Instrumental Variables Estimation and Two Stage Least Squares

Econometric Methods, ECON 370

Let's get back to the thiking in terms of cross sectional (or pooled cross sectional) data again. Recall the endogenous explanatory variables in multiple regressions problem which we had argued and showed that it is most likely to occur when we have misspecification errors, measurements errors, and most commonly omitted variables. We had examined a possible solution of using proxies for the last case, but of which a good proxy is not always available. We will now examine the use of **Instrumental Variables, IV** to solve the problem of endogeneity, and the technique used in its estimation, Two Stage Least Squares (2SLS).

1 Motivation for Instrumental Variable (IV) Regression

In cross sectional analysis, when faced with omitted variable bias, we have two options,

1. Ignore the problem → biased and inconsistent estimators.
2. Use proxy for unobserved variable.

which we had previously discussed at some length about when we might be able to use them, and yet be able to learn about the case at hand. Another approach is to permit the unobservable to remain in the error term, and instead of using OLS, we use another technique that recognizes that the unobservable variable captured in the error term, which is the Method of Instrumental Variables.

Consider a simple example where we are trying to understand how inherent ability, Ab of individuals affect their SAT scores. What other covariates do you think might affect this scores? Let's for the sake of simplicity assume that besides ability, the child's socioeconomic status, Inc fully determines how well she does. Then the population regression relationship can be written as,

$$SAT = \beta_0 + \beta_1 Inc + \beta_2 Ab + \epsilon$$

We had suggested that we could proxy ability with IQ scores, which if a good proxy would provide a consistent estimator of β_1. However, the fact of the matter is how many of you

took a IQ test? This is very typical in the sense that a good proxy is hard to come by. If we ignore the fact that ability is not observed and perform the regression,

$$SAT = \beta_0 + \beta_1 Inc + \nu$$

There is in truth nothing wrong with the dependent variable being correlated with the error term in cross sectional analysis, however the problem arises when the unobserved variable is correlated with those that are observed **(Can you remember how to show the bias? Which assumption is violated?)**. In that case,β_1 is no longer unbiased. However, it turns out that we can still estimate the effect of socioeconomic status on SAT scores if we can find a Instrumental Variable for Socioeconomic Status.

Let's turn to a general structure, and rewrite the above equation as,

$$y = \beta_0 + \beta_1 x_1 + \nu \tag{1}$$

But unlike in our previous discussions of OLS, we know that x_1, the endogenous variable, is correlated with ν, that is

$$cov(x_1, \nu) \neq 0$$

Although IV works whether the independent variable and the error terms are correlated, because the technique is motivated by omitted variables, when we know that our regression model is fully specified, we should use OLS instead.

The idea with IV is as follows, what if I can find a variable that is highly correlated with the covariate of interest, but uncorrelated with the original error term, ϵ, and the unobserved variable, let's call it x_2. In that case, since there is nothing wrong with the error term being correlated with the error term, ν, it might be possible for us to find out the true effect x_1 has on y.

Restating the above assumptions or requirements for a Instrumental Variable, z what we need is then,

1. $cov(z, \nu) = 0$. This assumption or requirement is typically assumed without testing if it is true. The argument is that we do not observe the unobserved variable, and consequently cannot test this. But if we have a good proxy, we might be able to ensure that it is true. Why is this true? By assumption, covariates must be uncorrelated with the population error terms, ϵ, then what is left is the covariance between z and the unobserved variable. And yet if we do have a good proxy, what may be useful is to

run the original model with the proxy. Often, what is done is to rely on intuition and a priori economic explanations to justify this assumption.

2. $cov(z, x_1) \neq 0$. This assumption should be tested by running the following regression.

$$x_1 = \alpha_0 + \alpha_1 z + \phi \tag{2}$$

so that if a_1 is statistically significant, then the second assumption holds.

