Assignment 1

The goal of this assignment is to familiarize you with three fundamental tasks in Stata: (1) Loading data, programming commands, and basic data manipulation; (2) Reshaping a dataset; and (3) Merging datasets. Your best resource for help will be the Internet.

You should turn in three things: **(1) a** **typed sheet of answers** to each of the questions below, **(2) your** **annotated .do file(s)** to perform the operations and answer the questions, and **(3) the .log file(s)** of Stata output.

1. The file **census13.dta** contains information on marriage and divorce. Load the file into Stata (Hint: the file is already in Stata format).
   1. What are the units of observation?
   2. How many variables are in the dataset?
   3. Summarize the variables
      1. What is the average birthrate in the sample?
      2. What is the maximum divorce rate?
      3. What is the standard deviation of median age?
   4. Tabulate the census divisions by region.
   5. Create a table of the mean birthrate, marriage rate, and divorce rate by division.
      1. What division is the outlier in marriage rates?
      2. What is the reason? (Hint: Make a table of marriage rate by state)
   6. Create a variable that is the ratio of the marriage rate to divorce rate. Generate, label, and save a histogram of this new variable.
2. The file **wage\_educ.dct** contains education and salary data for respondents of the National Longitudinal Study of Youth 1997 (NLSY97). The data are in a dictionary file (.dct), which is a text format. **wage\_educ.cdb** is the codebook. Import the data into Stata (Hint: use the *infile* command).
   1. Generate new names for all of the variables (Hint: rename “T3602100” as “region”).
   2. Create a variable named “year” and set it equal to the year before the respondents were surveyed about their income and education.
   3. How many respondents did not provide answers to the questions about their father’s and mother’s highest grades completed, respectively? (Hint: use the codebook)
   4. Create a table of the mean and standard deviation of income by highest degree ever received. Which degree level has the largest variance in income? Does this make sense?
   5. Save the dataset as a Stata file.
3. The file **unemp\_rate.csv** contains historic information on unemployment rates by region. The data are in a *wide* format. The goal of this problem is to reshape and merge the data with the previous dataset. Read the data into Stata.
   1. Reshape the data so that it is *long*. Each row should contain a region-year observation of the unemployment rate.
   2. To merge these data with the dataset in question 2., you need a numeric variable that identifies the region in both datasets *which has the same name and is coded the same way*.

Using the NLSY97 codebook, rename (if required) and recode “region” so that it is coded the same way as the region variable you renamed in 2.a. in the NLSY97 (e.g. 1 = “Northeast”).

* 1. Merge the NLSY97 data from question 2. into the dataset on “region” and “year.” To check whether the merge did what you expect, tabulate “\_merge.”
     1. What does 3 represent?
     2. What do 1 and 2 represent? Do the totals make sense? For the NLSY97 respondents for which “\_merge” = 2, why?
     3. Keep only the observations matched in both datasets.
  2. Create a table displaying the means of the unemployment rates and income by region.