Recitation 4: Weak IV, AR test, LIML and LATE

Seung-hun Lee

Columbia University
Introduction to Econometrics II Recitation

February 7th, 2022

Weak IV

The origin of the cutoff value 10

- Suppose X_2 is our endogenous variable and we have the following first-stage equation

$$X_2 = X_1\pi_{21} + Z_2\pi_{22} + v_2$$

- We usually test the F-statistic values for π_{22} parameter. The idea is that we can approximately write

$$\hat{\beta}_{IV} = \beta_0 + \frac{\hat{\beta}_{OLS} - \beta_0}{E(F) - 1}$$

- If $E(F)$ is large, then $\hat{\beta}_{IV}$ approximates β_0 through the decreasing role of $\hat{\beta}_{OLS}$, If it is small, then $\hat{\beta}_{IV}$ is very much closer to $\hat{\beta}_{OLS}$.
- When $F = 10$, we can write from the above equation that

$$\hat{\beta}_{IV} - \beta = \frac{1}{9}(\hat{\beta}_{OLS} - \beta)$$

  so when $F > 10$, then the bias of the IV estimator is around 10% of that for the OLS estimator.

Bias and Size correction incorporated: Stock-Yogo test

- Stock and Yogo (2005) extends on this framework by pointing out that weak IV induces bias and also size[1] distortion for the test.
- Weak IV test is defined to include two aspects:
  - One is to identify the critical values such that the bias of the 2SLS is less than 10% of the OLS bias.
  - The other is to find the critical value for the 2SLS estimators for the 5% test that will have a test size no larger than x%.
- In STATA, if you do a weak iv test with homoskedasticity, there will be a cutoff for $x = 10, 15, 20, 25$. This cutoff is very sensitive to the number of the instruments.
- Above assume homoskedasticity. For robust version, see Montiel-Olea, Pflueger (2013) and Andrews, Stock, Sun (2019)

[1] Probability of wrongly rejecting the null when it is correct.

Inference robust to weak IV: Anderson-rubin test

- Assume

$$y = X_1\beta_1 + X_2\beta_2 + e$$
$$X_2 = X_1\pi_{21} + Z_2\pi_{22} + v_2$$

- We are testing $H_2 : \beta_2 = \beta_2^0$. The problem here is that π_{22} represents the coefficients from the weak IVs.
- Replace X_2 in the structural equation with its reduced form equivalent to get

$$y = X_1\beta_1 + (X_1\pi_{21} + Z_2\pi_{22} + v_2)\beta_2 + e$$

- Then we subtract both sides by $X_2\beta_2^0$, which results in

$$\begin{aligned} y - X_2\beta_2^0 &= X_1\beta_1 + (X_1\pi_{21} + Z_2\pi_{22} + v_2)\beta_2 + e - X_2\beta_2^0 \\ &= X_1[\underbrace{\beta_1 + \pi_{21}(\beta_2 - \beta_2^0)}_{\theta_1}] + Z_2[\underbrace{\pi_{22}(\beta_2 - \beta_2^0)}_{\theta_2}] + [\underbrace{e + v_2(\beta_2 - \beta_2^0)}_{w}] \end{aligned}$$

Inference robust to weak IV: Anderson-rubin test

- So testing for $\beta_2 = \beta_2^0$ is equivalent to testing for

$$\theta_1 = \beta_1, \theta_2 = 0$$

- To test this, we use the following test statistic

$$AR(\beta_2^0) = \frac{[(y - X_2\beta_2^0)' M_{X_1}(y - X_2\beta_2^0) - (y - X_2\beta_2^0)' M_Z(y - X_2\beta_2^0)]/L_2}{(y - X_2\beta_2^0)' M_Z(y - X_2\beta_2^0)/(n-k)} \sim F_{L_2, n-k}$$

  where $Z = [X_1 Z_2]$ and L_2 refers to the exogenous IVs (not counting X_1's), and k is the total dimension of the reduced form estimation.
- Pivotal: No role of π_{22} here
- Confidence set: Invert critical value test for various β_2^0 values to get a non-rejection set

Too much IV is not good either

- Let the number of instruments $l = \alpha n$ or $l/n \to \alpha$
- When α is not zero, then it can be said that this setup has many IVs. This could cause the 2SLS estimators to be inconsistent as well.
- Consider the setup where x_i is endogenous and is a scalar.

