**EPI 510 -- Stata assignment 1**

For this assignment, please submit two files: (1) **a file containing text and tables** (.doc, .docx, .pdf or .txt) answering questions posed in the assignment, and (2) **a Stata do file** that performs the requested operation (.do).

**50 points possible**

Make sure that you include comments using at least 3 different types of comment delimiters in your code **(2 points for each type of delimiter you use for a maximum of 6 points)**, and whitespace **(2 points)** in your code

1. Let’s setup our session:
2. Change the working directory to be your Epi 510 Stata directory (e.g. C:/epi510/stata) **(2 points)**
3. Open the vipcls.dta data file in Stata **(2 points)**.   
     
   Note that a description of the data and data dictionary are given on the course website (<https://canvas.uw.edu/courses/1478728/pages/r-data-and-program-files-for-download>). Read through this!
4. Let’s look at the dataset to see what we have, look at variable types, and ensure that variables are saved in the most efficient format:
   1. describe the dataset. How many variables and observations are in the dataset, and what is the storage type of the variables? **(2 points)**
      * Variables: 23
      * Observations: 13,285
      * Storage type of variables: double
   2. We can use Stata’s commandmemory(not covered in class) to see the size of the dataset in bytes. Type memory and look in the first row and column to see the amount of memory being used for data. That tells us the size of the dataset. How large is the dataset? **(2 points)**
      * The dataset is 2,444,440 bytes.
   3. compress the dataset. **(2 points)**
   4. Run the memory command again. What is the size of the dataset now? **(2 points)**
      * 398,550 bytes
   5. describe the dataset again. What are the storage types of the variables now? **(2 points)**
      * The data is stored in long, byte, and int formats.
   6. Based on the storage types of the compressed dataset, do you think that the variables in this dataset are strings, integers, or continuous variables? **(2 points)**
5. Now, let’s examine the data:
   1. Browse the dataset to visually inspect the data. **(2 points)**
   2. Use the misstable summarize command (with the all option) to determine the number of missing observations for each variable, and to see the range of values for each variable. **(2 points)**
      * For which variables are there missing values? How many observations are missing for these variables? **(2 points)**
      * Look at the range of values for each variable. You should see a value that is impossible for most of the variables. What is that value? **(2 points)**
6. Clean the data:
   1. Set -1 values to missing for each variable (note: we’ll look at a faster way to do this in a couple of weeks – just do the slow way now and write a line of code for each variable) **(6 points)**
   2. Set mother’s age (momage) of less than 16 to missing **(2 points)**
   3. Set gestational age (delges) above 52 to missing **(2 points)**
   4. Set birth weight of babies (bw) that are either less than 300 or greater than 6000 to missing **(2 points)**
7. Use Stata commands tabulate, summarize, and tabstat as appropriate to answer the following questions:
   1. What percentage of mothers drank alcohol exactly one time per week during their second trimester, excluding those with missing values of this variable? **(2 points)**
   2. What are the median and 95th percentile of years of education? **(2 points)**
   3. What are the mean, standard deviation, and sample sizes for mother’s age in each of the five marital groups? **(2 points)**

BONUS QUESTION: What are the patient IDs of the 5 subjects whose children had the smallest birthweights? **(5 points extra credit)**