Single-Equation GMM

Ping Yu

School of Economics and Finance
The University of Hong Kong

GMM Estimator

Linear GMM Estimator

- Suppose

$$\begin{aligned} y_i &= \mathbf{x}_i'\beta + u_i \\ E[\mathbf{x}_i u_i] &\neq \mathbf{0}, E[\mathbf{z}_i u_i] = \mathbf{0}, \end{aligned}$$

  then the moment conditions are

$$E[g(\mathbf{w}_i, \beta)] = E\left[\mathbf{z}_i \left(y_i - \mathbf{x}_i'\beta\right)\right] = 0, \quad (1)$$

  where $g(\cdot,\cdot)$ is a set of moment conditions, and $\mathbf{w}_i = \left(y_i, \mathbf{x}_i', \mathbf{z}_i'\right)'$.
- Define the sample analog of (1)

$$\overline{g}_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} g_i(\beta) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_i \left(y_i - \mathbf{x}_i'\beta\right) = \frac{1}{n}\left(\mathbf{Z}'\mathbf{y} - \mathbf{Z}'\mathbf{X}\beta\right).$$

- When $l > k$, we cannot solve $\overline{g}_n(\beta) = \mathbf{0}$ exactly as intuitively shown in Figure 1.
- The idea of the GMM is to define an estimator which sets $\overline{g}_n(\beta)$ "close" to zero.

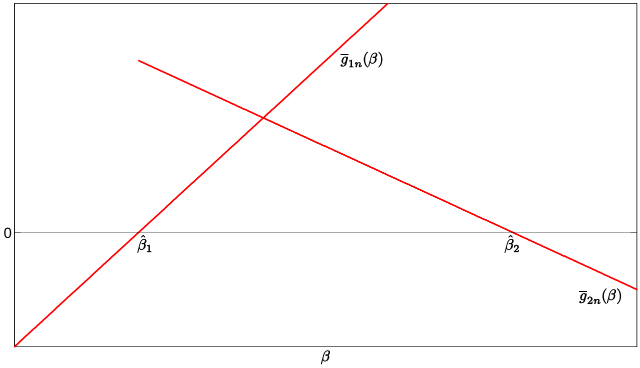

Figure: $\overline{g}_n(\beta) = 0$ Can Not Hold Exactly for Any β: $k = 1, l = 2$

continue...

- For some $l \times l$ weight matrix $\mathbf{W}_n > 0$, let

$$J_n(\beta) = n \cdot \overline{g}_n(\beta)' \mathbf{W}_n \overline{g}_n(\beta).$$

- This is a non-negative measure of the "length" of the vector $\overline{g}_n(\beta)$ under the inner product $\langle \cdot, \cdot \rangle_{\mathbf{W}_n}$.

  - If $\mathbf{W}_n = \mathbf{I}_l$, then, $J_n(\beta) = n \cdot \overline{g}_n(\beta)' \overline{g}_n(\beta) = n \left\| \overline{g}_n(\beta) \right\|^2$, the square of the Euclidean length.
- The GMM estimator minimizes $J_n(\beta)$.
- The first order conditions for the GMM estimator are

$$\begin{aligned} \mathbf{0} &= \frac{\partial}{\partial \beta} J_n\left(\widehat{\beta}\right) = 2n \frac{\partial}{\partial \beta} \overline{g}_n'(\widehat{\beta}) \mathbf{W}_n \overline{g}_n(\widehat{\beta}) \\ &= -2n \left(\frac{1}{n} \mathbf{X}'\mathbf{Z} \right) \mathbf{W}_n \left(\frac{1}{n} \left(\mathbf{Z}'\mathbf{y} - \mathbf{Z}'\mathbf{X}\widehat{\beta} \right) \right), \end{aligned}$$

so

$$\widehat{\beta}_{GMM} = \left[(\mathbf{X}'\mathbf{Z}) \mathbf{W}_n (\mathbf{Z}'\mathbf{X}) \right]^{-1} \left[(\mathbf{X}'\mathbf{Z}) \mathbf{W}_n (\mathbf{Z}'\mathbf{y}) \right]. \tag{2}$$