$$y_i = x_i'\beta + e_i \iff Y = X\beta + e$$
$$x_i = \pi' z_i + u_i \iff X = Z\Pi + u \ \ (z_i \in \mathbb{R}^l)$$

  - z_i is still a valid IV (relevant, exogenous) and
  - $var \begin{pmatrix} e_i \\ u_i \end{pmatrix} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} = \Sigma.$
  - $\frac{1}{n} \sum_{i=1}^{n} \pi' z_i z_i' \pi \xrightarrow{p} c > 0$
  - $var(x_i) = var(z_i'\pi) + var(u_i)$
  - Variance of x_i and u_i are unchanging with respect to l.

Inconsistency from too many IVs

- The 2SLS estimator can be characterized by

$$\hat{\beta}_{2SLS} - \beta = (X'P_Z X)^{-1}(X'P_Z e)$$
$$= \left[\frac{\pi' Z' Z\pi}{n} + \frac{\pi' Z' u}{n} + \frac{u' Z\pi}{n} + \frac{u' P_Z u}{n}\right]^{-1} \left[\frac{\pi' Z' e}{n} + \frac{u' P_Z e}{n}\right]$$

- Note that $\frac{\pi' Z' Z\pi}{n} \xrightarrow{p} c$, $\frac{\pi' Z' u}{n} \xrightarrow{p} 0$, $\frac{\pi' Z' e}{n} \xrightarrow{p} 0$.
- Using the fact that $u' P_Z e$ and $u' P_Z u$ are scalars,

$$E\left[\frac{1}{n}u' P_Z e\right] = \frac{1}{n}E[tr(u' P_Z e)] = \frac{1}{n}E[tr(P_Z e u')] = \frac{1}{n}tr[E(P_Z e u')]$$
$$= \frac{1}{n}tr[E(P_Z)\rho] = \frac{1}{n}E[tr(P_Z)]\rho = \frac{l}{n}\rho \; (\because tr(E(X)) = E(tr(X)) \text{ and } tr(AB) = tr(BA))$$

- In a similar fashion,

$$E\left[\frac{1}{n}u' P_Z u\right] = \frac{l}{n}$$

Inconsistency from too many IVs

- Based on the two facts and $l/n \to \alpha$, $\frac{1}{n}u'P_Z u \xrightarrow{p} \frac{l}{n}$ and $\frac{1}{n}u'P_Z e \xrightarrow{p} \frac{l}{n}\rho$.
- Since $l/n \to \alpha$, I can write

$$\frac{u'P_Z e}{n} \xrightarrow{p} \alpha\rho, \frac{u'P_Z u}{n} \xrightarrow{p} \alpha$$

- Therefore,

$$\hat{\beta}_{2SLS} - \beta \xrightarrow{p} \frac{\alpha\rho}{c+\alpha}$$

- Workaround: LASSO on X, principal component on Z, or LIML

Alternative estimators to deal with finite sample problems

- Split sample IV (SSIV): Split the sample into two - A and B. In sample A, we derive the reduced form estimate of the first stage estimator for π_{22} so $\hat{\pi}_{22} = (Z_A' Z_A)^{-1}(Z_A' X_A)$. The predicted values are created with sample B, so $\widehat{X}_B = Z_B \hat{\pi}_{22} = Z_B (Z_A' Z_A)^{-1}(Z_A' X_A)$. The resulting estimator is $\hat{\beta}_{SSIV} = (\widehat{X}_B X_B)^{-1}(\widehat{X}_B y_B)$
- Jackknife IV (JIVE): It involves a 'leave-one-out' method where the first-stage and the final estimator is created by leaving one observation out. The predicted value of X in this case is $\widehat{X}_i = Z_i \hat{\pi}_{2,-i}$ where $\hat{\pi}_{2,-i} = (Z'Z)^{-1}(Z'X - Z_i' X_i)$.
- Two-sample IV and 2SLS: Split the sample to two - Sample 1 with n_1 obervations and the other with n_2 observations. The two estimators take the following form

$$\hat{\beta}_{TSIV} = \left(\frac{Z_2' X_2}{n_2}\right)^{-1} \left(\frac{Z_1' Y_1}{n_1}\right)$$

$$\hat{\beta}_{TS2SLS} = (\widehat{X}_1' \widehat{X}_1)^{-1}(\widehat{X}_1' y_1)\ (\widehat{X}_1 = Z_1 (Z_2' Z_2)^{-1} Z_2' X_2)$$

Limited information maximum likelihood

What is LIML?

