Instrumental Variables & 2SLS

- $y_1 = b_0 + b_1 y_2 + b_2 z_1 + \ldots b_k z_k + u$
- $y_2 = p_0 + p_1 z_{k+1} + p_2 z_1 + \ldots p_k z_k + v$

Why Use Instrumental Variables?

- Instrumental Variables (IV) estimation is used when your model has endogenous *x*'s
- i.e. whenever $\text{Cov}(x,u) \neq 0$
- Thus, IV can be used to address the problem of omitted variable bias
- Also, IV can be used to solve the classic errors-in-variables problem

What Is an Instrumental Variable?

- In order for a variable, *z*, to serve as a valid instrument for *x*, the following must be true

1. The instrument must be exogenous

- i.e. $Cov(z,u) = 0$

-more specifically z should have no "partial" effect on y and should be uncorrelated with u

2. The instrument must be correlated with the endogenous variable *x*

- i.e $Cov(z,x)$ 0

Difference between IV and Proxy?

- With IV we will leave the unobserved variable in the error term but use an estimation method that recognizes the presence of the omitted variable
- With a proxy we were trying to remove the unobserved variable from the error term e.g. IQ
- IQ would make a poor instrument as it would be correlated with the error in our model (ability in *u*)
- Need something correlated with education but uncorrelated with ability (parents education?)

More on Valid Instruments

- We can't test if Cov(z,u) = 0 as this is a population assumption
- Instead, we have to rely on common sense and economic theory to decide if it makes sense
- However, we can test if Cov(z,x) ≠ 0 using a random sample
- Simply estimate $x = p_0 + p_1 z + v$, and test $H_0: p_1 = 0$
- Refer to this regression as the "first-stage regression"

IV Estimation in the Simple Regression Case

- We can show that if we have a valid instrument we can get consistent estimates of the parameters
- For $y = b_0 + b_1 x + u$
- $\text{Cov}(z,y) = b_1 \text{Cov}(z,x) + \text{Cov}(z,u)$, so
- Thus, $b_1 = \text{Cov}(z,y) / \text{Cov}(z,x)$, since $\text{Cov}(z,u)=0$ and $\text{Cov}(z,x) \neq 0$
- Then the IV estimator for b_1 is

$$\hat{b}_1 = \frac{\sum (z_i - \bar{z})(y_i - \bar{y})}{\sum (z_i - \bar{z})(x_i - \bar{x})}$$

- Same as OLS when $z = x$

Inference with IV Estimation

- The IV estimator also has an approximate normal distribution in large samples
- To get estimates of the standard errors we need a slightly different homoskedasticity assumption: $E(u^2|z) = s^2 = \text{Var}(u)$ (conditioning on z here)
- If this is true, we can show that the asymptotic variance of β1-hat is:

$$Var(\hat{b}_1) = \frac{s^2}{n s_x^2 r_{x,z}^2}$$

  - σ_x^2 is the pop variance of x
  - σ^2 is the pop variance of u
  - ρ^2_{xz} is the square of the pop correlation between x and z

Inference with IV Estimation

- Each of the elements in the population variance can be estimated from a random sample
- The estimated variance is then:

$$Var(\hat{b}_1) = \frac{\hat{s}^2}{SST_x R^2_{x,z}}$$

- Where σ^2 = SSR from IV divided by the df, SST_x is the sample variance in *x* and the R^2 is from the first stage regression (*x* on *z*)
- The standard error is just the square root of this

IV versus OLS estimation

- Standard error in IV case differs from OLS only in the R^2 from regressing x on z
- Since $R^2 < 1$, IV standard errors are larger
- However, IV is consistent, while OLS is inconsistent, when $\text{Cov}(x,u) \neq 0$
- Notice that the stronger the correlation between z and x, the smaller the IV standard errors

