**Session 9 and 10 (class 4)**

1. **Motivation**

Consider the model below:

-------------------------(1)

Suppose . and are uncorrelated by construction. The variables in are the only reason why and are correlated. E[] = 0. If we can observe we would happily include that in equation 1 to obtain the following long regression.

--------------------------(2)

If we do not observe , we face an omitted variable bias in estimating , the parameter of interest. Consider our usual example. Let be education in years (schooling), be ability/motivation, and be log wages. The regression estimating the effects of education on wages does not yield causal effects, if we do no observe and control for the individual’s ability or motivation levels. If we ignore ability/motivation, from equation 1 is biased upwards. What methods have we seen until now to deal with the omitted variable bias:

1. Ignore the missing variable

This approach can be satisfactory if the nature of biases and the estimates we find reinforce our hypothesis. Suppose we hypothesize that skills training positively affects wages. By ignoring ability, we find a positive coefficient on the training variable. But since we know that the direction of bias is negative, we can be assured that the estimates we obtain by ignoring ability are a lower bound for the actual magnitude of positive effects. Even though we do not obtain accurate estimates, the regression tells us something important. Unfortunately, ignoring ability in estimating the effects of education on wages yields a positive bias, so does not assure us of the sign of the coefficient on education.

1. Proxy variable

We can try to find a suitable proxy for the missing variable and include that in the regression. For ability, proxies such as a score in a competitive examination may be used. But it is hard to find proxies for missing variables.

1. Individual Fixed effects

Assuming the missing variable does not change over time, include person fixed effects using panel data. However, this approach requires panel data which may not be easily available in many settings. The assumption that the missing variable does not change over time may be strong. Ability may increase with age.

1. **Instrumental variable (IV) regression**

We have another method to tackle the missing variable bias. Let us say that the researcher has access to a variable , which is correlated with the causal variable of interest, , but uncorrelated with any other determinants of the dependent variable. Then, is called an instrumental variable for or instrument to .

1. The first criterion means that cov(,) ≠ 0, and is also called the relevance-criterion.
2. The second is called the exclusion criterion and means cov (,) = 0, or equivalently is uncorrelated with and .

The estimator: The instrumental variable estimate using z as instrument for x, if we run the regression in equation (1), is the following:

The second equality is useful because it is easier to think in terms of regression coefficients than in terms of covariances. is simply the ratio of coefficient of regression of on (reduced form regression), and coefficient of regression of on (first stage). The estimate is predicted on the assumption (no.1) that the first stage is non-zero (denominator); that is, the coefficient of on is non-zero; in other words, the instrument *should* have a relevant effect on the endogenous variable of interest. If the denominator is zero, notice that is undefined. The second assumption (no.2) indicates that the instrument is as good as randomly assigned. That is, it is not correlated with the error term. That is, after controlling for and other explanatory variables, is not correlated with .

OLS, a special case of IV: Substitute z for x in the formula for. Notice that the above becomes the OLS estimator, which is Cov(y,x)/Var(x).

Testing each criterion: You can easily test the relevance criterion, as to whether and are correlated. This is simply by checking if the first stage regression coefficient is significant. But notice that the exclusionary criterion is not testable when there is just one instrument. One needs to go by intuition, introspection or theory.

Example 1: Let us take an example related to wage regressions.

We do not observe ability/motivation which creates a missing variable bias. What can be potential instruments for ability?

1. How about your aadhar card number? It is not correlated with your education, of course. So it satisfies the exclusion criteria. But does it satisfy the relevance criteria? NO. Your aadhar number is not related to your wages at all. So this cannot be a good instrument.
2. How about your IQ score as an instrument for education? This satisfies the relevance criteria. That is, education could be correlated with your IQ score. But does it satisfy the exclusionary criteria? No, because IQ score is obviously correlated to the missing variables such as ability and motivation level, sitting in . But note that IQ can be used directly as a proxy (rather than an IV). See above notes.
3. Mother’s or father’s education. This satisfies the relevance criteria because obviously parents’ education affects child education. More educated parents may want their kids to be more education, compared to less educated parents. This also satisfied the exclusion criteria because in many settings, it may be reasonable to assume that parents’ education does not affect their children’s wages. However, this may not be true if parents’ education is related to their child’s social network, which can indirectly affect their children’s job opportunities and their wages. Also, mother’s education is not a valid instrument if mother education can influence children’s ability through better nourishment, nutrition and care in the children’s younger days.
4. Number of siblings. This is a good instrument if number of siblings is related to lower average education per child in the household (relevance criteria) and if number of siblings is not necessarily related to your innate ability (exclusionary criteria). Explore if this is true.

Example 2: what is the causal effect of skipping classes on final exam score?

where score is the final exam score, and skipped is the total number of lectures missed during the academic year.

* 1. What are the missing variables in in the above regression? Ability and motivation because more able and highly motivated students may miss fewer classes. So, a simple regression of skipped on score may not give causal effects.
  2. What is a good Instrument Variable for ? We need a variable that has no effect on score but has an effect on skipping classes. What may be such a variable? How about distance from home to class? In rural settings, where schools are fewer and spread out, distance to school is a major determinant of attendance. Hence, distance could be a good instrument. But what if people who live far have lower income, and income is related to motivation? This is probably true, but if we control for income in the regression, distance is a good IV.

Standard error of an IV estimate

The variance of the instrument variable estimate is given below, and the standard error is simply the square root of this.

is the residual variance, is the sum of squares deviation of x, and is the from the regression of x on z. Compare this to the variance of , the OLS estimate.