- Assume a data generating process

$$y_i = \beta_1' x_{1i} + \beta_2' x_{2i} + e_i, \ (x_{1i} \in \mathbb{R}^{k_1}, x_{1i} \in \mathbb{R}^{k_2})$$

  where $E(x_{1i} e_i) = 0$, but $E(x_{2i} e_i) \neq 0$.
- A **limited information maximum likelihood (LIML) estimator** derives the maximum likelihood estimator for the joint distribution of (y_i, x_{2i}) using structural equation of y_i and the reduced form equation for x_{2i}.
- This differs from the full information maximum likelihood (FIML) in the sense that the FIML requires structural equation for x_{2i} as well.

Why use LIML?

- When the number of the instruments are fixed, 2SLS and LIML have the same asymptotic distribution, with the former not requiring the assumption of normality (Greene, 2012).
- However, when there is a problem of weak instrument variable or too many instrumental variables, it can be shown that 2SLS becomes biased towards OLS.
- Others have shown that LIML estimators perform better in the presence of weak IVs and/or too many IVs.[2].

[2] For further explanation: http://econ.lse.ac.uk/staff/spischke/ec533/Weak%20IV.pdf

k-class estimators

- k-class estimator is defined as

$$\hat{\beta}_k = \arg\min_{\beta}(y - X\beta)'(I_n - kM_Z)(y - X\beta)$$

- The other way to express this, after some matrix differentiation, is

$$\begin{aligned} -2X'y + 2kX'M_Z y + 2(X'X)\beta - 2k(X'M_Z X)\beta = 0 \\ \iff (X'(I_n - kM_Z)X)\beta = X'(I_n - kM_Z)y \\ \implies \hat{\beta}_k = (X'(I_n - kM_Z)X)^{-1}(X'(I_n - kM_Z)y) \end{aligned}$$

- OLS ($k = 0$) and 2SLS ($k = 1$) are k-class estimators.
- LIML is a k-class estimator with a parameter choice of some $k > 1$, a minimum eigenvalue of $(W'M_1 W)(W'M_Z W)^{-1}$ where $W = [y\ X_2]$

Local average treatment effect

Potential outcome framework when D_i is a binary treatment variable

- Suppose our setup now involves a binary variable D_i. Then we can write

$$y_i = \alpha + \beta D_i + e_i$$

- To analyze this framework, we define a potential outcome framework $Y_i(D_i)$.
  - $Y_i(1)$ for those entering treatment
  - $Y_i(0)$ for those not entering treatment
- For a given individual i, we only observe only one of the two (*fundamental problem of causal inference*)

What does our regression parameters mean?

- With this framework, we can write

$$
\begin{aligned}
y_i &= D_i Y_i(1) + (1 - D_i) Y_i(0) \\
&= Y_i(0) + D_i(Y_i(1) - Y_i(0)) \\
&= \underbrace{E[Y_i(0)]}_{\alpha} + D_i(\underbrace{Y_i(1) - Y_i(0)}_{\beta}) + \underbrace{Y_i(0) - E[Y_i(0)]}_{e_i}
\end{aligned}
$$

- Our goal is to identify a treatment effect (TE): $E[y_i|D_i = 1] - E[y_i|D_i = 0]$, where

$$
\begin{aligned}
E[y_i|D_i = 1] &= \alpha + \beta + E[e_i|D_i = 1] \\
E[y_i|D_i = 0] &= \alpha + E[e_i|D_i = 0] \\
E[y_i|D_i = 1] - E[y_i|D_i = 0] &= \beta + E[e_i|D_i = 1] - E[e_i|D_i = 0]
\end{aligned}
$$

- identifying TE comes down to whether $E[e_i|D_i = 1] - E[e_i|D_i = 0]$ can be negated.
- If $E[e_i|D_i]$ is purely random in that $E[e_i|D_i = 1] = E[e_i|D_i = 0]$, OLS would identify the treatment effects. This is the average treatment effect (ATE).