More on $\mathbf{W}_n$ and the GMM Estimator

- If $l = k$, then $\overline{g}_n(\beta) = \mathbf{0}$. The GMM estimator reduces to the MoM estimator (the IV estimator) and $\mathbf{W}_n$ is not required.
- While the estimator depends on $\mathbf{W}_n$, the dependence is only up to scale, for if $\mathbf{W}_n$ is replaced by $c\mathbf{W}_n$ for some $c > 0$, $\widehat{\beta}_{GMM}$ does not change.
- In Section 4 of Chapter 7, β is identified as $(\Gamma'\mathbf{A}\Gamma)^{-1}\Gamma'\mathbf{A}\lambda =$ $\left(E\left[\mathbf{x}_i\mathbf{z}_i'\right]E\left[\mathbf{z}_i\mathbf{z}_i'\right]^{-1}\mathbf{A}E\left[\mathbf{z}_i\mathbf{z}_i'\right]^{-1}E\left[\mathbf{z}_i\mathbf{x}_i'\right]\right)^{-1}E\left[\mathbf{x}_i\mathbf{z}_i'\right]E\left[\mathbf{z}_i\mathbf{z}_i'\right]^{-1}\mathbf{A}E\left[\mathbf{z}_i\mathbf{z}_i'\right]^{-1}E\left[\mathbf{z}_i y_i\right]$, so there, $\mathbf{W}_n$ is the sample analog of $E\left[\mathbf{z}_i\mathbf{z}_i\right]^{-1}\mathbf{A}E\left[\mathbf{z}_i\mathbf{z}_i\right]^{-1}$.
- When $\mathbf{A} = E\left[\mathbf{z}_i\mathbf{z}_i\right]$, we obtain the 2SLS estimator, that is, $\mathbf{W}_n = (\mathbf{Z}'\mathbf{Z})^{-1}$.

- From the FOCs of GMM estimation, we can see that although we cannot make $\overline{g}_n(\beta) = \mathbf{0}$ exactly, we could let some of its linear combinations, say $\mathbf{B}_n\overline{g}_n(\beta)$, be zero, where $\mathbf{B}_n$ is a $k \times l$ matrix.
- For a weight matrix $\mathbf{W}_n$, $\mathbf{B}_n = \left(\frac{1}{n}\mathbf{X}'\mathbf{Z}\right)\mathbf{W}_n$. If $\mathbf{W}_n \xrightarrow{p} \mathbf{W} > 0$, and $\frac{1}{n}\mathbf{X}'\mathbf{Z} \xrightarrow{p} E\left[\mathbf{x}_i\mathbf{z}_i'\right] = \mathbf{G}'$, $\mathbf{B}_n$ converges to $\mathbf{B} = \mathbf{G}'\mathbf{W}$. So $\widehat{\beta}$ is as if defined by a MoM estimator such that $\mathbf{B}\overline{g}_n(\widehat{\beta}) = \mathbf{0}$.

Distribution of the GMM Estimator

Distribution of the GMM Estimator

- Note that

$$\left(\frac{1}{n}\mathbf{X}'\mathbf{Z}\right)\mathbf{W}_n\left(\frac{1}{n}\mathbf{Z}'\mathbf{X}\right)\xrightarrow{p}\mathbf{G}'\mathbf{W}\mathbf{G}$$

  and

$$\left(\frac{1}{n}\mathbf{X}'\mathbf{Z}\right)\mathbf{W}_n\left(\frac{1}{\sqrt{n}}\mathbf{Z}'\mathbf{u}\right)\xrightarrow{d}\mathbf{G}'\mathbf{W}N\left(\mathbf{0},\Omega\right),$$

  where $\Omega = E\left[\mathbf{z}_i\mathbf{z}_i'u_i^2\right] = E\left[g_i g_i'\right]$ with $g_i = \mathbf{z}_i u_i$.
- So

$$\sqrt{n}\left(\widehat{\beta}_{GMM}-\beta\right)\xrightarrow{d}N\left(\mathbf{0},\mathbf{V}\right),$$

  where

$$\mathbf{V}=\left(\mathbf{G}'\mathbf{W}\mathbf{G}\right)^{-1}\left(\mathbf{G}'\mathbf{W}\Omega\mathbf{W}\mathbf{G}\right)\left(\mathbf{G}'\mathbf{W}\mathbf{G}\right)^{-1}. \qquad (3)$$

- In general, GMM estimators are asymptotically normal with "sandwich form" asymptotic variances.
- It is easy to check this asymptotic distribution is the same as the MoM estimator defined by $\mathbf{B}\overline{g}_n(\widehat{\beta}) = \mathbf{0}$.