Since is always >0, > . If x and z are only slightly correlated, then is small, and is much larger compared to . The more correlated x and z are, the closer is to 1, and smaller is the sampling variance of . In case z and x are the same, , and we just get the OLS variance.

Example implemented in stata: Open MROZ.dta (W, page 386). We want to estimate returns to education by running a regression of log wage on education. Run this regression.

The OLS estimates indicate 8% return on education.

(.129) (.010)

N = 428; R2 = .141

Use father’s education now as an instrumental variable (*fathereduc*). We have to assume and maintain that father’s education is not correlated with . But we can check if the relevance criterion is satisfied. Run the regression of *educ* on *fathereduc.* The results are as follows:

(.198) (.020)

N = 428; R2 = .195

The t-statistic on *fathereduc* is 13.52, indicating that is it statistically significantly different from zero, and serves as a relevant instrument. The instrumental variables regression results of educ on log wages using fathereduc as instrument is as follows:

(.309) (.024)

N = 428; R2 = .12

The returns to education is 5.3%, suggesting that there was positive bias in the OLS regression because of omitted variables. The standard error of IV is 2.4 times that of the OLS estimate, for reasons we discussed earlier.

Using binary variables as instruments

Simplest IV estimator using a single dummy instrument can be used to estimate a model with one endogenous regressor and no covariates. Without covariates the causal regression model is:

Now assume is a binary instrument for , which takes the value 1 with probability p. we can show that:

We have an analogous formula for

**Based on the standard IV formula,**

The above is called the Wald estimator.

Example from quarter of the birth study (Angrist and Krueger, 1991): Good instruments come from institutional knowledge and ideas about the process determining the variable of interest. Angrist and Krueger (1991) gives an example of how one can find instruments arising from institutions such as compulsory schooling laws and school age-start policies in the US. In the US, most states require students to enter school in the calendar year in which they turn 6. School start age is therefore a function of date of birth. Specifically, those born late in the year are young for their age. If August 15th is the school start date, those born late in the year join a little before they turn 6. Those born earlier in the year, say first quarter are typically more than 6.5 years old. Compulsory schooling laws stipulate that kids should definitely go to school until their 16th birthday. After that, there is no legal requirement on behalf of parents to have their kids go to school. There is neat random variation created in the number of months kids obtain education based on which quarter they were born, driven basically by those who do not finish high school but drop out on the 16th birthday.

Examine the paper’s Figure I and note that years of education follows a certain pattern based on the quarter of birth, and those born in higher quarters end up with more education. This graph is the author’s justification for using quarter of birth as instrument for years of education. Figure II shows the reduced form relationship between wages and quarter of birth. Men born in earlier quarters earn less, similar to Figure 1’s relationship between education and quarter of birth. Because individual’s date of birth is probably unrelated to his or her innate ability, motivation, or family connections, Angrist and Kruger argues that it is credible to argue that quarter of birth is a good instrument for education. Examine Table III for the results, which uses the Wald estimator.

Example from Angrist (1990)’s study on Vietnam War veterans

Angrist looks at the effects of Vietnam era military service on the earnings of veterans using a Wald estimator. In the 1960s and 1970s, young American men were at the risk of being drafted into military service. To make it a fair process, a draft lottery was set up and conscription was based on the lottery results. For an entire cohort of 19 year olds, specifically, for each year between 1970 and 1972, random numbers were assigned to each birth date. Men with lottery numbers below a cut off were eligible for the draft, and those above the cut off were not required to. The lottery thus serves as a good instrument to determine participation in the war, to study the effects on earnings.

What are some reasons because of which the lottery may not be a good instrument? In practice, many draft eligible men were still exempted for health reasons, while many who were draft exempt still participated in the war. So veteran status was not still completely randomly decided based on lottery results, but the two were highly correlated and draft eligibility proves to be a good instrument. See table 2 and table 3 in the paper for the Wald estimates, the ratio of eligibility effect on earnings to eligibility effect on conscription.

**Weak instruments**

**Hands on**

Card (1993) uses College Proximity as an Instrument for Educational Attainment. The data for the paper comes from the National Longitudinal Survey of Youth, and is available in his website (<http://davidcard.berkeley.edu/index.html>) and in the shared dataset, carddata.dta. Open the data, and download the paper to understand it’s main idea.

1. What does Figure 1 demonstrate, and why is it important to the credibility of proximity to college as an instrumental variable?
2. Compute an OLS regression estimate of the returns to schooling following the specification in Table 2, column 2.
3. Compute the 1st Stage, reduced form, and the indirect least squares estimate of the return to schooling.
4. Compute the IV estimator using the ivregress command
5. Use the results of ivregress to perform a Hausman test on whether the OLS and 2SLS estimates are significantly different from each other. What does this regression tell you about whether education is endogenous? Explain.
6. Is the evidence consistent with the view that the OLS estimates of the return to schooling are upward biased because the more “able" (in unobservable ways) are more likely to have more school?
7. Is the evidence consistent with the view that the OLS estimates of the return to schooling are downward biased because only the “less" able get lots of schooling?

**References**

[Angrist and Krueger, November 1991](http://www.sciencedirect.com/science/article/pii/0305750X94900078#BIB4). “Does compulsory school attendance affect schooling and earnings?”. The Quarterly Journal of Economics, Vol. 106 (No. 4) (1991)

Card, D. (1993). Using Geographic Variation in College Proximity to Estimate the Return to Schooling. Working Paper 4483, National Bureau of Economic Research.

Angrist, Joshua D.1990. "Lifetime Earnings and the Vietnam-era Draft Lottery: Evidence from Social Security Administrative Records." American Economic Review 80(3), 313-36