Introduction to Econometrics II: Recitation 4[*]

Seung-hun Lee[†]

February 12th, 2020

1 Topics on Instrumental Variables

1.1 Problem with Many IVs

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)[1] 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.

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

To make the analysis more 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 $var(x_i) = var(z_i'\gamma) + var(u_i)$ and

[*] This is based on the lecture notes of Professors Jushan Bai and Bernard Salanie. I was also greatly helped by previous recitation notes from Paul Sungwook Koh and Dong Woo Hahm. The remaining errors are mine.

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

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

assume that

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

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

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

  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

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

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

We know from the above discussion about $\frac{1}{n}\sum_{i=1}^{n}\gamma' z_i z_i' \gamma$ and the properties of Z that $\frac{\Gamma' Z' Z \Gamma}{n} \xrightarrow{p} c$, $\frac{\Gamma' Z' u}{n} \xrightarrow{p} 0$, $\frac{\Gamma' Z' e}{n} \xrightarrow{p} 0$. 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.

Theorem 1.1 (Inconsistency in Large Scale IVs)**.** *If above assumptions hold, together with $E(e_i^2|z_i) < \infty, E(u_i^4|z_i) < \infty$. Then*

$$\hat{\beta}_{OLS} - \beta \xrightarrow{p} \frac{\rho}{c+1}, \hat{\beta}_{2SLS} - \beta \xrightarrow{p} \frac{\alpha\rho}{c+\alpha}$$

Hansen (2019, pp. 447-448) shows why LIML is immune to this problem.

2 Generalized Method of Moments (GMM)

GMM methods utilize the method of moments estimators to identify the values of the parameters of interest. It can be generalized in the sense that the number of moment conditions can be greater than the number of unknown parameters. In fact, we can show that OLS, IV, MLE estimators are part of GMM.

2.1 Framework

The moment equation takes the following form: Let w_i be IID across $i = 1, .., n$, $g_i(w_i, \theta)$ be a $l \times 1$ function of the ith observation, and $\theta \in \mathbb{R}^{k\times 1}$ be the parameter of interest. Note that $l \geq k$. Then, the **moment equation model** is characterized by

$$E[g(w_i, \theta)] = 0$$

We say θ is identified if there is a unique mapping from the data distribution to θ, or that there is a unique θ satisfying $E[g(w_i, \theta)] = 0$. When $l = k$, then we are in a just-identified case. If $l > k$, then we are in the over-identified case in the sense that there is excess information. We will not bother with the case of $l < k$, as they are under-identified and the moment equation model is unsolvable.

- **Just-identified**: In this case, we can work with the sample analogue of $g(w_i.\theta)$ straight away. Define $\bar{g}_n(\theta)$ as

  $$\bar{g}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} g_i(\theta)$$

  The **method of moments estimators** $\hat{\theta}_{MM}$ is defined as the parameter value which sets $\bar{g}_n(\theta) = 0$. In other expression:

  $$\bar{g}_n(\hat{\theta}_{MM}) = \frac{1}{n}\sum_{i=1}^{n} g_i(\hat{\theta}_{MM}) = 0$$

- **General Case**: If we can generalize to include the case where $l > k$, we are likely to run into a situation where $\hat{\theta}_{MM}$ cannot be found. This is because there may be no choice of θ that sets the moment equations to 0, implying the nonexistence of the method of moments estimator. In such case, we work with a different approach.

  Define $J(\theta)$ as

  $$J(\theta) = n\bar{g}_n(\theta)'W\bar{g}_n(\theta)$$

  where $W \in \mathbb{R}^{l\times l}$ is a positive definite weight matrix that is given. n does not really affect our estimation, but it makes the analysis of the asymptotic features much easier. The **generalized method of moments estimator** is defined as the minimizer of the GMM criterion above, or

$$\hat{\theta}_{GMM} = \arg\min_{\theta} J_n(\theta)$$

Note that when $l = k$, then method of moments estimator solve $\bar{g}_n(\hat{\theta}_{MM}) = 0$. Given that $J(\theta)$ is a positive definite matrix, the method of moments estimator in this case also minimizes $J(\theta)$. Thus, method of moments estimator is a special case of GMM estimator.

If $l > k$, then the problem we should solve is

$$\begin{aligned} &\min_{\theta} J_n(\theta) = n\bar{g}(\theta)'W\bar{g}(\theta) \\ \implies &\frac{\partial J_n(\theta)}{\partial \theta} = 2n\frac{\partial \bar{g}(\theta)}{\partial \theta}' W\bar{g}(\theta) = 0 \end{aligned}$$

To obtain this, I have used the following properties of matrix differentiation.