Optimal Weight Matrix

- A natural question is what is the optimal weight matrix $\mathbf{W}_0$ that minimizes $\mathbf{V}$. This turns out to be Ω^{-1} (exercise).
- This yields the efficient GMM estimator:

$$\widehat{\beta} = \left(\mathbf{X}'\mathbf{Z}\Omega^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Z}\Omega^{-1}\mathbf{Z}'\mathbf{y},$$

  which has the asymptotic variance $\mathbf{V}_0 = \left(\mathbf{G}'\Omega^{-1}\mathbf{G}\right)^{-1}$. This corresponds to the linear combination matrix $\mathbf{B} = \mathbf{G}'\Omega^{-1}$.
- $\mathbf{W}_0 = \Omega^{-1}$ is usually unknown in practice, but it can be estimated consistently.
- In the homoskedastic case, $E\left[u_i^2|\mathbf{z}_i\right] = \sigma^2$, then $\Omega = E\left[\mathbf{z}_i\mathbf{z}_i'\right]\sigma^2 \propto E\left[\mathbf{z}_i\mathbf{z}_i'\right]$ suggesting the weight matrix $\mathbf{W}_n = \left(\mathbf{Z}'\mathbf{Z}\right)^{-1}$, which generates the 2SLS estimator.
- So the 2SLS estimator is the efficient GMM estimator under homoskedasticity

Optimal Weight Matrix - An Illustration

- Suppose $E[x_i] = E[y_i] = \mu$ and $Cov(x_i, y_i) = 0$. We try to find an efficient GMM estimator for μ - the common mean of x and y.
- The moment conditions are $E[g(\mathbf{w}_i, \mu)] = \mathbf{0}$, where $\mathbf{w}_i = (x_i, y_i)'$:

$$g(\mathbf{w}_i, \mu) = \begin{pmatrix} x_i - \mu \\ y_i - \mu \end{pmatrix}.$$

- Since μ appears in both moment conditions, we hope to find a better estimator than $\overline{x}$ or $\overline{y}$ which uses only one moment condition.
- Suppose $\widehat{\mu} = \omega\overline{x} + (1 - \omega)\overline{y}$; then the asymptotic distribution of $\widehat{\mu}$ is

$$\sqrt{n}(\widehat{\mu} - \mu) \xrightarrow{d} N\left(0, \omega^2\sigma_x^2 + (1 - \omega)^2\sigma_y^2\right).$$

- Minimizing the asymptotic variance, we have

$$\omega = \frac{\sigma_y^2}{\sigma_x^2 + \sigma_y^2}.$$

- The sample (of x and y) with a larger variance is given a smaller weight, and the sample with a smaller variance is given a larger weight.

continue...

- The asymptotic variance under this optimal weight is $\frac{\sigma_x^2\sigma_y^2}{\sigma_x^2+\sigma_y^2} \leq \min\left\{\sigma_x^2, \sigma_y^2\right\}$.
- Note that

$$\begin{aligned} \mathbf{W}_0 &= E\left[g(\mathbf{w}_i,\mu)g(\mathbf{w}_i,\mu)'\right]^{-1} \\ &= \begin{pmatrix} E\left[(x_i-\mu)^2\right] & E\left[(x_i-\mu)(y_i-\mu)\right] \\ E\left[(x_i-\mu)(y_i-\mu)\right] & E\left[(y_i-\mu)^2\right] \end{pmatrix}^{-1} = \begin{pmatrix} \sigma_x^{-2} & 0 \\ 0 & \sigma_y^{-2} \end{pmatrix}. \end{aligned}$$

- So

$$J_n(\mu) = n\cdot\overline{g}_n(\mu)'\mathbf{W}_0\overline{g}_n(\mu) = n\left(\frac{(\overline{x}-\mu)^2}{\sigma_x^2} + \frac{(\overline{y}-\mu)^2}{\sigma_y^2}\right),$$

  and

$$\widehat{\mu} = \omega\overline{x} + (1-\omega)\overline{y}$$

  is the same as the weighted average above.
- In practice, σ_x^2 and σ_y^2 are unknown. In this simple example, they can be substituted by their sample analog. The next section deals with the general case.

