Lab Assignments

ECON 103, SJSU, Spring 2015

Lab Assignment, Week One

(Due to Canvas by 11:59p.m. on Thursday, 1/22/2015)

To complete this assignment, read Chapter 1, Section 3 of Stock and Watson's textbook. Then, find the data online that is discussed in Chapter 1. You can find these data on the publisher's Student Resources page for our textbook (http://wps.aw.com/aw_stock_ie_3).[1] From the main menu, choose Replication Files > and then click the second option, "Replication Files for Empirical Results (Original Edition)". Please familiarize yourself with this webpage; you will be visiting it often this semester.

After you access the data and complete the relevant reading, please prepare your answers using a word processing program of your choice, and then upload your answers to Canvas in either PDF or DOC format. Here are the questions:

1. Is the unit of observation of the data in Table 1.1 an individual student, a school district or a state?
2. Is the data in Table 1.1 cross sectional, time series, or a panel?
3. If you sort the data by average test score, what happens to the variable "Observation Number"? Why does this happen? Is there anything you can do to prevent this from happening?
4. What is the highest value of TESTSCR in the sample? The lowest?
5. Is the data in Table 1.2 cross sectional, time series, or a panel?
6. What was the average quarterly unemployment rate in the U.S. in the 1960s? What was the standard deviation of quarterly unemployment in the 1960s? (*Hint: type =AVERAGE(and =STDEV(into your spreadsheet cell to get started, and search your program's help file if you are having trouble with the syntax. For the statistics, see Angrist and Pischke, page 36, for a refresher on standard deviation*.)
7. Is the data in Table 1.3 cross sectional, time series, or a panel?
8. What was the average number of cigarettes consumed per person in the state of Alabama from 1985 to 1995? In California?
9. Which spreadsheet program did you use to complete this assignment?
10. What type of access do you have to a computing device at home?

[1] http://wps.aw.com/wps/media/objects/11422/11696965/replicationfiles3e/replicationfiles/ch1/Table11.xlsx
http://wps.aw.com/wps/media/objects/11422/11696965/replicationfiles3e/replicationfiles/ch1/Table12.xlsx
http://wps.aw.com/wps/media/objects/11422/11696965/replicationfiles3e/replicationfiles/ch1/Table13.xlsx

ECON 103, SJSU, Spring 2015

Lab Assignment, Week Two

(Due to Canvas by 11:59p.m. on Thursday, 1/29/2015)

In this assignment, you will prepare a software file which consists of variables that can be used to calculate statistics (e.g. means and standard deviations) using a statistical software programs (such as Stata or R), and then carry out hypothesis tests with the data. Please prepare your answers using a word processing program of your choice, and then upload your answers to Canvas in either PDF or DOC format.

To begin, download the file containing some of the data that was collected during our first meeting here: http://www.sjsu.edu/faculty/matthew.holian/misc/Data_Econ_103_SP15.xls You'll find it in the worksheet titled "syllabus_quiz". As a reminder, half of the students had a quiz where the last two questions were: (9.) Is the population of Istanbul greater or less than 1.4 million? _____ and (10.)What do you think is the population of Istanbul?

Meanwhile, the other half of the students had a question where 1.4 million was replaced by 14 million in question 9, but was otherwise identical in all respects. Given the best estimate of the population of Istanbul is much closer to 14 million, we could describe students in the former group as being in the "treatment" group (because we tried to manipulate this group), and the students in the latter group as being in the "control" group (because we told them the truth.)

Questions:

1. Open the data file in a spreadsheet program and generate an dummy variable named TREATMENT that is equal to one if the student is in the treatment group, and zero if they are in the control group, using two methods. In the first method, use the "sort" and "drag and drop" features to generate the variable "by hand". Next, generate this dummy variable using the "=if()" syntax, along with drag and drop. Briefly describe each method (to prove that you did it both ways) and state which method you prefer in this situation *and why*.

