14.74

Lecture 10: The returns to human capital: education

Esther Duflo

March 7, 2011

Education is a form of *human capital.* You *invest* in it, and you get *returns*, in the form of higher earnings, etc...

Today, we will try to measure returns to education

- First, there is a correlation between education and earnings:

- By now, we have learnt to beware of correlations...

-What are the possible bias if we interpret the relationship between education and earnings causally? (recall the model we used)

Randomly assigning "education" to people is not possible: one's education are closely linked to other aspects' of one's person.

At best, you can randomly assign them to a program which will improve her her education. Example: Indonesia school construction programme.

Let's first look at the effect of the program on wages. We can use exactly the same methods we used for education. Note we are comparing everybody's wage *in 1995*, so we are comparing the wages of different cohort in the same year.

- Difference in difference

- Using the regional variation

- Using all the time variation

Look at the data: what is our conclusion when we do this?

So:

- The program improved education of children exposed to it
- The program improved wages of the same children once they are adult
- Do we have reasons to think that the program would have a direct effect of wages?
- If not, we can conclude that it is through its impact of education that the program must have affected wages: education affects wages.
- But HOW much? We would like to calculate a returns, or an elasticity... not just a qualitative answer.

Today, we are going to learn a new tool, based on the idea we just described, to estimate a relationship of interest: the methods of instrumental variable.

1 An example starting from a randomized evaluation

1.1 Basic set up

- The question: Does teacher absense has effect on children presence
- Notation:

$$Y_i = \alpha + \beta H_i + \epsilon_i$$

(Note that this formulations assumes that the effect of absence is the same for all people: we also have some results on how to estimate a relationship where we don't make this assumption, but we will not cover them now)

We have a randomized experiment in India which affects teacher absence: teachers are paid more if they come to school more (Rs 500 for 10 days or more, Rs 50 extra for each extra day).

- Note Z_i a dummy variable equal to 1 if one is assigned to the treatment group, 0 otherwise.

Look at the table in the paper. being in a treatment school

-lower absence

-higher test scores.

1.2 Combining the two: instrumental variable estimate of the effect of health on labor market outcomes

Effect of treatment on absence could be measured by:

$$E[H_i|Z_i = 1] - E[H_i|Z_i = 0] \tag{1}$$

Effect of treatment on test scores could be measured by:

$$E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0] \tag{2}$$

Using our expression for Y_i, we have:

$$E[Y_i|Z_i = 1] = \alpha + \beta E[H_i|Z_i = 1] + E[\epsilon_i|Z_i = 1]$$

and:

$$E[Y_i|Z_i = 0] = \alpha + \beta E[H_i|Z_i = 0] + E[\epsilon_i|Z_i = 0]$$

Therefore

$$E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0] = \beta(E[H_i|Z_i = 1] - E[H_i|Z_i = 0]) + E[\epsilon_i|Z_i = 1] - E[\epsilon_i|Z_i = 0]$$

What can we assume about $E[\epsilon_i|Z_i = 1] - E[\epsilon_i|Z_i = 0]$?

What underlies this assumption, and is this justified:

-

-

$$\hat{\beta} = \frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[H_i|Z_i = 1] - E[H_i|Z_i = 0]} \tag{3}$$

Careful: never forget to check *both* condition when thinking about using an instrument. The second condition is often not verified even when the first is. Note that if the second condition is violated, we can still interpret the reduced form as the effect of the instrument (e.g., here, the cameras). However, we cannot use it to learn about the relationship between the instrument and the outcome of interest (e.g. here, the relationship between absence and test scores). This is because the instrument affect the outcome in more than one way.

For example, in this example, why could the camera have an effect on test scores.

We obtain the effect of health on labor market outcomes by dividing the effect of the program on labor market outcomes by the effect of the program on health.