Property 2.1 (Matrix Differentiation Tricks). *Let Let $x \in \mathbb{R}^{k\times 1}$ and $A \in \mathbb{R}^{k\times l}, y \in \mathbb{R}^{l\times 1}$. Then we have*

$$\frac{\partial x'Ay}{\partial x} = Ay, \frac{\partial x'Ay}{\partial y} = A'x$$

Let $x \in \mathbb{R}^{k\times 1}$ and $A \in \mathbb{R}^{k\times k}$. Then the following holds

$$\frac{\partial x'Ax}{\partial x} = (A + A')x$$

Example 2.1 (Various Examples of GMM). *These are some types of a GMM estimator.*

- *OLS: In a data generating process $y_i = x_i'\beta + e_i,\ x_i \in \mathbb{R}^k$ and the moment condition $E(x_i e_i) = 0$, we have k parameters β_k and k equations for each of the k variables. We can rewrite the moment condition as*

  $$E(x_i(y_i - x_i'\beta)) = 0$$

  And the method of moments estimators imply that we should solve

  $$\frac{1}{n}\sum_{i=1}^{n} x_i(y_i - x_i'\beta) = 0 \iff \hat{\beta}_{OLS} = \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} x_i y_i$$

- *MLE: If w_i is IID across $i = 1, ..., n$, then we can write the joint likelihood function as*

$$\prod_{i=1}^{n} f(w_i|\theta)$$

and thus, the log-likelihood function

$$\sum_{i=1}^{n} \log f(w_i|\theta)$$

When we take partial differentiation w.r.t θ,

$$\sum_{i=1}^{n} \frac{\partial \log f(w_i|\theta)}{\partial \theta} = 0 \implies \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \log f(w_i|\theta)}{\partial \theta} = 0$$

which is equivalent to $E\left(\frac{\partial \log f(w_i|\theta)}{\partial \theta}\right) = 0$. Practically, $E\left(\frac{\partial \log f(w_i|\theta)}{\partial \theta}\right) = 0$ becomes the moment condition applicable to MLE.

- *IV: Suppose we have a data generating process $y_i = x_i'\beta + e_i$ with x_i being k dimensions. Suppose we have an $l > k$ dimensional IV with $E(z_i e_i) = 0$ $\bar{g}(\beta)$ in our context would be*

$$\frac{1}{n}\sum_{i=1}^{n}(z_i y_i - z_i x_i'\beta) = \frac{Z'y}{n} - \frac{Z'X\beta}{n}$$

Then, we can write our $J_n(\beta)$ as

$$n\left(\frac{Z'y}{n} - \frac{Z'X\beta}{n}\right)' W \left(\frac{Z'y}{n} - \frac{Z'X\beta}{n}\right)$$

By solving the minimization problem we can obtain

$$\begin{aligned}\frac{\partial J_n(\beta)}{\partial \beta} &= -\frac{2}{n}(X'ZWZ'y) + \frac{2}{n}(X'ZWZ'X)\beta = 0\\ &\implies \hat{\beta} = (X'ZWZ'X)^{-1}(X'ZWZ'y)\end{aligned}$$

In fact, we can show that by setting $W = \left(\frac{Z'Z}{n}\right)^{-1}$, the above estimator is in fact, a 2SLS estimator.

2.2 Limiting Distribution of GMM

From above, we know that $\hat{\beta}_{GMM} = (X'ZWZ'X)^{-1}(X'ZWZ'y)$ for the case of an overidentified IV model. We can rewrite this by replacing y with $X\beta + e$, As a result, the limiting distribution of $\hat{\beta}_{GMM}$ is characterized by

$$\sqrt{n}(\hat{\beta}_{GMM} - \beta) = \left(\frac{X'Z}{n}W\frac{Z'X}{n}\right)^{-1}\left(\frac{X'Z}{n}W\frac{Z'e}{\sqrt{n}}\right)$$

For convenience, I assume the following

Assumption 2.1. *Assume that*

1. $E(z_i x_i') = Q$, *and that* $\frac{Z'X}{n} \xrightarrow{p} Q$
2. $\frac{Z'e}{\sqrt{n}} \xrightarrow{d} N(0, \Omega)$, *where* $\Omega = E(z_i z_i' e_i^2)$
3. *(If we are willing to assume W depends on n, thus W_n): $W_n \xrightarrow{p} W$, where W is a positive definite weight matrix*

Then we can say the following about the distribution of the GMM estimator

Theorem 2.1 (Limiting Distribution of the GMM estimator)**.** *If the above assumptions are satisfied, the limiting distribution of the GMM estimator can be characterized by*

$$\sqrt{n}(\hat{\beta}_{GMM} - \beta) \xrightarrow{d} N(0, (Q'WQ)^{-1}(Q'W\Omega W'Q)(Q'WQ)^{-1})$$

Where $(Q'WQ)^{-1}$ can be replaced with A, if we stick to the notation in class.