Estimation of the Optimal Weight Matrix

Estimation of the Optimal Weight Matrix

- Given any weight matrix $\mathbf{W}_n > 0$, the GMM estimator $\widehat{\beta}_{GMM}$ is consistent yet inefficient.
- For example, we can set $\mathbf{W}_n = \mathbf{I}_l$. In the linear model, a better choice is $\mathbf{W}_n = (\mathbf{Z}'\mathbf{Z})^{-1}$ which corresponds to the 2SLS estimator.
- Given any such fist-step estimator, we can define the residuals $\widehat{u}_i = y_i - \mathbf{x}_i'\widehat{\beta}_{GMM}$ and moment equations $\widehat{g}_i = \mathbf{z}_i\widehat{u}_i = g\left(\mathbf{w}_i, \widehat{\beta}_{GMM}\right)$. Construct

$$\begin{aligned}\overline{g}_n &= \overline{g}_n(\widehat{\beta}_{GMM}) = \frac{1}{n}\sum_{i=1}^{n}\widehat{g}_i, \\ \widehat{g}_i^* &= \widehat{g}_i - \overline{g}_n,\end{aligned}$$

and define

$$\mathbf{W}_n = \left(\frac{1}{n}\sum_{i=1}^{n}\widehat{g}_i^*\widehat{g}_i^{*\prime}\right)^{-1} = \left(\frac{1}{n}\sum_{i=1}^{n}\widehat{g}_i\widehat{g}_i' - \overline{g}_n\overline{g}_n'\right)^{-1}. \tag{4}$$

- $\mathbf{W}_n \xrightarrow{p} \Omega^{-1}$, and GMM using $\mathbf{W}_n$ as the weight matrix is asymptotically efficient.

An Alternative Estimator

- A common alternative choice is to set

$$\mathbf{W}_n = \left(\frac{1}{n} \sum_{i=1}^{n} \widehat{g}_i \widehat{g}_i' \right)^{-1}, \tag{5}$$

  which uses the uncentered moment conditions.
- Since $E[g_i] = \mathbf{0}$, these two estimators are asymptotically equivalent under the hypothesis of correct specification.
- However, Alastair Hall (2000) has shown that the uncentered estimator is a poor choice.
- When constructing hypothesis tests, under the alternative hypothesis the moment conditions are violated, i.e. $E[g_i] \neq \mathbf{0}$, so the uncentered estimator will contain an undesirable bias term and the power of the test will be adversely affected.

Routine to Compute the Linear Efficient GMM Estimator

1. set $\mathbf{W}_n = (\mathbf{Z}'\mathbf{Z})^{-1}$, estimate $\widehat{\beta}$ using this weight matrix, and construct the residual $\widehat{u}_i = y_i - \mathbf{x}_i'\widehat{\beta}$.
2. set $\widehat{g}_i = \mathbf{z}_i\widehat{u}_i$, and let $\widehat{g}$ be the associated $n \times l$ matrix.
3. the efficient GMM estimator[1] is

$$\widehat{\beta} = \left(\mathbf{X}'\mathbf{Z}\left(\widehat{g}'\widehat{g} - n\overline{g}_n\overline{g}_n'\right)^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Z}\left(\widehat{g}'\widehat{g} - n\overline{g}_n\overline{g}_n'\right)^{-1}\mathbf{Z}'\mathbf{y}.$$

4. set

$$\widehat{\mathbf{V}} = n\left(\mathbf{X}'\mathbf{Z}\left(\widehat{g}'\widehat{g} - n\overline{g}_n\overline{g}_n'\right)^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1},$$

and asymptotic standard errors are given by the square roots of the diagonal elements of $\widehat{\mathbf{V}}/n$.

