Notes on Simultaneous Equations and Two Stage Least Squares Estimates

Copyright - Jonathan Nagler; April 19, 1999

1. Basic Description of 2SLS
   - The endogeneity problem, and the bias of OLS.
   - The mechanics of the solution.
   - The consistency property of 2sls
2. Presenting Results with 2SLS: the first stage.
3. Tests for endogeneity: knowing you need to use 2SLS.
4. Tests for exogeneity of instruments: making sure you have used 2SLS legitimately.

1 The Problem: Endogeneity

There are two kinds of variables in our models: exogenous variables and endogenous variables.

Endogenous Variables: These are variables determined within the system of equations which represent the true world. This means that they are functions of other variables present in the system. Up till now (in the single equation world) the only endogenous variable we have dealt with has always been the dependent variable.

Exogenous Variables: These are variables determined *outside* the system. Up till now (in the single equation world) we have treated all of our independent variables as exogenous.

As a general rule, when a variable is endogenous, it will be correlated with the disturbance term, hence violating the GM assumptions and making our OLS estimates biased. This is easily seen in the following example of two equations where Y and X_1 are both endogenous.

$$\begin{aligned} Y_i &= \alpha_{10} + \beta_{11}X_{1i} + \beta_{12}X_{2i} + \beta_{13}X_{3i} \\ &\quad + \ldots + \beta_{1k}X_{ki} + u_i \end{aligned} \tag{1}$$

$$\begin{aligned} X_{1i} &= \alpha_{20} + \beta_{21}Y_i + \beta_{22}Z_{2i} + \beta_{23}Z_{3i} \\ &\quad + \ldots + \beta_{2k}Z_{ki} + v_i \end{aligned} \tag{2}$$

Now substitute the first equation into the second:

$$\begin{aligned} X_{1i} &= \alpha_{20} + \beta_{21}(\alpha_{10} + \beta_{11}X_{1i} + \beta_{12}X_{2i} + \beta_{13}X_{3i} \\ &\qquad + \ldots + \beta_{1k}X_{ki} + u_i) \\ &\quad + \beta_{22}Z_{2i} + \beta_{23}Z_{3i} + \ldots + \beta_{2k}Z_{ki} + v_i \end{aligned}$$

We can see that X_1 is a linear function of u (among other things), and hence will be correlated with u. This violates the GM assumptions, and the OLS estimator $\hat{\beta}_{11}$ will be biased.

2 Just-Identified Equations

If the set of equations is exactly identified, then we can solve for the reduced-form parameters, and then compute the structural parameters from the reduced form parameters.

3 Over-Identification: Use Two Stage Least Squares

The goal is to find a proxy for X, that will not be correlated with u. This proxy is going to be called $\hat{X}$.

The first stage of 2SLS is to generate the proxy, the second stage is to simply substitute the proxy for X, and estimate the resulting equation using OLS.

The trick to generating a proxy is to find a variable that belongs in the second equation (the one predicting X_1), but does **not** belong in the first equation (the one predicting Y).

In other words, we want to find a variable Z that determines X_1 in the world, but that does *not* influence Y.

This is generally not an easy task.

A slightly sloppy statement (actually all that is required is that in the limit these quantities approach 0 and 'not 0', respectively) of the technical conditions on Z are as follows:

$$\begin{aligned} corr(Z, u) &= 0 \\ corr(Z, x) &\neq 0 \end{aligned}$$

So say equation (1) above is the true model [we are now ignoring equation (2) from above]:

$$\begin{aligned} Y_i &= \alpha_{10} + \beta_{11}X_{1i} + \beta_{12}X_{2i} + \beta_{13}X_{3i} \\ &\quad + \ldots + \beta_{1k}X_{ki} + u_i \end{aligned}$$

And say we find some variable Z that influences X_1 but does not influence Y. Note that we only need to find *one* Z. Then we would estimate the following equation using OLS:

$$\begin{aligned} X_{1i} &= \alpha_{30} + \beta_{31}Z_i + \beta_{32}X_{2i} + \beta_{33}X_{3i} \\ &\quad + \ldots + \beta_{3k}X_{ki} + v_i \end{aligned} \tag{3}$$

What we have done in this equation is to include all of the *exogenous* variables from the first equation on the RHS, and added Z. These estimates would allow us to generate a new set of values for the variable $\hat{X}_1$:

$$\begin{aligned} \hat{X}_{1i} &= \hat{\alpha}_{30} + \hat{\beta}_{31}Z_i + \hat{\beta}_{32}X_{2i} + \hat{\beta}_{33}X_{3i} \\ &\quad + \ldots + \hat{\beta}_{3k}X_{ki} \end{aligned} \tag{4}$$

And:

$$X_1 = \hat{X}_1 + \hat{v} \tag{5}$$

Now, $\hat{X}_1$ can be substituted for X_1 in equation (1):

$$\begin{aligned} Y_i &= \alpha_{10} + \beta_{11}(\hat{X}_{1i} + \hat{v}_i) + \beta_{12}X_{2i} + \beta_{13}X_{3i} \\ &\quad + \ldots + \beta_{1k}X_{ki} + u_i \end{aligned}$$

Rewrite this:

$$\begin{aligned} Y_i &= \alpha_{10} + \beta_{11}\hat{X}_{1i} + \beta_{12}X_{2i} + \beta_{13}X_{3i} \\ &\quad + \ldots + \beta_{1k}X_{ki} + (u_i + \beta_{11}\hat{v}_i) \end{aligned}$$

The new equation estimated using OLS. This will produce **consistent** estimates of all the parameters, including β_{11}. Consistency is an assymptotic property, it requires large samples. The estimates will not be unbiased.

The final thing to do however is to correct the standard errors that would be generated this way as they would be incorrect. This is simply a technical correction.

4 2SLS in Practice

If the first stage produces poor predictors of X, then the 2SLS procedure will not perform very well in terms of the precision of the estimates generated. In other words, if $\hat{X}_1$ is a noisy predictor of X_1, then it is the same as estimating an equation using OLS when there is a large amount of measurement error.

Since you want to observe the first-stage estimates, you probably want to run 2SLS manually as well as using your stat package's canned routine.

If you use **STATA**, **ivreg** will compute two-state least squares estimates; and including the ', first' option will display the first-stage results.

5 Associated Tests

5.1 Testing for Endogeneity of X (Heckman)

We can test to see if endogeneity is a problem as follows:

1. Find Z, compute our clean version of X: $\hat{X}$.
2. Estimate the original model (equation 1) with both X **and** $\hat{X}$ on the right hand side.
3. If the coefficient of $\hat{X}$ is significant, then we have an endogeneity problem.

Obviously it is not a great feature of this test that we can only apply the diagnostic test *after* we have a cure.

5.2 Testing for Exogeneity of Z (Hausman)

1. Estimate the model via 2SLS, and compute the residuals e.
2. Run a regression of these residuals on all exogenous variables (included and excluded).
3. Compute a test statistic of nR^2; using the R^2 from the regression of the residuals and n is the number of observations.
4. The nR^2 statistic will have a χ^2 distribution with degress of freedom equal to the number of excluded exogenous variables (the number of Zs) minus the number of endogenous variables explained by the instruments.
5. If nR^2 is too big you can reject the assumption that Z is exogenous.