ØAMET4100 · Spring 2019
Lecture Note 7

Instructor: Fenella Carpena

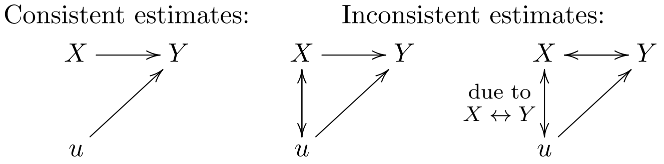
March 7, 2019

This lecture note provides a review of regression with instrumental variables (Stock & Watson, Chapter 12). This lecture note is not intended to be a comprehensive review of lecture or the textbook, since there is a lot more material than we have time to cover. However, I have tried to focus on the concepts which I believe are necessary to be successful in our class.

1 Introduction

In previous lectures, we discussed several threats to internal validity that led to a correlation between the regressor X and the population regressor error u. These threats include OVB, measurement error, and simultaneity. We learned that when there is OVB, we can address the bias by: (1) including the omitted variable in the regression, but this is only possible if you have data on the omitted variable, and (2) using panel data and fixed effects, but doing so accounts for only those omitted variables that are constant across time within entities or those that vary over time but is constant across entities.

The following diagrams show cases where OLS yields consistent or inconsistent estimates of the causal effect (here, the arrows indicate the direction of the relationship between the variables):



In this lecture, we will discuss a method called **instrumental variables (IV) regression.** This method allows us to obtain consistent estimators when the regressor X is correlated with the population error term u. Thus, IV regression allows us to overcome threats to internal validity that arise not only because of OVB, but also because of measurement error and simultaneity. So how does IV regression work? Intuitively, it allows us to separate the variation in X into two parts: (1) a part that is correlated with u (which is the part that causes the problem), and (2) a part that is uncorrelated with u. The **instrument** isolates the second part, which in turn permits consistent estimation of the regression coefficients.

2 IV Regression: 1 endogenous regressor, 1 instrument

2.1 Overview

Consider the simple linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + u_i \tag{1}$$

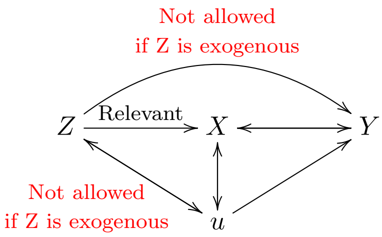
where as usual, u_i contains all omitted factors affecting Y_i. If X_i and u_i are correlated, the OLS estimator is inconsistent. Because X_i is correlated with the error term, we say that X_i is an **endogenous variable**. In contrast, variables that are not correlated with the error term are called **exogenous variables**.

To implement an IV regression, we need a third variable Z_i, which represents the **instrumental variable** or **instrument** for short. A **valid** instrumental variable must satisfy two conditions:

1. Relevance: $corr(Z_i, X_i) \neq 0$. If an instrument is relevant, this means that the variation in the instrument is related to the variation in X_i.

2. Exogeneity: $corr(Z_i, u_i) = 0$. If an instrument is exogenous, then the part of the variation in X_i that is captured by the instrument is not correlated with u_i.

If the above two conditions hold, then the instrument captures movements in X_i that is exogenous. This exogenous variation can then be used to estimate β_1. Understanding the above two conditions is **very important** in IV regression, and note that in practice, exogeneity is often very hard to satisfy. The following diagram illustrates the above two conditions:



As we can see in the above diagram, if the instrument Z is relevant, then Z affects Y only through X.

2.2 Estimation

If the instrument Z_i is relevant and exogenous, we can estimate β_1 using an IV estimator called **two stage least squares (TSLS or 2SLS)**. As the name suggests, TSLS proceeds in two stages.

First Stage The first stage begins with a population regression linking X_i and Z_i, written as:

$$X_i = \pi_0 + \pi_1 Z_i + v_i. \tag{2}$$

We estimate the above regression using OLS and obtain the predicted values $\widehat{X}_i = \widehat{\pi}_0 + \widehat{\pi}_1 Z_i$. Note that because Z_i is exogenous, $cov(Z_i, u_i) = 0$. Therefore, the predicted values $\widehat{X}_i$ contain the component of X that is uncorrelated with u_i.

