummy variable, i.e. a variable that takes the values 0 and 1, that is 1 if a firm has at least 1 employee that is a union member and 0 otherwise. Hint use generate and replace in combination with "if".
- g) Create a variable that is the sum of all individual management scores. Compare it to the existing variable that combines the management scores ("management"). Why do they differ? Account for the reason for the difference and check again whether scores differ.

## Exercise 3

Keep using the same do-file and add the commands to reproduce the following exercises.

- a) Create the conditional expectation function (CEF), i.e. the sample analog of  $E(Y_i|X_i)$ , where  $Y$  is log sales and  $X$  is the total servers using site file info.
- b) Create a graph that consists of a scatterplot of the above variables overlaid by the CEF.
- c) To save memory we want to replace the string variable 'analyst' with a byte variable. Create this variable without loss of information.
- d) Generate a variable that is 1 if a firms management score compared to the previous interview.
- e) Generate a ranking of firms by management score for each year and country.
- f) Generate a variable that is equal to 1.1 and summarize that data for all observations where the variable is 1.1.
- g) Generate a variable that is the sum of 'totserver' for each country and year.
- h) Generate a dummy variable for each country.