- **Iterative Estimator**: Given the efficient estimator $\widehat{\beta}$, we can continue to reestimate $\mathbf{V}$ by replacing $\widehat{g}_i$ by $g\left(\mathbf{w}_i, \widehat{\beta}\right)$ and construct a new estimator of β. This is repeated until the β estimator converges or enough iterations are conducted.

[1] In most cases, when we say "GMM" we actually mean "efficient GMM". There is little point in using an inefficient GMM estimator when the efficient estimator is easy to compute.

Nonlinear GMM

Nonlinear GMM

- Suppose the moment conditions are

$$E\left[g\left(\mathbf{w}_i,\theta_0\right)\right]=\mathbf{0},$$

where $g(\cdot,\cdot)\in\mathbb{R}^l$ is a general nonlinear function of $\theta\in\mathbb{R}^k$, $l\geq k$.

- The GMM estimator $\widehat{\theta}$ minimizes

$$J_n(\theta)=n\cdot\overline{g}_n(\theta)'\mathbf{W}_n\overline{g}_n(\theta),$$

where $\mathbf{W}_n$ is a consistent estimator of $\Omega^{-1}\equiv E\left[g_i(\theta_0)g_i(\theta_0)'\right]^{-1}$.

- Define $\mathbf{G}=E\left[\partial g_i(\theta_0)/\partial\theta'\right]$,

$$\sqrt{n}\left(\widehat{\theta}-\theta_0\right)\xrightarrow{d}N\left(\mathbf{0},\left(\mathbf{G}'\Omega^{-1}\mathbf{G}\right)^{-1}\right)\equiv N\left(\mathbf{0},\mathbf{V}\right). \quad (6)$$

- $\widehat{\mathbf{V}}\equiv\left(\widehat{\mathbf{G}}'\widehat{\Omega}^{-1}\widehat{\mathbf{G}}\right)^{-1}$, where $\widehat{\Omega}=n^{-1}\sum_{i=1}^n g_i^*(\widehat{\theta})g_i^*(\widehat{\theta})'$ with $g_i^*(\theta)=g_i(\theta)-\overline{g}_n(\theta)$, and $\widehat{\mathbf{G}}=n^{-1}\sum_{i=1}^n\partial g_i(\widehat{\theta})/\partial\theta'$.

Hypothesis Testing

Testing Overidentifying Restrictions: The J Test

- The hypotheses are

$$H_0 : \exists\, \beta_0 \text{ s.t. } E\left[g\left(\mathbf{w}_i, \beta_0\right)\right] = \mathbf{0} \tag{7}$$

  versus

$$H_1 :\ \forall\, \beta \in \mathscr{B},\ E\left[g\left(\mathbf{w}_i, \beta\right)\right] \neq \mathbf{0},$$

  where $\mathscr{B}$ is the parameter space.

- When $l = k$, there always exists a $\beta_0 \in \mathscr{B}$ such that $E\left[g\left(\mathbf{w}_i, \beta_0\right)\right] = \mathbf{0}$. So only if $l > k$, we need this test - to test whether the overidentifying restrictions are valid.
- For example, take the linear model $y_i = \mathbf{x}_{1i}'\beta_1 + \mathbf{x}_{2i}'\beta_2 + u_i$ with $E\left[\mathbf{x}_{1i}u_i\right] = \mathbf{0}$ and $E\left[\mathbf{x}_{2i}u_i\right] = \mathbf{0}$. It is possible that $\beta_2 = \mathbf{0}$, so that the linear equation may be written as $y_i = \mathbf{x}_{1i}'\beta_1 + u_i$. However, it is possible that $\beta_2 \neq \mathbf{0}$, and in this case it would be impossible to find a value of β_1 so that $E\left[\mathbf{x}_{1i}\left(y_i - \mathbf{x}_{1i}'\beta_1\right)\right] = \mathbf{0}$ and $E\left[\mathbf{x}_{2i}\left(y_i - \mathbf{x}_{1i}'\beta_1\right)\right] = \mathbf{0}$ hold simultaneously. In this sense an exclusion restriction ($\beta_2 = \mathbf{0}$) can be seen as an overidentifying restriction.

continue...