Equation 1 is the *first stage* relationship. Equation 2 is the *reduced form* relationship. $\hat{\beta}$ given by equation 3 is the *Wald estimate* of the effect of absence. It is the simplest form of the instrumental variable estimator (Z_i is our instrument).

Let us calculate the Wald estimator ourselves:

-mid-line math:

-mid-line language:

-end-line math:

-end-line english:

NB: You can see that even a "small" violation of either of the conditions for the validity of the instrument can result in very large bias. Any bias in the reduced form will be "blown up" when I divide by the first stage difference (e.g. even if there is a small difference between treated and untreated schools on test , we are going to divide this small difference by the fraction of people whose anemia status changed, which is very small: the bias will be large!).

2 The returns to education: using an instrument that is not randomly assigned

The case we just studied was particularly favorable: we had access to a randomized experiment that affected teacher presence, and formed a perfect instrument. Often, we don't have such instrument. For example, we do not have a randomized experiment that affected education levels: Can we use the instrumental variable method in this case?

- Imagine we start from the following wage equation, where each year of education improves wage by a constant percentage.

$$\ln(w_i) = a_i + bS_i$$

- More schools → more schooling → higher earnings.

2.1 Wald estimate

Note

$$a_{11} = E[a_i|\text{high=1, young=1}],$$

and use a similar notation for all the other groups.

Let us express analytically the difference in differences for wages:

	HIGH	LOW	Difference
YOUNG	$a_{11} + bS_{11}$	$a_{12} + bS_{12}$	$a_{11} - a_{12} + b(S_{11} - S_{12})$
OLD	$a_{21} + bS_{21}$	$a_{22} + bS_{22}$	$a_{21} - a_{22} + b(S_{21} - S_{22})$
Difference	$a_{11} - a_{21} + b(S_{11} - S_{21})$	$a_{12} - a_{22} + b(S_{12} - S_{22})$	

- What was our assumption for measuring the effect of INPRES on education.

- Two additional assumptions are now necessary:

(1) Wages did not move according to different trends from one region to another

(2) The only effect of the program was to changes wages?

-Can we test it (partly at least)?

-What is an assumption that is hidden in this way to write the wage function? (hint: this is an assumption about b). Is it likely to be satisfied? Can we test it?

How can we calculate the wald estimate (supposing the assumption is satisfied?)

What do we find? Is it plausible?

An Aside:

How to run Difference in difference in a regression format:

call S_{ijt} the years of education of person i born in region j in year t. Define a dummy YO_t= 1 if born after 1967, 0 otherwise and H_j= 1 if born in high program region, zero otherwise.

Run the regression

$$S_{ijt} = \pi_1 + \pi_2 YO_t + \pi_3 H_j + \pi_4 YO_t * H_j + \epsilon_{ijt}$$

What is:

-π_1:

-π_2:

-π_3:

-π_4:

How to run a Wald estimate as a regression:

First stage:

$$S_{ijt} = \pi_1 + \pi_2 YO_t + \pi_3 H_j + \pi_4 YO_t * H_j + \epsilon_{ijt}$$

Reduced form:

$$y_{ijt} = \lambda_1 + \lambda_2 YO_t + \lambda_3 H_j + \lambda_4 YO_t * H_j + \upsilon_{ijt}$$

Wald estimate is equal to?

Can also be obtained in stata by running two stage least square:

$$y_{ijt} = \lambda_1 + \lambda_2 YO_t + \lambda_3 H_j + bS_{ij} + \upsilon_{ijt}$$

using $YO_t * H_j$ as instrument for education.

