Introduction to Econometrics II: Recitation 4

Seung-hun Lee*

February 14th, 2022

1 Various tests in IV and GMM framework

1.1 Theories behind the weak IV testing

Theories behind the testing of the presence of weak IV depends on the local to zero framework. Suppose X_2 is our endogenous variable and we have the following first-stage equation

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

Our focus would be on the excluded instruments Z_2. So we usually test the F-statistic values for π_{22} parameter. The idea is that we can approximately write

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

The idea here is that $E(F)$ 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 $\hat{\beta}_{OLS}$.

This also shows where the rule of thumb cutoff value of 10 comes from. 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.

*Contact me at sl4436@columbia.edu if you spot any errors or have suggestions on improving this note.

Stock and Yogo (2005) extends on this framework by pointing out that weak IV induces bias and also size[1] distortion for the test. In this paper, 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.

Note that the above test assumes homoskedasticity. Weak IV tests that are robust to assumptions on the error variances are very much dealt in frontier works. For instance, there is Montiel-Olea, Pflueger (2013, This is our Pepe!!) and Andrews, Stock, Sun (2019) who proposed a weak IV test robust to hheteroskedasticity.

1.2 Tests that are robust to weak IV

We now talk about inference problems where we acknowledge that our instruments are possibly weak. Assume

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

And assume that we are testing $H_2 : \beta_2 = \beta_2^0$. The problem here is that π_{22} represents the coefficients from the weak IVs.

To get to the testable regression, we 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

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

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

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

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

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.

The interesting features of this estimation is that there is no role for π_{22} parameter (or any parameter for that matter). Thus, it is pivotal and robust to weak IV. It does assume normal distribution of the errors though. Note that the power of this test falls with the number of instruments.

The AR test statistic depends on β_2^0 though. So to create a confidence set, we use an inversion approach in the sense that we test all hypothesis of the form $\beta_2 = \beta_2^0$ for different values of β_2^0 and then check for regions where the null is not rejected.

1.3 Many IV framework

In some instances, we may work with "many" IVs. This indicates a situation where the number of IV is large relative to the sample size. To formalize the idea, Bekker (1994)[2] introduces an asymptotic approximation which treats the number of instruments l as proportional to the sample size, so that $l = \alpha n$. This is equivalent to

$$l/n \to \alpha$$

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.

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

To make the analysis simpler, I assume that z_i is still a valid IV (relevant, exogenous) and that $var\begin{pmatrix} e_i \\ u_i \end{pmatrix} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} = \Sigma$. In addition, note that $var(x_i) = var(z_i'\pi) + var(u_i)$ and assume

[2]Bekker, Paul A. (1994), "Alternative Approximations to the Distributions of Instrumental Variable Estimators", Econometrica 62(3), 657-681

that

$$\frac{1}{n}\sum_{i=1}^{n}\pi' z_i z_i'\pi \xrightarrow{p} c > 0$$

and that variance of x_i and u_i are unchanging with respect to l. This implies that the variance of $var(z_i'\pi)$ is not changing as well an that that R^2 of the reduced form[3] converges to a constant.

To understand the behavior of some of the estimators in many IV situations, I will first start with OLS as a benchmark.

- OLS: We know that $\hat{\beta}_{OLS}$ can be written as

$$\hat{\beta}_{OLS} - \beta = \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i e_i\right)$$

  Applying our setup, we can re-write this as

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} x_i x_i' &= \frac{1}{n}\sum_{i=1}^{n}\pi' z_i z_i'\pi + \frac{1}{n}\sum_{i=1}^{n} u_i u_i' + \frac{2}{n}\sum_{i=1}^{n}\pi' z_i u_i' \\ &\xrightarrow{p} c+1 \\ \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1} &\xrightarrow{p} (c+1)^{-1} \\ \frac{1}{n}\sum_{i=1}^{n} x_i e_i &\xrightarrow{p} \rho\end{aligned}$$

  Therefore,

$$\hat{\beta}_{OLS} - \beta \xrightarrow{p} \frac{\rho}{c+1}$$

  This also captures the idea that if x_i is endogenous (regardless of size), OLS estimators fail to be consistent.

