

1 Instrumental Variables

Last Updated: April 15, 2009

1.1 Introduction

In this section, we will discuss the use of instrumental variables as a solution to the endogeneity problem. To motivate matters, we consider the structural model given by

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + u_i$$

where $E[u_i x_{i,j}] = 0$ for $j = 1, \dots, k-1$. Note that we are now allowing for the possibility that $E[u_i x_{i,k}] \neq 0$ in which case OLS will not recover the structural parameters.

Suppose that we observe a variable z_{ik} which is not contained in the structural model that satisfies

$$E[u_i z_{ik}] = 0 \tag{A1}$$

and

$$x_{ik} = \delta_0 + \delta_1 x_{i1} + \dots + \delta_{k-1} x_{i,k-1} + \theta z_{ik} + r_{ik} \text{ for } \theta \neq 0. \tag{A2'}$$

Assumption A1 states that z_{ik} is orthogonal to the error term in the structural model and assumption A2' states that z_{ik} and x_{ik} are correlated *once we partial out the remaining covariates*. Also, we assume that the equation in A2' is a linear projection so that r_{ik} is orthogonal to all of

the right-hand side covariates. If we substitute A2' into the structural model, then we obtain

$$y_i = \alpha_0 + \alpha_1 x_{i1} + \dots + \alpha_{k-1} x_{ik-1} + \lambda z_{ik} + v_i$$

where $\alpha_j \equiv \beta_j + \beta_k \delta_j$, $\lambda \equiv \beta_k \theta$ and $v_i \equiv u_i + \beta_k r_i$. We call this equation the **reduced form** of the model.

Assumptions A1 and A2' guarantee that the parameters of the model are identified. To see this, we will write the structural model in its more compact form:

$$y_i = x_i \beta + u_i.$$

Next, we define $z_i \equiv (1, x_{i1}, \dots, x_{ik-1}, z_{ik})$ and note that

$$E[z_i' u_i] = 0 \Leftrightarrow E[z_i'(y_i - x_i \beta)] = 0 \Leftrightarrow \beta = E[z_i' x_i]^{-1} E[z_i' y_i].$$

We were able to take the inverse of $E[z_i' x_i]$ because of Assumption A2'.

To better see this, define

$$x_{ik}^* \equiv \delta_0 + \delta_1 x_{i1} + \dots + \delta_{k-1} x_{ik-1} + \theta z_{ik}$$

and

$$x_i^* \equiv (1, x_{i1}, \dots, x_{ik-1}, x_{ik}^*).$$

We can then write $x_i = x_i^* + r_i$ where $E[z_i' r_i] = 0$. Next, note that $x_i^* = z_i \Pi$ where

$$\Pi \equiv \begin{bmatrix} 1 & 0 & \dots & 0 & \delta_0 \\ 0 & \ddots & & & \vdots \\ \vdots & & \ddots & & \vdots \\ \vdots & \dots & 0 & 1 & \delta_{k-1} \\ 0 & \dots & \dots & 0 & \theta \end{bmatrix}$$

and, thus, we will have that

$$E[z_i' x_i] = E[z_i' x_i^*] = E[z_i' z_i] \Pi.$$

We can then conclude that $E[z_i' x_i]$ will be invertible if and only if $E[z_i' z_i]$ and Π are of full rank.

The latter can only occur when θ is different from zero (and, of course, if $E[z_i' z_i]$ is full rank, but this is a minor point).

Using the analogy principle, the Instrumental Variables (IV) estimator is then

$$\widehat{\beta} = \left[\sum_{i=1}^n z_i' x_i \right]^{-1} \left[\sum_{i=1}^n z_i' y_i \right].$$

It is important to note that the A1 can never be tested. However, the rank condition condition *i.e.* $\theta \neq 0$ is testable.

