Introduction to Endogeneity-related Methods

Xiang Ao

July 17, 2009

1 Endogeneity

When we discuss the linear model we assume that the regressors are exogenous, meaning that they are independent of or uncorrelated with the error term. Often there are reasons to believe that some regressors are correlated with the error term. In that case we call those regressors endogenous.

Under the classical assumptions OLS estimators are Best Linear Unbiased Estimator (BLUE). One key assumption is that the regressors have to be uncorrelated with the error term. If this condition does not hold, OLS estimators are biased and inconsistent.

When one independent variable does not satisfy this condition, we say this variable is endogenous.

The most popular cure for endogeneity is to use instrumental variables.

2 Instrumental Variables

Suppose the linear regression model

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u}, \quad \mathrm{E}(\mathbf{u}\mathbf{u}') = \sigma^{\mathbf{2}}\mathbf{I}, \tag{1}$$

at least one of the explanatory variables in the $n \times k$ matrix $\mathbf{X}$ is assumed not to be predetermined with respect to the error terms, or say, endogenous.

If we have only one endogenous variable, the moment condition is

$$\mathbf{Z}'(\mathbf{y} - \mathbf{X}\beta) = \mathbf{0}. \tag{2}$$

since there are k equations and k unknowns, we can solve it to obtain the simple IV estimator

$$\hat{\beta}_{\mathbf{IV}} \equiv (\mathbf{Z}'\mathbf{X})^{\mathbf{-1}}\mathbf{Z}'\mathbf{y}. \tag{3}$$

Suppose we have a set of variables $\mathbf{Z}$, an $n \times l$ matrix of instruments , which satisfies the moment condition

$$\mathbf{Z}'(\mathbf{y} - \mathbf{X}\beta) = \mathbf{0}. \tag{4}$$

Here is what we do for two-stage least squares (2sls):

Stage 1: Regress each of the variables in the $\mathbf{X}$ matrix on $\mathbf{Z}$ to obtain a matrix of fitted values $\hat{\mathbf{X}}$,

$$\hat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-\mathbf{1}}\mathbf{Z}'\mathbf{X} = \mathbf{P_Z X} \tag{5}$$

Stage 2: Regress $\mathbf{y}$ on $\hat{\mathbf{X}}$ to obtain the estimated β

$$\hat{\beta}_{\mathbf{2sls}} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-\mathbf{1}}(\hat{\mathbf{X}}'\mathbf{y}) = (\mathbf{X}'\mathbf{P_Z X})^{-\mathbf{1}}(\mathbf{X}'\mathbf{P_Z y}) = \hat{\beta}_{\mathbf{IV}} \tag{6}$$

Geometry Illustration:

Suppose the simplest case: $y = \beta_0 + \beta_x + u$, where x can be decomposed into x_1, which is exogenous and x_2, which is endogenous. Therefore x_2 is parallel to u and x_1 perpendicular to u. Suppose z is a vector that is perpendicular to x_2 or u, but not perpendicular to x_1. Then z can be an instrument for x. The way instrumental variable works: Regress x on z, suppose $\hat{x}_1$ is the projection. Regress y on z, $\hat{y}_2$ be the projection. Then the result of $\beta_2 = \hat{y}_2/\hat{x}_1$ is the same as $\beta_1 = \hat{y}_1/x_1$ (since the two triangles are similar). Here $\hat{y}_1$ is the projection of y on x_1, which is hypothetical since we have no way to decompose x into x_1 and x_2; otherwise, we would not need instrumental variables.

3 Durbin-Wu-Hausman Test

3.1 Idea

In econometric modeling, there are often questions on endogeneity. Do we know how to test whether an independent variable is endogenous statistically? The answer is: sort of, but not really. We cannot do endogeneity test without a valid instrument. Therefore, we have to have strong argument for a valid instrument first before we can do endogeneity test.

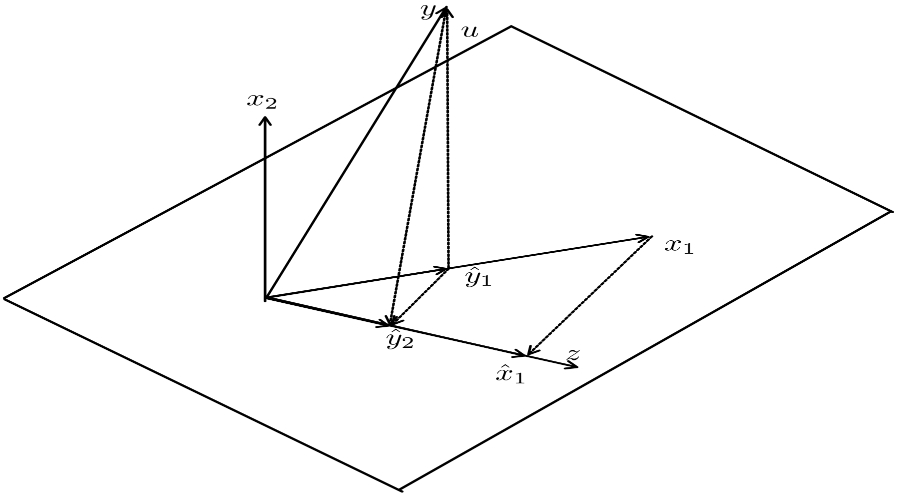
With endogenous variables on the right-hand side of the equation, we need to use instrumental variable (IV) regression for consistent estimation. However, with IV regression, we lose efficiency: the asymptotic variance of the IV estimator is larger, and can be much larger than the OLS estimator. Therefore, we gain consistency, but lose efficiency, by using IV estimator when there is an endogeneity problem.