- Note that $\overline{g}_n(\widehat{\beta}) \xrightarrow{p} E[g_i(\beta_0)]$, and thus $\overline{g}_n(\widehat{\beta})$ can be used to assess whether or not the hypothesis that $E[g_i(\beta_0)] = \mathbf{0}$ is true or not.
- The test statistic is the criterion function at the parameter estimates

$$J_n = J_n\left(\widehat{\beta}\right) = n\overline{g}_n(\widehat{\beta})'\mathbf{W}_n\overline{g}_n(\widehat{\beta}) = n^2\overline{g}_n(\widehat{\beta})'\left(\widehat{g}'\widehat{g} - n\overline{g}_n\overline{g}_n'\right)^{-1}\overline{g}_n(\widehat{\beta}).$$

- Under the hypothesis of correct specification,

$$J_n \xrightarrow{d} \chi^2_{l-k}.$$

- The degrees of freedom of the asymptotic distribution are the number of over-identifying restrictions.
- If the statistic J_n exceeds the chi-square critical value, we can reject the model.

Alternative Way to Understand the J Test (I)

- The J test is actually an F test in the homoskedastic linear model

$$\begin{aligned} y_i &= \mathbf{x}_{1i}'\beta_1 + \mathbf{x}_{2i}'\beta_2 + u_i, \\ E[\mathbf{z}_i u_i] &= 0, \, E[u_i^2|\mathbf{z}_i] = \sigma^2, \end{aligned} \tag{8}$$

  where $\mathbf{z}_i = (\mathbf{x}_{1i}', \mathbf{z}_{2i}')'$.
- Exogeneity of the instruments means that they are uncorrelated with u_i, which suggests that the instruments should be approximately uncorrelated with $\widehat{u}_i$, where $\widehat{u}_i = y_i - \mathbf{x}_{1i}'\widehat{\beta}_1 - \mathbf{x}_{2i}'\widehat{\beta}_2$ with $\widehat{\beta} = \left(\widehat{\beta}_1', \widehat{\beta}_2'\right)'$ being the 2SLS estimator.
- So we expect in the regression

$$\widehat{u}_i = \mathbf{x}_{1i}'\delta_1 + \mathbf{z}_{2i}'\delta_2 + v_i, \tag{9}$$

  the estimate of $\delta \equiv (\delta_1', \delta_2')'$ is close to zero.
- Let F denote the homoskedasticity-only F statistic testing $\delta_2 = \mathbf{0}$; then $l_2 F$ converges to $\chi^2_{l_2 - k_2} = \chi^2_{l-k}$.

Alternative Way to Understand the J Test (II)

- In the linear model (8), suppose we have one endogenous variable x_{2i} and two instruments $\mathbf{z}_{2i}$, and then we can use either instrument to estimate $\beta \equiv (\beta_1', \beta_2')'$.
- If H_0 holds, we expect that these two instruments will generate similar estimates. If the two estimates are very different, then we suspect H_0 fails.
- The J test implicitly makes this comparison.

- The J test is also called the *Sargan-Hansen test* due to a special case established by Sargan (1958) and the general case by Hansen (1982).
- The GMM over-identification test is a very useful by-product of the GMM methodology, and it is advisable to report the statistic J_n as a general test of model adequacy whenever GMM is used.

Three Asymptotically Equivalent Tests (I) - The Wald Test

- Suppose we want to test

$$H_0 : \underset{(q\times 1)}{\mathbf{r}(\beta)} = \mathbf{0} \text{ vs } H_1 : \underset{(q\times 1)}{\mathbf{r}(\beta)} \neq \mathbf{0}.$$

- The Wald statistic:

$$\mathbf{W}_n = n \cdot \mathbf{r}\left(\widehat{\beta}\right)' \left[\widehat{\mathbf{R}}'\widehat{\mathbf{V}}\widehat{\mathbf{R}}\right]^{-1} \mathbf{r}\left(\widehat{\beta}\right),$$

where $\widehat{\beta} = \arg\min_{\beta} J_n(\beta)$ is the unrestricted estimator and $\widehat{\mathbf{R}} = \partial \mathbf{r}\left(\widehat{\beta}\right)' / \partial \beta$.