2. The data file contains a variable named "q9". This variable consists of written responses indicating how the data entry technician interpreted the student's answer. These words are referred to as "strings" in statistical computing. Generate a new variable named "GREATER" that is numeric (i.e. not a string) and is equal to one if the student indicated "greater", zero if the student answered "less" and equal to "." if the student either provided the answer in an incorrect format or did not answer the question. (STATA will interpret numeric variables with "." values as missing values. What do you think would be another reasonable way to indicate a missing value in a spreadsheet program?
3. Generate a variable named "ESTIMATE" which is equal to the student's answer to question 10. The quiz question attempted to illicit the student's estimate of the

population of Istanbul. As you will see, many students did not follow directions as intended. Generate a variable equal to the value reported by the student for answers in correct format, and equal to "." for answers in incorrect format. Why do you think there were so many answers in incorrect format? What can be done differently in the future to mitigate this problem?

4. After completing questions 1-3, you will have added three new variables to the spreadsheet file. Save your spreadsheet as an XLS file under a new name. Why is it a good idea to change the file name after modifying a data file?
5. Calculate the sample average and sample standard deviation of the three variables TREATMENT, GREATER and ESTIMATE. Then, calculate the sample average and sample standard deviations of ESTIMATE for the treatment and control subgroups.
6. Write out a mathematical equation for the sample standard deviation, and provide a reference to this equation in your textbook. Describe the sample standard deviation in words, and note how it differs from the population standard deviation.
7. Describe in detail the steps involved in importing data that is in XLS format into a statistical software program (such as Stata, R, Gretl or SPSS).
8. Describe one method for using a statistical software program (such as Stata, R, Gretl or SPSS) for generating the variable TREATMENT. Either report the syntax if you are using the program's command line interface, or describe in detail how to navigate the graphical user interface if that is what you're using. Is it easier to use the spreadsheet or the statistical package to generate this variable?
9. As in the previous question, describe one method for using a statistical software program for generating the variable ESTIMATE. Is it easier to use the spreadsheet or the statistical package to generate this variable?
10. Someone suggested that the average value of ESTIMATE in the sample should be the midpoint between 1.4 and 14 (i.e. 15.4/2, or 7.7). Conduct a test of the hypothesis that the population mean of ESTIMATE is 7.7.
11. Why do you think Angrist and Pischke write on p. 43 that "One…is the loneliest number that you'll ever do."?
12. Using the sample and subsample averages and standard deviations you reported in question 5, conduct a hypothesis test that there is a statistically significant difference between the means for the treatment and control groups. Calculate the standard errors using a spreadsheet program.
13. Describe how to use a statistical software package to carry out the difference in means test of the previous question.
14. Discuss practical versus statistical significance in the context of the difference in means test of question 12.
15. Describe one other hypothesis test that would be interesting to carry out using data generated from the syllabus quiz.

ECON 103, SJSU, Spring 2015
Lab Assignment, Week Three
(Due to Canvas by 11:59p.m. on Thursday, 2/5/2015)

In this assignment, you will work with a software file that has already been prepared for analysis. (This is in contrast to last week's assignment, where you had to prepare a software file for analysis.) Read section 3.4 of Stock and Watson, including the case study, "The Gender Gap of Earnings of College Graduates in the United States." You can find the data used there on the publisher's website but for this assignment, please use the modified version of this data, in the worksheet titled "modified_CPS08": www.sjsu.edu/people/matthew.holian/docs/Data_Econ_103_SP15.xls Please do, however, download *and read* the "Data Description" file (DOCX format) for the original data file (http://wps.aw.com/wps/media/objects/11422/11696965/datasets3e/datasets/cps_ch3.docx) and also the description of the modified data in the "NOTES" worksheet in the XLS file linked above. Type your answers using a word processing program and upload a DOC or PDF file.