Before we provide a general treatment of IV, we will work through the “classic” example of endogeneity. Consider a simple supply and demand model given by

$$q_i^d = \alpha_0 + \alpha_1 p_i + u_i$$

and

$$q_i^s = \beta_0 + \beta_1 p_i + \beta_2 \tilde{z}_i + v_i.$$

As we have already shown, market equilibrium implies that $E[p_i u_i] \neq 0$. However, provided that $Cov(\tilde{z}_i, p_i) \neq 0$ and $E[\tilde{z}_i u_i] = 0$, we can use \tilde{z}_i as an instrument for the price in the demand equation. To see how this can be done, define $z_i = (1, \tilde{z}_i)$ and $x_i = (1, p_i)$. We will then have that

$$E[z_i' x_i] = \begin{bmatrix} 1 & E[p_i] \\ E[\tilde{z}_i] & E[p_i \tilde{z}_i] \end{bmatrix}.$$

This matrix will have full rank provided that

$$|E[z_i' x_i]| = E[p_i \tilde{z}_i] - E[\tilde{z}_i] E[p_i] = Cov(\tilde{z}_i, p_i) \neq 0.$$

This is testable by regressing p_i onto \tilde{z}_i and a constant.

Sometimes, we call \tilde{z}_i a supply-shifter because it shifts the supply curve while leaving the demand curve alone. In actuality, supply-shifters are hard to come by. Some people might argue that rainfall is a valid supply shifter.

1.2 General Treatment of 2 Stage Least Squares (2SLS)

Let's consider 2SLS which is a more general case of the IV estimator that we discussed earlier.

The structural model is given by

$$y_i = x_i \beta + u_i$$

where x_i is a $1 \times K$ vector. Let z_i be a $1 \times L$ vector which includes all of the exogenous elements of x_i plus any excluded exogenous variables that are not contained in x_i . Note that this is more general than the IV example discussed above because we are allowing for more excluded exogenous variables than we have endogenous variables. In addition, to assuming A1, we will also assume that

$$\text{rank} E [z_i' z_i] = L \tag{A2}$$

and

$$\text{rank} E [z_i' x_i] = K. \tag{A3}$$

We will call A3 “the rank condition.” The other assumption A2 is a technicality that will almost always hold. Note that a necessary condition for A3 to hold is that $L \geq K$ which means that we must have more excluded exogenous variables (*i.e.* variables that are in z_i but in x_i) than we have endogenous variables. A general test of the rank condition is difficult. However, in practice, researchers regress the endogenous elements of x_i onto z_i and check that the excluded exogenous variables are significant by using an F -test.

The 2SLS estimator can be derived as follows. The “first stage” can be written as

$$\underbrace{x_i}_{1 \times K} = \underbrace{z_i}_{1 \times L} \underbrace{\Pi}_{L \times K} + r_i$$

where $E [z_i' r_i] = 0$. We will call $x_i^* = z_i \Pi$. Note that

$$\Pi = E [z_i' z_i]^{-1} E [z_i' x_i].$$

2SLS uses x_i^* as an instrument for x_i . To see this, note that

$$E[x_i^{*'}u_i] = E[x_i^{*'}(y_i - x_i\beta)] = 0 \Leftrightarrow \beta = E[x_i^{*'}x_i]^{-1} E[x_i^{*'}y_i].$$

Our assumptions guarantee that $E[x_i^{*'}x_i]$ is full rank because

$$E[x_i^{*'}x_i] = \Pi' E[z_i'x_i] = E[x_i'z_i] E[z_i'z_i]^{-1} E[z_i'x_i].$$

We will also have that

$$E[x_i^{*'}y_i] = E[x_i'z_i] E[z_i'z_i]^{-1} E[z_i'y_i]$$

which then gives us that

$$\beta = \left(E[x_i'z_i] E[z_i'z_i]^{-1} E[z_i'x_i] \right)^{-1} E[x_i'z_i] E[z_i'z_i]^{-1} E[z_i'y_i].$$