- 2SLS: The 2SLS estimator can be characterized by

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

[3] In this context, it would be $\frac{var(z_i'\pi)}{var(x_i)}$

We know from the above discussion about $\frac{1}{n}\sum_{i=1}^{n}\pi' z_i z_i'\pi$ and the properties of Z that $\frac{\pi' Z' Z\pi}{n} \xrightarrow{p} c$, $\frac{\pi' Z' u}{n} \xrightarrow{p} 0$, $\frac{\pi' Z' e}{n} \xrightarrow{p} 0$. We need to know what happens to the rest of them. Using the fact that $u' P_Z e$ and $u' P_Z u$ are scalars and the properties of the trace operators,

$$E\left[\frac{1}{n}u' P_Z e\right] = \frac{1}{n}E[tr(u' P_Z e)] = \frac{1}{n}E[tr(P_Z e u')] = \frac{1}{n}tr[E(P_Z e u')]$$
$$= \frac{1}{n}tr[E(P_Z)\rho] = \frac{1}{n}E[tr(P_Z)]\rho = \frac{l}{n}\rho$$

where I use the fact that $tr(E(X)) = E(tr(X))$ and $tr(AB) = tr(BA)$ multiple times. In a similar fashion,

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

Based on the two facts and $l/n \to \alpha$, I can make use of Markov inequality to show that $\frac{1}{n}u' P_Z u \xrightarrow{p} \frac{l}{n}$ 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,

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

If we do not have many IVs, $\alpha = 0$ and 2SLS estimator is consistent. Otherwise, inconsistency of the above form occurs.

One way around this is to conduct regularization (e.g LASSO or Ridge) in the first stage or a principle component on Z. This may induce some bias but does reduce the variance of the estimators (bias-variance tradeoff). The other solution is to run a limited information maximum likelihood method.

1.4 (Lack of) discussions about small sample properties and adjustments

When we discussed IV, we really did not discuss its finite sample properties. This is mostly due to the following

- There are biases (note that consistency and biases are a different property)
- $E[(X' P_Z X)^{-1}(X' P_Z y)]$ is not easy to calculate in a small sample (i.e. not taking $n \to \infty$)

- Finite moment may not exist in the sense that

$$E||\hat{\beta}_{IV}||^r < \infty \text{ if } r < L_2 - l_2 - 1$$

  which means for a just identified models, no integer moment may exist

As far as bias is concerned, there are some other estimators that do better than a generic IV/2SLS estimators.

- 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

$$\hat{\beta}_{SSIV} = (\widehat{X}_B X_B)^{-1}(\widehat{X}_B y_B)$$

  This has lower bias than 2SLS estimators but it is sensitive to how we split the sample.

- 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: This is from Inoue and Solon (2010). 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

$$\hat{\beta}_{TSIV} = \left(\frac{Z_2'X_2}{n_2}\right)^{-1}\left(\frac{Z_1'Y_1}{n_1}\right)$$
$$\hat{\beta}_{TS2SLS} = (\widehat{X}_1'\widehat{X}_1)^{-1}(\widehat{X}_1'y_1)\ (\widehat{X}_1 = Z_1(Z_2'Z_2)^{-1}Z_2'X_2)$$

  Note that both are asymptotically equal but may have different numerical values at the finite samples.

1.5 FIML, LIML and k-class estimators

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.

Briefly speaking, the reason we prefer LIML is as follows. 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.[4].

1.5.1 Derivation

The structural equation for y_i is introduced in the previous page. Assume that there is a $z_i = \begin{pmatrix} x_{1i} \\ z_{2i} \end{pmatrix} \in \mathbb{R}^l$ where $E(z_i e_i) = 0$. I now define a reduced form equation for x_{2i}, written as

$$x_{2i} = \Gamma_{12}' x_{1i} + \Gamma_{22}' z_{2i} + u_{2i}, \ (\Gamma_{12}' \in \mathbb{R}^{k_2 \times k_1}, \Gamma_{22}' \in \mathbb{R}^{k_2 \times (l-k_1)}, z_{2i} \in \mathbb{R}^{l-k_1})$$

