Lecture 8

Endogeneity 2 of 4: IV & 2SLS (ii)

Michael Curran

Trinity College Dublin

JS Econometrics

Lecture 8 Outline

2SLS

Introduction

- In the last lecture we assumed there was a single endogenous explanatory variable y_2 and one IV for y_2.
- Sometimes we have more than one exogenous variable that is excluded from the structural model and might be correlated with y_2, which means that they are valid IVs for y_2.
- So, in this lecture, we will discuss how to use multiple IVs.

2SLS

Single Endogenous Explanatory Variable

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1 \tag{1}$$

Suppose *two* exogenous variables are excluded from (1), z_2 and z_3. This assumption and the assumption that z_2 and z_3 are uncorrelated with u_1 are called **exclusion restrictions**. If z_2 and z_3 are both correlated with y_2, could just use each as an IV as we did before: we would have 2 IV estimators. Since each of z_1, z_2 and z_3 is uncorrelated with u_1, any linear combination is also uncorrelated with u_1 and so any linear combination of exogenous variables is a valid IV. To find the best IV, choose the linear combination that is most highly correlated with y_2, which turns out to be given by the reduced form equation for y_2:

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + v_2 \tag{2}$$

$$E(v_2) = 0, Cov(z_1, v_2) = 0, Cov(z_2, v_2) = 0, Cov(z_3, v_2) = 0$$

2SLS

Single Endogenous Explanatory Variable

The best IV for y_2:

$$y_2^* = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3$$

For this IV not to be perfectly correlated with z_1, need at least one of π_2 or π_3 to be different from zero:

$$\pi_2 \neq 0 \vee \pi_3 \neq 0 \tag{3}$$

Key identifying assumption: structural equation (1) is not identified if $\pi_2 = 0$ and $\pi_3 = 0$. Can test $H_0 : \pi_2 = 0$ and $\pi_3 = 0$ against (3) using an F statistic. Useful way to think about (2) is that it breaks y_2 into two pieces:

1. y_2^*: part of y_2 uncorrelated with error term u_1.
2. v_2: part that's possibly correlated with u_1, which is why y_2 is possibly endogenous.

2SLS

Single Endogenous Explanatory Variable

Given data on z_j, we can compute y_2^* for each observation if we know population parameter π_j – never true in practice, but we an always estimate the reduced form by OLS, so using sample, regress y_2 on z_1, z_2 and z_3 and obtain fitted values:

$$\hat{y}_2 = \hat{\pi}_0 + \hat{\pi}_1 z_1 + \hat{\pi}_2 z_2 + \hat{\pi}_3 z_3 \tag{4}$$

Verify z_2 and z_3 jointly significant in (2); else IV estimation is a waste of time. Use $\hat{y}_2$ as the IV for y_2. To estimate β_0, β_1 and β_2, use:

$$\sum_{i=1}^{n} (y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) = 0$$

$$\sum_{i=1}^{n} z_{i1}(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) = 0$$

$$\sum_{i=1}^{n} \hat{y}_{i2}(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) = 0$$

Solving the three equations in three unknowns yields IV estimators.

2SLS

Single Endogenous Explanatory Variable

With multiple instruments, IV estimator using $\hat{y}_{i2}$ as the instrument is called the **two stage least squares (2SLS) estimator** since using $\hat{y}_2$ as IV for y_2, IV estimates $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\beta}_2$ are *identical* to OLS of

$$y_1 \text{ on } \hat{y}_2 \text{ and } z_1 \tag{5}$$

i.e. can obtain 2SLS estimator in two stages:

1. Run regression in (4) to obtain fitted values $\hat{y}_2$.
2. OLS reg (5).

Since use $\hat{y}_2$ in place of y_2, 2SLS estimates can differ substantially from OLS estimates. Some economists like to interpret regression in (5) as follows. Fitted value $\hat{y}_2$ is estimated version of y_2^* and y_2^* is uncorrelated with u_1. So, 2SLS first 'purges' y_2 of its correlation with u_1 before doing the OLS reg in (5). Can show this by plugging $y_2 = y_2^* + v_2$ into (1):

$$y_1 = \beta_0 + \beta_1 y_2^* + \beta_2 z_1 + u_1 + \beta_1 v_2 \tag{6}$$

Composite error $u_1 + \beta_1 v_2$ has zero mean and is uncorrelated with y_2^* and z_1, which is why OLS regression (5) works.