We can estimate β via

$$\widehat{\beta}_{2SLS} = \left(\sum_{i=1}^n x_i'z_i \left(\sum_{i=1}^n z_i'z_i \right)^{-1} \sum_{i=1}^n z_i'x_i \right)^{-1} \sum_{i=1}^n x_i'z_i \left(\sum_{i=1}^n z_i'z_i \right)^{-1} \sum_{i=1}^n z_i'y_i.$$

If there are as many excluded exogenous variables in z_i as there are endogenous variables in x_i

then 2SLS is IV since

$$\beta = E[z_i'x_i]^{-1} E[z_i'z_i] E[x_i'z_i]^{-1} E[x_i'z_i] E[z_i'z_i]^{-1} E[z_i'y_i] = E[z_i'x_i]^{-1} E[z_i'y_i].$$

1.3 Asymptotic Properties of 2SLS

We will now show the 2SLS estimator is consistent and asymptotically normal. To see consistency, note that

$$\widehat{\beta}_{2SLS} = \beta + \left(\sum_{i=1}^n x_i' z_i \left(\sum_{i=1}^n z_i' z_i \right)^{-1} \sum_{i=1}^n z_i' x_i \right)^{-1} \sum_{i=1}^n x_i' z_i \left(\sum_{i=1}^n z_i' z_i \right)^{-1} \sum_{i=1}^n z_i' u_i$$

and that the second term will converge to zero in probability because $E[z_i' u_i] = 0$. Thus, we will have that $\widehat{\beta}_{2SLS} \xrightarrow{p} \beta$. The proof of asymptotic normality of 2SLS is just like the proof for OLS and will follow from

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n z_i' u_i \xrightarrow{d} N(0, E[u_i^2 z_i' z_i]).$$

If we are willing to further assume that

$$E[u_i^2 z_i' z_i] = \sigma^2 E[z_i' z_i] \tag{A4}$$

where $\sigma^2 = E[u_i^2]$ then the limiting distribution of 2SLS will be given by

$$\sqrt{n} \left(\widehat{\beta}_{2SLS} - \beta \right) \xrightarrow{d} N \left(0, \sigma^2 (E[x_i' z_i] E[z_i' z_i]^{-1} E[z_i' x_i])^{-1} \right).$$

We can calculate the standard errors via

$$\widehat{\sigma}^2 \left(\sum_{i=1}^n \widehat{x}_i' \widehat{x}_i \right)^{-1}$$

where $\hat{\sigma}^2 \equiv \frac{1}{n} \sum_{i=1}^n (y_i - x_i \hat{\beta})^2$ and $\hat{x}_i = z_i \hat{\pi}$. This formula works because

$$\frac{1}{n} \sum_{i=1}^n \hat{x}_i' \hat{x}_i = \frac{1}{n} \sum_{i=1}^n \hat{\pi}' z_i' z_i \hat{\pi} \xrightarrow{p} \pi' E[z_i' z_i] \pi = E[x_i' z_i] E[z_i' z_i]^{-1} E[z_i' x_i]$$

since $\pi = E[z_i' z_i]^{-1} E[z_i' x_i]$. If assumption A4 fails so that there is heteroskedasticity, then we simply use

$$\left(\sum_{i=1}^n \hat{x}_i' \hat{x}_i \right)^{-1} \left(\sum_{i=1}^n \hat{u}_i^2 \hat{x}_i' \hat{x}_i \right) \left(\sum_{i=1}^n \hat{x}_i' \hat{x}_i \right)^{-1}$$

to calculate the standard errors.