By putting the structural and reduced form equation together in a matrix form, we get

$$\underbrace{\begin{pmatrix} 1 & -\beta_2' \\ 0 & I_{k_2} \end{pmatrix}}_{=A} \underbrace{\begin{pmatrix} y_i \\ x_{2i} \end{pmatrix}}_{=w_i} = \underbrace{\begin{pmatrix} \beta_1' & 0 \\ \Gamma_{12}' & \Gamma_{22}' \end{pmatrix}}_{=B} \underbrace{\begin{pmatrix} x_{1i} \\ z_{2i} \end{pmatrix}}_{z_i} + \underbrace{\begin{pmatrix} e_i \\ u_{2i} \end{pmatrix}}_{=\eta_i}$$

This now puts all endogenous variables on the left hand side and the exogenous variables on the right. To use the maximum likelihood approach, we need some assumptions on η_i. Specifically, we assume that η_i is conditionally IID normal and

$$\eta_i | z_i \sim N(0, \Sigma_\eta)$$

If we apply these facts, then we can derive the log-likelihood function for η_i

[4] I refer further explanation to Jorn-Steffen Pischke's notes, found at http://econ.lse.ac.uk/staff/spischke/ec533/Weak%20IV.pdf

Lemma 1.1 (Multivariate Normal Density). *Let $X \in \mathbb{R}^k$ be a multivariate normal distribution s.t. $X \sim N(\mu, \Sigma)$. Then the density function can be written as*

$$f_X(x) = (2\pi)^{-\frac{k}{2}} \det(\Sigma)^{-\frac{1}{2}} \times e^{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)}$$

Lemma 1.2 (Jacobian Transformation). *If $X \in \mathbb{R}^k$ is a continuous random vector and $g : \mathbb{R}^k \to \mathbb{R}^k$ is one-to-one and onto function, the density of $Y = g(X)$ is given by*

$$f_Y(y) = f_X(g^{-1}(y)) \, |J|$$

where $J = \det \begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \dots & \frac{\partial x_1}{\partial y_k} \\ \dots & \dots & \dots \\ \frac{\partial x_k}{\partial y_1} & \dots & \frac{\partial x_k}{\partial y_k} \end{pmatrix}$. *Alternatively, we can write*

$$f_X(x) = f_Y(g(x))|H|, H = \det \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \dots & \frac{\partial y_1}{\partial x_k} \\ \dots & \dots & \dots \\ \frac{\partial y_k}{\partial x_1} & \dots & \frac{\partial y_k}{\partial x_k} \end{pmatrix}$$

So to work through the likelihood function, we take these steps

1. **Re-express the PDF**: The PDF pf η_i is

   $$(2\pi)^{-\frac{k_2+1}{2}} \det(\Sigma)^{-\frac{1}{2}} \times e^{-\frac{1}{2}(\eta_i)'\Sigma_\eta^{-1}(\eta_i)}$$

   Note that $\eta_i = Aw_i - Bz_i$, Then we can write the pdf of w_i in terms of A and B as [5]

   $$(2\pi)^{-\frac{k_2+1}{2}} \det(\Sigma)^{-\frac{1}{2}} \times e^{-\frac{1}{2}(Aw_i-Bz_i)'\Sigma_\eta^{-1}(Aw_i-Bz_i)}|A|$$

   by Applying a Jacobian Transformation. Note that the determinant of the A matrix is 1. So this term would be dropped. By IID assumption, the log likelihood for $i = 1, ..., n$ becomes

   $$C - \frac{n}{2}\log(\det(\Sigma_\eta)) - \frac{1}{2}\sum_{i=1}^{n}(Aw_i - Bz_i)'\Sigma_\eta^{-1}(Aw_i - Bz_i)$$

2. **Concentrating out**: To find A, B that maximizes the log - likelihood, we concentrate out

[5] Note that $\chi_i = w_i - A^{-1}Bz_i$ is also another normal distribution. You can notice that $\eta_j = A\chi_i$. Try matching this with the second version of the Jacobian transformation.