2SLS

Single Endogenous Explanatory Variable

- Most metrics packages have special commands for 2SLS so no need to perform the 2 stages explicitly.
- Most cases you should avoid doing the second stage manually as the standard errors and test statistics obtained in this way are *not* valid (because the error term in (6) includes v_2 but the standard errors involve the variance of u_1 only).
- Any regression software that supports 2SLS asks for the dependent variable, list of explanatory variables (both exogenous and endogenous) and entire list of IV (all exogenous variables).
- Output is similar to that of OLS typically.

2SLS

Single Endogenous Explanatory Variable

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \cdots + \beta_k z_{k-1} + u_1 \tag{7}$$

With a single IV for y_2, the IV estimator is identical to 2SLS. So, with one IV for each endogenous explanatory variable, we call the estimation method IV or 2SLS. Adding more exogenous variables changes very little, e.g.:

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + u_1$$

where u_1 is uncorrelated with *exper* and $exper^2$. Suppose we also think mother's and father's education are uncorrelated with u_1. Then, we can use both of these as IVs for *educ*. Reduced form equation for *educ* is:

$$educ = \pi_0 + \pi_1 exper + \pi_2 exper^2 + \pi_3 motheduc + \pi_4 fatheduc + v_2$$

Identification requires that $\pi_3 \neq 0 \vee \pi_4 \neq 0$ (or both of course). Example 15.5.

2SLS

Single Endogenous Explanatory Variable

- Assumptions for 2SLS to have desired sample properties – chapter 15 appendix.
- Summarising: write structural equation as in (7) and assume each z_j to be uncorrelated with u_1.
- Need at least one exogenous variable *not* in the structural equation that is partially correlated with y_2; this ensures consistency.
- For usual 2SLS standard errors and t statistics to be asymptotically valid, need homogeneity assumption: variance of structural error u_1 can't depend on any of the exogenous variables.
- More assumptions required for time series applications (not on course).

2SLS

Multicollinearity

Multicollinearity can lead to large standard errors for OLS estimates (chapter 3). Multicollinearity can be even more serious with 2SLS. (Asymptotic) variance of 2SLS estimator for β_1 can be approximated as:

$$\sigma^2/[S\hat{S}T_2(1-\hat{R}_2^2)] \tag{8}$$

Variance of the 2SLS estimator is larger than that for OLS because:

1. $\hat{y}_2$ by construction has less variation than y_2 (remember: TSS $=$ ESS $+$ RSS; variation in y_2 is TSS, while variation in $\hat{y}_2$ is ESS from first stage regression).
2. Correlation between $\hat{y}_2$ and exogenous variables in (7) is often much higher than correlation between y_2 and these variables.

This defines the multicollinearity problem in 2SLS. Example.

2SLS

Multiple Endogenous Explanatory Variables

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 y_3 + \beta_3 z_1 + \beta_4 z_2 + \beta_5 z_3 + u_1 \qquad (9)$$

2SLS requires *at least two* exogenous variables that don't appear in (9) but that are correlated with y_2 and y_3. Suppose we have 2 excluded exogenous variables, say z_4 and z_5. Need either z_4 or z_5 to appear in each reduced form for y_2 and y_3 – can use F statistics to test – necessary but not sufficient for identification. Suppose z_4 appears in each reduce form but z_5 appears in neither. Then, we don't really have two exogenous variables partially correlated with y_2 and y_3. 2SLS won't produce consistent estimators of the β_j. Generally, when we've more than 1 endogenous explanatory variable in a regression model, identification can fail in several complicated ways, but we can easily state a necessary condition for identification, the **order condition**.