So far, we haven't said too much about W. We just assumed that they are fixed. Even if we suppose that W depends on n somehow, the above theorem still holds, provided that W_n converges in probability to W. Then, the natural question should be "What is the optimal selection of W?". This naturally moves us to...

2.3 Efficient GMM

To select an optimal W matrix, it must be that the resulting variance should be the smallest. If we let $W = \Omega^{-1}$ and work with $(Q'WQ)^{-1}(Q'W\Omega W'Q)(Q'WQ)^{-1} - (Q'\Omega Q)^{-1}$, we can see that it is positive semidefinite[3]. If we put $W = \Omega^{-1}$ and recalculate the variance, we get that the efficient GMM has a limiting distribution characterized by

$$\sqrt{n}(\hat{\beta}_{GMM} - \beta) \xrightarrow{d} N(0, (Q'WQ)^{-1})$$

Note that this weighting matrix, which can be rewritten as

$$W = \Omega^{-1} = E(z_i z_i e_i^2)^{-1}$$

is not exactly same as the weighting matrix we used for deriving the 2SLS estimator from GMM, which is $\left(\frac{Z'Z}{n}\right)^{-1}$. In the $W = E(z_i z_i e_i^2)^{-1}$ setup, we allowed for heteroskedasticity. If we impose conditional homoskedasticity in the sense that $E(e_i^2|z_i) = \sigma^2$, we can rewrite $E(z_i z_i' e_i^2)$ as

$$\begin{aligned} E(z_i z_i' e_i^2) &= E(E(z_i z_i' e_i^2|z_i)) = E(z_i z_i' E(e_i^2|z_i)) \\ &= E(z_i z_i' \sigma^2) = E(z_i z_i')\sigma^2 \end{aligned}$$

Thus, $E(z_i z_i' e_i^2)^{-1} = (Z'Z)^{-1}(\sigma^2)^{-1}$. However, scaling does not affect how the estimator is defined (by calculating $\hat{\beta}_{GMM}$ the scalar scales disappear). So $E(z_i z_i' e_i^2)^{-1}$ is effectively $(Z'Z)^{-1}$ and we get the conclusion that in a conditional homoskedastic setting, 2SLS estimator is the efficient GMM.

2.4 Two-step Optimal GMM

Since we have no idea what the true value of e_i is, the true value of Ω is naturally unknown. Therefore, we require a consistent estimator, denoted as $\widehat{W}$, for $W = \Omega^{-1}$. To do so, we need a preliminary consistent estimator for the true θ, which I will denote with $\tilde{\theta}$. You can use any W matrix for this step (I or $(Z'Z)^{-1}$). Then, define

$$\widehat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} g(w_i.\tilde{\theta})g(w_i.\tilde{\theta})'$$

since $\Omega = E[g(w_i, \theta)g(w_i, \theta)']$. Then we can identify what $\widehat{\Omega}^{-1}$ is. This will allow us to compute the efficient GMM estimator $\hat{\theta}_{GMM}$

[3] This is your homework, so I will not proceed further with any more explanations.

Definition 2.1 (Optimal Two-step GMM Estimator)**.** We can compute Optimal Two-step GMM in these steps

1. Compute a preliminary, but consistent estimator for the true θ. Denote this as $\tilde{\theta}$.
2. Using $\Omega = E[g(w_i,\theta)g(w_i,\theta)']$, create a sample analogue of this, defined as $\widehat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} g(w_i.\tilde{\theta})g(w_i.\tilde{\theta})'$. We can find our $\widehat{\Omega}^{-1}$ here.
3. Using this $\widehat{\Omega}^{-1}$, construct an efficient GMM estimator $\hat{\theta}_{GMM}$