Second Stage In the second stage, we regress Y_i on $\widehat{X}_i$ to obtain the TSLS Estimators $\widehat{\beta}_0^{TSLS}$ and $\widehat{\beta}_1^{TSLS}$ where

$$\widehat{\beta}_1^{TSLS} = \frac{s_{Y\widehat{X}}}{s^2_{\widehat{X}}} = \frac{s_{ZY}}{s_{ZX}} \tag{3}$$

How can we tell whether the Z_i satisfies relevance and exogeneity? Relevance can be tested directly: after estimating the first stage regression, we can conduct the hypothesis test $H_0 : \pi_1 = 0$ vs. $H_1 : \pi_1 \neq 0$. The rule of thumb is that the F-stat for the (joint) significance of the instrument(s) in the first-stage should be greater than 10. If the F-stat is greater than 10, relevance is satisfied and we say we have a **strong instrument**. If the F-stat is less than 10, we say we have a **weak instrument**. In this case, TSLS no longer gives us reliable estimates.

Why are weak instruments a problem? Consider the extreme case where there is no correlation between Z and X, then from the equation for β_1^{TSLS} above, we see that β_1^{TSLS} is undefined because the denominator is zero. More precisely, when the instruments are weak, the sampling distribution of the TSLS estimator is no longer approximately normal even if the sample size is large.

For exogeneity, there is no direct test because u_i is unobserved. Hence, we need to use economic theory, expert knowledge, and intuition. Nevertheless, there is a special case where we can provide evidence of exogeneity; this can be done in when we have more instruments than endogenous variables, which is discussed later in Section 4.2.

2.3 Example: Estimating the Price Elasticity of Demand for Cigarettes

To see how IV regressions work in practice, let's consider the following example. Smoking is a big public health problem, and the government wants to tax cigarettes to reduce smoking. What tax should be put in place if we want to decrease cigarette consumption by 20%? To answer this question, we need to know the price elasticity of demand for cigarettes (i.e., $\%\Delta Q \div \%\Delta P$), and our goal in this exercise is to estimate this elasticity. For example, if the elasticity of demand is -0.5, this means that to decrease consumption by 20%, prices must increase by 40%.

To find the elasticity, we need to estimate the demand equation

$$ln(Q_i^D) = \beta_0 + \beta_1 ln(P_i) + u_i \tag{4}$$

where we expect that $\beta_1 < 0$ (i.e., law of demand, as the price of the good increases, demand decreases). The coefficient we are interested in is β_1, which represents the price elasticity of demand.

Now, suppose that we have data from 48 U.S. states, and the data contain the average price per pack of cigarettes including all taxes (our measure of P_i) and the number of packs of cigarettes sold per capita in the state (our measure of Q_i^D). Would we obtain a consistent estimate of β_1 by simply regressing $ln(Q_i^D)$ on $ln(P_i)$? No, because the data we observe are prices and quantities in equilibrium, which result from the interactions between supply and demand. In particular, there is simultaneity bias in equation (4) because there is also a supply equation

$$ln(Q_i^S) = \gamma_0 + \gamma_1 ln(P_i) + v_i \tag{5}$$

where we expect that $\gamma_1 > 0$ (i.e., law of supply) and v_i is assumed to be uncorrelated with u_i. Together, equations (4) and (5) simultaneously determine the prices and quantities we observe in the data (i.e., $P \to Q$ and $Q \to P$).

The simultaneity bias in equation (4) leads to a correlation between $ln(P_i)$ and u_i, so our estimates are biased and inconsistent. More specifically, β_1 will be biased upwards. To see this, consider a negative demand shock where u_i falls. When u_i falls, Q_i^D must fall by equation (4). But in equilibrium $Q_i^D = Q_i^S$, so Q_i^S must also fall. Then, from equation (5), we see that γ_1 is positive, so P_i must fall. Thus, P_i and u_i are positively correlated, and β_1 is biased upward.

