Problem Set 2: Estimating the Returns to Education Using Instrumental Variables

Many of the papers we have been studying in class recently are concerned with the following question: “How much more income can a person expect to earn if they obtain an extra year of education?” A related question is “If we build more schools or otherwise encourage people to get more education, by how much can we expect poverty to be alleviated?” Unfortunately, obtaining an accurate answer to these questions can be difficult. In the Duflo, Jensen, and Foster & Rosenzweig papers, we saw a few ways to think about the problems surrounding these questions.

In this problem set, we will be using data collected in one province of the Philippines in the 1970’s, 80’s, and 90’s. One nice feature of this data is that we can link data on a person’s household with the outcomes experienced by that same person 10 years later. Since the data includes the characteristics of the households where people lived when they were school-aged and their labor outcomes in adulthood, we can try to use household characteristics to help us get a reliable measure of the relationship between an individual’s level of education and their wage.

Dataset:

Download the dataset “ps2data” from Canvas and open it in stata. Each unit of observation in this data is an individual person. For the purposes of this problem set, the key variables found in it are defined below:

wkwage: average daily wage earned in the week before the survey (in pesos)

distance: distance from the barangay where the household lives to the nearest secondary school (a *barangay* is small administrative unit in the Philippines, similar to a village)

edatnm: years of educational attainment

The other variables are defined in the stata datafile.

Exercises:

1.
    a. Run a simple OLS regression of daily wages on educational attainment. What is the relationship between schoolings and earnings in this data?

    b. Generate a variable equal to ln(wages), where ln(.) is the natural logarithm function. Regress ln(wages) on educational attainment. Is the relationship different than in part A? How is the interpretation different than in Part A?

    c. Economists think of ‘the return to schooling’ as the amount of extra income that a given individual (or type of individual) would earn if they went to school for one additional year. Does your answer from parts A and B tell you the return to education? Why or why not? Are there any other factors that you think could influence both education and wages?

d. Suppose that the dictatorial leader of a country wanted to get the simplest and most reliable estimate of the returns to education possible and was not concerned about the ethical implications of how the estimate was obtained. What "experiment" would this person do? How do you think the return you estimated from this policy would differ from the one you got in part A and B?

e. Now suppose that the dictator in part D is not willing to force people to make certain decisions about going to school, but *is* willing to experiment with the opportunities offered to people and has unlimited authority. What would be a policy "experiment" this kind of person might try in order to estimate the returns to education? (Hint: think about the policy that Duflo studied)

2. Although the policies you discussed in parts d and e would give reliable estimates of the returns to education, many people would consider such policies unethical and inefficient. In the following exercises, you'll use the technique of instrumental variables to estimate the returns to education. In the dataset, there is a variable called 'distance'. This variable gives the distance in km from the village where the person lived to the nearest secondary school (measured in 1983). The data include people who were of the age to be in school in 1983 and their outcomes measured in 1994.

   a. In a rural, developing country setting where most people are engaged in agriculture, like this part of the Philippines in 1983, how to do you think distance from school would affect years of education attained and why?

   b. Regress educational attainment on distance from school in 1983. This is what we refer to as the "First Stage". Are the results what you expected in 2.A.?

   c. Do you think that the correlation between distance from school and years of education attained is linear? Why or why not?

   d. Regress educational attainment on distance from school and distance squared. How are your results different from 2.b? Draw a graph of the relationship between distance and educational attainment whose shape is consistent with the coefficients (it doesn't have to be exact, just the general shape).

e. Given your answers to 2.A. and 2.B., what do you expect to be relationship between distance you lived from school in 1983 and the log of your wages in 1994?

f. Regress ln(wages) on distance and distance squared. This is the "Reduced Form" equation. How is distance from school related to wages?

g. Is it plausible that distance from school is randomly assigned with regard to the characteristics that affect schooling and wages? Think about your answer to 1.C.

h. What are the characteristics of a good instrument? Do your answers to 2.a. – 2.g. suggest that distance to school is a good instrument for educational attainment?

i. Assume that distance to school during childhood and adolescence is randomly assigned. Use Two-Stage Least Squares (2SLS) to estimate the return to educational attainment on ln(wage) using distance and distance squared as instruments for years of education. Do your results suggest that staying in school increases wages in this population?

j. How is the coefficient on educational attainment in the wage equation different when you use IV than it was when you used OLS in part 1.A.? Is the change in the coefficient in the direction that you expected?

3. There are many reasons why instrumental variables estimates can differ from OLS. In the remainder, we will develop a simple model of how the coefficients are different which depends on an interpretation of instrumental variables under certain conditions as giving an estimate called the Local Average Treatment Effect (LATE).

The model: Suppose people in this area have 10 years in which they can either go to school or go to work. They can choose 3 different levels of schooling:
1. They don't go to school at all, and work for 10 years
2. They go to school for 1 year and then work for 9 years
3. They can go to school for 2 years and then work for 8 years.

If they go to school, they get a better job and earn more per year according to the following rule:

Years of school:	Yearly Salary
0	$2
1	$4
2	$5

a. In terms of yearly salary, what is the return to education of going to 1 year of school (as opposed to 0)? What is the return to going to 2 years (as opposed to 1)? At which level of schooling is the return highest?

b. What is the socially optimal level of education for everyone to get? (in terms of creating the biggest pot of available resources for society)

c. What are the lifetime earnings of people who don't go to any school, people who go for 1 year, and people who go for 2 years? Assume: 1. There is no discounting and 2. The only cost of school is that you don't earn wages for every year you are in school. If everyone faces this same problem, how many years of schooling will everyone get?

d. Now suppose that going to school costs $1 per year. What is the net lifetime earnings associated with each choice now? Still assuming no discounting, does this cost change the number of years of education people choose?

e. Suppose that school still costs $1 per year, and there are two types of people, patient and impatient people. Patient people have a discount rate of 0, so they value their lifetime earnings exactly as in part C. But impatient people have a discount rate of .6 (meaning that after the first year, they deflate future earnings by a term equal to 1/(1.6^x) where x is the number of years after the first. What is the lifetime value of each schooling choice for the impatient people? What level of education will each type of people choose?

f. Finally, suppose that going to school costs now costs $2 per year, and there are impatient and patient people as before. Will either type of person choose a different amount of schooling than they did when school cost $1 as in parts d and e? If so, which type of person changes their amount of education?

g. OLS calculates the correlation between years of education and wages for everyone in the sample. IV calculates the return only for people who are affected by the instrument. Our instrument, distance from school, mainly affects the upfront cost of going to school instead of working during school age. Given your answers to 3.a-f., can this model explain why you found that the coefficient changed the way it did when you used IV relative to OLS in questions 1 and 2?