Testing whether an X is Endogenous

Tyler Ransom

Univ of Oklahoma

Mar 28, 2019

Today's plan

1. Review reading topics on when $E(u|\mathbf{x}) \neq 0$
   1.1 Review of IV topics
   1.2 Testing whether an x is Endogenous
2. In-class activity: Work on project

Quick Review

Instrumental Variables (IV)

- An IV (call it z) is a variable correlated with x, but not with u
  - IV's typically come out of so-called **natural experiments**
  - e.g. exogenous change in laws; school choice lotteries; military conscription
- Allows us to estimate a causal effect, even when A4 is violated

IV conditions

- The instrument z must satisfy

  1. z is **exogenous** to the equation:

  $$Cov(z, u) = 0$$

  2. z is **relevant** for explaining x:

  $$Cov(z, x) \neq 0$$

Practical considerations in IV

- How “relevant” is relevant?
  - 1st stage $F > 10$ or $|t| > \sqrt{10} \approx 3.2$
  - i.e. we need z to be strongly correlated with x
  - Otherwise, our estimation suffers from **weak instruments**
- Can we test exogeneity condition?
  - No; typically need to appeal to theory or qualitative evidence
  - We can, however, test if we even need to do IV

Multiple instruments, multiple endogenous x's

- With multiple z's and endogenous x's, must satisfy the **order condition**:
- Must exclude at least as many z's from our equation as endogenous x's
- Example: Female labor supply (from a few times ago)
- `KIDCOUNT` is endogenous x, `SAMESEX` is excluded z
- Could also come up with additional instruments for `KIDCOUNT`

Testing whether an *x* is Endogenous

Setting

- Consider the model
$$y_1 = \alpha_1 y_2 + \mathbf{z}_1 \delta_1 + u_1,$$
where $\mathbf{z}_1 \delta_1$ represents a set of exogenous rhs vars & coeffs (incl. constant)

- We want to test whether y_2 and u_1 are uncorrelated

- i.e. $H_0 : y_2$ is exogenous (which means we could use OLS rather than IV)

- Assume: elements of $\mathbf{z}_1$ are exogenous (so they act as their own IVs)

- Need at least one outside exogenous variable (could have more; call it $\mathbf{z}_2$)

Reduced-form Equation

- The **reduced form** for y_2 is

$$y_2 = \mathbf{z}_1\pi_1 + \mathbf{z}_2\pi_2 + v_2$$

- $\mathbf{z}_1\pi_1$ is shorthand for $z_{11}\pi_{11} + \cdots + z_{L1}\pi_{L1}$

- $\mathbf{z}_2\pi_2$ is shorthand for $z_{12}\pi_{12} + \cdots + z_{M2}\pi_{M2}$

- All $\mathbf{z}$'s are exogenous

- With y_2 written this way, it is endogenous if and only if

$$Cov(v_2, u_1) \neq 0$$

Setting up the test

- To test $H_0 : Cov(v_2, u_1) = 0$, we can write

$$u_1 = \rho_1 v_2 + e_1$$

where the new error e_1 is uncorrelated with $\mathbf{z}$ and v_2, and therefore y_2

- Plug in for u_1 into the original equation:

$$\begin{aligned} y_1 &= \alpha_1 y_2 + \mathbf{z}_1 \delta_1 + u_1 \\ &= \alpha_1 y_2 + \mathbf{z}_1 \delta_1 + \rho_1 v_2 + e_1 \end{aligned}$$

which is an equation that, if we observed v_2, could be estimated by OLS

The testing procedure

- In the equation
$$y_1 = \alpha_1 y_2 + \mathbf{z}_1 \delta_1 + \rho_1 v_2 + e_1,$$
e_1 is uncorrelated with y_2, $\mathbf{z}_1$, and v_2

- we can estimate v_2 using the first-stage regression

- The two-step testing procedure is

1. Regress y_2 on $\mathbf{z}_1$ and $\mathbf{z}_2$ to obtain the residuals, $\hat{v}_2 = y_2 - \mathbf{z}\hat{\pi}_2$
2. Regress y_1 on y_2, $\mathbf{z}_1$, and $\hat{v}_2$; use a robust t-test on $\hat{\rho}_1$

More on the test

- If we do not have an instrument for y_2, this model is perfectly collinear
- $\hat{v}_2$ would be an exact linear function of y_2 and $\mathbf{z}_1$
- Interestingly, OLS estimates from step (2) for the coefficients on y_2 and $\mathbf{z}_1$ ...

  ... are always identical to the 2SLS estimates of the structural equation
- Including the first-stage residuals "controls" for the endogeneity of y_2

Example: School spending and test scores

```
df <- as_tibble(meap_94_95) %>% filter(enroll>2000)

est.reduced.form <- lm(lavgrexp ~ lfound + math4_94, data=df)

df %<>% mutate(v2h = est.reduced.form$residuals)

est.residualized <- lm(math4 ~ lavgrexp + math4_94 + v2h, data

coeftest(est.residualized, vcov=hccm)

               Estimate  Std. Error     t value  Pr(>|t|)
(Intercept)   -27.02431    23.00407       -1.17     0.241
 lavgrexp      6.639055     2.71628        2.44     0.015 **
 math4_94      .6392136    .0324096       19.72     0.000 ***
      v2h     -14.95647    8.179467       -1.83     0.068 *
```

Example: School spending and test scores

2SLS estimation:

```
est.iv <- ivreg(math4 ~ lavgrexp + math4_94 |
                math4_94 + lfound, data=df)

summary(est.iv)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -27.02431   22.9201   -1.18    0.239
   lavgrexp   6.639055  2.709613    2.45    0.015 **
   math4_94   .6392136  .0328142   19.48    0.000 ***
```

Discussion

- Coefficient on `v2h` was significant at 10% level, but not 5% level
- Implies a (marginal) violation of exogeneity condition
- The standard errors are also slightly different between the two
- Should use the ones from `ivreg`