# TA section 4

May 1st, 2009 by Yang

## 1. Asymptotic Normality of the TS estimator

Consider the two step (TS) estimator that was mentioned in the problem set 2. The stochastic criterion function is given in the LHS of the following equation.

$$Q_n(\theta) = \|A_n G_n(\theta, \widehat{\tau}_n)\|^2/2 \xrightarrow{p} Q(\theta) = \|AG(\theta, \tau_0)\|^2/2$$

The above pointwise convergence holds if $A_n \xrightarrow{p} A$, $\widehat{\tau}_n \xrightarrow{p} \tau_0$, and $G_n(\theta,\tau) \xrightarrow{p} G(\theta,\tau)$ for any $\theta \in \Theta$ and $\tau \in B(\tau_0, \varepsilon)$ for some $\varepsilon$. Then the TS estimator has the asymptotic normal distribution under the following assumptions.

1. [CF i] $\theta_0 \in int\Theta$
2. $G_n(\theta,\tau)$ is twice continuously differentiable with respect to $\theta$ in $\Theta_0$ for any $\tau \in \mathcal{T}_0$ (wp1)
3. $G_n(\theta_0,\tau)$ is once continuously differentiable with respect to $\tau$ in $\mathcal{T}_0$ (wp1)
4. $G_n(\theta,\tau) \xrightarrow{p} G(\theta,\tau)$ for any $\theta \in \Theta$ and $\tau \in \mathcal{T}_0$
5. $\sqrt{n}\begin{pmatrix} G_n(\theta_0,\tau_0) \\ \widehat{\tau}_n - \tau_0 \end{pmatrix} \xrightarrow{d} \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim N\left(0, \begin{pmatrix} V_1 & V_2 \\ V_2' & V_3 \end{pmatrix}\right)$
6. $A_n \xrightarrow{p} A$
7. $\sup_{\theta\in\Theta_0} \|\frac{\partial^2}{\partial\theta\partial\theta'} Q_n(\theta) - B(\theta)\| \xrightarrow{p} 0$ for some nonstochastic $B(\theta)$
8. $\Gamma_0 := \frac{\partial}{\partial\theta'} G(\theta_0,\tau_0)$ and $A$ are full rank matrices.
9. [EE2 i] $\widehat{\theta}_n \xrightarrow{p} \theta_0$
10. [EE2 ii] $\frac{\partial}{\partial\theta} Q_n(\widehat{\theta}_n) = o_p(n^{-1/2})$

Note that 2 implies CF ii since only $G_n(\cdot)$ has $\theta$ as its argument. Also 2-6 imply CF iii. To verify, consider

$$\sqrt{n}\frac{\partial}{\partial\theta} Q_n(\theta_0) = \frac{\partial}{\partial\theta} G_n(\theta_0, \widehat{\tau}_n)' A_n' A_n \sqrt{n} G_n(\theta_0, \widehat{\tau}_n)$$

Note that $\frac{\partial}{\partial\theta'} G_n(\theta_0,\widehat{\tau}_n) \xrightarrow{p} \Gamma_0$ by 2, 4 and 5, and that $A_n'A_n \xrightarrow{p} A'A$ by 6. Apply the mean value theorem to the last part, with respect to $\tau$ around $\tau_0$, then

$$G_n(\theta_0,\widehat{\tau}_n) = \left[G_n(\theta_0,\tau_0) + \frac{\partial}{\partial\tau'} G_n(\theta_0, \tau_n^*)(\widehat{\tau}_n - \tau_0)\right]$$

where $\tau_n^*$ is between $\widehat{\tau}_n$ and $\tau_0$, and thus converges in probability to $\tau_0$. It follows that

$$\sqrt{n}G_n(\theta_0, \widehat{\tau}_n) = \left(I_k \vdots \frac{\partial}{\partial \tau'} G_n(\theta_0, \tau_n^*)\right) \sqrt{n} \left( \begin{array}{c} G_n(\theta_0, \tau_0) \\ \widehat{\tau}_n - \tau_0 \end{array} \right)$$