It is important to note that a proxy variable by virtue of its high correlation with the unobservable is a poor IV since it will defintely violate the first assumption. For our current example, IQ, parental education would then all make poor IVs. How about the number of siblings (There has been some research that the ability of the first born is the highest among a family of children)? How about where the child lives, i.e. a high end neighborhood or ghetto etc. Assuredly it is correlated with socioeconomic status, and seem to have little to do with a child's ability which he is born with. How about attendance rate? What you should get out of these considerations is that a IV is not easy to come by. Note that in our question at hand, it is likely possible for us to find a good proxy for ability using say cumulative GPA, and run the complete model as in the first equation. But again, since subjects chosen in high school are personal choices, a student can choose easier subjects so that they may have a high GPA, but may reveal a negative effect on SAT scores.

We will now show that if the assumptions for a good IV is observed, that β_1 is identified, identified in the sense that we can write the formula for β_1 in terms of population moments that can be estimated from a sample from the population.

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \nu \\ \Rightarrow cov(z, y) &= \beta_1 cov(z, x_1) + cov(z, \nu) \\ \Rightarrow cov(z, y) &= \beta_1 cov(z, x_1) \\ \Rightarrow \beta_1 &= \frac{cov(z, y)}{cov(z, x_1)} \end{aligned}$$

Where the second equality follows from the second assumption for a good IV **(You can prove the same using the method of moments. Try it!)**. Consequently, β_1 is identified. Computing the estimate using sample analogs gives us the formula for β_1,

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(z_i - \overline{z})(y_i - \overline{y})}{\sum_{i=1}^{n}(z_i - \overline{z})(x_{1,i} - \overline{x})}$$

while the estimator for the intercept β_0 is

$$\beta_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$$

which looks similar to the OLS estimator, but note that we are using $\widehat{\beta}_1$ obtained from IV instead of the regular OLS. This estimator is consistent. Note that whenever, we have endogeneity due to omitted variable, and use IV estimation, the estimator is not unbiased, consequently, we have to ensure that we have large samples when using IV. Can you proof that the estimator is only consistent?

$$\begin{aligned}
p\lim \widehat{\beta}_1 &= \frac{cov(z,y)}{cov(z,x_1)} \\
&= \frac{cov(z,y)}{cov(z,\alpha_0 + \alpha_1 z + \phi)} \\
&= \frac{cov(z,\beta_0 + \beta_1 x_1 + \nu)}{\alpha_1 \sigma_z^2} \\
&= \frac{cov(z,\beta_0 + \beta_1\alpha_0 + \beta_1\alpha_1 z + \beta_1\phi + \nu)}{\alpha_1 var(z)} \\
&= \frac{\beta_1\alpha_1 var(z)}{\alpha_1 var(z)} = \beta_1
\end{aligned}$$

Can you see why it is not unbiased?

1.1 Statistical Inference with the IV Estimator

The IV estimators has an approximate normal distribution in large samples. To construct standard errors for inference, we assume homoskedasticity, $E(\epsilon^2|z) = \sigma^2 = var(\epsilon)$, noting that the expectation is conditioning on the instrumental variable.

$$p\lim var(\widehat{\beta}_1) = \frac{\sigma^2}{n\sigma_x^2 \rho_{x,z}^2}$$

Note that the rate of convergence is $\frac{1}{n}$. It is easy to see that all the components to the asymptotic variance has easy sample counterparts.

$$\begin{aligned}
\widehat{\sigma}^2 &= \frac{1}{n-2}\sum_{i=1}^{n} \widehat{e}_i^2 = \frac{1}{n-2}\sum_{i=1}^{n} \left(y - \widehat{\beta}_0 - \widehat{\beta}_1 x_{1,i}\right)^2 \\
\widehat{\sigma}_x^2 &= \frac{1}{n}\sum_{i=1}^{n} (x_{1,i} - \overline{x})^2 \\
\widehat{\rho}_{x,z}^2 &= R_{x,z}^2
\end{aligned}$$

where $R^2_{x,z}$ is the goodness of fit measure for the regression,

$$x_1 = \alpha_0 + \alpha_1 z + \nu$$

Note the very important distinction that β_0 and β_1 are the IV estimates, i.e. y on z, since σ^2 is conditioning on z not x_1. That is we can reexpress the formula for the variance of $\widehat{\beta}_1$ as

