Instrumental Variables & 2SLS

\[ y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \ldots + \beta_k z_k + u \]

\[ y_2 = \pi_0 + \pi_1 z_{k+1} + \pi_2 z_1 + \ldots + \pi_k z_k + \nu \]

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
  2. The instrument must be correlated with the endogenous variable \( x \)

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
More on Valid Instruments

- We can’t test if $\text{Cov}(z,u) = 0$ 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 $\text{Cov}(z,x) \neq 0$ using a random sample
- Simply estimate $x = \pi_0 + \pi_1 z + \nu$, and test $H_0: \pi_1 = 0$

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 $y = \beta_0 + \beta_1 x + u$
- $\text{Cov}(z,y) = \beta_1 \text{Cov}(z,x) + \text{Cov}(z,u)$, so
- Thus, $\beta_1 = \frac{\text{Cov}(z,y)}{\text{Cov}(z,x)}$, since $\text{Cov}(z,u)=0$ and $\text{Cov}(z,x) \neq 0$
- Then the IV estimator for $\beta_1$ is
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:

If this is true, we can show that the asymptotic variance of $\beta_1$-hat is:

$$Var(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{SST_x R_{x,z}^2}$$
### 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.

### The Effect of Poor Instruments

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

\[
\text{IV: } \text{plim} \hat{\beta}_1 = \beta_1 + \frac{\text{Corr}(z,u)}{\text{Corr}(z,x)} \frac{\sigma_u}{\sigma_x} \\
\text{OLS: } \text{plim} \tilde{\beta}_1 = \beta_1 + \text{Corr}(x,u) \frac{\sigma_u}{\sigma_x}
\]
Effect of Poor Instruments (cont)

- So, it is not necessarily better to use IV instead of OLS even if \( z \) and \( u \) are not “highly” correlated.
- Also notice that the inconsistency gets really large if \( z \) and \( x \) are only loosely correlated.

A Note on \( R^2 \) in IV

- \( R^2 \) after IV estimation can be negative.
- Recall that \( R^2 = 1 - \frac{SSR}{SST} \) 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.
IV Estimation in the Multiple Regression Case

- IV estimation can be extended to the multiple regression case
- Estimating: \( y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1 \)
- 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 \( y_2 \)

Multiple Regression IV (cont)

- No, because it appears in the structural model
- Instead, we need an instrument, \( z_2 \), that:
  1. 
  2. 
  3. 
- Now because of \( z_1 \) we need a partial correlation
- i.e. for the “reduced form equation”
  \( y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \nu_2 \)
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:
  
  $y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + v_2$ or $y_2 = y_2^* + v_2$
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 $\hat{y}_2$ 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
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 $x_2$.

Example of Instrument

- Suppose that we have a second measure of $x_1^*(z_1)$.
- Examples:
  - $z_1$ 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:
\[ y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + u \]

We suspect that \( y_2 \) is endogenous and we have instruments for \( y_2 \) (\( z_3, z_4 \))

Testing for Endogeneity (cont)

1. Hausman Test
   - If all variables are exogenous both OLS and 2SLS are consistent

2. Regression Test
   - In the first stage equation:
     \[ y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + \pi_4 z_4 + v_2 \]
   - Each of the \( z \)’s are uncorrelated with \( u_j \)
Testing for Endogeneity (cont)

- 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

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
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:

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.  

2.  

3. Under the null that all instruments are uncorrelated with the error, \( LM \sim \chi^2_q \) 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 ($\rho$) is statistically different than zero
- Can also correct for serial correlation by doing 2SLS on a quasi-differenced model, using quasi-differenced instruments