1. Which bureau or department of the U.S. government collected this data?
2. What is the unit of observation?
3. How many variables are in the data file and what do each represent?
4. Would you describe this data as a panel, or does it consist of repeated cross-sections?
5. Use a software package of your choice to calculate sample average earnings, standard deviation of earnings, and the observation counts for males and females for all years ($\bar{Y}_M^{1992}$, $\bar{Y}_F^{1992}$, s_M^{1992}, s_F^{1992}, n_M^{1992}, n_F^{1992}, $\bar{Y}_M^{1996}$, …) Describe how you did this and report each value. (Hint: this can be most easily accomplished using the command line interface in Stata or R. It is possible to accomplish this using a spreadsheet program, but it will very time consuming!)
6. Describe the difference between a command line and graphical user interface.
7. Using the statistics reported in Table 3.1, calculate and report test-statistics for difference in means tests, using SW's Equations 3.19 & 3.20 (for d_0=0), for each year. Which other standard error formula might you use here? Which do you prefer and why? What are the p-values associated with each test statistic, for a two-tailed test?
8. Using the statistics reported in your answers to question 5, calculate and report test-statistics and p-values for difference in means tests for each year, as above.
9. What distribution did you use to calculate p-values above? Which other distribution might one use? Which do you prefer and why?
10. Using all years of data, calculate the average hourly earnings, the standard deviation of hourly earnings, and the number of observations, without regard to gender. In other words, calculate the average and standard deviation of earnings, using the entire sample. Use these statistics to test the null hypothesis that average hourly earnings were $100 per hour over this period. What is the value of your test statistic? Can you reject the null? Is this a one-tailed test or a two-tailed test? What is the p-value?
11. Again using all years of data, calculate the average hourly earnings, the standard deviation of average hourly earnings, and the number of observations, for two groups--males and females. Report these numbers, and use them to calculate a test statistic for a difference in means test of the null hypothesis that there is no difference between the average hourly earnings of men and women. Report all statistics you calculate. Interpret the test statistic you calculated. Is the difference statistically significant? Is the difference economically significant? Why or why not?
12. Describe the Stata resources provided by the textbook publisher that help you complete this assignment. (Hint: look under "Replication Files for Empirical Results (Original Edition)") What is a do file? What is a log file? How does the do file provided by Professor Holian differ from the one provided by the publisher?

ECON 103, SJSU, Spring 2015
Lab Assignment, Week Four
(Due to Canvas by 11:59p.m. on Thursday, 2/12/2015)

Access the course data: www.sjsu.edu/people/matthew.holian/docs/Data_Econ_103_SP15.xls
For this assignment, you will use data in the worksheet titled "modified_CPS12" and also "caschool". You can find links to descriptions of these data in the worksheet titled "NOTES". Please type your answers using a word processing program and upload a DOC or PDF file.

1. Fill in the missing information in "Table 4.0: Data Description" that appears below:

Table 4.0: Variable Descriptions

Variable	Description
testscr	
str	
avginc	

2. Create a table named "Table 4.1a" that contains summary statics, including the number of observations, sample average, sample standard deviation, and minimum and maximum value. (Hint: Stata's summarize command will return all necessary info. After running the command, highlight the results, right-click, and select "Copy Table". Then paste the results into a spreadsheet. In Excel you can add borders by right click-> Format Cells -> Borders.)

Table 4.1: Variable Descriptions

Variable	Obs	Mean	Std. Dev.	Min	Max
testscr					
Str					
Avginc					

3. Using a spreadsheet file, produce a scatterplot like Figure 4.3 "The Estimated Regression Line for the California Data" but include the R2 as well as the regression equation and line. (Hint: In Excel, arrange the data so the two variables are next to each other, with str first. Then highlight the two columns of data, click the Insert tab, and then the Scatter icon…choose the first option that appears. The scatterplot will appear. Please your cursor over the dots, right click and select "add trendline". Remember to select the options to show the equation and R2 value. You can make the figure look better if you place the cursor over the x-axis values, right click and select "format axis" and then change the minimum value of the x-axis from "auto" to "fixed" and set the value at 10.)

