# Education: returns to education evaluation

## Lecture 4/2 - 2/2013

Chayanee Chawanote

February 13, 2014

# Returns to education and OLS estimates

- Classic method: ordinary least-squares (OLS) regression
- Mincer equation for returns to education:
  $log(wage) = \beta_0 + \beta_1 S + \beta_2 Exper + \beta_3 Exper^2 + u$
- OLS: estimating a correlation, not necessarily a causal relationship
- If the 'other stuff' in the estimating equation affects both schooling and earnings, then the OLS estimates will be biased
  - High-ability people or more wealthy people both go to school and have higher earnings

# Returns to education and OLS estimates

- From the basic regression, we can add controls to help improve the causal interpretation of the estimates
- Some useful controls would be
  - Age
  - Income (or consumption)
  - A measure of ability (standardized test score)
- By including control variables, we can measure the relationship between schooling and earnings, *holding these factors fixed*.
  - These controls can help isolate the causal effects, but it is difficult to control for everything

# School construction in Indonesia: Duflo (2001)

- The paper presents both a program evaluation of school construction and a clever way to identify returns to education
- Between 1973 and 1979, Indonesia undertook a massive primary school construction program - more than 1 school was built for every 500 children across the country
- Question 1: What was the effect of the program on
  - Educational attainment
  - Wages
- Regress outcomes (Y) on number of schools built per 1,000 students (P)?
  $Y_{ij} = \alpha_0 + \alpha_1 P_j + \epsilon_{ij}$, for individual i and region j

# Regression of school allocation on regional characteristics

TABLE 2—THE ALLOCATION OF SCHOOLS

| | Log(INPRES schools)[a] |
|---|---|
| Log of number of children aged 5–14 in the region | 0.78 |
| | (0.027) |
| Log(1 − enrollment rate in primary school in 1973)[b] | 0.12 |
| | (0.038) |
| Number of observations | 255 |
| $R^2$ | 0.78 |

*Notes*: Standard errors are in parentheses.

[a] The dependent variable is the log of the number of INPRES schools built between 1973 and 1978.

[b] The enrollment rate in primary school is the number of children enrolled in primary school in 1973 (obtained from the Ministry of Education and Culture) divided by the number of children aged 5–14 in the region in 1973.

- Concern: regions would have had differences in enrollment (and wages) in the absence of the program.

# Controlling for differences across regions

- Control for 'pre-existing' differences in enrollment across regions using older individuals.
- Children who were 2-6 years old in 1974 were likely exposed to the program for a number of years.
- Children who were 12-17 years old in 1974 were unlikely to have been in primary school when the new schools were constructed.
- Thus, individuals aged 12-17 in 1974 can serve as a proxy for baseline enrollment differences across regions.

# Causal Effects

- Let there are 2 treatments categories we care about: low intensity school construction ares ($P_j = 0$), and high-intensity school construction areas ($P_j = 1$).
- We are interested in estimating the effect of high-intensity construction, relative to low-intensity construction.
- $Y_{0i}$ = the outcome for individual i in the state where she receives low-intensity construction,
- $Y_{1i}$ = the outcome in the state where she receives high-intensity construction
- Define the causal effect of high intensity construction on those who actually received high-intensity construction as $E(Y_{1i}|P = 1) - E(Y_{0i}|P = 1)$ where the conditioning in the expectation indicates the category that the individuals actually falls into.
  - The latter term is not observed, called the counterfactual
- What we observe is $E(Y_{1i}|P = 1) - E(Y_{0i}|P = 0)$.

# Causal Effects

- Therefore, the assumption that you have to make for this estimator to be valid is
  $E(Y_{0i}|P = 0) - E(Y_{0i}|P = 1)$
- However, in this case, we can observe only one of the events.
- We need to use the differences across regions among the older cohort as a control.
- Then, we can estimate the differences across regions in the younger cohort relative to the differences in the older cohort.
- Define $DD = E(Y_{m1}|P = 1) - E(Y_{m0}|P = 0) - [E(Y_{n1}|P = 1) - E(Y_{n0}|P = 0)]$
- m = individual in the younger group (exposed to the program)
- n = inidividual in the older group (not exposed to the program)