2SLS

Multiple Endogenous Explanatory Variables

- Order condition (necessary condition) for identification of an equation: need at least as many excluded exogenous variables as there are included endogenous explanatory variables in the structural equation.
- Simple to check: only involves counting endogenous and exogenous variables.
- Sufficient condition for identification is **rank condition**, but general statement of rank condition requires matrix algebra and is beyond this course.
- Question 15.3.

2SLS

Testing

- When testing multiple hypotheses after 2SLS estimation, it may be tempting to use either SSR or R-squared form of F statistics.
- However, as suggested by the fact that R-squareds in 2SLS can be negative, usual way of computing F statistics is inappropriate.
- Using 2SLS residuals to compute SSRs for restricted and restricted models, there's no guarantee $SSR_r \geq SSR_{ur}$; if reverse is true, F statistic would be negative.
- Can combine SSR from 2^{nd} stage regression, e.g. (5) with SSR_{ur} to get statistic with approximate F distribution in large samples.
- Many metrics packages have simple-to-use test commands to test multiple hypotheses after 2SLS estimation.

Lecture 8 Outline

Errors-in-Variables

IV Solutions

Seen IV solves omitted variables problem – IV can also deal with measurement error problem.

$$y = \beta_0 + \beta_1 x_1^* + \beta_2 x_2 + u \tag{10}$$

where y and x_2 observed but x_1^* is not. Let x_1 be an observed measurement of x_1^*: $x_1 = x_1^* + e_1$ where e_1 is measurement error. Recall that correlation between x_1 and e_1 causes OLS where x_1 is used in place of x_1^* to be biased and inconsistent:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + (u - \beta_1 e_1) \tag{11}$$

If classical errors-in-variables (CEV) assumptions hold, bias in OLS estimator of β_1 is toward 0. Can't do anything without further assumptions.

Errors-in-Variables

IV Solutions

Can use IV to solve measurement error problem. In (10) assume u uncorrelated with x_1^*, x_1 and x_2; in CEV case, assume e_1 uncorrelated with x_1^* and x_2. These imply x_2 exogenous in (11) but x_1 is correlated with e_1. Need IV for x_1: correlated with x_1, uncorrelated with u so can be excluded from (10) and uncorrelated with the measurement error e_1. One possibility: obtain a second measurement on x_1^* say z_1. Since x_1 affects y, natural to assume z_1 uncorrelated with u. With $z_1 = x_1^* + a_1$ where a_1 is the measurement error in z_1, assume a_1 and e_1 uncorrelated, i.e. x_1 and z_1 both mismeasure x_1^* but their measurement errors are uncorrelated. Can use z_1 as an IV for x_1 as x_1 and z_1 are correlated through their dependence on x_1^*. Where might we get 2 measurements on a variable? Employers can provide a second measure of annual salary for a group of workers. Each spouse can independently report the level of savings or family income. Ashenfelter and Krueger (1994): each twin was asked about his or her sibling's years of education; gives a 2nd measure to be used as an IV for self-reported education in a wage equation.

Errors-in-Variables

IV Solutions

- Having 2 measures of an explanatory variable is rare.
- Alternative: use other exogenous variables as IVs for a potentially mismeasured variable, e.g. using *motheduc* and *fatheduc* as IVs for *educ*.
- If we think that $educ = educ^* + e_1$, then the IV estimates in e.g. 15.5 don't suffer from measurement error if *motheduc* and *fatheduc* are uncorrelated with measurement error e_1.
- Probably more reasonable than assuming *motheduc* and *fatheduc* are uncorrelated with ability, which is contained in u in (10).
- IV methods can also be adopted when using things like test scores to control for unobservable characteristics. Under certain assumptions, proxies can be used to solve omitted variables problem, e.g.
- IQ as proxy variable for unobservable ability.