To estimate equation (4), we need to find a third variable that is shifting the supply curve but not the demand curve. This third variable will be our instrument: it must be correlated with prices (so it is shifting the supply curve) but is uncorrelated with u (so the demand curve remains stable). Our candidate instrument is $SalesTax_i$, which is the general sales tax, measured in dollars per pack. Is $SalesTax_i$ potentially relevant? A high sales tax increases the after-tax prices, P_i, so the relevance condition is plausibly satisfied. Is $SalesTax_i$ potentially exogenous? For this condition to be satisfied, $SalesTax$ must be uncorrelated with u_i; this means that sales tax must affect demand for cigarettes only indirectly through prices. This is plausible if, for example, states choose their sales tax based only on politics, unrelated to any factors affecting demand for cigarettes. We assume for now that this is the case, and we examine the exogeneity condition in more detail later.

Using TSLS, we estimate the first stage regression (see Stata code):

$$\widehat{ln(P_i)} = \widehat{\pi}_0 + \widehat{\pi}_1 SalesTax_i$$

and obtain the predicted values $\widehat{ln(P_i)}$. In the second stage regression, $ln(Q_i)$ is regressed on $\widehat{ln(P_i)}$. The resulting estimate is

$$\widehat{ln(Q_i)} = \widehat{\beta}_0 + \widehat{\beta}_1 \widehat{ln(P_i)}$$

Note that it is also customary to write the endogenous variable in the second stage regression without the "hat," i.e. writing $ln(P_i)$, when it is understood that the estimates come from TSLS estimation.

Returning to the objective of this analysis, what does our estimate show? We find that demand for cigarettes is elastic: a 1% increase in the price reduces consumption by 1.08%. However, these estimates would only be consistent if the IV is relevant and exogenous. As mentioned above, exogeneity is often very difficult to satisfy. For example, one reason why $SalesTax_i$ might not be exogenous is the average household income in the state. In equation (1), this variable is absorbed in the error term because the demand for cigarettes depends on income. Furthermore, income and sales taxes may be correlated if states that are richer (and receive a lot of tax revenue from income taxes) charge lower sales taxes.

3 IV Regression: 1 endogenous regressor, 1 instrument, and control variables

3.1 Overview

In the previous section, we considered a regression where we have one Y and one X variable. However, it is also possible to extend this regression to include control variables. In particular, suppose we have r control variables denoted as $W_{1i}, W_{2i}, \ldots, W_{ri}$. Then our equation of interest (i.e., the population regression model we are trying to estimate) becomes

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{1i} + \ldots + \beta_{1+r} W_{ri} + u_i \tag{6}$$

The population first stage regression then relates X to the instrument Z and all the control variables W:

$$X_i = \pi_0 + \pi_1 Z_i + \pi_2 W_{1i} + \ldots + \pi_{1+r} W_{ri} + v_i \tag{7}$$

Important: In the first stage regression, all control variables are included.

Why do we care about including control variables in the regression? Doing so allows us to relax the exogeneity assumption. Specifically, the exogeneity assumption now becomes $cov(Z_i, u_i | W_1, \ldots, W_r) = 0$.

3.2 Example: Estimating the Price Elasticity of Demand for Cigarettes

Our conclusion from the previous section suggests that we should include income as a control variable in the regression. Hence, our equation of interest becomes

$$ln(Q_i^D) = \beta_0 + \beta_1 ln(P_i) + \beta_2 ln(Inc_i) + u_i \tag{8}$$

where $ln(Inc_i)$ is the natural logarithm of the state's per capita income. By adding this control variable, we are making the exogeneity assumption less strict: now, we only need $cov[StateTax_i, u_i | ln(Inc_i)] = 0$. In words, this means that conditional on the natural log of state per capita income, the state tax and the population regression error term are uncorrelated. TSLS estimation proceeds in two stages as before, but note that the control variables are all included in the first stage regression. See Stata code for how to implement the IV regression.

