Professor Debraj Ray
19 West 4th Street, Room 608
email: debraj.ray@nyu.edu, homepage: https://debrajray.com/
Webpage for course: https://pages.nyu.edu/debraj/Courses/26UGDev/index.html

ECON-UA 323

Development Economics

Problem Set 6

(1) A researcher wants to measure the return to schooling in the province of Vidyapur. For each of 1000 young adults she observes years of schooling s, monthly earnings y, and whether a secondary school was built in the person's village during their childhood ($z = 1$ if so, $z = 0$ if not). She plans to run the regression

$$y_i = A + bs_i + \epsilon_i.$$

What she does *not* observe is whether a person comes from a well-connected family ($a = 1$) or not ($a = 0$). Unknown to her, the truth in Vidyapur is

$$s = 10 + 2z + 2a \quad \text{and} \quad y = 5 + 2s + 4a,$$

so that the true return to a year of schooling is $b = 2$. Well-connected families keep their children in school for longer, *and* they find them better-paying jobs. The schools were assigned to villages by lottery. The population falls into four groups of 250 people each:

Group	School in village (z)	Well-connected (a)	Schooling (s)	Earnings (y)
1	0	0	10	25
2	1	0	12	29
3	0	1	12	33
4	1	1	14	37

Because the four groups are equally large, you can do all the calculations below by treating them as four observations.

(a) Using only what the researcher can see, calculate the OLS estimate of the return to schooling with the formula from class,

$$\hat{b} = \frac{\sum_i (s_i - \bar{s})(y_i - \bar{y})}{\sum_i (s_i - \bar{s})^2}.$$

How does it compare with the truth?

(b) Which of the endogeneity problems that we discussed in class is at work here? Write down the error term ϵ_i of the researcher's regression in terms of the true model, and verify from the table that it is correlated with s. Then use the decomposition from class,

$$\hat{b} = b + \frac{\sum_i (s_i - \bar{s})(\epsilon_i - \bar{\epsilon})}{\sum_i (s_i - \bar{s})^2},$$

to account *exactly* for the gap between your answer to (a) and the truth. Would the problem go away if she had data on a million people instead of a thousand?

(c) The researcher proposes to use z as an instrument for s. State the two requirements for a good instrument, and check whether each one holds in Vidyapur. Which of the two can she verify with her own data? What must she do about the other, and what in the story helps her?

(d) Calculate the IV estimate

$$\hat{b}_{\text{IV}} = \frac{\sum_i (z_i - \bar{z})(y_i - \bar{y})}{\sum_i (z_i - \bar{z})(s_i - \bar{s})}.$$

Next, calculate the difference in average earnings between villages with a school and villages without one, and the same difference for average schooling, and divide the first by the second. Explain in words why this ratio measures the return to schooling. (Compare your explanation with the quotation from Angrist and Krueger in Slides 12.)

(e) Now do the same thing by two-stage least squares. In the first stage, predict each group's schooling $\hat{s}$ from z alone. In the second stage, calculate the OLS coefficient in a regression of y on $\hat{s}$. Do you get the answer of part (d)?

(f) Suppose, instead, that the schools had not been assigned by lottery. They were built in villages that lie on the main road, and living on the main road raises a person's earnings by 1, whatever her schooling. So the truth is now $y = 5 + 2s + 4a + z$, with schooling determined as before. (i) Recalculate the earnings column and the IV estimate. (ii) Which requirement for a good instrument has failed? Could the researcher discover this from her data?

(2) This question is about the two studies of "long shadows" in Slides 13: Acemoglu, Johnson and Robinson (AJR) on colonial institutions, and Alesina, Giuliano and Nunn (AGN) on the plough. Both papers are linked from the slides and are worth looking at. A few careful sentences on each part are enough.

(a) Put the central question of each study in the form "does x cause y?" What are x and y in each case, and how are they measured?

(b) In each study, explain why the OLS coefficient cannot simply be read as a causal effect. For AJR, give a concrete example of each of reverse causation, an omitted variable, and measurement error. For AGN, give two reasons why the historical adoption of the plough might be correlated with the error term.

(c) Describe the instrument used in each study, and the chain of reasoning that connects it to the explanatory variable x. Two further questions on AGN. Why do they use the *suitability* of the land for plough-positive crops, and not whether those crops were actually grown? And why do they set wheat, barley and rye against millets and sorghum, instead of simply using the suitability for wheat, barley and rye?

(d) State in words what the exclusion restriction requires in each study. What is the most serious threat to it, and how do the authors respond? Are you convinced?