Since 3 and 4 imply $\frac{\partial}{\partial \tau'} G_n(\theta_0, \tau_n^*) \xrightarrow{p} \frac{\partial}{\partial \tau'} G(\theta_0, \tau_0) =: \Lambda_0$, apply 5 to get

$$\sqrt{n}G_n(\theta_0, \widehat{\tau}_n) \xrightarrow{d} \left(I_k \vdots \Lambda_0\right) \left( \begin{array}{c} Z_1 \\ Z_2 \end{array} \right) = Z_1 + \Lambda_0 Z_2$$

In other words,

$$\sqrt{n}G_n(\theta_0, \widehat{\tau}_n) \xrightarrow{d} N\Big(0, \underbrace{V_1 + \Lambda_0 V_2' + V_2 \Lambda_0' + \Lambda_0 V_3 \Lambda_0'}_{=: V_0}\Big)$$

It then follows that

$$\sqrt{n}\frac{\partial}{\partial \theta} Q_n(\theta_0) \xrightarrow{d} \Gamma_0' A' A (Z_1 + \Lambda_0 Z_2) \sim N(0, \Gamma_0' A' A V_0 A' A \Gamma_0)$$

Define

$$\Omega_0 := \Gamma_0' A' A V_0 A' A \Gamma_0 = \Gamma_0' A' A (V_1 + \Lambda_0 V_2' + V_2 \Lambda_0' + \Lambda_0 V_3 \Lambda_0') A' A \Gamma_0$$

Finally, 2, 4 and 6-8 imply CF iv. Find $B(\theta)$ first. Since

$$\left[\frac{\partial^2}{\partial \theta \partial \theta'} Q_n(\theta)\right]_{mj} = \frac{\partial}{\partial \theta_m} G_n(\theta, \widehat{\tau}_n)' A_n' A_n \frac{\partial}{\partial \theta_j} G_n(\theta, \widehat{\tau}_n) + \frac{\partial^2}{\partial \theta_m \partial \theta_n} G_n(\theta, \widehat{\tau}_n)' A_n' A_n G_n(\theta, \widehat{\tau}_n)$$

by 6 and 7,

$$[B(\theta)]_{mj} = \text{plim} \left[\frac{\partial^2}{\partial \theta \partial \theta'} Q_n(\theta)\right]_{mj} = \frac{\partial}{\partial \theta_m} G(\theta, \tau_0)' A' A \frac{\partial}{\partial \theta_j} G(\theta, \tau_0) + \frac{\partial^2}{\partial \theta_m \partial \theta_n} G(\theta, \tau_0)' A' A G(\theta, \tau_0)$$

This is continuous at $\theta_0$ by 2, 4 and 7. Moreover,

$$B_0 := B(\theta_0) = \frac{\partial}{\partial \theta} G(\theta_0, \tau_0)' A' A \frac{\partial}{\partial \theta'} G(\theta_0, \tau_0) = \Gamma_0' A' A \Gamma_0$$

is nonsingular by 8. Therefore,

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} N\Big(0, (\Gamma_0' A' A \Gamma_0)^{-1} \Gamma_0' A' A (V_1 + \Lambda_0 V_2' + V_2 \Lambda_0' + \Lambda_0 V_3 \Lambda_0') A' A \Gamma_0 (\Gamma_0' A' A \Gamma_0)^{-1}\Big)$$

### 2. Bias and Consistency of 2SLS estimator

Let $\theta_0$ be the true parameter. An estimator $\widehat{\theta}_n$ is said to be

- unbiased if $E[\widehat{\theta}_n] = \theta_0$, and
- consistent if $\widehat{\theta}_n \xrightarrow{p} \theta_0$

Which property is more desirable? We care more about consistency. The following example makes it clear.