4. Run a regression of TESTSCR on STR in a statistical software package and confirm you get the same results. (Hint: regress testscr str; or, see SW Ch 4 lecture slides, slide 14)

5. Use a spreadsheet program to calculate predicted values of TESTSCR (that's "TESTSCR-hat") for each district.

6. Use a spreadsheet program to calculate the ESS (see SW equation 4.14)

7. Use a spreadsheet program to calculate the TSS (see SW equation 4.15)

8. What is the value of ESS / TSS ?

9. Does district student teacher ratio account for a large fraction of the variance of district average test scores?

10. Use a spreadsheet program to calculate the SSR (see SW equation 4.17)

11. What is the value of 1 – (SSR/TSS)?

12. What is the value of SSR / (n-2)?

The worksheet "modified_cps12" contains data for full-time, full-year workers, age 25–34, with a high school diploma or B.A./B.S. as their highest degree. See the "NOTES" worksheet for a detailed description . In this exercise, you will investigate the relationship between a worker's age and earnings. (Generally, older workers have more job experience, leading to higher productivity and earnings.) You could

13. Run a regression of average hourly earnings (*AHE*) on age (*Age*). What is the estimated intercept? What is the estimated slope? Use the estimated regression to answer this question: How much do earnings increase as workers age by 1 year?

14. Bob is a 26-year-old worker. Predict Bob's earnings using the estimated regression. Alexis is a 30-year-old worker. Predict Alexis's earnings using the estimated regression.

15. Does age account for a large fraction of the variance in earnings across individuals? Explain.

ECON 103, SJSU, Spring 2015
Lab Assignment, Week Five
(Due to Canvas by 11:59p.m. on Thursday, 2/19/2015)

This week, you'll be solving some of the Additional Empirical Exercises distributed through our textbook publisher's website. For question 1, there is a slight modification, which is that you'll use a different version of the data than is indicated in the problem. For problem two, please download the data from the publisher's webpage. You'll also be revisiting a causal question we have previously examined; Data for Problem 3 can be found on the course data file.

1. Please complete the problem outlined here:

   http://wps.aw.com/wps/media/objects/11422/11696965/aee/Stock_Watson_3U_AEE_5_1.pdf

However, please make the following modification to this problem. Instead of using the full CPS12 data, use the same modified version you used last week. You can access the course data here: www.sjsu.edu/people/matthew.holian/docs/Data_Econ_103_SP15.xls and find the data under the worksheet titled "modified_CPS12"

2. Please complete the problem outlined at the link below, using data described therein:

   http://wps.aw.com/wps/media/objects/11422/11696965/aee/Stock_Watson_3U_AEE_5_2.pdf

3. Recall Question 12 from your Week 2 Lab Assignment:

   12. Using the sample and subsample averages and standard deviations you reported in question 5, conduct a hypothesis test that there is a statistically significant difference between the means for the treatment and control groups. Calculate the standard errors using a spreadsheet program.

Test the null hypothesis that there is no effect of priming in a survey asking students to estimate the population of Istanbul, using a regression framework. (Hint: Regress average hourly earnings on a dummy treatment variable, recover the estimated coefficient on the dummy variable and its estimated standard error, and carry out a "t-test".) Show your regression output (The formatting doesn't have to be perfect but please make it clear) and interpret the results of the hypothesis test in terms of statistical and practical significance. Discuss the options available to an analyst regarding deciding how to estimate standard errors; name and describe two ways of estimating standard errors for a standard difference in means test, and two ways of estimating standard errors in a regression framework. Which method do Stock and Watson seem to suggest is preferable for difference in means tests at the end of Section 3.6? Provide a quote and page number. Which practical advice on deciding do they provide at the end of Section 5.4? Again, provide a quote and page number.

