STATA Assignment

Due June 19th, by 11:15 AM

Download the Sachs and Warner data set from the class homepage[1]. Follow the steps outlined below. Use a log file to record your results, and print out all STATA output. Note that for most problems you need to provide a short answer or explanation, I would suggest after completing your log file you copy it into a text or Word document (with nice font so the result come out) then type your answers after you have preformed the command. (Note it is a comma-delimited file).

1)

a. Get summary statistics of all the data and become familiar with what this data set contains. A codebook has been provided to explain what the variables are.

b. Drop lgdpea65, openy, and life2.

c. Rename those variables that do not make sense to you (for the assignment, rename at least 3 variables).

2)

a. The life expectancy variable is in natural logs, generate a new variable that has life expectancy in common form (take the exponential of it).

b. Drop the original logged life expectancy value (LIFE) and then drop the missing values in your new life expectancy variable.

c. Look at the summary statistics and create a histogram of life expectancy (be sure to title all of your graphs). What do you notice?

d. List the countries that have a life expectance higher than 69 years old and then those countries that have a life expectancy less than 40 years old. Are there any surprises in the group?

e. Is the country with the highest life expectancy an outlier? – use a z-score

f. Pick another variable that interests you and find interesting things about it.

---

[1] This data set is basically the data and Jeffery Sachs and Andrew Warner used in "Fundamental Sources of Long-Run Growth" American Economic Review May 1997, 184-188. Parts of it were also used in numerous other publications.

3)

a. Find the correlation between average annual growth and the share of exports of primary products in GNP (SXP). This variable indicates the proportion of a country's exports that is unprocessed food, minerals, or timber; not manufactured goods. Interpret the correlation coefficient.

b. Generate a scatter plot graph with a fitted line that has growth on the y-axis and SXP on the x. Does there appear to be a relationship here, if so what is it (positive or negative; linear or not)?

c. Repeat a and b for average annual growth and another variable of your choice.

4)

a. You want to try and describe what leads to growth. Of the available variables which ones would you think would have an effect on growth? Justify your choices from a theoretical standpoint and predict what sign the estimated coefficient should have (example: not having access to a waterway limits export by making it more expensive to ship goods, thus limits opportunities to grow – so I think ACCESS should be included and that it will have a negative sign).

b. If you have not already addressed the issue, do so now. Should we include a life expectancy squared in the regression (use the life variable that calculated in part 2). Plot a scatter graph with a fitted line with growth on the x axis and life expectancy on the y. Does it appear that you need to include squared life expectancy, why or why not?

c. Run a regression with average annual growth as the dependent variable and your choices for determinates of growth as independent variables– include at least one of the two dummy variables in the data set, and no matter what your answer was to 4b, include your generated life expectancy value but do not include a squared term for life expectancy.
    i. Interpret the coefficients that you come up with, and list which coefficients are significantly different that zero at a 95% confidence level.
    ii. Do the results match what you had thought from part a.
    iii. Interpret the overall fit of your regression using the R squared value.

d. Generate a variable that is your life expectancy value squared. Include this variable in the same regression you ran in part c. Use the adjusted R squared, should life expectancy squared have been included?

e. Use an F-test to determine if life expectancy and life expectancy squared should be included in the regression you ran in part d.