*Example.* Let $X_i \sim iidN(\theta_0, 1)$. Consider the following 2 estimators: $\widehat{\theta}_n = X_n$ and $\overline{\theta}_n = \frac{1}{n-1}\sum_{i=1}^{n} X_i$. $E[\widehat{\theta}_n] = \theta_0$ but $\widehat{\theta} \xrightarrow{p} \theta_0$ does not hold. Lots of data would not help us figure out $\theta_0$. On the contrary, $E[\overline{\theta}_n] \neq \theta_0$, but $\overline{\theta}_n \xrightarrow{p} \theta_0$. In the finite sample, there is always a bias, but we get less variance as the sample size grows. ∎

Consider the linear model.

$$y_i = x_i'\beta_0 + \varepsilon_i, \quad i = 1, \cdots, n$$

Let us investigate the OLS estimator $\widehat{\beta}_{OLS} = (X'X)^{-1}X'Y$ first.

- $\widehat{\beta}_{OLS}$ is unbiased if either
  1. $x_i$ is nonstochastic and $E[\varepsilon_i] = 0$, or
  2. $E[\varepsilon|X] = 0$.
- $\widehat{\beta}_{OLS}$ is consistent if either
  1. $x_i$ is nonstochastic and $E\varepsilon_i = 0$, or
  2. $(x_i, \varepsilon_i)$ are iid, $Ex_i\varepsilon_i = 0$ and $E\|x_i\|^2 < \infty$.
- $\widehat{\beta}_{OLS}$ has minimum variance among all linear unbiased estimators if either
  1. $x_i$ is nonstochastic, $E\varepsilon_i = 0$, and $E[\varepsilon\varepsilon'] = \sigma^2 I_n$, or
  2. $E[\varepsilon|X] = 0$ and $E[\varepsilon\varepsilon'|X] = \sigma^2 I_n$.

In the linear IV model, none of the above assumptions are satisfied. We assume

$$(x_i, z_i, \varepsilon_i) \text{ are } iid$$
$$Ex_i\varepsilon_i \neq 0$$
$$Ez_i\varepsilon_i = 0$$

Recall that

- the 2SLS estimator $\widehat{\beta}_{2SLS} = (X'P_zX)^{-1}X'P_zY$ is consistent under the assumptions that
  1. $E\|x_iz_i'\| < \infty$
  2. $E\|z_i\|^2 < \infty$
  3. $Ex_iz_i'(Ez_iz_i')^{-1}Ez_ix_i'$ is nonsingular

Note that $\widehat{\beta}_{2SLS}$ is never unbiased by the assumption $Ex_i\varepsilon_i \neq 0$. When $E[\varepsilon|X,Z]=0$, $\widehat{\beta}_{2SLS}$ would be unbiased, but less efficient than $\widehat{\beta}_{OLS}$ in the sense that the variance of $\widehat{\beta}_{2SLS}$ is always greater than that of $\widehat{\beta}_{OLS}$. We have seen that

- $\widehat{\beta}_{2SLS}$ has minimum asymptotic variance among all GMM estimators that use the moment condition $Ez_i\varepsilon_i = 0$ if $E[u_i^2|z_i] = \sigma^2$ (conditional homoskedasticity) holds.

### 3. Adding more instruments

As long as $z_{ij}$ satisfies $Ez_{ij}\varepsilon_i = 0$, it seems better to include $z_{ij}$ as instruments. Is it true? Yes and no. Consider the following example.

*Example.* Suppose we have the same number of instruments as the number of observations. All the instruments satisfy $Ez_i\varepsilon_i = 0$. However, it is easy to see that $P_z = Z(Z'Z)^{-1}Z' = I_n$ so that

$$\widehat{\beta}_{2SLS} = (X'P_zX)^{-1}X'P_zY = (X'X)^{-1}X'Y = \widehat{\beta}_{OLS}$$

We know that $\widehat{\beta}_{OLS}$ is biased in the linear IV model. It is not consistent either. ∎

From this example, we can infer that using too many instruments may not be the best way to follow. Indeed, given the number of instruments fixed, the 2SLS estimator is consistent, but if the number of instruments grows as the sample size increases, it may not be. But the following result supports using more instruments.

