2. Jumping into Cross-Country Data

This question is an empirical exercise using cross-country data. Completion of this exercise will require you to manipulate data, compute descriptive statistics, use OLS regression, and plot data. The dataset is posted on *Blackboard* in the file `cross_country_data.csv` in CSV format. This dataset is drawn from the *Penn World Tables* version 8.1. An observation consists of the following information for a given country and year in the sample:

- `Country`: the name of the country
- `Year`: the year of the observation
- `SavingsRate`: the ratio of investment to GDP
- `RealGDP`: the value of gross domestic product (millions of 2005 US dollars)
- `LaborForce`: the size of the labor force (millions of people)
- `HumanCapital`: an index of human capital per person based on years of schooling and Mincerian returns to schooling
- `PhysicalCapital`: the capital stock (millions of 2005 US dollars)

To complete this assignment, use any data analysis software or language you wish. Deciding which tool to use, and learning how to use that tool independently, is part of the assignment. Several options include:

- The statistical programming language *R* – my suggestion. *R* is open source, i.e. free. *R* is available at the link here. *R* is the most common statistical analysis tool for data scientists and statisticians in private industry and academics. All of the tools – data manipuation, OLS regression, scatterplots, etc... – needed for this assignment are readily available within *R*. *R* is easier to use together with the additional software package *RStudio*, available for free here. *RStudio* functions as what is known as an integrated development environment for *R*, which is a fancy way of saying that by using *RStudio* you can easily see and manipulate the data and results from your *R* code. You might be interested in the tutorial *Econometrics in R*, available here.
- The econometrics programming language *Stata*. *Stata* is popular with economists. *Stata* can be bought at a discounted price for students at the link here. All of the tools needed for this assignment are also readily available within *Stata*. You might be interested in the *Stata* tutorial available at the link here.
- *Microsoft Excel*. Using the data analysis, manipulation, and plotting tools in *Excel*, it is technically possible to complete this assignment. Doing so would involve much more time-consuming manual work than one of the statistical programming language options. However, *Excel* is conceptually simpler. You should feel free to weigh the tradeoffs among the various options and come to your own choice.

Important Note: To receive credit for completing this question you must turn in a writeup of your answers which states the requested information in each part in **summarized, easily readable form**. You will receive no credit for raw output from a statistical package. Although statistical software can easily be used to churn out pages and pages of results, it is your responsibility – not the TA's – to extract answers from that output! If you choose to work in a group – of up to five students – remember to turn in a separate writeup of your analysis and include your team members' names at the top of your writeup.

(a) **Descriptive Statistics**
For each nation in the sample, compute the following values:

- GDP per worker in 1985 and 2005, $\frac{Y}{L}_{1985}$ and $\frac{Y}{L}_{2005}$
- the average savings rate from 1985 to 2005, s
- the average growth rate of the labor force from 1985 to 2005, n. To compute the average growth rate of a variable X from year t to year $t+T$, use the log approximation formula
$$\frac{1}{T}\log\left(\frac{X_{t+T}}{X_t}\right).$$
- the average growth rate of GDP per worker from 1985 to 2005, $g_{Y/L}$. Use the log approximation formula above. Remember, economists use $\log$ to refer to natural logarithms!

Report the following information:

- the number of countries in the sample of data
- the mean – across countries – of $\log\left(\frac{Y}{L}_{1985}\right)$, $\log\left(\frac{Y}{L}_{2005}\right)$, s, n, and $g_{Y/L}$
- the standard deviation – across countries – of $\log\left(\frac{Y}{L}_{1985}\right)$, $\log\left(\frac{Y}{L}_{2005}\right)$, s, n, and $g_{Y/L}$

(b) **Unconditional Convergence**
Produce and report a scatterplot of $\frac{Y}{L}_{1985}$ (on the horizontal axis, in logs) and $g_{Y/L}$ (on the vertical axis, using exactly the formula from above). Then, estimate the following OLS regression
$$g_{Y/L} = \beta_0 + \beta_1 \log\left(\frac{Y}{L}_{1985}\right) + \varepsilon.$$
Report the value of the estimated coefficient $\hat{\beta}_1$. Does the estimated value of $\hat{\beta}_1$ suggest that living standards in nations which are initially poorer catch up, fall further behind, or remain the same relative to wealthier nations?

(c) **MRW Revisited**
Following the analysis in MRW, estimate the following OLS regression
$$\log\left(\frac{Y}{L}_{2005}\right) = \beta_0 + \beta_1 \log(s) + \beta_2 \log(n+g+\delta) + \varepsilon.$$
To implement this regression, you should assume that $g+\delta = 0.05$ is constant across nations. Report the following information:

- The estimated value $\hat{\beta}_1$. Does this value suggest that living standards in nations which save more are higher, lower, or the same as in nations which save less?
- The value of α, the capital elasticity of output in the Solow model **with physical capital only**, is implied by the estimate $\hat{\beta}_1$. Is this implied value of α consistent with a labor share of around $2/3$?
- The R^2 of the regression. What fraction of the variation in living standards in 2005 is explained by the variables in this regression?

(d) **HJ Levels Accounting**
Following the analysis in HJ, assume that $\alpha = \frac{1}{3}$. Then, for each nation i in the dataset,

compute the implied value of productivity $A_{2005,i}$ in the year 2005 using the following decomposition

$$\frac{Y}{L}_{2005,i} = \left(\frac{K}{Y}_{2005,i}\right)^{\frac{\alpha}{1-\alpha}} \left(\frac{H}{L}_{2005,i}\right) A_{2005,i}.$$

Above, the subscript $2005, i$ refers to the value for country i in the year 2005 of the following quantities:

- $\frac{Y}{L}_{2005,i}$: GDP per worker
- $\frac{K}{Y}_{2005,i}$: capital intensity, the ratio of the capital stock to GDP
- $\frac{H}{L}_{2005,i}$: human capital per person, i.e. the raw index from the dataset
- $A_{2005,i}$: implied productivity or technology

Report the mean and standard deviation – across countries – of $\log(A_{2005,i})$. Produce and report a scatterplot of $A_{2005,i}$ (on the horizontal axis, in logs) and $\frac{Y}{L}_{2005,i}$ (on the vertical axis, in logs). Then, estimate the following OLS regression

$$\log\left(\frac{Y}{L}_{2005}\right) = \beta_0 + \beta_1 \log(A_{2005,i}) + \varepsilon.$$

Report the value of the estimated coefficient $\hat{\beta}_1$. Does the estimated value of $\hat{\beta}_1$ suggest that living standards are higher or lower in nations with higher productivity? Report the R^2 of the regression. What fraction of the variation in living standards in 2005 is explained by productivity?