There is another way to come up with a $\widehat{\Omega}^{-1}$ in this context. Define $\bar{g}(\theta) = \frac{1}{n}\sum_{i=1}^{n} g(w_i,\tilde{\theta})$. Then, an alternative definition of $\widehat{\Omega}$ can be written as

$$\widehat{\Omega}^{+} = \frac{1}{n}\sum_{i=1}^{n}(g(w_i,\tilde{\theta}) - \bar{g}(\theta))(g(w_i,\tilde{\theta}) - \bar{g}(\theta))'$$

So why use this? For one thing, both $\widehat{\Omega}$ and $\widehat{\Omega}^{+}$ converge in probability to $E[g(w_i,\theta)g(w_i,\theta)']$. This is because

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(g(w_i,\tilde{\theta}) - \bar{g}(\theta))(g(w_i,\tilde{\theta}) - \bar{g}(\theta))' \\
&= \frac{1}{n}\sum_{i=1}^{n} g(w_i,\tilde{\theta})g(w_i,\tilde{\theta})' - \frac{1}{n}\sum_{i=1}^{n} g(w_i,\tilde{\theta})\bar{g}(\theta)' - \frac{1}{n}\sum_{i=1}^{n} \bar{g}(\theta)g(w_i,\tilde{\theta})' + \frac{1}{n}\sum_{i=1}^{n} \bar{g}(\theta)\bar{g}(\theta)' \\
&= \frac{1}{n}\sum_{i=1}^{n} g(w_i,\tilde{\theta})g(w_i,\tilde{\theta})' - \bar{g}(\theta)\bar{g}(\theta)' \quad \left(\because \bar{g}(\theta) = \frac{1}{n}\sum_{i=1}^{n} g(w_i,\tilde{\theta})\right)
\end{aligned}$$

Under the assumption that $E[g(w_i,\theta)] = 0$, (or $E(z_ie_i) = 0$) is a correct specification, we have $\bar{g}(\theta) \xrightarrow{p} E(g(w_i,\theta)) = 0$. So $\widehat{\Omega}^{+}$ and $\widehat{\Omega}$ both converge to $E[g(w_i,\theta)g(w_i,\theta)']$.

The difference? If it is the case that $E[g(w_i,\theta)] \neq 0$ we view $\widehat{\Omega}^{+}$ as a robust estimator. Note that

- $\widehat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} g(w_i.\tilde{\theta})g(w_i.\tilde{\theta})' \xrightarrow{p} E[g(w_i,\tilde{\theta})g(w_i,\tilde{\theta})']$
- $\widehat{\Omega}^{+} = \frac{1}{n}\sum_{i=1}^{n} g(w_i,\tilde{\theta})g(w_i,\tilde{\theta})' - \bar{g}(\theta)\bar{g}(\theta)' \xrightarrow{p} E[g(w_i,\tilde{\theta})g(w_i,\tilde{\theta})'] - E[g(w_i,\tilde{\theta})]E[g(w_i,\tilde{\theta})]' = var[g(w_i,\tilde{\theta})]$

In other words, $\widehat{\Omega}$ is inconsistent in case where $E[g(w_i,\tilde{\theta})] = 0$ is not guaranteed. Therefore, for tests, such as overidentification tests, it is much more desirable to use $\widehat{\Omega}^{+}$.

Since we know how to find the optimal $\widehat{W}$, we can estimate the asymptotic variance of the GMM estimators. This can be done by replacing matrices in the original variance with their sample counterparts. In general, we can estimate by

$$\widehat{V}_{GMM} = \left(\widehat{Q}'\widehat{W}\widehat{Q}\right)^{-1}\left(\widehat{Q}'\widehat{W}\widehat{\Omega}\widehat{W}\widehat{Q}\right)\left(\widehat{Q}'\widehat{W}\widehat{Q}\right)^{-1}$$

where $\widehat{Q} = \frac{1}{n}\sum_{i=1}^{n} z_i x_i' = \frac{Z'X}{n}$, $\widehat{W}$ is expressed by either $\widehat{\Omega}$ or $\widehat{\Omega}^{+}$. The residuals used in this estimation is defined as $\hat{e}_i = y_i - x_i'\hat{\beta}_{GMM}$

One alternative to the two-step GMM estimator is contructed by letting weight matrix be an explicit function θ. The criterion function is now

$$J(\theta) = n\bar{g}_n(\theta)'\left(\frac{1}{n}\sum_{i=1}^{n} g(w_i,\theta)g(w_i,\theta)'\right)^{-1}\bar{g}_n(\theta)$$

The $\hat{\theta}$ that minimizes this function is the continuously-updated GMM estimator. However, this setup is nonlinear, implying that acquiring this estimator requires numerical methods.