ECO 380 Economic Statistics and Data Analysis
Problem Set 1
Due September 16, 2019 (beginning of class)

Instruction: The problem sets are designed to be challenging (especially if you are new to data analysis) and very time-intensive, so plan ahead. In general, the problem sets consist of both solving theoretical problems, and analyzing and interpreting real data. You may discuss the questions with your classmates, but you are required to hand in your own independently written solutions, do-files, and log-files. No late work will be accepted and I do NOT accept any electronic copy. Please do not email me your assignments as you will not receive any credit. All the data necessary for the problem set is available under UBlearns.

Important: It is extremely important to write a clean well-commented program for transparency and replication purposes in *all* empirical work. You should always be able to reproduce your result from raw data to support your claim.

There are 3 items to hand in: (1) Typed write-up (i.e., word-file) answering the assigned questions, reporting your results, and interpreting your findings; if the question asks for graphs or tables, these must be in the word-file in an organized manner with your interpretation, (2) do-file (i.e., program text-file), and (3) log-file (i.e., output text-file that shows the results). You MUST use Stata. For questions involving data analysis, you will NOT get any credit if you do not provide a program code and the output. You may not use Excel. Do not submit any undigested log-file that contains errors.

1. [Empirical Exercise] (10 points) What types of jobs are available for students who graduate with a business degree? The website careerbuilder.com lists job opportunities classified in a variety of ways. A recent posting had 25,120 jobs. BUSJOBS data on UBlearns show types of jobs and the numbers of postings listed under the classification "business administration" on a recent day. Describe these data using the methods you learned in Chapter 1, and write a short summary about jobs that are available for those who have a business degree. Include comments on the limitations that should be kept in mind when interpreting this particular set of data.

2. [Empirical Exercise] (40 points) This exercise focuses on data management using a dataset that is downloaded directly from an original source. First, you will learn how to download a dataset from the Bureau of Labor Statistics (BLS). We are going to use the aggregate Current Population Survey data prepared by the BLS to compute unemployment rate by different demographic groups. The data is available on https://www.bls.gov/data/ under the unemployment selection. Select "Top Picks" of Labor Force Statistics including the National Unemployment Rate (Current Population Survey - CPS). Select overall unemployment rate as well as unemployment rates by gender, race/ethnicity, and education. Select "Retrieve data." Then you will use formatting option to have the data available in "column format" and all years, and all time periods. (If you do not select this format, you will have a coding challenging to set up the data appropriate for a time-series analysis. You will download 11 Excel files from the website. Second, you need to learn how to merge the datasets. Finally, plot four sets of graphs: (1) overall unemployment rate over time since 1948, (2) unemployment rate over time by gender, (3) unemployment rate over time by race/ethnicity, and (4) unemployment rate over time by education. Describe your findings in words (max ½ page). Hint: (1) Convert the Excel file to csv file. (2) Use Stata's `merge` command to merge all 11 files into one single file. (3) Sort the data by year and month, then create a time variable to use `tsset` command. (4) Use `tsline` command.

3. [Empirical Exercise] (10 points) Use EDUSEV to answer the following questions.

    a. Which variables are categorical? (1 point)
    b. How many percent is female? (1 point)
    c. Make a suitable graph that describes the shape, center, and spread of the distribution of students' IQ scores. (2 points)
    d. In general, IQ scores are usually said to be centered at 100. Is this true for this data? (2 point)
    e. Make a suitable graph that describes the shape, center, and spread of the distribution of self-concept scores. (2 points)
    f. Can you identify any suspected outliers? Why? (2 point)

4. [Empirical Exercise] (10 points) Use TALK to answer the following questions. People often generalize that women are more talkative than men. Is this supported by data? One study designed to examine this stereotype collected data on the speech of 42 women and 37 men in the U.S.

    a. Calculate the mean and standard deviation of number of words spoken per day by gender. Report the results by gender. (2 points)
    b. Use the 68-95-99.7 rule to describe the distribution by gender. Report the results. (4 points)
    c. Describe the skewness of the distribution by gender. Support your statement by constructing an appropriate graph of your choice. (2 points)
    d. Do you think that applying the rule in this situation is reasonable? Do you think that the data support the generalization that women are more talkative than men? Explain your answer. (2 points)

5. [Empirical Exercise] Use COLLEGE to answer the following questions. (Total 20 points)

    a. Report the basic descriptive statistics of all the variables that is contained in the dataset (i.e., mean, standard deviation, and median). (2 points)
    b. Make a scatterplot of undergraduate population and population <u>with</u> the least-squares regression line. (4 point) [Hint: explore 'lfit' command]
    c. Focus on California, the states with the largest population. Is this state an outlier when you consider only the distribution of population? Why? (2 points)
    d. Is California an outlier when viewed in terms of the relationship between number of undergraduate college students and population? Why? (2 points)
    e. Repeat (c) and (d) using the logs of both variables (4 points)
    f. Delete four largest states and run your own regression and report the results. (2 points)
    g. What is the equation of your least-squares regression line? (2 point)
    h. Interpret the value of r^2 from your regression. (2 point)

6. [Empirical Exercise] (10 points) The table below is adopted from Stock and Watson (2015) that summarizes average hourly earnings by education and gender. Using MARCHCPS2013 replicate the results reported in the table. In other words, use the data to see if you can independently produce the numbers reported in the table, then report your estimates in a table.

Description of Variables

- a_age = age
- a_hga = education level
  0 "Children"
  31 "Less Than 1st Grade"
  32 "1st,2nd,3rd,or 4th grade"
  33 "5th Or 6th Grade"
  34 "7th and 8th grade"
  35 "9th Grade"
  36 "10th Grade"
  37 "11th Grade"
  38 "12th Grade No Diploma"
  39 "High school graduate-high school diploma"
  40 "Some College But No Degree"
  41 "Assc degree-occupation/vocation"
  42 "Assc degree-academic program"
  43 "Bachelor"s degree (BA,AB,BS)"
  44 "Master"s degree (MA,MS,MENG,MED,MSW,MBA)"
  45 "Professional school degree (MD,DDS,DVM,L"
  46 "Doctorate degree (PHD,EDD)"

- a_sex = gender
  1 male;
  2 female;

Table 2.4 Summaries of the Conditional Distribution of Average Hourly Earnings of U.S. Full-Time Workers in 2012 Given Education Level and Gender

			Percentile			
	Mean	Standard Deviation	25%	50% (median)	75%	90%
(a) Women with high school diploma	$15.49	$8.42	$10.10	$14.03	$18.75	$24.52
(b) Women with four-year college degree	25.42	13.81	16.15	22.44	31.34	43.27
(c) Men with high school diploma	20.25	11.00	12.92	17.86	24.83	33.78
(d) Men with four-year college degree	32.73	18.11	19.61	28.85	41.68	57.30

Average hourly earnings are the sum of annual pretax wages, salaries, tips, and bonuses divided by the number of hours worked annually.