(3) Draw Lorenz curves and calculate the Gini coefficient and the coefficient of variation for the income distributions (a)–(d). In each situation, the first set of numbers represents the

various incomes, whereas the second set of numbers represents the number of people earning each of these incomes:

(a) $(100, 200, 300, 400)$; $(40, 50, 60, 50)$

(b) $(200, 400, 600, 800)$; $(80, 10, 80, 30)$

(c) $(100, 200, 300, 400)$; $(100, 0, 0, 100)$

(d) $(100, 200, 300, 400)$, $(0, 100, 100, 0)$.

[Try to understand the implicit transfers that move you from one income distribution to the other.]

(4) The economy of Nintendo has ten people. Three of them live in the modern sector of Nintendo and earn \$3000 per month. The rest live in the traditional sector and earn only \$1000 per month. One day, two new modern sector jobs open up and two people from the traditional sector move to the modern sector.

(a) Show that the Lorenz curves of the income distributions before and after must cross. Do this in two ways: (i) by graphing the Lorenz curves and (ii) by first expressing both income distributions as divisions of a cake of size 1, and then showing that the two distributions are linked by "opposing" Dalton transfers.

(b) Calculate the Gini coefficients and the coefficients of variation of the two distributions.

(5) The land of Oily is filled with inhabitants that only hold shares of the national oil resource, but do so unequally. Number the individuals in $[0, 1]$ from poorest to richest, so that x_i is the amount of oil held by person i, where $x_i \leq x_j$ if $i \leq j$. The international price of oil is given by p, so for each person i,

$$\text{Wealth } = w_i = px_i$$

(a) Show that no matter how oil prices p move around at any date t, wealth inequality is unchanged in Oily.

Now we complicate matters a bit by giving everyone a fixed handout $y > 0$ over and above their oil holdings, so that for each person i, wealth is now given by

$$\text{Wealth } = w_i = px_i + y$$

(b) Show that if oil prices p rise, inequality must go up in Oily. [Try and do this in two ways for practice: by showing how the Lorenz curves change, or deriving the result from the axioms of inequality measurement. Either approach, if done right, would be valid for full credit.]

(c) How would your answer to (b) change if y were a tax instead, so that it is negative, not positive? How would you answer change if the tax or the handout were to be exactly proportional to oil holdings?

Three more problems to think about, which do not need to be handed in. Questions 7 and 8 are not relevant for the exam, but they will help your understanding of the material.

(6) Are the following statements true or false? In each case, back up your answer with a brief, but precise explanation.

(a) If an explanatory variable x is correlated with the error term of a regression, the OLS estimate of its coefficient will be biased in small samples, but the bias will disappear as the sample becomes large.

(b) If z is a valid instrument for x, then the IV estimate is right on average in samples of any size, just as the OLS estimate is when x is uncorrelated with the error.

(c) The settler mortality rates used by AJR were determined two centuries ago or more, so they cannot possibly have been caused by per-capita incomes today. This is enough to make settler mortality a valid instrument for institutions today.

(d) AGN find that among women born and raised in the United States, those whose parents came from countries with a tradition of plough agriculture are less likely to work outside the home. This suggests that the plough affects women's participation mainly through the laws and policies of the ancestral country.

(e) The mean absolute deviation satisfies the Dalton transfer principle.

(f) If the Lorenz curve of two situations cross, the Gini coefficient and the coefficient of variation must disagree.

(g) If a relatively poor person loses income to a relatively rich person, the Gini coefficient *must* rise.

(h) The Lorenz curve must *necessarily* be convex; that is, going from left to right on the Lorenz diagram, the slope of the Lorenz curve can never fall, and will generally rise.

(i) The ethical principles of inequality measurement — anonymity, population, relative income, and transfers — are enough to compare any two income distributions in terms of relative inequality.

(j) If everybody's income decreases by a constant dollar amount, inequality *must* rise.

(7) A well-known measure of economic inequality is the log variance. It is measured by

$$\text{LV} = \frac{1}{n}\sum_{i=1}^{m} n_i[\ln y_i - \ln \mu]^2.$$

Show that this measure satisfies anonymity, the population principle and the relative income principle, but fails to be Lorenz-consistent.

(8) A beautiful property of the Gini coefficient is that it turns out to be proportional to the *area* between the Lorenz curve of an income distribution and the 45^0 line. (It is actually equal to the ratio of this area relative to the area of the triangle formed below the 45^0 line.

Try and prove this for the case in which there are just two income classes, with a proportion n in the poorer class who earn a fraction a of total income, and the remaining richer fraction $1 - n$ earning the remaining fraction $1 - a$ of income.