1.4 Weak Instruments

Unlike OLS, the rank condition plays a crucial role when employing instrumental variables. To see this, we consider a simple regression with a single endogenous covariate denoted by x_i and one excluded IV denoted by z_i . We have shown that the probability limit of the IV estimator will be

$$p \lim \hat{\beta}_{IV} = \beta + \frac{Cov(u_i, z_i)}{Cov(x_i, z_i)}$$

which contrasts with the probability limit of the OLS estimator

$$p \lim \hat{\beta}_{OLS} = \beta + \frac{Cov(u_i, x_i)}{Var(x_i)}.$$

The problem here is that if $Cov(x_i, z_i)$ is small then even small correlations between u_i and z_i will get magnified. Because $Var(x_i)$ will always be bigger than $Cov(x_i, z_i)$, it may actually be the case that OLS is *less* biased than IV when the instrument is weak.

Now, we consider the multivariate regression

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + u_i$$

where z_i is a set of instruments. We can show that, under some homoskedasticity assumptions, that the asymptotic variance of the 2SLS estimator will be given by

$$AVar(\hat{\beta}_k) \approx \frac{\hat{\sigma}^2}{SSR_k}$$

where SSR_k is the sum of squared residuals of a regression of $\hat{x}_{ik} \equiv z_i \hat{\pi}$ onto $(1, \hat{x}_{i1}, \dots, \hat{x}_{ik-1})$.

Now, recall from the definition of R^2 that

$$SSR_k = TSS_k(1 - R_k^2)$$

where $TSS_k \equiv \sum_{i=1}^n (\hat{x}_{ik} - \overline{\hat{x}_{ik}})^2$ and R_k^2 is the “R-squared” of a regression of \hat{x}_{ik} onto $(1, \hat{x}_{i1}, \dots, \hat{x}_{ik-1})$.

This calculation shows us two channels through which weak instruments will increase the variance of the 2SLS estimator. The first is by reducing TSS_k . To see how this occurs, note that because $\hat{x}_{ik} \equiv z_i \hat{\pi}$, a weak instrument will reduce the variability of \hat{x}_{ik} . In fact, if $\pi = 0$ so that the instrument is as weak as possible, then there will be no variability in \hat{x}_{ik} . The second channel is that weak instruments will force R_k^2 to be close to unity. To see this, suppose that x_{ik} is the only endogenous variable and that z_i is the only instrument. If the instrument is very weak then we will have that

$$\hat{x}_{ik} \approx \hat{\alpha}_0 + \hat{\alpha}_1 x_{i1} + \dots + \hat{\alpha}_{k-1} x_{ik-1}$$

and, thus, a regression of \hat{x}_{ik} onto $(1, x_{i1}, \dots, x_{ik-1})$ will yield an R^2 that will be very close to unity.

The discussion, thus far, has suggested that weak instruments is a finite sample issue that could, at least in principle, be corrected by larger sample sizes. To some extent, this is true if the population matrix, $E[z_i'x_i]$, still has full rank, but just barely. However, what if the population matrix does not have full rank, but in the sample the rank condition is barely satisfied? In this scenario, weak instruments *is* a population issue.

To fix ideas, we will work through an example based on Staiger and Stock (1996). Consider the simple model:

$$y_i = x_i\beta + \varepsilon_i$$

and

$$x_i = z_i\pi + v_i$$

where all of the variables are scalars. The second equation is a linear projection. We will allow x_i to be endogenous and we will assume that $E[z_i\varepsilon_i] = 0$. Next, we will assume that $\pi = \frac{1}{\sqrt{n}}$. This assumption allows the rank condition to barely pass in the sample, but implies that it will fail in the population. We will always have that

$$\hat{\beta} - \beta = \frac{\frac{1}{n} \sum_{i=1}^n \varepsilon_i z_i}{\frac{1}{n} \sum_{i=1}^n z_i x_i}.$$

Note that the rank condition will fail because

$$\frac{1}{n} \sum_{i=1}^n z_i x_i = \frac{1}{\sqrt{n}} \bar{z}_i^2 + \frac{1}{n} \sum_{i=1}^n v_i z_i \xrightarrow{p} 0.$$

