**<u>Econ 172:</u> Issues in African Economic Development**
**Problem Set 1 (Due in class Thursday February 8, 2007)**
**[**Please hand in your own solutions – do not solve in teams]

**<u>1. African versus South Asian Economic Development [3 points]</u>**
In roughly 1 to 1.5 pages, double-spaced, discuss the relative economic development trajectories of Sub-Saharan Africa versus South Asia since approximately 1950. In what main ways is Africa a laggard and in what ways a leader, compared to South Asia?

**<u>2. Child Health and Education</u>**
(Download the MS-Excel dataset "PS1.XLS" from the course page.)
a) Why was the health intervention in the Primary School Deworming Project (Miguel and Kremer 2001) randomized across schools? How does the randomized design affect the estimation of treatment effects, and how does it help address omitted variable bias? (Please illustrate your points with the discussion presented in class.) [2 points]

b) Using the "LINEST" command in EXCEL, determine the average difference between 1998 treatment schools (GROUP1=1) and 1998 comparison schools (GROUP1=0) in the following two pre-treatment characteristics:

Proportion of female children (FEMALE)
Average child year of birth (YOB)

Report the regression output for the two regressions – you should hand in the actual MS-Excel print-out – and interpret the coefficients. Did the randomization succeed in creating comparable groups? At what level of confidence? Please present the standard errors and the t-statistics. [1 point]

c) Determine the average difference between treatment and comparison schools in:

Average school participation in 1998 after the program (PART98)

Report the regression results – you should again hand in the print-out – and interpret the coefficients. What is the impact of attending a treatment school on average school participation? Is this significantly different than zero at 95% confidence? 1 [point]

d) Re-run the regression in part (c), but include FEMALE, and YOB as additional explanatory variables. Please report and interpret the regression results. Does the inclusion of these additional control variables change the conclusions in (c)? Is this a surprising result? Why or why not? [2 points]

e) In MS-Excel, graphically plot out the relationship between average year of birth of students in the school (X-axis) versus average 1998 school participation (Y-axis), using a "scatter chart". Include this scatter chart in your solution, and speculate on what factors might help explain the observed pattern. [1 point]

**BRIEF EXCEL LINEAR REGRESSION (OLS) TUTORIAL**

The most important MS-EXCEL command used in this problem set is "LINEST". LINEST carries out ordinary least squares (OLS) regressions. OLS coefficient estimates constitute a "best-fit" to the data, and give us insight into the relationships among variables. It is probably worthwhile to read the EXCEL "Help" description of this command. The GSI's will also discuss linear regression in section.

The following example illustrates how to use LINEST in MS-EXCEL. Consider the following regression equation, where $Y$ is the outcome of interest, $X_1$ and $X_2$ are explanatory variables, $e$ is the error term, and $i$ denotes a particular observation.

$$Y_i = a + b_1X_1 + b_2X_2 + e_i$$

If OLS coefficient estimates for $b_1$ and $b_2$ are positive, this would suggest that $X_1$ and $X_2$ are positively correlated with $Y$ in the data.

In this example, imagine that the $Y$ values were contained in cells *A2:A50* in the spreadsheet, $X_1$ values were contained in cells *B2:B50*, and $X_2$ values were contained in cells *C2:C50* in the spreadsheet. On the same EXCEL sheet as the data, entering the following command generates the desired OLS coefficient estimates and standard errors:

*LINEST(A2:A50, B2:B50:C2:C50, TRUE, TRUE)*

The *A2:A50* component lets EXCEL know where to find the values of the dependent variable. The *B2:B50* component contains the first explanatory variable, and the *C2:C50* component contains the second explanatory variable. $a$ is estimated automatically. (Similar syntax can be used to include additional explanatory variables.)

In order to generate the OLS output, first highlight a block of cells on the same sheet as the data, with dimension two (2) cells in height, and dimension three (3) cells in width since, in this case, three coefficients will be estimated, $a$, $b_1$, and $b_2$. (In general, the width of the highlighted block of cells should be the number of coefficients being estimated). Then press the equal sign ("="). Then type the appropriate LINEST formula, as above. Do not press enter once you have typed the required formula! Rather, press "CTRL-SHIFT-ENTER" (all three keys simultaneously). This should generate the following OLS output:

| | | |
|---|---|---|
| Estimated $b_2$ | Estimated $b_1$ | Estimated $a$ |
| Standard error $b_2$ | Standard error $b_1$ | Standard error $a$ |

The standard error indicates how uncertain the coefficient estimate is for that variable; smaller standard errors suggest that the coefficient estimate is known with greater precision. Roughly, when the absolute value of the ratio of the coefficient estimate to the standard error (also known as the *t-statistic*) is greater than 2, the coefficient is considered to be different than zero with greater than 95 percent confidence. When the absolute value of the *t-statistic* is at least 1.6, the coefficient is considered to be different than zero with approximately 90 percent confidence.