Errors-in-Variables

IV Solutions

Alternative when IQ doesn't satisfy proxy variable assumptions:

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + abil + u \quad (12)$$

again have omitted ability problem. But have two test scores that are *indicators* of ability. Assume test scores can be written as

$$test_1 = \gamma_1 abil + e_1 \quad test_2 = \delta_1 abil + e_2$$

As ability affects wage, assume $test_1$ and $test_2$ are uncorrelated with u.

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \alpha_1 test_1 + (u - \alpha_1 e_1) \quad (13)$$

Assume e_1 uncorrelated with all explanatory variables in (12) including *abil*, then e_1 and $test_1$ *must* be correlated. *educ not* endogenous in (13) but $test_1$ is. So estimating (13) by OLS produces inconsistent estimators of β_j and α_1. Under our assumptions, $test_1$ doesn't satisfy proxy variable assumptions. Assume e_1 uncorrelated with all explanatory variables in (12) *and* e_1 and e_2 are uncorrelated. Then e_1 uncorrelated with $test_2$. Thus, $test_2$ can be used as an IV for $test_1$. Example 15.6.

Lecture 8 Outline

Testing for endogeneity

Single Explanatory Variable

- 2SLS estimation is less efficient than OLS when explanatory variables are exogenous: 2SLS estimators can have very large standard errors.

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + u_1 \tag{14}$$

  2 additional exogenous variables, z_3 and z_4 don't appear in (14).
- If y_2 is uncorrelated with u_1, we should estimate (14) by OLS.
- How can we test this? Hausman (1978): see if OLS & 2SLS estimates are statistically significantly different – conclude that y_2 must be endogenous (maintaining that the z_j are exogenous).

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + \pi_4 z_4 + v_2 \tag{15}$$

- Since each z_j is uncorrelated with u_1, y_2 is uncorrelated with $u_1 \iff v_2$ uncorrelated with u_1; this is what we wish to test.
- $u_1 = \delta_1 v_2 + e_1$ where e_1 is uncorrelated with v_2 and has 0 mean.
- Then u_1 and v_2 are uncorrelated $\iff \delta_1 = 0$.

Testing for endogeneity

Single Explanatory Variable

- Easiest way to test this is to include v_2 as an additional regressor in (14) and to do a t test.
- Problem with implementing this: v_2 is unobserved since it is the error term in (15).
- Since we can estimate the reduced form for y_2 by OLS, we can obtain the reduced form residuals $\hat{v}_2$. Thus, we estimate

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + \delta_1 \hat{v}_2 + error \tag{16}$$

  by OLS and test $H_0 : \delta_1 = 0$ using a t stat.
- If we reject H_0 at a small significance level, we conclude that y_2 is endogenous since v_2 and u_1 are correlated.

Testing for endogeneity

Single Explanatory Variable

1. Estimate reduced form for y_2 by regressing it on *all* exogenous variables (including those in the structural equation and the additional IV). Obtain residuals $\hat{v}_2$.
2. Add $\hat{v}_2$ to the structural equation (includes y_2) and test for significance of $\hat{v}_2$ using OLS regression. If coefficient on $\hat{v}_2$ is statistically significantly different from 0, conclude y_2 is endogenous. Might want to use heteroscedastic-robust t test.

