Handout 4 – Instrumental Variables Estimation and Two Stage Least Squares

Chunyu Qu

Oct 25, 2021

1. Fundamentals for IV Estimate

1.1. Three alternative ways

- **Ignore the problem and suffer the consequences of biased and inconsistent estimators**
  Seldom works if the estimates are coupled with the direction of the biases for the key parameters.
  Ex: For example, if we can say that the estimator of a positive parameter, say, the effect of job training on subsequent wages, is biased toward zero and we have found a statistically significant positive estimate, we have still learned something: job training has a positive effect on wages, and it is likely that we have underestimated the effect.

- **Try to find and use a suitable proxy variable for the unobserved**
  Not always possible to find a good proxy

- **Assume that the omitted variable does not change over time and use the fixed effects or first-differencing methods**
  Be careful with the parallel trend for DiD

1.2. IV Approach

(1) Argument

The availability of an instrumental variable can be used to estimate consistently the parameters in linear regression, and thus to identify the parameters. Identification of a parameter in this context means that we can write β_i in terms of population moments that can be estimated using a sample of data.
Ex: Consider the wage example

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 abil + e$$

Suppose, however, that a proxy variable is not available (or does not have the properties needed to produce a consistent estimator Then, we put abil into the error term, and we are left with the simple regression model

$$\log(wage) = \beta_0 + \beta_1 educ + u$$

where u contains abil. This is gives a biased and inconsistent estimator of educ if educ and abil are correlated.

In order to obtain consistent estimators of when x and u are correlated, we need some additional information. The information comes by way of a new variable z that satisfies
(1) z is uncorrelated with u $cov(z,u) = 0$ (**instrument exogeneity**),
(2) z is correlated with x, $cov(z,x) \neq 0$ (**instrument relevance**).
Then, we call z an **instrumental variable** for x, or sometimes simply an instrument for x.

A simple inference given the model that $y = \beta_0 + \beta_1 x + u$ is that

$$cov(z,y) = \beta_1 cov(z,x) + cov(z,u)$$
$$\Rightarrow \beta_1 = \frac{cov(z,y)}{cov(z,x)}, given\ cov(z,u) = 0$$

Given a sample of data on *x*, *y*, and *z*, it is simple to obtain the IV estimator as above.

- **Examine instrument exogeneity**

Sometimes, we might have an observable proxy variable for some factor contained in u, in which case we can check to see if z and the proxy variable are roughly uncorrelated. Of course, if we have a good proxy for an important element of u, we might just add the proxy as an explanatory variable and estimate the expanded equation by ordinary least squares.

For the covariance between z and the unobserved error u, we cannot generally hope to test the u from OLS and z (i.e. obtain u from regressing y on x, then check $cov(z,u)$). The entire reason for moving beyond OLS is that we think the OLS estimators are inconsistent due to correlation between x and u. Therefore, in computing the OLS residuals $\hat{u} = y - \widehat{\beta_0} - \widehat{\beta_1} x$, we are not getting useful estimates of the u. Therefore, we can learn nothing by studying the correlation between z and u. The bottom line is that, in the current setting, we have no way of testing **instrument exogeneity** unless we use external information.

- **Examine instrument relevance**

Test the hypothesis

$$H_0: \pi_1 = 0$$

Where $x = \pi_0 + \pi_1 z + v$.

(2) IV proposal examples

- *IQ* is a good candidate as a proxy variable for *abil*, it is not a good instrumental variable for *educ* because it violates the instrument exogeneity requirement.
- Last digit of an individual's Social Security Number satisfies the first requirement but not the relevance requirement.
- Mother's education is positively correlated with child's education, as can be seen by collecting a sample of data on working people and running a simple regression of *educ* on *motheduc*, which means relevance is satisfied. The problem is that mother's education might also be correlated with child's ability.
- **Number of siblings while growing up** is associated with lower average levels of education. Thus, if it is uncorrelated with ability, it can act as an instrumental variable for *educ*.