The Effect of Poor Instruments

- What if our assumption that Cov(z,u) = 0 is false?
- The IV estimator will be inconsistent also
- We can compare the asymptotic bias in OLS to that in IV in this case:

$$\text{IV}: \operatorname{plim} \hat{b}_1 = b_1 + \frac{Corr(z,u)}{Corr(z,x)} \bullet \frac{s_u}{s_x}$$

$$\text{OLS}: \operatorname{plim} \tilde{b}_1 = b_1 + Corr(x,u) \bullet \frac{s_u}{s_x}$$

- Even if Corr(z,u) is small the inconsistency can be large if Corr(z,x) is also very small

Effect of Poor Instruments (cont)

- So, it is not necessarily better to us IV instead of OLS even if z and u are not "highly" correlated
- Instead, prefer IV only if $\text{Corr}(z,u)/\text{Corr}(z,x) < \text{Corr}(x,u)$
- Also notice that the inconsistency gets really large if z and x are only loosely correlated
- Best to test for correlation in the first stage regression

A Note on R^2 in IV

- R^2 after IV estimation can be negative
- Recall that $R^2 = 1 - SSR/SST$ where SSR is the residual sum of IV residuals
- SSR in this case can be larger than SST making the R^2 negative
- Thus, R^2 isn't very useful here and can't be used for F-tests
- Not important as we would prefer consistent estimates of the coefficients

IV Estimation in the Multiple Regression Case

- IV estimation can be extended to the multiple regression case
- Estimating: $y_1 = b_0 + b_1 y_2 + b_2 z_1 + u_1$
- Where y_2 is endogenous and z_1 is exogenous
- Call this the "structural model"
- If we estimate the structural model the coefficients will be biased and inconsistent
- Thus, we need an instrument for y_2
- Can we use z_1 if it is correlated with y_2 (we know it isn't correlated with u_1)?

Multiple Regression IV (cont)

- No, because it appears in the structural model
- Instead, we need an instrument, z_2, that:

1. Doesn't belong in the structural model

2. Is uncorrelated with u_1

3. Is correlated with y_2 in a particular way

- Now because of z_1 we need a partial correlation
- i.e. for the "reduced form equation"

$$y_2 = p_0 + p_1 z_1 + p_2 z_2 + v_2,\ p_2 \neq 0$$

- If we have such an instrument and u_1 is uncorrelated with z_1 the model is "identified"

Two Stage Least Squares (2SLS)

- It is possible to have multiple instruments
- Consider the structural model, with 1 endogenous, y_2, and 1 exogenous, z_1, RHS variable
- Suppose that we have two valid instruments, z_2 and z_3
- Since z_1, z_2 and z_3 are uncorrelated with u_1, so is any linear combination of these
- Thus, any linear combination is also a valid instrument

Best Instrument

- The best instrument is the one that is most highly correlated with y_2
- This turns out to be a linear combination of the exogenous variables
- The reduce form equation is:

  $y_2 = p_0 + p_1 z_1 + p_2 z_2 + p_3 z_3 + v_2$ or $y_2 = y_2^* + v_2$
- Can think of y_2^* as the part of y_2 that is uncorrelated with u_1 and v_2 as the part that might be correlated with u_1
- Thus the best IV for y_2 is y_2^*

More on 2SLS

- We can estimate y_2* by regressing y_2 on z_1, z_2 and z_3 – the first stage regression
- If then substitute $_2$ for y_2 in the structural model, get same coefficient as IV
- While the coefficients are the same, the standard errors from doing 2SLS by hand are incorrect
- Also recall that since the R2 can be negative F-tests will be invalid
- Stata will calculate the correct standard error and F-tests

More on 2SLS (cont)

- We can extend this method to include multiple endogenous variables
- However, we need to be sure that we have at least as many excluded exogenous variables (instruments) as there are endogenous variables
- If not, the model is not identified