**Theorem**. Consider a linear IV model and suppose more strict conditional homoskedasticity holds, that is $var[\varepsilon_i|x_i,z_i] = \sigma^2$ holds. Let $Z_2 = (Z_1, W)$ for some $W$. Define $\widehat{\beta}_1 = (X'P_{z1}X)^{-1}X'P_{z1}Y$ and $\widehat{\beta}_2 = (X'P_{z2}X)^{-1}X'P_{z2}Y$ where $P_{z1} = Z_1(Z_1'Z_1)^{-1}Z_1'$ and $P_{z2} = Z_2(Z_2'Z_2)^{-1}Z_2'$. Then,

$$var(\widehat{\beta}_1) \geq var(\widehat{\beta}_2)$$

in the sense that $var(\widehat{\beta}_1) - var(\widehat{\beta}_2)$ is positive semidefinite.

*Proof.* Note that under the assumptions, the variance of the 2SLS estimators are given by

$$\begin{aligned} var(\widehat{\beta}_1) &= \sigma^2(X'P_{z1}X)^{-1} \\ var(\widehat{\beta}_2) &= \sigma^2(X'P_{z2}X)^{-1} \end{aligned}$$

Note that

$$\begin{aligned} X'P_{z2}X - X'P_{z1}X &= X'P_{z2}P_{z2}X - X'P_{z2}P_{z1}P_{z2}X \\ &= X'P_{z2}(I-P_{z1})P_{z2}X \\ &= X'P_{z2}(I-P_{z1})(I-P_{z1})P_{z2}X \end{aligned}$$

The first equality follows from the fact that $P_{z2}$ is idempotent, and that $P_{z2}P_{z1} = P_{z1}P_{z2} = P_{z1}$ since $Z_2$ includes $Z_1$. The last equality holds since $(I - P_{z1})$ is idempotent. The expression has a quadratic form, so $X'P_{z2}X - X'P_{z1}X$ is positive semidefinite. Use the theorem that $A - B$ is positive semidefinite if and only if $B^{-1} - A^{-1}$ is positive semidefinite, then we get the desired result. ∎

To summarize, adding more valid instruments increases bias but reduces variance. Since the mean squared error is bias$^2$ + variance, we need to choose the number of instruments that minimizes it.

**4. The J-test**

The $J$-statistic is given as follows.

$$J_n := n \cdot \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i'\widehat{\beta}_n)z_i' \left(\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i'\widehat{\beta}_n)^2 z_i z_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i'\widehat{\beta}_n)z_i$$

The J-test tests the null hypothesis $H_0 : Ez_i\varepsilon_i = 0$. The intuition behind this is the following. If all the moment conditions that are used in the 2SLS estimation are true, the sample analogue of those will be close to 0, and thus $J_n$ will be bounded in probability. More specifically, $J_n \xrightarrow{d} \chi^2_{k-d}$ where $k$ is the number of instruments and $d$ is the dimension of $\beta_0$. If at least one of the moment conditions is false, their sample analogue would be different from 0, and thus $J_n$ will diverge to infinity in probability.

But there are two problems as mentioned in the lecture note. First, the J-test has very low power against some alternative hypotheses. For example, if $Ez_i\varepsilon_i = Ez_ix_i'\gamma \neq 0$ for some $\gamma$, we can show that $J_n \xrightarrow{d} \chi^2_{k-d}$. In this case, we would use incorrect moment conditions, so $\widehat{\beta}_n$ may not be consistent. We will verify it later. Second, the finite sample approximation of J-test is very poor. $J_n$ converges in distribution to $\chi^2_{k-d}$ but very slowly, so when the sample size is small, it may reject true $H_0$ too often, or may not reject false $H_0$ many times.