- Angrist and Krueger (1991), in their simplest analysis, came up with a clever binary instrumental variable for *educ*, using census data on men in the United States. Let *frstqrt* be equal to one if the man was born in the first quarter of the year, and zero otherwise. It seems that the error term should be unrelated to quarter of birth. But *frstqrt* also needs to be correlated with educ. It turns out that years of education do differ systematically in the population based on quarter of birth. Angrist and Krueger argued persuasively that this is due to compulsory school attendance laws in effect in all states. Briefly, students born early in the year typically begin school at an older age. Therefore, they reach the compulsory schooling age (16 in most states) with somewhat less education than students who begin school at a younger age. For students who finish high school, Angrist and Krueger verified that there is no relationship between years of education and quarter of birth. Because years of education varies only slightly across quarter of birth, which means $R^2_{x,z}$ is very small, Angrist and Krueger needed a very large sample size to get a reasonably precise IV estimate. Using 247,199 men born between 1920 and 1929, the OLS estimate of the return to Education was .0801 (standard error .0004), and the IV estimate was .0715 (.0219); these are reported in Table III of Angrist and Krueger's paper. Note how large the t statistic is for the OLS estimate (about 200), whereas the t statistic for the IV estimate is only 3.26. Thus, the IV estimate is statistically different from zero, but its confidence interval is much wider than that based on the OLS estimate.
- An interesting finding by Angrist and Krueger is that the IV estimate does not differ much from the OLS estimate. In fact, using men born in the next decade, the IV estimate is somewhat higher than the OLS estimate. One could interpret this as showing that there is no omitted ability bias when wage equations are estimated by OLS. However, the Angrist and Krueger paper has been criticized on econometric grounds. As discussed by Bound, Jaeger, and Baker (1995), it is not obvious that season of birth is unrelated to unobserved factors that affect wage. As we will explain in the next subsection, even a small amount of correlation between z and u can cause serious problems for the IV estimator.
- Consider the following example:

  $$score = \beta_0 + \beta_1 skipped + u$$

  Let's consider distance between living quarters and classrooms as a candidate for IV. Especially at large universities, some living quarters will be further from a student's classrooms, and this may essentially be a random occurrence. Some students live off campus while others commute long distances. Living further away from classrooms may increase the likelihood of missing lectures due to bad weather, oversleeping, and so on. Thus, *skipped* may be positively correlated with *distance*; this can be checked by regressing *skipped* on *distance* and doing a *t* test, as described earlier. For exogeneity, some factors in *u* may be correlated with *distance*. For example, students from low-income families may live off campus; if income affects student performance, this could cause *distance* to be correlated with *u*.
- For policy analysis, the endogenous explanatory variable is often a binary variable. For example, Angrist (1990) studied the effect that being a veteran of the Vietnam War had on lifetime earnings. A simple model is

  $$\log(earns) = \beta_0 + \beta_1 veteran + u$$

  where veteran is a binary variable. The problem with estimating this equation by OLS is that there may be a self-selection problem, perhaps people who get the most out of the military choose to join, or the decision to join is correlated with other characteristics that affect earnings. These will cause veteran and u to be correlated. Angrist pointed out that the Vietnam draft lottery provided a natural experiment that created an instrumental variable for \
  veteran. Because the numbers given were (eventually) randomly assigned, it seems plausible that draft lottery number is uncorrelated with the error term u. But those with a low enough number had to serve in Vietnam, so that the probability of being a veteran is correlated with lottery number. If both of these assertions are true, draft lottery number is a good IV candidate for veteran.