Addressing Errors-in-Variables with IV Estimation

- Recall the classical errors-in-variables problem where we observe x_1 instead of x_1^*
- Where $x_1 = x_1^* + e_1$, we showed that when x_1 and e_1 are correlated the OLS estimates are biased
- We maintain the assumption that u is uncorrelated with x_1^*, x_1 and x_2 and that and e_1 is uncorrelated with x_1^* and x_2
- If we can find an instrument, z, such that Corr(z,u) = 0 and Corr(z,x_1) $\neq$ 0, then we can use IV to remove the attenuation bias

Example of Instrument

- Suppose that we have a second measure of $x_1^*(z_1)$

Examples:

- both husband and wife report earnings
- both employer and employee report earnings
- z_1 will also measure x_1^* with error
- However, as long as the measurement error in z_1 is uncorrelated with the measurement error in x_1, z_1 is a valid instrument

Testing for Endogeneity

- Since OLS is preferred to IV if we do not have an endogeneity problem, then we'd like to be able to test for endogeneity
- Suppose we have the following structural model:

  $y_1 = b_0 + b_1 y_2 + b_2 z_1 + b_3 z_2 + u$
- We suspect that y_2 is endogenous and we have instruments for y_2 (z_3, z_4)
- How do we determine if y_2 is endogenous?

Testing for Endogeneity (cont)

1. Hausman Test

- If all variables are exogenous both OLS and 2SLS are consistent
- If there are statistically significant differences in the coefficients we conclude that y_2 is endogenous

2. Regression Test

- In the first stage equation:

  $y_2 = p_0 + p_1 z_1 + p_2 z_2 + p_3 z_3 + p_3 z_3 + v_2$

- Each of the z's are uncorrelated with u_1

Testing for Endogeneity (cont)

- Thus y_2 will only be correlated with u_1 if v_2 is correlated with u_1
- The test is based on this observation
- So, save the residuals from the first stage
- Include the residual in the structural equation (which of course has y_2 in it)
- If the coefficient on the residual is statistically different from zero, reject the null of exogeneity
- If multiple endogenous variables, jointly test the residuals from each first stage

Testing Overidentifying Restrictions

- How can we determine if we have a good instrument -correlated with y_2 uncorrelated with u?
- Easy to test if z is correlated with y_2
- If there is just one instrument for our endogenous variable, we can't test whether the instrument is uncorrelated with the error (u is unobserved)
- If we have multiple instruments, it is possible to test the overidentifying restrictions
- i.e. to see if some of the instruments are correlated with the error

The OverID Test

- Using our previous example, suppose we have two instruments for y_2 (z_3, z_4)
- We could estimate our structural model using only z_3 as an instrument, assuming it is uncorrelated with the error, and get the residuals:

$$\hat{u}_1 = y_1 - \hat{b}_0 - \hat{b}_1 y_2 - \hat{b}_2 z_1 - \hat{b}_3 z_2$$

- Since z_4 hasn't been used we can check whether it is correlated with *u_1-hat*
- If they are correlated z_4 isn't a good instrument

The OverID Test

- We could do the same for z_3, **as long as we can assume that z_4 is uncorrelated with u_1**
- A procedure that allows us to do this is:

1. Estimate the structural model using IV and obtain the residuals
2. Regress the residuals on all the exogenous variables and obtain the R^2 to form nR^2
3. Under the null that all instruments are uncorrelated with the error, LM ~ χ_q^2 where q is the number of "extra" instruments

Testing for Heteroskedasticity

- When using 2SLS, we need a slight adjustment to the Breusch-Pagan test
- Get the residuals from the IV estimation
- Regress these residuals squared on **all** of the exogenous variables in the model (including the instruments)
- Test for the joint significance
- Note: there are also robust standard errors in the IV setting

Testing for Serial Correlation

- Also need a slight adjustment to the test for serial correlation when using 2SLS
- Re-estimate the structural model by 2SLS, including the lagged residuals, and using the same instruments as originally
- Test if the coefficient on the lagged residual (ρ) is statistically different than zero
- Can also correct for serial correlation by doing 2SLS on a quasi-differenced model, using quasi-differenced instruments