Now we have a familiar scenario (if you are familiar with Hausman test for fixed effect and random effect estimator for panel data): Suppose we have the null hypothesis as the regressor being exogenous. We have an efficient estimator under null hypothesis yet inconsistent under alternative hypothesis (OLS estimator). We also have a consistent estimator under both null and alternative (IV estimator).

Similar to panel data setting, we have the Hausman test statistic as:

$$H = (\hat{\beta}_c - \hat{\beta}_e)'D^{-}(\hat{\beta}_c - \hat{\beta}_e)$$

Figure 1: Instrumental Variable geometry illustration



where $D = \text{Var}[\hat{\beta}_c] - \text{Var}[\hat{\beta}_e]$, $^-$ is the generalized inverse, $\hat{\beta}_c$ is the consistent estimator (in this case the IV estimator) and $\hat{\beta}_e$ is the efficient estimator (in this case OLS estimator).

H conforms to χ^2_k asymptotically, where k is the number of endogenous variables.

This test is to compare the IV estimator and the OLS estimator: if it's close, then OLS estimator is fine (fail to reject null that OLS is consistent, or say the variable is exogenous). If it's large, then IV estimator is needed, although we lose some efficiency. This test is also based on the assumption that the instruments are exogenous. If that is in question, then it's pointless to do the test, since the IV estimator cannot guarantee consistency either.

3.2 Implementation in Stata

In Stata, there are different ways to do it:

1. Do a regular Hausman test:
   - ivreg y x1 (x2=x3 x4)
   - estimates store iv
   - reg y x1 x2
   - hausman iv ., constant sigmamore
2. Use ivendog

4 Identification

Identification in a regression equation means that all parameters can be uniquely estimated. A necessary condition of that is to have at least as many instruments as the number of endogenous variables. That is, $l \geq k$ in our example above. If $l = k$ we have exact-identification. If $l > k$, we have over-identification.

Over-identification generates more efficient estimates, given the assumption of instruments being exogenous. The other advantage of over-identification is that over-identification tests can be done to test the adequacy of instruments.

Under the null hypothesis that all the instruments are uncorrelated with the error term, an LM statistic $N \times R^2$ conforms to $\chi^2(r)$ distribution, $r = l - k$, the number of excess instruments, or say, the number of excluded restrictions. If we reject the null, then we should be concerned about the exogeneity of the whole set of the instruments. This test is called Sargan's test in IV context, and (Hansen's) J test in GMM context.

What the J test or Sargan's test does is to test the whole set of instruments being exogenous or not. There is another test for testing exogeneity for a subset of instruments. It's call a C test or a difference-in-Sargan test. The idea is to calculate the difference between two Sargan's statistics (or Hansen's J in GMM setting); one is with the whole set of instruments, the other one without the suspected instruments. The null is that the suspect instruments are exogenous; or orthogonal to the error term. Obviously to conduct the C test, we'll have to have at least one extra instrument more than the number of endogenous variables.

4.1 Implementation in Stata

In Stata, there are different ways to do over-identification test, *ivreg2* reports a comprehensive set of tests; *overid* command does the over-identification test after the *ivreg* command.

ivreg2 with *gmm* option returns J test; it reports Sargan's test without this option.

ivreg2 also reports C test statistic, with *ortho()*. If the C test rejects the null, and J test without the suspect instruments fail to reject null, then the suspect instruments are indeed the ones are not exogenous.

5 Weak Instruments

5.1 Problem with the cure

An instrument needs to satisfy to criteria: orthogonality and relevance. We need instruments to be orthogonal to the error term. We can verify the orthogonality condition by Sargan's test if there are extra instruments.

It turns out instrument relevance is important too: if instruments are weak, then the regular large sample properties of IV or GMM estimators do not hold any more. The estimators are inconsistent or biased.