the Σ_η term by using its estimator,

$$\widehat{\Sigma}_\eta = \frac{1}{n}\sum_{i=1}^{n}(Aw_i - Bz_i)(Aw_i - Bz_i)'$$

Then plug this estimator into the log-likelihood function to get

$$C - \frac{n}{2}\log(\det(\widehat{\Sigma}_\eta)) - \frac{1}{2}\sum_{i=1}^{n}(Aw_i - Bz_i)'\widehat{\Sigma}_\eta^{-1}(Aw_i - Bz_i)$$

Since the last term is a scalar, we can take

$$(Aw_i - Bz_i)'\widehat{\Sigma}_\eta^{-1}(Aw_i - Bz_i) = tr\left((Aw_i - Bz_i)'\widehat{\Sigma}_\eta^{-1}(Aw_i - Bz_i)\right)$$

Also, using the fact that $tr(AB) = tr(BA)$,

$$tr\left((Aw_i - Bz_i)'\widehat{\Sigma}_\eta^{-1}(Aw_i - Bz_i)\right) = tr\left(\widehat{\Sigma}_\eta^{-1}(Aw_i - Bz_i)(Aw_i - Bz_i)'\right)$$

This reduced the likelihood function to

$$C - \frac{n}{2}\log(\det(\widehat{\Sigma}_\eta)) - \frac{1}{2}tr\left(\widehat{\Sigma}_\eta^{-1}\sum_{i=1}^{n}(Aw_i - Bz_i)(Aw_i - Bz_i)'\right) = C - \frac{n}{2}\log(\det(\widehat{\Sigma}_\eta)) - \frac{n(k_2+1)}{2}$$

3. **Maximize**: The only thing left to maximize over now is $-\frac{n}{2}\log(\det(\widehat{\Sigma}_\eta))$. This is equivalent to maximizing

$$-\det\left(\frac{1}{n}\sum_{i=1}^{n}(Aw_i - Bz_i)'(Aw_i - Bz_i)\right) = -\det\left(\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} y_i - \beta_1' x_{1i} - \beta_2 x_{2i} \\ x_{2i} - \Gamma_{12}' x_{1i} - \Gamma_{22}' z_{2i}\end{pmatrix}\begin{pmatrix} y_i - \beta_1' x_{1i} - \beta_2 x_{2i} \\ x_{2i} - \Gamma_{12}' x_{1i} - \Gamma_{22}' z_{2i}\end{pmatrix}'\right)$$

The combination of β and Γ that maximizes this is the LIML estimator. [6]

1.5.2 *k*-class Estimators

Another way to compute the LIML estimator is to use a k-class estimator with a particular choice for k. In general, k-class estimator is defined as

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

[6] Further steps involve heavy amounts of algebra involving multi-dimensional matrices. For details, I will have you refer to the recitation notes from Dong Woo in Spring 2019, which I will upload.

where $M_Z = I - Z(Z'Z)^{-1}Z' = I - P_Z$. The other way to express this, after some matrix differentiation, is

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

In fact, we can show that OLS ($k = 0$) and 2SLS ($k = 1$) are k-class estimators. LIML is a k-class estimator with a parameter choice of some $k > 1$. What we need to do find the associated value of k. To do so, define

$$W = (y \;\; X_2) \in \mathbb{R}^{n \times (k_2+1)}, \; M_1 = I - X_1(X_1'X_1)^{-1}X_1'$$

Then we compute the minimum eigenvalue of [7]

$$(W'M_1W)(W'M_ZW)^{-1}$$

This minimum eigenvalue will be our choice of k, which will be denoted as $\hat{k}$, and the LIML estimator would be

$$\hat{\beta}_{LIML} = (X'(I_n - \hat{k}M_Z)X)^{-1}(X'(I_n - \hat{k}M_Z)y)$$

From here, we can also find that LIML is also a type of an IV estimator. We can see that by rewriting above equation as

$$\hat{\beta}_{LIML} = (\tilde{X}'X)^{-1}(\tilde{X}'y)$$

where $\tilde{X} = (I_n - \hat{k}M_Z)X$. This also hints that the asymptotic properties of LIML and some other types of IV estimators are similar.

