Problem Set 1
Global Poverty and Economic Development

Lorenzo Casaburi & Manon Delvaux

NB: This problem set is due on the 25^{th} of October by 5.00pm. Please submit your work by email to natascha.baer@econ.uzh.ch with the title of the class and the problem set number in the subject line. Please also indicate your name on the first page of the document.

1 Education & Poverty

A team of researchers are interested in studying the impact of education on poverty. They have survey data available on 200 individuals.

They run the following initial regression:

$$ln(inc)_i = \theta_0 + \theta_1 educ_i + \epsilon_i$$

Where i indexes the individuals, *educ* is education measure in years of schooling, and *inc* is income per year.

1. The researchers find that $\hat{\theta}_1 = 0.05$ with significance. How should they interpret the magnitude of the coefficient? What could be the underlying mechanism behind the result?

2. The researchers decide to re-run their regression to include only individuals above the age of 45. They end up with a sample size of 30, find $\hat{\theta}_1 = 0.07$, and find a standard error of 0.4. Is this result significantly different from 0 at the $\alpha = 0.05$ level?

3. Returning to the specification in (1.), will the researchers' estimates for θ_1 be unbiased? Why or why not? Show it analytically.

4. Suppose now that the researchers are unable to access income data, so instead they ask people directly to tell them their income per year. What could be a possible concern with this methodology? Is this a problem for our estimates?

2 Data Analysis

This is a data exercise using World Bank Health data. You will need data on educational attainment for this, use the dataset *Health_WB.csv*, available on OLAT. Make sure you have a dataset at the country level. For this exercise, please do not submit your code.

1. Generate a table of summary statistics averaged across all countries (mean, standard deviation, min, max etc.) for each of the following sets of variables separately:
   (a) Immunizations (all)
   (b) Life Expectancy at birth (male, female and total)
   (c) School enrollment, net (Total) for primary and secondary education.
2. Regression output:
   (a) Run a regression of life expectancy as a function of public expenditure on healthcare (% of GDP) and private expenditure on healthcare (% of GDP).
   (b) Run a regression of life expectancy on access to clean water and access to sanitation.
   (c) Finally, regress life expectancy on the log of primary school enrollment (net & total) and the log of secondary school enrollment (net & total).
   (d) Combine and present all of your regression results in a well-formatted, easily-readable table. (*Hint for those using stata: outreg or esttab*)
3. Generate a scatter plot of average life expectancy by country against average primary school education (by country) (net, total)

3 Fixed Effects

Use the database *data_FE.csv*. The database includes 100 households for 10 periods each. There is one outcome variable y and one explanatory variable x_1

For the questions below, you don't need to export regression tables, just report the coefficient of interest in the text of your answers.

1. Summarize variable x_1. Report mean and standard deviation.
2. Run an OLS regression of y on x_1 (and a constant).
3. Now run the same regression with household fixed effects.
4. Generate a variable $\hat{x}_1$ which equals x_1 for households 1-60 and x_1+100 for households 61-100. Summarize the variable $\hat{x}_1$. Report mean and standard deviation.

5. Run an OLS regression of y on $\hat{x}_1$ and a constant. Does the coefficient of $\hat{x}_1$ differ from the coefficient of x_1 in point 2 above? Explain the intuition behind your answer.

6. Run a regression of y on $\hat{x}_1$ with household fixed effects. Does the coefficient of $\hat{x}_1$ differ from the coefficient of x_1 in point 3 above? Explain the intuition behind your answer.

4 Omitted Variable Bias

Take the true model to be:

$$Y_i = \theta_0 + \theta_1 T_i + \theta_2 X_i + \epsilon_i$$

with

$$E[X_i|T_i] = \alpha_0 + \alpha_1 T_i$$

1. Assume that X_i is unobserved, and we run an OLS of Y on T. Derive the omitted variable bias formula.

2. How would randomization address OVB?

   Now consider a case where the treatment T is *Conflict* and the outcome Y is *Income*

3. (a) Which sign would you expect for the treatment coefficient of the true model θ_1?
   (b) Another driver of income may be local mineral prices. Assume we don't observe mineral prices at the local level, so they induce omitted variable bias (i.e. mineral prices are X in the above notation). Which signs would you expect for θ_2 and α_1?
   (c) Based on your answer to the previous question, would you expect that omitting prices would lead to upward or downward bias in the estimation of the effect of conflict of on income? Provide an example of a case with an upward bias and one with a downward bias.