To see the problem, suppose

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u}, \quad \mathrm{E}(\mathbf{uu}') = \sigma^{\mathbf{2}}_{\mathbf{u}}\mathbf{I}, \tag{7}$$

$$\mathbf{X} = \mathbf{Z\Pi} + \mathbf{v}, \quad \mathrm{E}(\mathbf{vv}') = \sigma^{\mathbf{2}}_{\mathbf{v}}\mathbf{I}, \tag{8}$$

and

$$\mathrm{E}(\mathbf{Zu}) = \mathbf{0}. \tag{9}$$

We can see here $\mathbf{Z}$ is exogenous. However, the model does not say anything about relevance. To illustrate the problem caused by weak instrument, suppose we have only one endogenous variable and one instrument.

$$\hat{\beta}_{2sls} = \frac{\mathbf{Z}'\mathbf{y}}{\mathbf{Z}'\mathbf{X}} = \frac{\mathbf{Z}'(\mathbf{X}\beta + \mathbf{u})}{\mathbf{Z}'\mathbf{X}} = \beta + \frac{\mathbf{Z}'\mathbf{u}}{\mathbf{Z}'\mathbf{X}}. \tag{10}$$

If $\mathbf{Z}$ is irrelevant, or, $\Pi = 0$, then

$$\hat{\beta}_{2sls} - \beta = \frac{\mathbf{Z}'\mathbf{u}}{\mathbf{Z}'\mathbf{v}} = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{N} Z_i u_i}{\frac{1}{\sqrt{n}}\sum_{i=1}^{N} Z_i v_i} \xrightarrow{d} \frac{z_u}{z_v}, \tag{11}$$

where

$$\begin{bmatrix} z_u \\ z_v \end{bmatrix} \sim N(0, \begin{bmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{bmatrix}). \tag{12}$$

Therefore, if $\mathbf{Z}$ is irrelavent, β_{2sls} is inconsistent. Also, the distribution of the bias is Cauchy-like (the ratio of correlated normals).

This is a case where the cure might be worse than the disease itself: the bias can be big comparing to the bias an OLS estimate suffers.

5.2 Tests

There are a variety of weak-instruments tests proposed. Most of them are based on so-called weak-instruments asymptotics and a new parameter called *concentration parameter* $\mu^2 = \Pi' Z' Z \Pi / \sigma_v^2$. Sample size only enters the distribution through μ^2.

With weak-instruments asymptotics, IV estimators are no longer consistent, and they are not normal asymptotically. Most test statistics (J test, etc.) do not have normal or χ^2 distributions anymore.

Now I list the following tests in the order of recommended level by James Stock:

1. Moreira (2003) conditional likelihood ratio test (CLR).

   Advantages of this test:

   (a) Uniformly most powerful tests among valid tests.

   (b) Implemented in Stata as *condivereg*.

   Disadvantages:

   (a) Complicated.

   (b) Only developed so far for one endogenous variable case.

2. Stock-Yogo bias method and size method.

   Stock and Yogo (2005) provide critical values for both methods: one is to control the size of bias, the other one is to control the size of a Wald test of $\beta = \beta_0$. Bias method is more frequently used. In the case of multiple endogenous variables, the Craigg-Donald statistics is used to compare with the critical values. It is implemented in Stata as part of the *ivreg2* command, but it's only available for the situation there are at least two excluded variables (meaning the number of instruments minus the number of endogenous variables).

3. Anderson-Rubin confidence intervals.

   In the model of

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u}, \quad \mathrm{E}(\mathbf{u}\mathbf{u}') = \sigma_{\mathbf{u}}^{\mathbf{2}}\mathbf{I}, \tag{13}$$

$$\mathbf{X} = \mathbf{Z\Pi} + \mathbf{v}, \quad \mathrm{E}(\mathbf{vv'}) = \sigma^{\mathbf{2}}_{\mathbf{v}}\mathbf{I}, \tag{14}$$

The null hypothesis $H_0 : \beta = \beta_0$. Anderson-Rubin statistic is the F statistic in the regression of $y - X\beta_0$ on Z, the F test on Π being zero:

$$AR(\beta_0) = \frac{(y - X\beta_0)' P_Z (y - X\beta_0)/k}{(y - X\beta_0)' M_Z (y - X\beta_0)/(N - k)} \tag{15}$$

The idea of AR confidence interval is to construct an interval for all possible values of β to fail to reject $\Pi = 0$.

4. First-stage F test.

   The rule of thumb for first-stage F test is $F > 10$ for a single instrument case, the more instruments, the higher it gets.

5. Kleibergen's LM test.

   This test is dominated by the CLR test, thus no longer the optimal test to use.

6. First-stage R^2, or partial R^2, etc., are not recommended.

References

[1] Woodridge, J. (2001) *Econometric Analysis of Cross Section and Panel Data* The MIT Press

[2] Baum, K (2006) *An Introduction to Modern Econometrics using Stata* Stata Press