Lecture 7: The effect of school buildings on schooling:
A “natural experiment”

Esther Duflo
February 25-March 1

1 The question we try to answer

Does the availability of more schools cause an increase in schooling levels?
Recall previous lecture:

\[ S^* = \frac{b - r}{\phi} \] (1)

Education chosen by the parents is a negative function of the cost of education and a positive function of the returns to education.

We have seen that a significant source of cost is likely to be the cost of transportation. This suggests that by increasing the number of schools, governments could increase education levels. Is this true in practice?

Many studies have shown that there is a correlation between school availability and educational outcomes: Children are more likely to attend school, and for a longer period of time in places where there are more schools.

However, as we discussed last time, correlation does not mean causality: These estimates can be biased upwards or downwards.

We have seen that the ideal case for evaluation is when we have access to a randomized experiment, but we do not have access to an experiment for all types of problems. Can we use real world, non-random, variation to try to answer this question (and other policy questions?).
2 School construction in Indonesia: Set up

2.1 The INPRES school construction program

Second five year plan (1974-79)—Oil shock.

- A large program:
    Number of schools multiplied by 2. 1 school for every 500 children.
  - A change in policy: Before 1973, no construction, ban on recruiting for public service positions.

- A program meant to favor low enrollment regions.

Allocation rule: Number of schools constructed in a district proportional to the number of children (ages 7 to 12) not enrolled in primary school.

2.2 Data

SUPAS 95. A survey done in 1995: After the children educated in these schools have completed their schooling, and have started working.

- 150,000 men born 1950-1972

- Variables: education, year and region of birth, wages.

2.3 Sources of variation

Two factors affect the intensity of the program.

- Year of birth: Examples
  - Born in 1962 or earlier: 12 or older in 1974. Not exposed to the program.
What would we find if we compare the education of those born before and after 1962?
Would this be a good measure of the impact of the program? Why?

- Region of birth: The government was targeting low enrollment regions ⇒ substantial
variation in program intensity across districts.

What would we find if we compare average education in regions that received more schools
to that in regions that receives few schools? Would this be a good measure of the program?
Why?

3 The “Difference in differences” methodology

- Basic idea
Suppose that there are two regions in the data: A “high program” region, and a “low program”
region.

Suppose that we have two age groups of individuals: “young people”, born after 1967 and who
could fully benefit from the schools; and “old people” born before 1962, and who could not
benefit at all from the schools.

So in total, we have four groups: YOUNG and High program, OLD and high program,....

Let us construct the average education for each of this group, and put them into a box. Use the
stata handout, and the template (table 3).

- Calculate, $D_{11}$, the difference between the “HIGH” and “LOW” average among the young:
  What do we find and why?

- Calculate $D_{21}$ the difference between the YOUNG and the OLD in the high program
  region: What do we find and why?

- Calculate, $D_{12}$ the difference between the “HIGH” and “LOW” average among the old:
  How does it compare to $D_{11}$? Why?

- Calculate the difference $DD_a = D_{11} - D_{12}$. How do you interpret it?

- Calculate the difference $DD_b = D_{21} - D_{22}$. How does it compare to $DD_a$? How do you
  interpret it?
• Could $DD_a$ or $DD_b$ be a good measure of the program?
  - Under what assumption?
  - Is assumption likely to be satisfied?

• Control experiment
We have a possibility to check that the assumption is not rejected in the available data.
Suppose we fill the same boxes, but we now compare the “OLD” to the “VERY OLD”. Neither of them benefited from the program: What do we expect to see if the assumption is satisfied?
What do we expect to see if the assumption is not satisfied?
Do it: What do we see?

4 Extending difference in differences

4.1 Using all the regional variation
There are 280 districts in Indonesia, and we know how many schools each district has received:
Grouping the region into two groups is throwing away some information!
Before, we had 2 regional groups, and 2 age groups, we formed 4 age-region groups. Now we have 280 regional groups, 2 age groups, how many groups can we form? What are these groups?
First, we form the average for each group (we can use the stata command: collapse). See an extract of the data set in the handout. We will note $SY_j$ the average education of the young in any region $j$, and $SO_j$ the average education of the young in any region $j$.
What can we do next?
- Take the difference between young and old in all the regions
- Plot the differences against the number of schools constructed per 1,000 children during the INPRES program (see graph)
- What do we see? What does this suggest?
- Suppose we run the regression:
  \[ SY_j - SO_j = \alpha P_j + v_j \] (2)
Where can you see the slope of this regression?
- See stata handout: What is the result of running this regression? What can we conclude?
- Under what assumption is this conclusion valid?
- Any suggestion to test this assumption?
- Do you see this test anywhere in the handout?

### 4.2 Using regional and age variation

The last generalization (after that, we are done!) is that we don’t have only 3 age groups (young, old, and very old): we have 23 age groups (everybody born between 1950 and 1972).

How many groups can we now form?

Note $S_{j2}, S_{j3}, ..., S_{jk}, ... S_{j24}$ the average education of people born in region $j$, and who were of age $2, 3, ... k, ... 24$, when the program started.

Suppose we run the regression:

$$S_{j2} - S_{j24} = \alpha_2 P_j + v_{j2}$$

What is $\alpha_2$?

Suppose we run the regression:

$$S_{j23} - S_{j24} = \alpha_{23} P_j + v_{j23}$$

What is $\alpha_{23}$? What should $\alpha_{23}$ be equal to?

In general, suppose that for all ages $k$ we run the regression:

$$S_{jk} - S_{j24} = \alpha_k P_j + v_{jk}$$

For what values of $k$ should we see a positive $\alpha_k$? (remember that children attend primary school until age 12). Should we see the coefficient be larger for younger children or older children?

Look at figure 2 in the handout: What does each dot represent? Do the dots have the expected pattern?
Constructing the difference in differences estimate.
Running the control experiment

<table>
<thead>
<tr>
<th>PANEL B: Experiment of interest</th>
<th>Years of education</th>
<th>Log( hourly wage)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Level of program in</td>
<td>Level of program in</td>
</tr>
<tr>
<td></td>
<td>Region of birth</td>
<td>Region of birth</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>age in</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1974</td>
<td>2 to 6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>12 to 17</td>
<td></td>
</tr>
<tr>
<td>Difference</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PANEL C: Control experiment</th>
<th>Years of education</th>
<th>Log(hourly wage)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Level of program in</td>
<td>Level of program in</td>
</tr>
<tr>
<td></td>
<td>Region of birth</td>
<td>Region of birth</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>age in</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1974</td>
<td>12 to 17</td>
<td></td>
</tr>
<tr>
<td></td>
<td>18 to 24</td>
<td></td>
</tr>
<tr>
<td>Difference</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>