$$\widehat{\sigma}^2 = \frac{\widehat{\sigma}^2}{SST_x R^2_{x,z}} \tag{3}$$

Recall that the OLS estimator for β_1 is $\frac{\sigma^2}{SST_x}$, which means that the two differ only in $R^2_{x,z}$ (Also there is a distinct difference between the estimate for σ^2. Can you see?). Since the goodness of fit measure is less than 1, when OLS is valid, the variance from IV will always be larger than that of OLS. **Note: Read your text carefully about the paper by Angrist and Krueger (1991). Read the paper if you have to, or if you're really interested. It is a very interesting and insightful paper.** Note also that endogeneity most commonly occurs with a binary variable when dealing with policy analysis due to selection bias. Also there is nothing wrong with having a binary instrumental variable.

1.2 Properties of IV with a Poor Instrumental Variable

The IV estimate is consistent when z, ν are *uncorrelated* and z and x *have any correlation*, but as noted above can have large standard errors, especially when z and x are only weakly correlated. Further, when they are weakly correlated, the IV estimator can have large asymptotic bias even if z and ν are only moderately correlated. To see this, assume that z and ν are correlated, so that

$$\begin{aligned} p\lim \widehat{\beta}_1 &= \beta_1 + \frac{cov(z,\nu)}{cov(z,x_1)} \\ &= \beta_1 + \frac{\frac{cov(z,\nu)}{\sigma_z \sigma_\nu}}{\frac{cov(z,x_1)}{\sigma_z \sigma_{x_1}}} \frac{\sigma_z \sigma_{x_1}}{\sigma_z \sigma_\nu} \\ &= \beta_1 + \frac{corr(z,\nu)}{corr(z,x_1)} \frac{\sigma_\nu}{\sigma_x} \end{aligned}$$

What the above equation says is that even if the correlation between z and ν is small, if the correlation between z and x_1 is likewise small, there would be a substantial bias in the estimator for be_1. In which case, when would it be a good move to use IV instead of OLS?

Recall that we can also write the asymptotic OLS estimator as,

$$\begin{aligned} p\lim \widetilde{\beta}_1 &= \beta_1 + \frac{cov(x_1, \nu)}{\sigma_\nu \sigma_{x_1}} \frac{\sigma_\nu}{\sigma_{x_1}} \\ &= \beta_1 + corr(x_1, \nu) \frac{\sigma_\nu}{\sigma_{x_1}} \end{aligned}$$

Then we should use IV if and only if we believe $\frac{corr(z,\nu)}{corr(z,x_1)} < corr(x_1, \nu)$. It should be clear that if $corr(z, x_1)$ is not correlated at all, since the second term in the asymptotic IV estimator is not defined.

1.3 R^2 after IV

Read you text on this, page 520-521. Essentially, R^2 can be negative in IV estimation, which arises principally due to correlation between the endogenous variable and the error term. In any case, the primary reason for the use of IV is to obtain better estimates of the effect of the endogenous variable, and not the goodness of fit.

2 IV Estimation of the Multiple Regression Model

We will now consider the application of IV to multiple regressions, but still consider only one endogenous variable. Let the model we consider be,

$$y_1 = \beta_0 + \beta_1 x_1 + be_2 z_2 + \epsilon \tag{4}$$

Let x_1 remain as the endogenous variable, but z_2 is a strictly exogenous variable (which implies that it is not correlated with the error term. Based on our examination earlier on endogeneity, we know that all of the coefficient estimates will be biased if we use OLS. Consequently we have to use other techniques, and in keeping with our examination here, we'll think about using IV. Can we then use z_2 since it is exogenous. We cannot, since it is already a regressor, and its use would violate a critical requirement in performing regressions. **What is that**? Suppose we can find an instrument z_1, then based on our previous analysis, we need it to be uncorrelated with ϵ, but correlated with x_1, in other words we need,

$$\begin{aligned} E(\epsilon) &= 0 \\ cov(z_1, \epsilon) &= 0 \\ cov(z_2, \epsilon) &= 0 \end{aligned}$$