[7] This is shown in Anderson and Rubin (1949) and Amemiya (1985)

1.5.3 Asymptotics of LIML

Given the above definition of $\hat{\beta}_{LIML} = (X'(I_n - \hat{k}M_Z)X)^{-1}(X'(I_n - \hat{k}M_Z)y)$, we can use this to study the asymptotic properties of LIML estimators. Note that

$$\begin{aligned}
\hat{\beta}_{LIML} - \beta &= (X'(I_n - \hat{k}M_Z)X)^{-1}(X'(I_n - \hat{k}M_Z)e) \\
&= (X'(P_Z - (\hat{k}-1))M_Z)X)^{-1}(X'(P_Z - (\hat{k}-1))e) \\
&(\because I - M_Z = P_Z, \implies I - \hat{k}M_Z = P_Z - (\hat{k}-1)M_Z) \\
\implies \sqrt{n}(\hat{\beta}_{LIML} - \beta) &= \left(\frac{X'P_ZX}{n} - (\hat{k}-1)\frac{X'M_ZX}{n}\right)^{-1}\left(\frac{X'P_ZX}{\sqrt{n}} - (\hat{k}-1)\frac{X'M_Ze}{\sqrt{n}}\right)
\end{aligned}$$

Note that

- Anderson and Rubin (1949) shows that $\hat{k} - 1 \xrightarrow{p} 0$, which we accept as given.
- $\frac{X'M_ZX}{n} = \frac{X'X}{n} - \frac{X'P_ZX}{n} = \frac{X'X}{n} - \frac{X'P_Z'P_ZX}{n} = \frac{X'X}{n} - \frac{(P_ZX)'(P_ZX)}{n} \leq \frac{X'X}{n} \xrightarrow{p} E(x_ix_i')$. Therefore, $\frac{X'M_ZX}{n}$ is bounded (and thus $O_p(1)$)
- $\frac{X'M_Ze}{\sqrt{n}}$ converges in distribution to a normal distribution (CLT), so it is $O_p(1)$
- From the third and first points, we can infer that $(\hat{k}-1)\frac{X'M_Ze}{\sqrt{n}} = o_p(1)$

Combining these leads to the result that

$$\sqrt{n}(\hat{\beta}_{LIML} - \beta) = \left(\frac{X'P_ZX}{n}\right)^{-1}\frac{X'P_ZX}{\sqrt{n}} + o_p(1)$$

which is equivalent to $\sqrt{n}(\hat{\beta}_{2SLS} - \beta) + o_p(1)$. So for a fixed n, the 2SLS and LIML converge to a similar limiting distribution. However, when either instrument is weakly related to X or large in numbers, 2SLS bias becomes large, giving advantage to the LIML[8].

1.6 Potential outcome framework and local average treatment effect

Suppose our setup now involves a binary variable D_i. For instance, an indicator variable of receiving a treatment. Then we can write

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

[8]This is the key takeaway from Pischke's slides.

To analyze this framework, we define a potential outcome framework $Y_i(D_i)$. They are composed of

- $Y_i(1)$ for those entering treatment
- $Y_i(0)$ for those not entering treatment

A key feature is that for a given individual i, we only observe one of the two (what we call a *fundamental problem of causal inference* in treatment framework) - we will revisit this in the second half of the course.

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): $E[y_i|D_i = 1] - E[y_i|D_i = 0]$. More specifically,

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

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

But suppose not, then the OLS is not going to give us the true treatment effect and ATE is inconsistent with TE. Then the idea is to find a binary instrument Z_i[9] with which an exogenous variation in D_i can be induced. The new setup is now

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

[9]We can obviously do a continuous instrument Z, but this gives us something different - a marginal treatment effect (MTE). The idea between MTE and LATE is similar though.

The first stage regression regresses D onto Z, so that $\widehat{D} = Z\widehat{\pi}$. Then we regress Y onto $\widehat{D}$, the regression becomes

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

and β can be identified. The rationale is as follows: 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

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

This gives us the local average treatment effect (LATE). 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

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

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

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

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

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