Next, note that

$$\sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n z_i x_i \right) = \bar{z}_i^2 + \frac{1}{\sqrt{n}} \sum_{i=1}^n v_i z_i \xrightarrow{d} E[z_i^2] + a$$

where $a \sim N(0, E[v_i^2 z_i^2])$ and, thus, we will have that

$$\hat{\beta} - \beta = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i z_i}{\frac{1}{\sqrt{n}} \sum_{i=1}^n z_i x_i} \xrightarrow{d} \frac{b}{E[z_i^2] + a}$$

where $b \sim N(0, E[\varepsilon_i^2 z_i^2])$. This is a problematic result as it tells us that 2SLS is no longer \sqrt{n} consistent and, in fact, is not consistent at all!

1.5 Generated Regressors

We now consider the problem of generated regressors. Consider the model

$$y_i = x_i \beta + u_i$$

where $x_i = f(w_i; \delta)$. The idea here is that we do not observe the x_i 's, but that we have to generate them by estimating δ . Suppose that we have a \sqrt{n} -consistent estimator of δ . Assume that we also have generated instruments for x_i which are given by $\hat{z}_i = g(v_i; \hat{\lambda})$ where $\hat{\lambda}$ is a \sqrt{n} -consistent estimator of λ . We will assume that $E[u_i | v_i] = 0$.

We can estimate β via 2SLS:

$$\widehat{\beta} = \left[\sum_{i=1}^n \widehat{x}'_i \widehat{z}_i \left(\sum_{i=1}^n \widehat{z}'_i \widehat{z}_i \right)^{-1} \sum_{i=1}^n \widehat{z}'_i \widehat{x}_i \right]^{-1} \sum_{i=1}^n \widehat{x}'_i \widehat{z}_i \left(\sum_{i=1}^n \widehat{z}'_i \widehat{z}_i \right)^{-1} \sum_{i=1}^n \widehat{z}'_i y_i.$$

Next, note that

$$y_i = \widehat{x}_i \beta + (x_i - \widehat{x}_i) \beta + u_i$$

and, thus, we will have that

$$\sqrt{n} (\widehat{\beta} - \beta) = (\widehat{C}' \widehat{D} \widehat{C})^{-1} \widehat{C}' \widehat{D} \left[\frac{1}{\sqrt{n}} \sum_{i=1}^n \widehat{z}'_i [(x_i - \widehat{x}_i) \beta + u_i] \right]$$

where $\widehat{C} \equiv \frac{1}{n} \sum_{i=1}^n \widehat{z}'_i \widehat{x}_i$ and $\widehat{D} \equiv \frac{1}{n} \sum_{i=1}^n \widehat{z}'_i \widehat{z}_i$. It turns out that $\widehat{C} \xrightarrow{p} E[x'_i z_i]$ and $\widehat{D} \xrightarrow{p} E[z'_i z_i]$ provided that the functions $f(\cdot)$ and $g(\cdot)$ are well-behaved. A mean-value expansion yields

$$\begin{aligned} \frac{1}{\sqrt{n}} \sum_{i=1}^n \widehat{z}'_i u_i &= \frac{1}{\sqrt{n}} \sum_{i=1}^n z'_i u_i + \underbrace{\left[\frac{1}{n} \sum_{i=1}^n \nabla_{\lambda} g(v_i; \lambda) u_i \right]}_{o_p(1)} \underbrace{\sqrt{n} (\widehat{\lambda} - \lambda)}_{O_p(1)} + o_p(1) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n z'_i u_i + o_p(1). \end{aligned}$$

A similar calculation gives us that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \widehat{z}'_i (x_i - \widehat{x}_i) \beta = -G \sqrt{n} (\widehat{\delta} - \delta) + o_p(1)$$

where $G \equiv E[(\beta \otimes z_i)' \nabla_{\delta} f(w_i; \delta)]$. If we assume that

$$\sqrt{n}(\hat{\delta} - \delta) = \frac{1}{\sqrt{n}} \sum_{i=1}^n r_i(\delta) + o_p(1)$$

where $E[r_i(\delta)] = 0$ then we will obtain that

$$\sqrt{n}(\hat{\beta} - \beta) = (C'DC)^{-1} C'D \left[\frac{1}{\sqrt{n}} \sum_{i=1}^n (z_i' u_i - Gr_i(\delta)) \right] + o_p(1) \xrightarrow{d} N(0, V)$$

where $V \equiv (C'DC)^{-1} C'DMDC(CDC')^{-1}$ and $M \equiv \text{Var}(z_i' u_i - Gr_i(\delta))$.