2.2 Generalization 1: using all the regional variation– indirect least squares

Recall the example with two cohorts:

For education we have:

$$S_{Yj} - S_{Oj} = \alpha P_j + \upsilon_j \tag{4}$$

This is the first stage. Likewise for wages, we can run:

$$y_{Yj} - y_{Oj} = \gamma P_j + \epsilon_j \tag{5}$$

This is the reduced form:

We also know that $y_{Oj} = a_{Oj} + bS_{Oj}$ and $y_{Yj} = a_{Yj} + bS_{Yj}$

So we can express $y_{Yj} - y_{Oj}$ as a function of S, b, and α.

$$y_{Yj} - y_{Oj} = a_{Yj} - a_{Oj} - b(S_{Yj} - S_{Oj}) = a_{Yj} - a_{Oj} - b(\alpha P_j + \upsilon_j) \tag{6}$$

Write down the expression for b:

When we run this regressions in the data, we find:

-$\alpha = 0.196$ (Standard error: 0.0424)

-$\gamma = 0.012$ (Standard error: 0.0474)

What is our estimate of b? How does it compare to the Wald estimate? Which one do we prefer? This way of calculating b is called *indirect least squares.* It is a form of instrumental variables: we use the program as an *instrument* to predict the education variable. This prediction will only retain the variation in education which is not due to the individual's choices.

Doing this in a regression: we now have 281 differences (so we need to have those 281 dummies), which we are regressing on a continuous variable. So the equivalent of the DD is:

Replace H_j by P_j, a variable that indicate the number of schools constructed in the region of birth, and control for a dummy for each region

$$S_{ijt} = \pi_1 + \pi_2 YO_t + a_2 R_2 + a_3 R_3 + ... + \pi_4 YO_t * P_j + \epsilon_{ijt}$$

For wages we would have:

$$y_{ijt} = \lambda_1 + \lambda_2 YO_t + d_1 R_1 + d_2 R_2 + d_3 R_3 + ... + \lambda_4 YO_t * P_j + \epsilon_{ijt}$$

and the ratio λ over π is the indirect least squares estimate.

2.3 Generalization 2: using all the regional and cohort variation– two stage least squares

As for education, we can run the regression:

$$y_{jk} - y_{j24} = \gamma_k P_j + \upsilon_{j23}$$

for all cohorts k.

Or in a single regression, for education:

$$S_{ijt} = \pi_1 + b_2 y_2 + b_3 y_3 + ... + a_2 R_2 + a_3 R_3 + ... + c_2 y_2 * P_j + c_3 y_3 * P_j + ...c_{23} y_{23} * P_{23} + \epsilon_{ijt}$$

or wages

$$y_{ijt} = \pi_1 + e_2 y_2 + e_3 y_3 + ... + d_2 R_2 + d_3 R_3 + ... + \gamma_2 y_2 * P_j + \gamma_3 y_3 * P_j + ...\gamma_{23} y_{23} * P_{23} + \epsilon_{ijt}$$

If the program affected wages through its effects on education, we expect that the coefficient in the wage equation will "track" the coefficients in the education equation: See in the figure. Could you calculate b by indirect least squares using the estimates plotted in this graph?

How many ways do you have to calculate b? What is the best?

2SLS combines all the ways to calculate b in the "optimal" way (in the sense of reducing the variance of b).

We have 24"instruments" available (1 program, 24 cohorts of birth), or 12 if we decide to set to zero all the first 12 (where we have zero effect). With 2SLS we proceed in two stages:

1. Predict education using this 24 or 12 instruments (run an OLS regression of education on cohort of birth dummies multiplied by the program) by estimating equation (2.3)

2. Regress wages on this predicted value

3. This can be done in one single operation using the 2SLS command in stata (this is also the way to compute the right standard errors).

run

$$y_{ijt} = \pi_1 + e_2 y_2 + e_3 y_3 + ... + d_2 R_2 + d_3 R_3 + ... + bS_{ijt} + \epsilon_{ijt}$$

by 2SLS, using $y_2 .... y_2 4$, R_2, R_{281} and $y_2 * P_j$, $y_3 * P_j$, etc, as instruments.

The estimate of b using this method range from 0.0675 to 0.10, depending on the control variables we use: not very different from 0.075 (which we obtained running OLS). Each year of education increase wages on average by 7%.