4 IV Regression: 1 endogenous regressor, multiple instruments, and control variables

4.1 Overview

So far we have been discussing a case where we have one instrument and control variables. But there may also be cases where we might have m instruments, where $m > 1$ (i.e., we have more than one instrument). In this case, the first stage regression is the same as in equation (7) but we add all instruments in the first stage. Specifically, the population first stage regression becomes

$$X_i = \pi_0 + \pi_1 Z_i + \ldots + \pi_m Z_m + \pi_{m+1} W_{1i} + \ldots + \pi_{m+r} W_{ri} + v_i \tag{9}$$

Because we have one endogenous variable but multiple instruments, the relevance and exogeneity conditions are a bit different (though the intuition is the same). In the case with multiple instruments, the conditions are as follows:

1. Relevance: At least one of the Z's must have a nonzero coefficient in the population first stage regression. If this does not hold, then the X and W's are perfectly multicollinear in our equation of interest, i.e. equation (6).

2. Exogeneity: all instruments are uncorrelated with the error term, i.e., $cov(Z_{1i}, u_i) = 0$, $cov(Z_{2i}, u_i) = 0, \ldots, cov(Z_{mi}, u_i) = 0$.

4.2 J-test

As mentioned earlier, we cannot directly test whether the exogeneity condition holds because the population regression errors u_i are unobservable. However, if we have multiple instruments than we have endogenous regressors, then we can provide evidence for exogeneity. The intuition is as follows. Suppose that you have one endogenous regressor X but two instruments Z_1 and Z_2. Then you can compute two different TSLS estimators, one using Z_1 and another using Z_2. If both instruments are exogenous, then the two TSLS estimates should be similar to each other. If you find two estimates that are very different, this suggests that there is something wrong with one or both of the instruments.

The **test of overidentifying restrictions** or the J-test implicitly makes the above comparison. Basically, it tests whether the covariances $cov(Z_{1i}, u_i)$ and $cov(Z_{2i}, u_i)$ are jointly zero. Doing so is equivalent to testing whether the two TSLS estimates for β_1 (that we obtain by using each instrument Z separately, one at a time) are equal.

How is the J-test carried out? Let's first write down how the steps for implementing this test, and then we can discuss what it means.

1. Estimate equation (6) by TSLS using Z_1 and Z_2 as instruments.
2. Obtain the residuals $\widehat{u}_i^{TSLS} = Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i - \beta_2 W_{1i} - \ldots - \beta_{1+r} W_{ri}$. **Important**: we use X (i.e., the observe values for X in the data) and not the predicted values $\widehat{X}$.
3. Regress $\widehat{u}_i^{TSLS}$ on all instruments and all control variables.
4. Test the hypothesis that the coefficient on the instruments are jointly zero. Obtain the F-statistic.
5. Compute the J-statistic as $J = mF$, where $J\ sim\chi^2_{m-k}$, m is the number of instruments, and k is the number of endogenous regressors. In our case, we have $k = 1$.
6. Compare the J-statistic to the critical value (e.g., from the χ^2 table at the back of the textbook).

What exactly are we doing in this test? The idea is that if the instruments are exogenous, they are uncorrelated with u_i. This suggests that the instruments are approximately uncorrelated with the residuals $\widehat{u}_i^{TSLS}$. Hence, if the instruments are in fact exogenous, then if we regress $\widehat{u}_i^{TSLS}$ on the instruments, then the coefficients on the instruments should all be zero. This is exactly the hypothesis that we are testing in Step 4 above.

The J-test can only be carried out if $m > k$ (i.e., we have more instruments than endogenous regressors). Why? Consider the case where we have two instruments Z (so $m = 2$) and one endogenous regressor X (so $k = 1$). As explained above, because we have two instruments, we can compute two TSLS estimators (one using each instrument), and we can compare them to see if they are similar. But if we had $m = 1$ and $k = 1$, then we have only one instrument and we can only compute one TSLS estimator, and we have nothing to compare it to.

Although the J-test seems useful in practice, note that it is still a hypothesis test, so we are never certain whether H_0 or H_1 is true. Even if we fail to reject H_0, this does not mean the instruments are exogenous; we only failed to find evidence for it.

4.3 Example: Estimating the Price Elasticity of Demand for Cigarettes

Suppose that in addition to the general sales tax, $SalesTax_i$, we have another candidate instrument: cigarette-specific taxes, denoted as $CigTax_i$, which apply only to cigarettes and other tobacco products. $CigTax$ is plausibly relevant because the cigarette-specific tax increases the price paid by consumers. If $CigTax$ is uncorrelated with u_i, then it is exogenous.

