14.74 Foundations of Development Policy
Spring 2009

Feel free to work in groups. But please hand in individual problem sets. Keep your answers clear and concise.

1) Worms in Kenya. The dataset, **ted_miguel_worms.dta** is from Ted Miguel's (UC Berkeley) and Michael Kremer's (Harvard) de-worming project in Kenya. Please read the paper to familiarize yourself with the experiment before answering this question. You can find the paper on the course website under Assignment #3.

   Miguel and Kremer **randomize** over schools (not individuals) and to introduce de-worming drugs to randomly selected treatment group to estimate the effect of deworming on school attendance.

   a. First, why randomize at the school level? Think what issues might arise in your evaluation if you randomize at the individual level.
   b. Suppose you had **pre- and post- treatment** attendance records for all schools, describe the calculations/comparisons you do in order to estimate the effect of deworming on school attendance (in **words and with a DD grid**). Why might you prefer this DD estimate to a straightforward OLS estimate using only **post-treatment** data (i.e. mean attendance in program schools minus mean attendance in non-program schools after the program)?
   c. Write down the regression you want to run in to estimate the DD you described in part (b)? Explain the meaning of the interaction term.

Now look at the data set. *As before, you are only required to hand in your documented .do files along with the answers. You can also include relevant parts of your .log file.*

   totpar98 is school participation in 1998, during the first year of treatment (they do not have good pre-treatment data - it is from attendance registers only).

   pill98 is a dummy variable for whether or not the student took a pill.

   infect_early99 is the health outcome in early 1999 when only Group 1 schools have received deworming (Group 2 schools got them right after this data was collected)

   Use the **describe** command to see the meaning of other variables. The authors have pre-treatment health, but only for the Group 1 kids - since it was considered unethical to get health information for control group kids -and thus this data cannot be used in regressions. So we cannot run the DD proposed above

d. How many observations are there per pupil? What percentage of the pupils are boys? What percentage of pupils took the deworming pill in 98 and 99? What percentage of schools was assigned deworming in 98? Is this more or less than the percentage of pupils who actually took the pill in 98?

e. What does the mean of a **dummy variable** tell us? Give one example from the dataset. What are two advantages of using dummy variables?

f. Using the data, find the **difference in outcomes (Y: school participation)** between.

      i. Students that took the pill and students which did not in 1998.
      E(Y|X=1)-E(Y|X=0)
      Is this a good estimate of the effect of taking the pill on school attendance?
      ii. Students in treatment schools versus students not in treatment schools in 1998 (regardless of whether they actually took the pill).
      E(Y|Z=1)-E(Y|Z=0)

g. Using the data, calculate the difference in the probability of taking the pill given that a student was in a treatment school and the probability of taking it if a student was not in a treatment school.
      E(X|Z=1)-E(X|Z=0)

h. Derive the **Wald Estimator** and Explain the meaning of this result. What is the advantage of the Wald Estimator over what calculated in f(i)

2) Economist Jeffrey Sachs does a cross-country regression by regressing economic growth on the incidence of malaria while controlling for education, income level, macroeconomic policy, etc. He finds that countries with malaria grow more slowly.

a. Is this a causal relationship? Why or Why not? (Give an alternative interpretation to his regression results.)

b. Generally speaking, what do the coefficients of the right hand side controls tell us? It will help to write down the regression equation. Do you think the control variables resolve joint determination problems? Explain briefly.

c. Let's assume that education jointly determines income and malaria incidence and economic growth. In this case, what does the coefficient of education tell us? Would this cause us to over or underestimate the true relationship between malaria on income? Why? (hint: think of possible ways education jointly increases/decreases income and malaria). Next, assume that education decreases malaria and increases income. Explain how/why the regression might overestimate the effect of malaria on income.