**ISF110 Lab 4 – Calculating descriptive statistics**

In this lab, we will describe our data and calculate descriptive statistics. At the same time, we will interpret the results. Follow the steps below to complete the lab:

Open Stata. Open a new ‘do’ file. Write the following command to open a Stata system-installed sample dataset on automobiles sold in the United States in 1978. We will maintain a log file for all activities and UPLOAD BOTH THE DO FILE AND THE LOG FILE to *bCourses* as part of the lab.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*LAB 4 – SUMMARY STATISTICS\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*Written by \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

clear all

set more off

capt log close

local c\_date = c(current\_date)

local study "auto"

sysuse auto.dta //This command will open the Stata-installed sample dataset “1978 Automobile Data”. sysuse is a command that loads (uses) example (system) datasets.

\*If your system does not have the dataset embedded in, use the following command to open the dataset.

use "E:\ISF\ISF 110\Spring 2022\Labs\Lab4\auto.dta", clear //Replace the file location with your file location where you saved the data.

log using "E:\ISF\ISF 110\Spring 2022\Labs\Lab2\LOG `study' -`c\_date'.log", replace //This log file MUST BE uploaded to *bCourses*.

\*First, describe your data by using the following command:

describe

\*Look at the number of observations and variable labels. You should also notice that the dataset has notes attached: (\_dta has notes). Type the following to see the notes:

notes

\*Have a feel of the data by clicking on the Data Editor (Browse) button, or by selecting Data > Data Editor > Data Editor (Browse) from the menus or by typing the command:

browse

/\*A dataset is far more than simply the data it contains. It is also information that makes the data usable by someone other than the original creator.

Although describing the data tells us something about the structure of the data, it says little about the data themselves. For this, we need summary statistics. We can get them using different commands. Try this first: \*/

codebook

/\*Note if there are missing values in any variables. Also note the total number of observations, range, mean, SD, and median of each variable. Can you calculate the interpercentile or interquartile range from the data **mpg**? Answer the following questions based on your calculations.

No. of missing values for rep78: ?

Median mpg: ?

The range between 25% and 75% for price: ?

\*To use Stata as a calculator, type: display or the shortcut:

di //For example, di 50-20

\* Another useful command for getting a quick overview of a data file is the inspect command. It also shows frequency distributions in bar graphs.

inspect

/\*To get all the summary statistics in one table, simply type sum or on the menu bar, go to Statistics > Summaries, tables, and tests > Summary and descriptive statistics > Summary statistics and clicking on the OK button.

From this simple summary, we can learn a bit about the data. First of all, the prices are nothing like today’s car prices—of course, these cars are now antiques. We can see that the gas mileages are not particularly good—this is in 1978! \*/

\*We saw before that **rep78** has some missing values. We can browse only those observations for which **rep78** is missing, so we could type:

browse if missing(rep78) // This command uses two new concepts for Stata commands—the if qualifier and the missing() function.

\*The same ends could be achieved by typing:

list make if missing(rep78)

\*We cannot impute missing values because we do not see any patterns here.

/\* We saw above that the summarize command gave brief summary statistics about all the variables. Suppose now that we became interested in the prices while summarizing the data because they seemed fantastically low (it was 1978, after all). To get an in-depth look at the price variable, we can use the menus and a dialog:

1. Select Statistics > Summaries, tables, and tests > Summary and descriptive statistics >

Summary statistics.

2. Enter or select price in the Variables field.

3. Select Display additional statistics.

4. Click on Submit.

Alternatively, typing summarize price, detail will get the same result. The portion after the comma contains options for Stata commands; hence, detail is an example of an option. \*/

/\* From the output, answer the following questions:

The median price of the cars in the dataset is:

The four most expensive cars are all priced between:

If we wished to browse the most expensive, we could start by clicking on the Data Editor (Browse) button. Once the Data Editor is open, we can click on the Filter Observations button, to bring up the Filter Observations dialog. We can look at the expensive cars by putting price > 13000 in the ‘Filter by expression’ field.

Pressing the Apply Filter, list the four most expensive cars in 1978.

Comment on their gas mileage and repair records. \*/

\*We now decide to turn our attention to foreign cars, mileage and repairs.

\*Let’s start by looking at the proportion of foreign cars in the dataset along with the proportion of cars with each type of mileage and repair record. Type:

tab foreign

\*What percentage of cars is foreign made? What percentage of cars is domestic?

tab mpg

tab rep78

\*Note: 1 means a poor repair record and 5 means a good one.

\* These one-way tables do not help us compare the mileage and repair records of foreign and domestic cars. A two-way table would help greatly. Type:

tab rep78 foreign, row

tab rep78 foreign, col //See the difference with row percentages

tab mpg foreign, row

\*The tables are long and difficult to compare. Use:

sum rep78 if foreign==0

sum rep78 if foreign==1

sum mpg if foreign==0

sum mpg if foreign==1

\*Note: We can reduce the lines of Stata command by writing:

by foreign, sort : summarize mpg

by foreign, sort : summarize rep78

\*Compare foreign cars with domestic cars in terms of repair records and mpg. WRITE A FEW SENTENCES HERE.

\*Now, we will do cross tabulation to produce some formatted tables. Type:

tab foreign, summarize(mpg)

tab foreign, summarize(rep78)

tab rep78 foreign

tab rep78 foreign, column nofreq

\*Finally, we will see some correlations. Type:

cor

\*Interpret the correlation between **weight** and **mpg**, and **foreign** and **rep78**. What do the results say, when investigating collinearity among predictor variables for **price**?

\*End of Lab\*

log close

exit, clear