Adopting the usual assumption that the expected value of the error term, we can rewrite the conditions for IV as,

$$\begin{aligned} E(\epsilon) &= 0 \\ E(z_1\epsilon) &= 0 \\ E(z_2\epsilon) &= 0 \end{aligned}$$

Which are nothing but moments which we can easily find empirical counterpart, and from which we could obtain closed form solutions to the coefficients. Writing the sample counterparts to the above conditions, we have,

$$\begin{aligned} \sum_{i=1}^{n} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{1,i} - \widehat{\beta}_2 z_2 \right) &= 0 \\ \sum_{i=1}^{n} z_1 \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{1,i} - \widehat{\beta}_2 z_2 \right) &= 0 \\ \sum_{i=1}^{n} z_2 \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{1,i} - \widehat{\beta}_2 z_2 \right) &= 0 \end{aligned}$$

Since this is nothing but 3 simultaneous equations, and since we have three unknowns, given the regression equations, there is a unique solution to the coefficients in question, β_0, β_1, and β_2. **Solve for all the coefficients.** We call the solution to the above problem, $\widehat{\beta}_0$, $\widehat{\beta}_1$, and $\widehat{\beta}_2$ IV estimators. Notice further an interesting point, if x_1 were strictly exogenous, setting $z_1 = x_1$ and substituting this into the second condition creates the three first order conditions of the OLS 2 variable regression problem. Note that you can think of z_2 being its own instrument.

As in our discussion of OLS, we do allow the covariates to be correlated, which is to say that we allow our instrument, z_1 to be correlated with z_2. Suppose the relationship between the covariates can be written as follows,

$$x_1 = \alpha_0 + \alpha_1 z_1 + \zeta_2 + \nu$$

That is we have stated the relationship between the two exogenous variable and the endogenous x_1. One critical requirement we have learned previously is that we need the instrument for the endogenous variable to be correlated, in fact highly correlated for our estimator to be consistent, that is we need $\alpha_1 \neq 0$, and preferably high based on our previous analysis. This requirement is essentially saying that after allowing for z_1 and z_2 to be correlated, and after accounting for the effect of z_2 on x_1, z_1 would still be a significant contributor to how

x_1 behaves. Further, we can always test that this hypothesis is true, since based on all the assumptions, all we need to perform is OLS, since by definition, x_1 being the dependent variable in the last regression equation, would have to be correlated with the error term. We are however unable to test that z_1, and z_2 are uncorrelated with ν.

Generalizing the ideas to the k variable regression with **1 endogenous variable** is straight forward. You would however need to note the usual OLS assumption that all the other exogenous variables besides the endogenous variable cannot have a perfect linear relationship with each other. Further, as usual, the error term, ϵ, is assumed to be homoskedastic for statistical inference.

3 Two Stage Least Squares

It is likewise possible that there may be more than 1 excluded exogenous variable, i.e. more than 1 instrumental variable, all or some of which might be correlated with the endogenous variable. We will now examine how to include both instruments.

3.1 A Single Endogenous Explanatory Variables

Consider the same regression equation as before,

$$y_1 = \beta_0 + \beta_1 x_1 + be_2 z_2 + \epsilon$$

with x_1 being the endogenous variable. But now, we have two instruments q_1, and q_2 excluded from the above regression, and are uncorrelated with ϵ. These last assumptions are known as **exclusion restrictions**.

Given we have one problem, and two possible variables that might provide a solution, what do we do? If we use both independently, we would obtain two estimators using the previous IV technique, but neither of which might be efficient in themselves. But, note the following, by virtue that the two variables, q_1 and q_2 are instruments, by definition, they cannot be correlated with ϵ. Then any linear combination would still be uncorrelated, which suggests we could use a "weighted"combination of the two instruments. Great idea, but how do we decide which is the best combination. Well, what we want is a combination that yields the greatest correlation with the endogenous variable, x_1. There's the hint, we could find