(3) Key notes

- IV approach may not be necessary at all if a good proxy exists for student ability, such as cumulative GPA prior to the semester.
- It is important to address the discussion of exogeneity assumption, sometimes from references. Arguments for why a variable z makes a good IV candidate for an endogenous explanatory variable x should include a discussion about the nature of the relationship between x and z.
- If in the sample of data we find unexpected relevance, such as negative relationship between *edu* and *motheduc*, then the use of mother's education as an IV for child's education is likely to be unconvincing. In the example of measuring whether skipping classes has an effect on test performance, one should find a positive, statistically significant relationship between *skipped* and *distance* in order to justify using *distance* as an IV for skipped: a negative relationship would be difficult to justify [and would suggest that there are important omitted variables driving a negative correlation—variables that might themselves have to be included in the model.
- One feature of the IV estimator is that, when x and u are in fact correlated—so that instrumental variables estimation is actually needed—it is essentially never unbiased. This means that, in small samples, the IV estimator can have a substantial bias, which is one reason why large samples are preferred.
- When compare the IV estimate with the OLS estimate, if the IV confidence interval include the OLS estimate, even if they are largely different, we cannot say if the difference is statistically significant. The presence of larger confidence intervals is a price we must pay to get a consistent estimator of the return to education when we think educ is endogenous.
- Weak instrument: Weak correlation between z and x can have even more serious consequences, the IV estimator can have a large asymptotic bias even if z and u are only moderately correlated. In the Angrist and Krueger (1991) example mentioned earlier, where x is years of schooling and z is a binary variable indicating quarter of birth, the correlation between z and x is very small. Bound, Jaeger, and Baker (1995) discussed reasons why quarter of birth and u might be somewhat correlated.
- Card (1995) used wage and education data for a sample of men in 1976 to estimate the return to education. He used a dummy variable for whether someone grew up near a four-year college (*nearc4*) as an instrumental variable for education. In a log(*wage*) equation, he included other standard controls: experience, a black dummy variable, dummy variables for living in an SMSA and living in the South, and a full set of regional dummy variables and an SMSA dummy for where the man was living in 1966. In order for *nearc4* to be a valid instrument, it must be uncorrelated with the error term in the wage equation—we assume this—and it must be partially correlated with *educ*.
- Computing R-Squared after IV Estimation. Unlike in the case of OLS, the R-squared from IV estimation can be negative because SSR for IV can actually be larger than SST. Although it does not really hurt to report the R-squared for IV estimation, it is not very useful, either. If our goal was to produce the largest R-squared, we would always use OLS. IV methods are intended to provide better estimates of the ceteris paribus effect of x on y when x and u are correlated; goodness-of-fit is not a factor. A high R-squared resulting from OLS is of little comfort if we cannot consistently estimate the parameters.

2. IV Estimation Process

(1) 2SLS: IV Estimate of the Simple Linear Regression

For

$$y = \beta_0 + \beta_1 x + u,$$

we follow the steps below for the IV estimate

Step 0: Run the OLS.

Step 1: Regress the endogenous variable x on the IV z, generate the estimated x, and testify the relevance assumption

Step 2: Regress y on the estimated x.

(2) 2SLS: IV Estimate of the Multiple Linear Regression

For

$$y = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u,$$

We call this a **structural equation**, where we assume that y_1 is endogenous, z_1 is exogenous, y_2 is is suspected of being correlated with u. An example is

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + u_1$$

Where we assume that *exper* is exogenous, but we allow *educ* be correlated with u_1. We need another exogenous variable—call z_2, as an IV for y_2. The essential conditions to be satisfied is $E[z_1 u_1] = E[z_2 u_1] = 0 = E[u_1]$.

Step 0: Run the OLS.

Step 1: Regress the endogenous variable y_2 on the IV z_2 and the exogenous variable z_1, generate the estimated y_2, and testify the relevance assumption, i.e. we work on the **reduced form equation** as follows

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + v_2$$

Where by construction, $E[v_2] = cov(z_1, v_2) = cov(z_2, v_2) = 0$. And we test $H_0: \pi \neq 0$. In an other word, after partialling out z_1, y_2 and z_2 are still correlated.

Step 2: Regress y_1 on the estimated y_2 and z_1.

We can add more **exogenous explanatory variable**s similarly as

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \cdots \beta_k z_{k-1} + u,$$

Where we propose z_k is the IV for y_2, and $E[z_i u_1] = 0 = E[u_1], i = 1,2, \ldots k$

And the reduced form for it is

$$y_2 = \pi_0 + \pi_1 z_1 + \cdots \pi_{k-1} z_{k-1} + \pi_k z_k + v_2,$$

and we need some partial correlation between z_k and y_2. In this case, the 2SLS becomes

Step1: Regress y_2 on $z_1 \ldots z_k$, and generate estimated $\widehat{y_2}$

Step2: Regress y_1 on $z_1 \ldots z_{k-1}$,and $\widehat{y_2}$.