**5. Efficient GMM**

We have seen that in the GMM estimation, choosing the weight matrix so that $A'A = V_0^{-1}$ yields the minimum variance of the GMM estimator. Note that $V_0$ is defined as

$$V_0 = Eg(w_i, \theta_0)g(w_i, \theta_0)'$$

where $g(w_i, \theta_0)$ consists of moment equations. Given the moment conditions $Eg(w_i, \theta_0) = 0$, $V_0$ is the variance of $g(w_i, \theta_0)$. What would be the intuition that weighting each moment equation in reverse proportion to its variance yields the minimum variance? Take a linear IV example so that

$$E(y_i - x_i'\beta_0)z_i = 0$$

If $z_i$ is volatile, pertubation of $\beta$ around $\beta_0$ by a small amount makes the moment condition violated by a lot. So we do not want to weight much on such an equation. On the other hand, if $z_i$ does not vary much, we may be able to try many candidate $\beta$'s and find the exact $\beta_0$ that makes the moment condtion hold. The same logic applies to general moment equations $g(w_i, \theta_0)$. So we want to find a consistent estimator of $V_0$ and use it to perform an efficient GMM estimation.

In some cases, it can be analytically derived. For example, in the linear IV model, if we assume conditional homoskedasticity, we may use $A_n' A_n = \frac{1}{n}\sum_{i=1}^n z_i z_i'$ to get the efficient estimator, which is the 2SLS estimator. But in other cases, it may not be easily found. So an efficient GMM estimation is infeasible. There is, fortunately, the second best way. It uses the property that the GMM estimator is consistent even though it does not use the optimal weight matrix. The procedure is as follows.

- First step: Obtain the GMM estimator using any weight matrix. Denote it by $\widehat{\beta}_1$.
    1. Use $A_n' A_n = \frac{1}{n}\sum_{i=1}^n z_i z_i'$ in the IV model, then we get the 2SLS estimator.
    2. We may use $A_n = I_k$, or even other matrices.
    3. The resulting estimators are all consistent.
- Second step: Form $\widehat{V}_n = \frac{1}{n}\sum_{i=1}^n g(w_i, \widehat{\beta}_1) g(w_i, \widehat{\beta}_1)'$. This is a consistent estimator of $V_0$. Use $A_n' A_n = \widehat{V}_n^{-1}$ as a weight matrix to get the GMM estimator. Denote it by $\widehat{\beta}_2$.
- Third step: Repeat second step using $\widehat{\beta}_2$ and denote the resulting estimator by $\widehat{\beta}_3$, and so on.

It is proven that the second step estimator $\widehat{\beta}_2$ has the same asymptotic distribution as the efficient GMM estimator. Also, $\widehat{\beta}_2, \widehat{\beta}_3, \cdots$ have the same asymptotic distribution. So we may stop at the second step.[1] For example, in the linear IV example as in problem set 4, the (feasible) efficient GMM estimator can be obtained as follows.

- Calculate the 2SLS estimator $\widehat{\beta}_{2SLS} = (X' P_z X)^{-1} X' P_z Y$.
- Obtain the consistent estimator of $V_0$

$$\widehat{V}_n = \frac{1}{n}\sum_{i=1}^n (y_i - x_i' \widehat{\beta}_{2SLS})^2 z_i z_i'$$

- Use $\widehat{V}_n^{-1/2}$ as a weight matrix to get the feasible efficient GMM estimator. It is equivalent to minimize

$$Q_n(\beta) = \frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^n (y_i - x_i'\beta) z_i'\right) \widehat{V}_n^{-1} \left(\frac{1}{n}\sum_{i=1}^n (y_i - x_i'\beta) z_i'\right)$$

  Then we would get

$$\widehat{\beta}_{EFF} = (X' Z \widehat{V}_n^{-1} Z' X)^{-1} X' Z \widehat{V}_n^{-1} Z' Y$$

[1] Note that those finite step estimators have different finite sample distribution.