# Difference-in-differences

- Think about this as the young-old diffrence in the high-intensive areas relative to the young-old difference in the low-intensive program areas:
  $DD = E(Y_{m1}|P = 1) - E(Y_{n1}|P = 1) - [E(Y_{m0}|P = 0) - E(Y_{n0}|P = 0)]$
- We are now interested in estimating the yound-old difference in the presence of the treatment, relative to what it would have been without the treatment.
  $E(Y_{m1}|P = 1) - E(Y_{n1}|P = 1) - [E(Y_{m0}|P = 1) - E(Y_{n0}|P = 1)]$
  - The latter part is the young-old difference if they were without treatment (in low-intensity areas)

# Difference-in-differences

- We need the assumption that the young-old difference in the low-intensity areas is equivalent to the young-old difference in the high-intensity areas, had they received the low-intensity intervention.
- Set the true effect to the estimated effect and solve:
  $E(Y_{m1}|P=1) - E(Y_{n1}|P=1) - [E(Y_{m0}|P=1)$
  $-E(Y_{n0}|P=1)] = E(Y_{m1}|P=1) - E(Y_{n1}|P=1)$
  $-[E(Y_{m0}|P=0) - E(Y_{n0}|P=0)]$
- Result:
  $E(Y_{m0}|P=1) - E(Y_{n0}|P=1) = E(Y_{m0}|P=0)$
  $-E(Y_{n0}|P=0)]$
  - The young-old difference in the intensity areas had they received the low-intensity intervention has to equal the young-old difference in the low-intensity areas.
- See Duflo (2001) Table 3

# Diffrence-in-differences

DD in the regression model:

$Y_{ijk} = \alpha_0 + \alpha_1 P_j + \alpha_2 T_k + \alpha_3 (P_j T_k) + \epsilon_{ijk}$

- $Y_{ijk}$ = the outcome for student in i in region j in birth cohort k
- $P_j$ = a dummy variable for high-intensity areas
- $T_k$ = a dummy variable representing the younger cohort
- $\alpha_3$ is the DD estimate for the effect of the program

# Regression

- High-low comparisons is throwing away a lot of useful data. The main question is that we want to know the incremental effect one more school.
- Regression design: Measure the incremental effect of 1 more school per 1000 students for the younger cohort by comparing across regions, controlling for differences across regions using the older cohort.
- The simplified of the regression in the paper:
  $Y_{ir} = \gamma_0 + \gamma_1 P_{ir} + \gamma_2 C_{ir} + \epsilon_{ir}$
  - $Y_{ir}$ = outcome of interest for individual i in region r
  - $P_{ir}$ = the number of schools constructed per 1000 students in region r for the younger cohort
  - $C_{ir}$= controls for differences across regions among the older cohort
- See Duflo (2001) Table 4

# Instrumental Variables

- We can actually use the school construction experiment to construct estimates of the returns to education
- This involved using the number of constructed schools in a region as an instrument for the level of a child's schooling.
- Instrumental variables (IV) estimates a regression of wages on the level of schooling predicted by the number of schools
- School construction, conditional on observable factors, must be exogenous, that is, not related to any third factor associated with wages.
- IV has to satisfy 2 properties:
  - I: "First stage" IV has to be correlated with the variable that it instruments (i.e., the level of schooling)
  - II: "Exclusion restriction" IV can only affect the dependent variable (wages) through the instrumented variable (the level of schooling)

# IV Estimation (simplified)

- Main equation:
  $ln(wage_{ir}) = \alpha_0 + \alpha_1 \hat{S}_{ir} + \alpha_2 C_{ir} + \nu_{ir}$
  - $\hat{S}_{ir}$ are predicted values of the level of schooling from the first stage
- First stage:
  $S_{ir} = \gamma_0 + \gamma_1 P_{ir} + \gamma_2 C_{ir} + \epsilon_{ir}$
- To get the main equation, we run the first stage to get $\hat{S}_{ir}$ and using the estimated values with other controls to get $\gamma_0, \gamma_1, \gamma_2$
- See Duflo (2001) Table 7

# Conclusions

- Schooling program raised educational attainment by 0.15 years on average for each school built per 1000 students in the region, and raised wages by about 2 percent.
- Using IV, returns to education estimates are 0.7 to 0.11 per year of additional schooling; not much different compared to OLS estimates and in line with the Psacharopoulos and Patrinos (2004)
- Broadly, IV strategies tend to produce similar estimates of returns to education compared with OLS.