the following,

$$x_1 = \alpha_0 + \alpha_1 q_1 + \alpha_2 q_2 + \alpha_3 z_2 + \nu$$

Note that I have included the exogenous variable in the original equation that included the endogenous variable. Why? Well, it is, like q_1 and q_2, an exogenous variable, so that a combination between all of these variable would provide the best instrument. However, as you should have noted, a keep condition that would allow this idea to work is that we need either one or both coefficients, a_1 and a_2 to be statistically different from zero, failing which if indeed they are zero, we would be faced with effectively using z_2 as an instrument, which would give rise to perfect collinearity in the original regression! This is the key assumption or condition that would permit the identification when we actually use the instrument. Is it possible to test this condition? Well, notice that all the standard assumptions for OLS holds, which means we can perform and OLS, and use a F test on the joint restriction of $a_1 = 0$ and $\alpha_2 = 0$.

What are the other assumptions we need to use this idea of a linear combination of instruments? Like in our discussion of OLS, we require $E(\nu) = 0$, $cov(q_1, \nu) = 0$, $cov(q_2, \nu)$=0, and $cov(z_2, \nu) = 0$. Then given that the assumptions hold, the instrument we use is nothing but,

$$x_1^* = \alpha_0 + \alpha_1 q_1 + \alpha_2 q_2 + \alpha_3 z_2$$

The above discussion pertains to the use of population parameters as usual, which we never have. But we can always use a estimated version of the instrument, that is

$$\widehat{x}_1 = \widehat{\alpha}_0 + \widehat{\alpha}_1 q_1 + \widehat{\alpha}_2 q_2 + \widehat{\alpha}_3 z_2$$

That is the instrument is just the predicted OLS dependent variable of x_1. To use this instrument is as before, but to be concrete the moments are as follows,

$$\begin{aligned}
\sum_{i=1}^{n} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{1,i} - \widehat{\beta}_2 z_2 \right) &= 0 \\
\sum_{i=1}^{n} \widehat{x}_1 \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{1,i} - \widehat{\beta}_2 z_2 \right) &= 0 \\
\sum_{i=1}^{n} z_2 \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{1,i} - \widehat{\beta}_2 z_2 \right) &= 0
\end{aligned}$$

What the process has essentially done is to remove all elements of correlation that x_1 has with ϵ by creating this ultimate instrument. The importance of having a good instrument

has already be talked about prior. The process is know as 2 Stage Least Squares Estimation (2SLS) because the first stage is after all using OLS, and it turns out that using the estimated instrument in the original regression means that none of the Gauss Markov assumptions are violated, and can consequently be estimated using OLS. We have to be careful about the calculation of the standard errors. To see the reason, note first that

$$x_1 = \widehat{x}_1 + \nu \tag{5}$$

This means that in the final regression,

$$y_1 = \beta_0 + \beta_1 \widehat{x}_1 + be_2 z_2 + \beta_1 \nu + \epsilon$$

So that although the Gauss Markov assumptions are met, the standard error from the second stage OLS is incorrect, since the true standard error does not involve ν. Fortunately, this is calculated correctly in most statistical packages that provide 2SLS.

Another problem that often arises in using this technique is that of multicollinearity, i.e. that the covariates are highly correlated, which consequently raises the asymptotic variance estimated. To see this,

$$p\lim var(\widehat{\beta}_1) = \frac{\sigma^2}{\widehat{TSS}_2(1 - \widehat{R}_2^2)}$$

The intuition is that the first stage has been regressed on all the exogenous variables, and if the included exogenous variable is contributing the greatest to the first stage estimation of the instrument, it is natural to suffer from multicollinearity.

2SLS can just as well be used in cases when we have more than 1 endogenous variable. Consider the following example where we have 2 endogenous variables.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 z_3 + \beta_4 z_4 + \beta_5 z_5 + \epsilon$$

where x_1, and x_2 are endogenous variables, and z_3, z_4, and z_5 are exogenous. And as usual, $E(\epsilon) = 0$. To estimate this equation, or relationship, we need at least two exogenous variables that do not appear in the above regression, so that they are valid instruments. However, this in itself is not sufficient to guarantee identification (What do I mean by identification? Look back at the expression of 2SLS as moment conditions. For every unknown we need one equation. Yet each moment condition corresponds with one instrument. So to identify the 6 parameters we need two additional instruments to identify all 6 parameters.). The reason is that you recall a good instrument must be correlated with the endogenous variable

but uncorrelated with the errors. If one of the exogenous variables or instruments do not conform to this requirement then endogeneity remains a problem. In general for k endogenous variables, we need at least k instruments or excluded exogenous variables to solve the problem of endogeneity. This sufficient condition for identification is called the **rank condition**.