If ATE is infeasible

- Find a binary Z_i instrument for D_i
- The new setup is now

$$y = D\beta + e = \alpha + D_i\beta + e_i$$
$$D = Z\pi + v_2 = \phi + \pi Z_i + v_{2i}$$
$$(E[e|Z] = 0, \ E[v_2|Z] = 0)$$

- First stage: Regress D onto Z, so that $\widehat{D} = Z\widehat{\pi}$
- Regress Y onto $\widehat{D}$, the regression becomes

$$y = \widehat{D}\beta + e = Z\widehat{\pi}\beta + e$$

and β can be identified

How does Z work?

- Note that

$$
\begin{aligned}
E[D|Z=1] &= \pi + E[v_2|Z=1] \\
E[D|Z=0] &= E[v_2|Z=0] \\
E[D|Z=1] - E[D|Z=0] &= \pi + E[v_2|Z=1] - E[v_2|Z=0] = \pi \ (\because E[v_2|Z]=0)
\end{aligned}
$$

  and from the second stage regression where we have $Y = (\pi Z + v_2)\beta + e$

$$
\begin{aligned}
E[y|Z=1] &= \pi\beta + E[e|Z=1] + \beta E[v_2|Z=1] \\
E[y|Z=0] &= E[e|Z=0] + \beta E[v_2|Z=0] \\
E[y|Z=1] - E[y|Z=0] &= \pi\beta + E[e|Z=1] - E[e|Z=0] + \beta(E[v_2|Z=1] - E[v_2|Z=0]) \\
&= \pi\beta \ (\because E[e|Z]=0, E[v_2|Z]=0)
\end{aligned}
$$

- Thus, we get the local average treatment effect (LATE)

$$
\beta = \frac{E[y|Z=1] - E[y|Z=0]}{\pi} = \frac{E[y|Z=1] - E[y|Z=0]}{E[D|Z=1] - E[D|Z=0]}
$$

Wald estimator

- It can be shown that the If we have

$$\bar{Y}_1 = \frac{\sum_{i=1}^n Z_i Y_i}{\sum_{i=1}^n Z_i}, \bar{Y}_0 = \frac{\sum_{i=1}^n (1 - Z_i) Y_i}{\sum_{i=1}^n (1 - Z_i)}, \bar{D}_1 = \frac{\sum_{i=1}^n Z_i D_i}{\sum_{i=1}^n Z_i}, \bar{D}_0 = \frac{\sum_{i=1}^n (1 - Z_i) D_i}{\sum_{i=1}^n (1 - Z_i)}$$

we can write the Wald estimator for β

$$\hat{\beta}_W = \frac{\bar{Y}_1 - \bar{Y}_0}{\bar{D}_1 - \bar{D}_0}$$

Equivalence between Wald and IV

- Also note that the overall average can be obtained as

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Z_i \bar{Y}_1 + \frac{1}{n}\sum_{i=1}^{n}(1 - Z_i)\bar{Y}_0$$

  and that

$$\bar{Y}_1 - \bar{Y} = \frac{1}{n}\sum_{i=1}^{n}(1 - Z_i)(\bar{Y}_1 - \bar{Y}_0), \quad \bar{D}_1 - \bar{D} = \frac{1}{n}\sum_{i=1}^{n}(1 - Z_i)(\bar{D}_1 - \bar{D}_0)$$

- As a result, we can show that the IV estimator is equal to the Wald estimator. So the IV estimator returns the LATE.

$$\begin{aligned}\hat{\beta}_{IV} &= \frac{\sum_{i=1}^{n} Z_i(Y_i - \bar{Y})}{\sum_{i=1}^{n} Z_i(D_i - \bar{D})} = \frac{\sum_{i=1}^{n} Z_i(Y_i - \bar{Y})/\sum_{i=1}^{n} Z_i}{\sum_{i=1}^{n} Z_i(D_i - \bar{D})/\sum_{i=1}^{n} Z_i} \\ &= \frac{\bar{Y}_1 - \bar{Y}}{\bar{D}_1 - \bar{D}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(1 - Z_i)(\bar{Y}_1 - \bar{Y}_0)}{\frac{1}{n}\sum_{i=1}^{n}(1 - Z_i)(\bar{D}_1 - \bar{D}_0)} = \frac{\bar{Y}_1 - \bar{Y}_0}{\bar{D}_1 - \bar{D}_0} = \hat{\beta}_W\end{aligned}$$