NAME:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**STATA Assignment #4**

In this exercise, you will use data from the General Social Survey (GSS). The GSS interviews a random cross-section of about 3,000 Americans every other year. For an introduction to the GSS, go to its website ([here](http://www3.norc.org/gss%2Bwebsite/)).

For this exercise, you will analyze GSS data … specifically the variables [EDUC](http://sda.berkeley.edu/D3/GSS10/Doc/gs100002.htm#EDUC) and [WORDSUM](http://sda.berkeley.edu/D3/GSS10/Doc/gs100003.htm#WORDSUM). EDUC measures the total numbers of years of education that people have completed. WORDSUM is people’s score on a 10 item vocabulary test (where 10 is a perfect score).

We will treat both variables as though they are continuous variables in this example. Be sure to read the on-line documentation well enough to understand what survey questions were asked to generate these variables and to ascertain value labels and missing data codes. (Note that “IAP” means “inapplicable to this person,” “DK” means “don’t know,” and “NA” means “not applicable.”)

For this exercise, I have constructed a data file (“STATA Assignment 4.dat”) that has these two variables. EDUC is in columns 6 through 7 and WORDSUM is in columns 8 through 9.

Use STATA syntax files that you already have (from earlier assignments or from class examples) and modify them to accomplish the following goals. When you are done, type or paste your answers for questions #2 through #11 below into a word processor (e.g., Microsoft Word) and turn in a paper version of the assignment.

**PART I.**

1. Read the data file into STATA
2. Be sure to declare missing data codes to be missing
3. Select observations for which there are valid values on both EDUC and WORDSUM
4. Report the mean and variance of both EDUC and WORDSUM among people who have valid values on both variables
5. Report the correlation between EDUC and WORDSUM among people who have valid values on both variables
6. Carry out a regression of WORDSUM (as the Y variable) on EDUC (as the X variable) among people who have valid values on both variables

**PART II.** Using the information you obtained from #4 and #5 above about the means and variances of EDUC and WORDSUM, the correlation between EDUC and WORDSUM, and the number of observations with valid values on both EDUC and WORDSUM, do the following by hand (showing your work):

1. Compute the least squares prediction equation
2. Compute R2YX
3. Test the null hypothesis that 2YX equals zero in the population
4. Test the null hypothesis that YX equals zero in the population
5. Test the null hypothesis that YX equals zero in the population

Use =0.05 for the hypothesis tests in #3 through #5. Do #1 through #5 by hand, and show your work below. Confirm that your answers for #1 through #5 match output from STATA.

**HOW TO DO CORRELATION AND REGRESSION IN STATA**

Imagine that X and Y are continuous variables. How do you generate the correlation between X and Y? How do you regress Y on X? Below I provide STATA syntax that should work; the STATA syntax assumes you use a semi-colon after commands. You will need to modify this code to suit your purposes.

*STATA*

correlate X Y ;

regress Y X;