For the testing of multiple hypothesis, the same problem arises as discussed before since the R^2 cannot be used. Nonetheless, the STATA package has simple valid test commands. Refer to your text for references on page 529.

4 IV Solutions to Errors in Variables Problems

Instrumental Variables Regression can also be likewise used to solve endogeneity even if it arises from measurement errors. Consider the following regression relationship,

$$y = \beta_0 + \beta_1 x_1^* + \beta_2 x_2 + \epsilon$$

where x_1^* is an unobserved variable, but of which we have x_1 which is an observed measurement of x_1^*.

$$x_1 = x_1^* + \gamma$$

You should recall that because of this measurement error, estimates of the parameters to be biased. Under certain circumstances, we can use instrumental variable regression to solve the problem. Scanning the above relationships, you can guess that what we need is an exogenous variable that is uncorrelated with both ϵ, and γ, but correlated with x_1. The idea is rather convoluted, but will be clear if you think hard about it.

1. One possibility is to obtain a second measurement on the unobserved but measured with error variable, x_1^*. Let call that second variable that is likewise measured with error, z_1. It is natural to assume that z_1 is uncorrelated with the original error term ϵ since it is measuring a variable that is assumed to be uncorrelated with ϵ. Let $z_1 = x_1^* + \phi$, where ϕ is the measurement error of z_1. Because $z_1 \neq x_1$ neither ϕ and γ are correlated. But assuredly z_1 is correlated with x_1 since they are both measurements for x_1^*, which suggests that we can use z_1 as an instrument for x_1. Although this situation is rare, there are circumstances where it might occur. Read your text on page 530.

2. Another alternative is just to find an exogenous but excluded variable as an instrument for the variable that is measured with error, z_1.

5 Testing for Endogeneity and Testing Overidentifying Restrictions

5.1 Testing for Endogeneity

As we have initially found, because the standard errors under IV are larger, implying less efficient, it would be good if we have a method of testing for endogeneity to examine if it exists at all. If the evidence for endogeneity is small, it makes sense that we follow the usual prescribed procedure (which depends on the type of regression we would otherwise have performed).

The test is call the Hausman Test, and the procedure is as follows:

Consider the following model,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 z_2 + \beta_3 z_3 + \epsilon$$

where x_1 is the endogenous variable.

1. Estimate the reduced form for x_1 by regressing it on all the exogenous variables (all exogenous variables together with the instrumental variable). That is

   $$x_1 = \alpha_0 + \alpha_1 z_1 + \alpha_2 z_2 + \alpha_3 z_3 + \nu$$

   Where z_1 is the instrumental variable. Obtain the predicted residuals, $\widehat{\nu}$.

2. Add $\widehat{\nu}$ to the regression,

   $$y = \beta_0 + \beta_1 x_1 + \beta_2 z_2 + \beta_3 z_3 + \delta\widehat{\nu} + \epsilon$$

   and perform the regression to obtain OLS estimates. Test for significance of $\widehat{\nu}$.

3. If δ is statistically different from zero, conclude that x_1 is endogenous. (You should also use a heteroskedastic-robust t test. That is you should calculate the heteroskedasticity robust standard errors.)

