14.48, Economics of Education
Prof. Frank Levy

Lecture 7: Do our regression estimates overestimate the impact of education on earnings? The Case of Ability Bias

Why do we even need instrumental variables for return to education estimations?
Why don't we just include years of education in our regressions?
Education is correlated with ability so we might actually be measuring the return to ability rather than the return to education.

David Card's Paper:

Instrument: the geographical proximity to a 4-year college
Why might proximity to college be correlated with higher educational attainments?

1) Lower costs to attending college since students have the option of living at home
2) Informational availability/access about the benefits of attending education

Are these effects uniform across families? No. These effects should be the largest for lower income families.

How do we test whether these effects are greatest for lower income families?
We could ask people by taking a survey, but our best bet is a regression.

$$(ED, age26)_i = \hat{\alpha} + \hat{\beta}_1(familyinc)_i + \hat{\beta}_2(mothersED)_i + \hat{\beta}_3(fathersED)_i + \varepsilon_i$$

Next, we use data to predict someone's educational attainment by age 26 without accounting for geographical proximity to a 4-year college. We split the data into two groups: those near and those not near a 4-year college. Among those who received 12 years of education or less, the group that lived near a 4 year college received 1.1 more years of education that those who did not. Among those who received more than 16 years of education, those who lived near a 4-year college received .3 more years of education that those who did not live near a 4-year college.

Using the instrumental variables method, we set up two regressions:

1) The reduced form equation:

$$\ln(wage)_i = \alpha + \beta_1(Age)_i + \beta_2(Age)_i^2 + \beta_3 z_i + \delta_1(familyinc)_i + ... + \mu_i$$

z_i=1 if you live near a 4-year college, 0 otherwise. Normally, we'd include years of education instead of z_i.

2)

$$ED_i = \gamma + \omega_1(z_i) + \omega_2(Age)_i + \omega_3(Age)_i^2 + \varepsilon_i$$

We can choose whether to include family background variables in this regression equation or not.

Results:

	ω_1
without	0.320
	(0.088)
family background	0.322
	(0.083)

	β_3
without	0.042
	(0.018)
family background	0.045
	(0.018)

The null hypothesis: Does this variable belong in this equation? What is the chance that the true coefficient for a variable should be equal to 0? We test this with a t-test. t>2 means a 5% chance or less that the variable should be equal to 0.

Problem of multi-colinearity: Independent variables in a regression are highly correlated. If the standard errors of ω_1 and β_3 greatly increase when we include family background variables in the regression, then we have a multi-colinearity problem.

How do we calculate the marginal return to education with the above information?
-We know that when z=1, equation 2 tells us we add 0.32 years of schooling
-We know that when z=1, equation 1 tells us we add 0.043 to ln(wage)

Combining this information:
(β_3/ω_1)= the implied marginal rate of return
(1/.32)(.043)=13%

We used the instrumental variables method because we were worried that ability bias inflated standard estimates of the rate of return to education. However, using the IV method actually increased the rate of return to education. Therefore, can we conclude that ability bias isn't a big problem? Is proximity to college even a good IV? The IV might be correlated with the availability of higher jobs in the region. Card, however, controls for this by including a measure of local salaries in the regression. Alternatively, living near a college might be a motivational factor that affects school performance positively while not living near a college works in the opposite way. This problem is much harder to address.

Esther's Duflo's Paper:

This paper uses the difference-in-different approach.

In regressions, we try to include control variables to isolate the effect of our main variable of interest. How do you factor out the effect of school on other changes in the Indonesian economy:

Facts:

-the Indonesian school building project was implemented to alleviate the low average level of education in the country and, specifically, to help the many areas of the country that lacked schools.
-from 1973-1979, the government built and staffed 61,000 primary schools and increased the number of schools in the country to 2 per 1000 children.
-enrollment rates for 7-17 year olds: 69% in 1973, 83% in 1978.
-school shortages were concentrated in certain areas of the country

Theory:
More school lead to more education, which, in turn, leads to high wages. For that to be true, schools have to be built in area that needed them the most. Otherwise, the program would just be lowering class sizes in areas that already have schools and wouldn't be helping the needy areas.

To test this, we use the following regression:

$$\ln(\#ofschoolsbuilt,1973-1978)_i = \alpha + \beta_1 \ln(children)_i + \beta_2 \ln(1-EnrollmentRate,1973)_i + \varepsilon_i$$

where i corresponds to the region in question.

Results:
β_1: 0.78, (0.027)
β_2:0.12, (0.038)---this shows that the program money was allocated correctly to regions in need

Set up:
We look at provinces that got school and compare 1975 and 1989 wages.
If 1975 wages < 1989 wages, are we done with out analysis? No. There are many other factors that could be affecting wages in that time period outside of the construction of new school.

How do we isolate the effect of the additional schools?
We need to use a control group of people who just missed qualifying for the new schools because they were too old. We separate those who were aged 2-6 in 1974 from those who were aged 12-17 in 1974 and just missed being affect by the education intervention.

Educational attainment in 1988:
If the younger group of people got more education that the older group, are we done with our analysis? No. Again, schooling might not be the only cause for increased educational attainment. The government could have launched a national ad campaign at the time of the school construction encouraging attending school, which could have also increased educational attainment in all regions.
Therefore, we need to look at both regions that did receive schools and that did not receive schools.

Results:

	Regions		
	Received schools	***Didn't receive schools***	***Difference***
Ages 2-6, 1974	8.49	9.76	-1.27
Ages 12-17, 1974	8.02	9.4	-1.39

The difference-in-difference method isolates the effect of new school construction on educational attainment since we compare both regions that received school with regions that didn't receive school and students that benefited from the school construction and students that just missed benefiting from the school construction.

Duflo performs a similar analysis on wages and finds a similar effect.

Signaling/Screen model:
Suppose you are assigned the job of selecting 7th graders to participate in an MIT summer program and you want to select the students that you think will do well in the program.

Given the following information on a student, would you accept, reject, or waitlist him?
-He attends a middle school near Spanish Harlem
-Has a Hispanic surname
-A high proportion of students at his school receive free or reduced priced lunch

Given this new information, would you change your decision?
-His school is home to a top 5 national middle school chess team that won the national championship the past 2 years

We have a signal that conveys positive academic information about this student.

AP classes also serve as a signal of academic achievement and potential in college admissions.