Lecture 10 - Instrumental Variables

Part III : Practice

1. Instrumental Variables and 2SLS

$$y_i = \alpha + \rho s_i + \beta X_i + \varepsilon_i$$

s_i is endogenous because some omitted variable A_i (ability) is both correlated with s_i and y_i. Therefore $E(s_i \varepsilon_i) \neq 0$ $\Rightarrow$ $Cov(s_i, \varepsilon_i) \neq 0$.

Instrument for s_i with z_{i1} and z_{i2}

$$s_i = \pi_0 + \pi_1 z_{i1} + \pi_2 z_{i2} + \pi_3 X_i + \eta_i$$

→ Two conditions for z to be a good instrument:

(i) strong. F-test of π_1 and $\pi_2 \neq 0$ is ≥ 10.

(ii) valid. z_{i1}, z_{i2} only affect y_i through s_i. No other, independent affect on y_i.

↳ Test for this in over identified case:

$$y_i = \alpha + \rho \hat{s}_i + \beta x_i + \varepsilon_i$$

$$\hat{\varepsilon}_i = \gamma_0 + \gamma_1 z_{i1} + \gamma_2 z_{i2} + \gamma_3 x_i + u_i$$

$$NR^2 \sim chisq\left(\#_{inst} - \#_{endog}\right)$$

H_0: instruments are not invalid
H_a: instruments are invalid.

We don't want to reject H_0 in favor of H_a.
Low chi-square statistic $\Rightarrow$ high p-value; cannot reject H_0.

This is 2SLS.

2. Forbidden Regressions
- If you ever run these regressions, I will find you and retroactively fail you.
(not really, but don't run these regressions).

a.
$$s_i = \pi_0 + \pi_1 z_{i1} + \pi_2 z_{i2} + \pi_3 x_{i1} + \eta_i$$
$$y_i = \alpha + \rho \hat{s}_i + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

NO!

x_{i2} not in first-stage and it should be.

b. suppose $s_i = \{0, 1\}$ i.e., Dummy variable for completing high school (=1) or not (=0).

Estimate by probit / logit:

$$s_i = \pi_0 + \pi_1 z_{i1} + \pi_2 z_{i2} + \pi_3 X_{i1} + \eta_i$$

Estimate by OLS:

$$y_i = \alpha + \rho \hat{s}_i + \beta X_{i1} + \varepsilon_i$$

NO!

↓

cannot do this! 2SLS must be two OLS regressions. Because $\hat{s}_i$ is estimated nonlinearly (with probit / logit), then $E(\hat{s}_i \varepsilon_i) \neq 0$, so our estimate of ρ will be biased!

i) You could do the following:

→ estimate by probit / logit:

$$s_{i1} = \pi_{01} + \pi_{11} z_i + \pi_{21} z_{i2} + \eta_{i1}$$

→ estimate by OLS:

$$s_{i2} = \pi_{02} + \pi_{12} \hat{s}_{i1} + \pi_{22} X_i + \eta_{i2}$$

→ estimate by OLS:

$$y_i = \alpha + \rho \hat{s}_{i2} + \beta x_i + \varepsilon_i$$

↳ This is technically fine, but economists don't like it. We're using non-linearities in the first-stage relationship as identifying information, but it's not clear what the underlying experiment is.

c. peer effects / social learning.

$$s_i = \alpha + \rho \bar{s} + \beta x_i + \varepsilon_i$$

$\bar{s}$ is other peers' avg. schooling

↳ schooling is a function of peers' schooling.

It's going to be really difficult to find an instrument that affects $\bar{s}$ only and not s_i.

$z_i \rightarrow \bar{s} \rightarrow s_i$

Therefore, we're still faced with endogeneity issues.

i) Best we can do in this case:

$$s_i = \alpha + \rho \bar{m} + \beta X_i + \varepsilon_i$$

↓

replace $\bar{s}$ with a proxy variable $\bar{m}$ that is determined <u>before</u> (not jointly) with $\bar{s}$. i.e. $\bar{m}$ is number of books in HH in infancy.

3. What happens if:

a. I add lots of <u>irrelevant</u> z's?

$$s_i = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 X_i + \eta_i \qquad \text{F-stat} \sim 10$$

$$s_i = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + \pi_4 z_4 + \pi_5 z_5 + \pi_6 z_6 + \pi_7 z_7 + \pi_8 X_i + \eta_i$$

If $\pi_3 \approx \pi_4 \approx \ldots \approx \pi_7 \approx 0$, F-stat ↓

Why? F-stat is a test that $\pi_1, \ldots \pi_7$ jointly different from zero. Even if π_1, π_2 are different from zero, it's now much more difficult to reject that π_1 <u>and</u> π_2 <u>and</u> π_3 <u>and</u> π_4 ... <u>and</u> π_7 are all different from zero.

When F-stat low:

$\rho_{2SLS} \rightarrow \rho_{OLS}$, i.e., ρ_{2SLS} becomes more like the biased ρ_{OLS}.

↳ What to do if instruments are weak:

LIML. ⇒ less biased but "noisier" estimates of ρ than 2SLS.

SUMMARY

(1) Report first-stage. Does it make sense? Are magnitudes and signs on z's as expected?

(2) Report F-stats. The bigger the better!

(3) In over-identified case (more z's than x's), pick the best instruments and show results from a just-identified case

(4) Report over-identification test results. If reject H_0, STOP! get rid of "invalid" z's. If do not reject H_0, try LIML. Hopefully $\rho_{LIML} \sim \rho_{2SLS}$. ρ_{LIML} less biased but also less precise than ρ_{2SLS}.

(5) Run reduced-form regression.
Look at coeffs, t-stats, F-stats for excluded instruments in the reduced form. Remember that reduced form estimates are proportional to the causal effect of interest and are unbiased (if your instruments are valid). If you can't already see a causal relationship in your reduced form, it's probably not there.