See Stata code for how to implement the IV regression. Since we now have two instruments, our first stage regression will include $SalesTax$, $CigTax$, and $ln(Inc)$ on the right-hand side. Furthermore, because we have two instruments, we can implement the J-test. Doing so, we obtain a J-stat of 0.3. For a test at 5% significance, the corresponding critical value (from the χ^2 table at the back of the textbook) is 3.84 because we have 1 degree of freedom. Since $0.3 < 3.84$, we fail to reject the null hypothesis that the instruments are jointly uncorrelated with u_i.

5 General IV Regression

We now consider the general IV regression where we have multiple endogenous regressors, multiple exogenous or control variables, and multiple instruments. In this general framework, our equation of interest is:

$$Y_i \ = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \ldots + \beta_{k+r} W_{ri} + u_i \tag{10}$$

where:

- $X_1, \ldots, X_k$ are endogenous variables (we have k of them)
- $W_1, \ldots, W_r$ are either exogenous variables or control variables (we have r of them)
- $Z_1, \ldots, Z_m$ are instruments (we have m of them)

What we want is to obtain consistent estimates of $\beta_1, \ldots, \beta_k$. How many instruments do we need? We need $m \geq k$; this means we need at least as many instruments as endogenous variables.

- If $m > k$, we say that the model is **overidentified**. As shown earlier, in this case, we can implement the J-test.
- If $m = k$, we say that the model is **exactly** or **just identified**.
- If $m < k$, we say that the model is **underidentified**. We **CANNOT** implement instrumental variables in this case.

How does TSLS work in the general IV model? It is proceeds as before, except that each endogenous variable X requires its own first stage regression.

First Stage We will have k first stage regressions, one for each of X_1, …, X_k.

1. Regress X_1 on $Z_1, \ldots, Z_m$, $W_1, \ldots W_r$, then obtain $\widehat{X}_1$.
2. Regress X_2 on $Z_1, \ldots, Z_m$, $W_1, \ldots W_r$, then obtain $\widehat{X}_2$.

$\vdots$

k. Regress X_k on $Z_1, \ldots, Z_m$, $W_1, \ldots W_r$, then obtain $\widehat{X}_k$.

Second Stage Regress Y on $\widehat{X}_1, \ldots, \widehat{X}_k$, $W_1, \ldots, W_r$. The resulting estimators for $\beta_0, \beta_1, \ldots, \beta_{k+r}$ are the TSLS estimates.

What are the requirements for a valid IV in the general framework? As before, we need relevance and exogeneity.

1. Relevance: There should be no perfect multicollinearity in the second stage population regression. In other words, $(1, \widehat{X}_1, \ldots, \widehat{X}_k, W_1, \ldots, W_r)$ should not be perfectly multicollinear. The intuition behind this is that if we have multiple X's, the instruments should provide enough information about the exogenous parts of each X to allow us to separate their effects on Y.
2. Exogeneity: all instruments are uncorrelated with the error term, so $cov(Z_{1i}, u_i) = 0$, $cov(Z_{2i}, u_i) = 0$, $\ldots$, $cov(Z_{mi}, u_i) = 0$.

6 IV Regression Assumptions

Finally, it is important to remember that for any model we implement, there are always underlying assumptions. Here, we discuss the IV regression assumptions. For simplicity, I focus on the case where the W's are all exogenous variables. The IV regression assumptions are as follows:

- $E(u_i|W_{1i}, \ldots, W_{ri}) = 0$
- $(X_{1i}, \ldots, X_{ki}, W_{1i}, \ldots, W_{ki}, Z_{1i}, \ldots, Z_{mi}, Y_i)$ are i.i.d. draws from their joint distribution. As we have seen before, this assumption holds with simple random sampling.
- Large outliers are unlikely.
- The instrument is relevant and exogenous.

Why do we care about these assumptions? We care because the TSLS estimator is **consistent** and is approximately normally distributed in large samples. Because we know it is approximately normal, we can use the normal distribution for hypothesis tests and confidence intervals. For example, to calculate the 95% confidence interval, we can use 1.96 which comes from the normal distribution, i.e., $\widehat{\beta} \pm 1.96 se(\widehat{\beta})$.