(3) A Single Endogenous Explanatory Variable

Suppose we have to exogenous variables z_2 and z_3 that can be candidate for IV for y_2, we could just use each as an IV, as in the previous section. But then we would have two IV estimators, and neither of these would, in general, be efficient. Since each of $z_1, z_2, and\ z_3$ is uncorrelated with u_1, any linear combination is also uncorrelated with u_1, and therefore any linear combination of the exogenous variables is a valid IV. To find the best IV, we choose the linear combination that is most highly correlated with y_2, This turns out to be given by the reduced form equation for y_2

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + v_2$$

Where $E[v_2] = cov(z_1, v_2) = cov(z_2, v_2) = cov(z_3, v_2) = 0$. Then the best IV is the linear projection of y_2 on $z_1, z_2, and\ z_3$,

$$y_2^* = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3$$

For this IV not to be perfectly correlated with z_1, we need at least one of $\pi_2\ or\ \pi_3$ to be different from zero.

(4) Multicollinearity and 2SLS

Multicollinearity can be even more serious with 2SLS. There are two reasons why the variance of the 2SLS estimator is larger than that for OLS. First $\widehat{y_2}$ by construction, has less variation than y_2 (SST=SSE+SSR), the variation in y_2 is the SST, while that for $\widehat{y_2}$ is the SSE from the first stage. Second, the correlation between $\widehat{y_2}$ and exogenous variables is often much higher than the correlation between y_2 and these variables. This essentially defines the multicollinearity problem in 2SLS. Consider Card (1995) example, when educ is regressed on the exogenous variables (excluding nearc4), $R^2 = 0.475$, this is a moderate degree of multicollinearity, but the important thing is that the OLS standard error on $\widehat{\beta_{educ}}$ is small. When we obtain the first stage fitted values, $\widehat{educ}$, and regress these on the exogenous variables $R^2 = 0.995$ which indicates a very high degree of multicollinearity between $\widehat{educ}$ and the remaining exogenous variables in the table.

3. Tests for IV Estimation

3.1. Testing for Endogeneity

The 2SLS estimator is less efficient than OLS when the explanatory variables are exogenous; as we have seen, the 2SLS estimates can have very large standard errors. Therefore, it is useful to have a test for endogeneity of an explanatory variable that shows whether 2SLS is even necessary. Obtaining such a test is rather simple. If, in the end, the 2SLS estimates are chosen, one should obtain the standard errors using built-in 2SLS routines. We can also test for endogeneity of multiple explanatory variables. For each suspected endogenous variable, we obtain the reduced form residuals. Then, we test for joint significance of these residuals in the structural equation, using an F test. Joint significance indicates that at least one suspected explanatory variable is endogenous. The number of exclusion restrictions tested is the number of suspected endogenous explanatory variables.

Here is the recipe of testing for Endogeneity of a Single Explanatory Variable:

Step 1. Estimate the reduced form for y_2 by regressing it on all exogenous variables (including those in the structural equation and the additional IVs). Obtain the residuals, $\widehat{v_2}$.

Step 2. Add $\widehat{v_2}$ to the structural equation (which includes y_2) and test for significance of $\widehat{v_2}$ using

an OLS regression. If the coefficient on $\widehat{v_2}$ is statistically different from zero, we conclude that y_2 is indeed endogenous. We might want to use a heteroskedasticity-robust t test.

3.2. Testing Overidentification Restrictions

When we introduced the simple instrumental variables estimator, we have now seen that, even in models with additional explanatory variables, the second requirement can be tested using a t test (with just one instrument) or an F test (when there are multiple instruments). In the context of the simple IV estimator, we noted that the exogeneity requirement cannot be tested. However, if we have more instruments than we need, we can effectively test whether some of them are uncorrelated with the structural error.

Testing Overidentifying Restrictions:
Step 1. Estimate the structural equation by 2SLS and obtain the 2SLS residuals, $\widehat{u_2}$.
Step 2. Regress $\widehat{u_1}$ on all exogenous variables. Obtain the R-squared, say, R_1^2.
Step 3. Under the null hypothesis that all IVs are uncorrelated with u_1, $nR_1^2 \sim \chi_q^2$, where q is the number of instrumental variables from outside the model minus the total number of endogenous explanatory variables. If nR_1^2 exceeds (say) the 5% critical value in the χ_q^2 distribution, we reject H_0 and conclude that at least some of the IVs are not exogenous.