A few points are worth mentioning. First, the fact that we had generated regressors is picked up the term $Gr_i(\delta)$. Thus, if G is zero, then the generated regressors will not affect the asymptotic properties of the estimator. However, they may impact the finite sample properties. Next, the fact the parameter λ was estimated does not matter. Thus, generated instruments, in contrast to generated regressors, will not impact the asymptotic properties of the procedure.

1.6 Testing for Endogeneity

Consider the model

$$y_{i1} = z_{i1}\delta + y_{i2}\alpha + u_{i2}.$$

Suppose that we have a vector of instruments that satisfies $E[z_i' u_{i2}] = 0$ and that the vector z_{i1} is a proper subset of z_i . The regressor, y_{i2} , may or may not be endogenous. It is important to know whether or not this regressor is endogenous before we use IV since IV is inefficient when compared to OLS.

We write

$$y_{i2} = z_i\pi + v_{i2}$$

where $E[z_i'v_{i2}] = 0$. Note that because the elements of z_i are valid instruments, we will have that y_{i2} is endogenous if and only if $E[v_{i2}u_{i2}] \neq 0$. We then write

$$u_{i2} = \rho v_{i2} + e_i$$

where the residual is orthogonal to v_{i2} . Note that $E[z_i e_i] = 0$ since $E[z_i u_{i2}] = E[z_i v_{i2}] = 0$.

Thus, we can write

$$y_{i1} = z_{i1}\delta + y_{i2}\alpha + \rho v_{i2} + e_i$$

where e_i is orthogonal to the regressors by construction. We can now test for whether or not y_{i2} is endogenous by testing $H_0 : \rho = 0$. Be aware, however, that v_{i2} cannot be observed directly and must be generated and, so an adjustment to the standard errors will be necessary.

It is important to know what this test is and what it is not. The test says that, provided that we have a vector of valid instruments, we can test for whether or not a regressor is exogenous. It is important to note that the test requires valid instruments. The test does not tell us whether or not our instruments are valid. In general, we can never know whether or not we have a valid instrument. Some people may claim that overidentification tests allow us to see if our instruments are valid, but as we will see this statement is a bit imprecise.

1.7 Overidentification Tests

We consider the model

$$y_i = z_{i1}\delta_1 + y_{i2}\delta_2 + u_{i1}$$

where z_{i1} is $1 \times L_1$ and y_{i2} is $1 \times G_1$. Let $z_i \equiv (z_{i1}, z_{i2})$ be a set of instruments where z_{i2} is $1 \times L_2$ and $L = L_1 + L_2$. Assume that the model is overidentified so that $L_2 > G_1$. The Hausman Principle says that we should compare the 2SLS estimator using all of the instruments to the 2SLS estimator that only uses a subset of the instruments. The idea is that if all of the instruments are valid then the two estimators should be similar and, in the limit, the same. Let R_u^2 be the R^2 from a regression of \hat{u}_{i1} onto z_i . It turns out that if the instruments are valid and given a homoskedasticity assumption that we will have

$$nR_u^2 \stackrel{A}{\sim} \chi_{L_2 - G_1}^2$$

where the degrees of freedom is given the number of overidentifying assumptions. The logic of the overidentifying test is that if at least one of the extra IV's is invalid then the tests will be rejected. It is important to note that this test is a necessary, but not sufficient condition, for the validity of a set of instruments.