Testing for overidentifying restrictions

Seen: test if IV correlated with endogenous explanatory variable via t or F test in reduced form regression. Claimed: can't test if IV uncorrelated with error since we don't observe the error; however, if we've more than 1 IV, we can effectively test whether some of them are uncorrelated with the structural error. Consider (14) with 2 additional IVs z_3 and z_4. Can estimate (14) using only z_3 as IV for y_2. Given IV estimates, compute residuals $\hat{u}_1 = y_1 - \hat{\beta}_0 - \hat{\beta}_1 y_2 - \hat{\beta}_2 z_1 - \hat{\beta}_3 z_2$. Since z_4 unused, can check whether z_4 and $\hat{u}_1$ are correlated in sample. If yes, then z_4 is not valid IV for y_2. Says nothing about whether z_3 and u_1 are correlated; to be a useful test, must *assume* that z_3 and u_1 are uncorrelated. However, if z_3 and z_4 are chosen using same logic (e.g. mother's educ and father's educ), finding that z_4 is correlated with u_1 casts doubt on using z_3 as IV. Since roles of z_3 and z_4 can be reversed, we can test whether z_3 is correlated with u_1 provided z_4 and u_1 are assumed to be uncorrelated. Which test should we use? Turns out that our test choice doesn't matter. Must assume that at least one IV is exogenous. Then we can test the **overidentifying restrictions** that are used in 2SLS.

Testing for overidentifying restrictions

Number of overidentifying restrictions = number of extra IVs. Suppose we've only one endogenous explanatory variable. If we've only a single IV for y_2, we've *no* overidentifying restrictions, and there's nothing that can be tested. If we've two IVs for y_2, we've one overidentifying restriction. If we've 3 IVs, we've 2 overidentifying restrictions, etc. Testing overidentifying restrictions is rather simple. We obtain the 2SLS residuals and run an auxiliary regression. Testing overidentying restrictions:

1. Estimate structural equation by 2SLS and obtain 2SLS residuals $\hat{u}_1$.
2. Regress $\hat{u}_1$ on *all exog* variables and obtain R-squared, say R_1^2.
3. Under H_0 that all IVs are uncorrelated with u_1, $nR_1^2 \overset{a}{\sim} \chi_q^2$ where q is number of IVs from outside the model minus the total number of endogenous explanatory variables. If nR_1^2 exceeds (say) 5% critical value in χ_q^2 distribution, reject H_0 and conclude that at least some of the IVs are not exogenous.

Example 15.8.

Testing for overidentifying restrictions

- The overidentifying test can be used whenever we've more instruments than we need.
- If we have just enough instruments, the model is said to be *just identified* and the R-squared in part 2 will be identically 0.
- As mentioned earlier, we cannot test exogeneity of instruments in the just identified case.
- Tests can be made robust to heterogeneity of arbitrary form (advanced Wooldridge).

Lecture 8 Outline

2SLS with Heteroscedasticity

- Similar issues as with OLS.
- Can obtain standard errors and test statistics (asymptotically) robust to heterogeneity of arbitrary and unknown form.
- Metrics packages do this routinely.
- Test for heterogeneity using analog of Breusch-Pagan test.
- Let $\hat{u}$ denote 2SLS residuals and let $z_1, z_2, \ldots, z_m$ denote all the exogenous variables (including those used as IVs for the endogenous explanatory variable). Then, under reasonable assumptions an asymptotically valid statistic is the usual F statistic for joint significance in a regression of $\hat{u}^2$ on $z_1, z_2, \ldots, z_m$. Null hypothesis of homogeneity is rejected if z_j are jointly significant.
- If know how error variance depends on exogenous variables, can use W2SLS procedure.

Lecture 8 Outline

Summary

- Multiple IVs necessary when more than one endogenous explanatory variable.
- Exclusion restrictions: explanatory variables uncorrelated with disturbances.
- 2SLS second most popular linear estimator after OLS.
  - Multicollinearity more serious issue with 2SLS than OLS.
  - Order and rank conditions important for identification.
  - Don't use SSR of R-squared form for F statistic.
- Along with omitted variables, IV also solves errors-in-variables / measurement error problem.
- Tests: Hausman for single explanatory variable, more than one: test for overidentifying restrictions.

References

- 2SLS: Wooldridge 15.3.
- IV solutions to errors-in-variables: Wooldridge 15.4.
- Testing: Wooldridge 15.5.
- 2SLS with heteroscedasticity: Wooldridge 15.6.