**Stata and Writing Prompt**

Part 1: Please begin by using the attached Stata data, MalariaDataAdapted.dta

The dataset is a subset of the data used in the paper by Jessica Cohen, Pascaline Dupas, and Simone Schaner, published in the American Economic Review. It examines behavior in response to various discounts (“subsidies”) – or no discount – for malaria treatment, called “artemisinin combination therapy,” or “ACT.”

Note that one variable has the description, “RCT arm: assigned ACT any subsidized price (less than 500 KSh) (indicator)” – this indicates which overarching study arm the household was randomly assigned to. (There were randomly assigned subsidy levels as well, but that is not the subject of the present task.)

Please make a .do file that opens the dataset, then does the following tasks. Please comment the .do file so that it is clear what you are doing at each step. Please also save the graphical output. Your work for the following items, 1.1 and 1.2, should include a single .do file and multiple graphical outputs, with reasonably interpretable file names.

[1.1] A number of baseline variables are included in the dataset. Among them, you should find one that has the number of years of the household head at baseline. Produce a two-panel histogram, or a pair of histograms, depicting the distribution of (baseline) years of schooling of the household head. Of the two panels (or the two histograms), one should depict the distribution of years of schooling of the household head only for those assigned to the subsidized price arm; the other should depict the same distribution but only for those households assigned to the full price arm (those not receiving any subsidy). The number of years of schooling is always a whole number, and there are not that many values, so the histogram should be done using the “discrete” option rather than having bins that combine multiple values.

[1.2] There are also post-randomization outcomes in the dataset. The descriptions of these variables begin with “episode care,” and the variable names themselves start with the letter “c.” As above, either make a two-panel histogram using the “by” option, or a pair of histograms, showing the distribution of the episode care outcome indicator described as “treated with ACT,” by treatment arm (those subsidized versus those not).

For Parts 2 and 3: Using web-downloadable but not necessarily well-formatted data:

[1] The Barro-Lee Educational Attainment Data Set: <http://www.barrolee.com/>

[2] Data on GDP per capita and male and female labor force participation from the World Bank’s World Development Indicators (WDI): <https://databank.worldbank.org/data/source/world-development-indicators>

Part 2: Write a .do file that does the following:

For any of the World Development Indicator variables you download, make a scatter plot of one against the other, using the options of the “scatter” command to label the graph reasonably well. (Title of the graph, axis labels, etc.)

Part 3: Write a .do file that does the following tasks:

[3.1] Prepare and merges the two country-level data sets described above.

You are free to choose the year or years to be used in your analysis, but please make sure you are merging data for comparable years. So, for example, you could focus on labor force participation and education in 2000, or on changes in education and changes in growth between 1980 and 2010. So, you should either merge the educational and WDI data by country and year or restrict attention to a single year or a specific year-to-year change (i.e., the change between 1980 and 2010).

Please make sure that your do file is well commented throughout. Specifically note any countries that are dropped from either data set because they do not merge.

[3.2] Makes an attractive scatter plot with education (in some form) on the x-axis and your WDI outcome on the y-axis.

[3.3] Runs at least two regressions and exports the coefficients to a well-formatted table (for example, in Excel or LaTeX). You should regress your WDI outcome of choice on education. One specification should not include any controls, and the other should include fixed effects for the World Bank regions (East Asia and Pacific, Europe and Central Asia, Latin America and the Caribbean, Middle East and North Africa, North American, South Asia, and Sub-Saharan Africa).

[3.4] Write anything from a few sentences to a few paragraphs describing the patterns you find in the regression work in part [3.3] above. Does education appear to increase GDP per capita or labor force participation? Why would you arrive at this conclusion, or why would you not?

Step 4: Submit your do-files, graphical output, and short written discussion from parts 1-3 above.