The intuition of the test is as follows. Consider the above regression model. We know that if x_1 is indeed exogenous, then both OLS and 2SLS produce consistent estimates. Then what we want to do is to see if the difference in the estimates is statistically significant. However, to do this comparison, it is easier to do a regression test, that is to include a variable within a regression, and see if the coefficient estimated is statistically significant. Consider the following regression,

$$x_1 = \alpha_0 + \alpha_1 z_1 + \alpha_2 z_2 + \alpha_3 z_3 + \nu$$

Next note that z_1 to z_3 are exogenous variables variables and are by assumption uncorrelated with ϵ. Then x_1 the suspected endogenous variable is exogenous if and only if it is uncorrelated with ϵ, which in turn is true if and only if ν is uncorrelated with ϵ (Since everything else is exogenous already). Then consider the relationship,

$$\epsilon = \delta\nu + \phi$$

where ϕ is uncorrelated with ν, and has zero mean. Then ϵ and ν are uncorrelated if and only if δ is zero. This can be easily achieved by including ν into

$$y = \beta_0 + \beta_1 x_1 + \beta_2 z_2 + \beta_3 z_3 + \epsilon$$

which is easily achieved by using $\widehat{\nu}$ in place of ν.

This test works as well when we have more than 1 suspected endogenous variable. For each suspected endogenous variable, obtain the reduced form residuals. Then test for joint significance of the residuals using an F test. Joint significance indicates at least one of the suspect variables is endogenous.

5.2 Testing Overidentifying Restrictions

A good instrument cannot be correlated with the original error, ϵ, but must be correlated with the endogenous variable it is instrumenting for. We have just provided for a test of the second requirement. But the first cannot, since ϵ is not observed. But if we have more than one instrument, we can test whether some of them are uncorrelated with ϵ.

Consider the same model of

$$y = \beta_0 + \beta_1 x_1 + \beta_2 z_2 + \beta_3 z_3 + \epsilon$$

and suppose you have two additional exogenous variables that could be used as instruments, q_1, and q_2.

The procedure is as follows,

1. Estimate the model,
$$y = \beta_0 + \beta_1 x_1 + \beta_2 z_2 + \beta_3 z_3 + \epsilon$$
using 2SLS. Obtain the 2SLS residuals, $\widehat{\epsilon}$

2. Regress $\widehat{\epsilon}$ on all exogenous variables, and obtain R_1^2.

3. Under H_0 that all IVs are uncorrelated with ϵ. nR_1^2 is asymptotically distributed as χ_p^2, where p is the number of instrumental variables the number of extra exogenous variables not used.

4. If nR_1^2 exceeds the predetermined critical value, we conclude that at least some of the IVs are not exogenous.

The intuition of this test is as follows, in relation to the model above, and the two exogenous variables q_1, and q_2. Suppose we believe that q_1 is the better instrument (suppose we can't use both or a combination of both), we can compute the 2SLS estimate for β_1. Since q_2 is not used as an instrument, we can check to see if q_2 is correlated with $\widehat{\epsilon}$. If it is, then q_2 is not a valid instrument (of course all this while assuming q_1 is a valid instrument). This tells us nothing about whether q_1 is a valid instrument. But if q_1 and q_2 are very related measures, then the fact that one of them is not a valid instrument hence also suggests that the one we're using isn't a good instrument as well. Of course you can always reverse the assumption that q_2 is the better instrument, and perform the test again, testing q_1 instead. But it has been found that which choice does not matter. All we need is to assume is one of them is exogenous, then testing the **Overidentifying Restrictions**, which is just the test above. This test hence cannot be performed if all we have is one instrument for one endogenous variable. In the case above, because we have two exogenous variables excluded for one endogenous variable, we say that we have one overidentifying restriction, and if we have three additional excluded exogenous variables, and one endogenous variable, then we have two overidentifying restriction... You would have to perform the above procedure for each restriction.

6 2SLS with Heteroskedasticity

Just because we are facing endogeneity of variables, does not mean we can ignore other lesser problems such as heteroskedasticity. But given current statistical packages, all we need is to calculate a heteroskedasticity-robust standard error (which is STATA involves using the ˜robust″command with ivreg.).

To test for heteroskedasticity, you could perform a procedure similar to Breusch-Pagan test (Read your text for a brief on the relevant references. The procedure does not differ much, all you need to note now is that you obtain the residuals from 2SLS and not OLS. And the regression on the square of the residuals in all the **exogenous** variables.).

Further, if you know how the error variance depends on the exogenous variables, you can apply a weighted 2SLS. All this involves is to transform all the variables with the weights and performing 2SLS.