ECON 103, SJSU, Spring 2015
Lab Assignment, Week Six
(Due to Canvas by 11:59p.m. on Friday, 2/27/2015)

Please complete ***parts a, b and d*** of the Empirical Exercise that you can find at the following link (you can skip part c):

http://wps.aw.com/wps/media/objects/11422/11696965/aee/Stock_Watson_3U_AEE_6_1.pdf

In addition, please prepare a document that lists each question you answered incorrectly on the midterm, briefly summarize the question and the correct answer, and then write at least one complete sentence that explains WHY the correct answer is correct. The goal of this additional aspect of this week's Lab Assignment is for students to review their answers and to learn from their mistakes. Requiring students to submit a once sentence explanation for each question answered incorrectly will help students learn, and will also allow me to verify you attempted to understand why some of your answers were wrong.

ECON 103, SJSU, Spring 2015
Lab Assignment, Week Seven
(Due to Canvas by 11:59p.m. on Thursday, 3/5/2015)

This question builds off of Empirical Exercises 4.3, 6.3 and 7.4

On the text Web site you will find a data file Growth that contains data on average growth rates from 1960 through 1995 for 65 countries along with variables that are potentially related to growth. A detailed description is given in the Growth_Description also available on the Web site. In this exercise, you will investigate the relationship between growth and trade.

a. Create a table of variable descriptions for the data set; for this assignment you are allowed to copy and paste the descriptions written by another author, if you cite the source.
b. Create a table of summary statistics for the data set.
c. Create a scatterplot of average annual growth rate (Growth) on the average trade share (TradeShare). Does there appear to be a relationship between the variables?
d. One country, Malta, has a trade share much larger than the other countries. Find Malta on the scatterplot. Does Malta look like an outlier?
e. Using all observations, run a regression of Growth on TradeShare. What is the estimated slope? What is the estimated interept? Use the regression to predict the growth rate for a country with a trade share of 0.5 and with a trade share equal to 1.0.
f. Estimate the same regression excluding the data from Malta. Answer the same questions in (e).
g. Where is Malta? Why is the Malta trade share so large? Should Malta be included or excluded from the analysis?
h. Run a regression of Growth on Trade Share, YearsSchool, Rev_Coups, Assassinations and RGDP60. What is the value of the coefficient on Rev_Coups? Interpret the value of this coefficient. Is it large or small in a real-world sense?
i. In the regression from part (h), construct a 95% confidence interval for the coefficient on TradeShare. Is the coefficient statistically significant at the 5% level?
j. Use the regression in (h) to predict the average annual growth rate for a country that has average values for all regressors.
k. Repeat (i) but now assume that the country's value for TradeShare is one standard deviation above the mean.
l. Why is Oil omitted from the regression? What would happen if it were included?

ECON 103, SJSU, Spring 2015
Lab Assignment, Week Eight
(Due to Canvas by 11:59p.m. on Thursday, 3/12/2015)

E.E. 7.1 (using modified CPS12)

http://wps.aw.com/wps/media/objects/11422/11696965/aee/Stock_Watson_3U_AEE_7_1.pdf

E.E. 7.4, (using Growth data; part B only)

http://wps.aw.com/wps/media/objects/11422/11696965/aee/Stock_Watson_3U_AEE_7_4.pdf

ECON 103, SJSU, Spring 2015
Lab Assignment, Week Nine
(Due to Canvas by 11:59p.m. on Thursday, 3/19/2015)

Note: We have a midterm earlier in the day when this assignment is due. Even though I have kept the deadline the same time as always (11:59p.m.) you should submit this BEFORE the midterm, because completing the assignment will help you prepare.

EE 8.1 (using modified CPS12…this may be numbered differently in your book…make sure it is the assignment involving the CPS data on earnings.)

See book (not available online)

EE 8.4 (from book; or see AEE 8.3, linked to below)

http://wps.aw.com/wps/media/objects/11422/11696965/aee/Stock_Watson_3U_AEE_8_3.pdf