- Advantage: it only requires the unconstrained estimator to compute it.
- Disadvantage: it is not invariant to reparametrization.
  - When the hypothesis is non-linear, a better approach is to directly use the GMM criterion function.

Three Asymptotically Equivalent Tests (II) - the Distance Test

- The idea was first put forward by Newey and West (1987a), so the test is also called the *Newey-West test*.
- Define the restricted estimator $\widetilde{\beta}$ as

$$\widetilde{\beta} = \arg \min_{\mathbf{r}(\beta)=\mathbf{0}} J_n(\beta).$$

  The two minimizing criterion functions for $\widehat{\beta}$ and $\widetilde{\beta}$ are $J_n(\widehat{\beta})$ and $J_n(\widetilde{\beta})$.
- The GMM distance statistic is the difference

$$D_n = J_n(\widetilde{\beta}) - J_n(\widehat{\beta}).$$

- Newey and West (1987a) suggested to use the same weight matrix $\mathbf{W}_n$ for both null and alternative, as this ensures that $D_n \geq 0$.
- This reasoning is not compelling, however, and some current research suggests that this restriction is not necessary for good performance of the test.
- This test shares the useful feature of likelihood ratio (LR) tests in that it is a natural by-product of the computation of alternative models.

Three Asymptotically Equivalent Tests (III) - the LM Test

- Another test is the Lagrange multiplier (LM) test or C.R. Rao's score test.
- Its test statistic is constructed as

$$LM_n = n\left[\overline{g}_n\left(\widetilde{\beta}\right)'\mathbf{W}_n\mathbf{G}_n\left(\widetilde{\beta}\right)\right]\widetilde{\mathbf{V}}\left[\mathbf{G}_n\left(\widetilde{\beta}\right)'\mathbf{W}_n\overline{g}_n\left(\widetilde{\beta}\right)\right],$$

  where

$$\widetilde{\mathbf{V}} = \left[\mathbf{G}_n\left(\widetilde{\beta}\right)'\mathbf{W}_n\mathbf{G}_n\left(\widetilde{\beta}\right)\right]^{-1},$$

  and $\mathbf{G}_n\left(\widetilde{\beta}\right)'\mathbf{W}_n\overline{g}_n\left(\widetilde{\beta}\right)$ is the first-order derivative of $J_n(\cdot)$ at $\widetilde{\beta}$ and plays the role of the score function in the likelihood framework.
- Advantage: we need only calculate the restricted estimator $\widetilde{\beta}$, while we need to calculate both $\widehat{\beta}$ and $\widetilde{\beta}$ in the distance statistic.

The Trinity in GMM

- **Proposition 1**: Under some regularity conditions, and the local alternatives $\beta_n = \beta + n^{-1/2}\mathbf{b}$,

$$W_n \xrightarrow{d} \chi_q^2(\lambda),$$

where $\lambda = \mathbf{b}'\mathbf{R}(\mathbf{R}'\mathbf{V}\mathbf{R})^{-1}\mathbf{R}\mathbf{b}$. In addition, $W_n - D_n = o_p(1)$ and $W_n - LM_n = o_p(1)$.

- The three tests are asymptotically equivalent even under the local alternatives and when the moment conditions are nonlinear in β.
- It should be emphasized that the optimal weight matrix is used in the construction of D_n
- Otherwise, D_n is not asymptotically chi-squared and is not asymptotically equivalent to W_n.
- Also, the form of the LM statistic would be more complicated, and would in general involve the Jacobian matrix $\mathbf{R}$ of the constraints.
- So it is strongly suggested to use the optimal weight matrix in the hypothesis testing of GMM.

Numerical Equivalence

- **Proposition 2**: (i) When the model is just-identified, $LM_n = D_n$. (ii) When $g(\mathbf{w},\beta) = g_1(\mathbf{w}) - g_2(\mathbf{w})\beta$, $D_n = LM_n$. (iii) When $g(\mathbf{w},\beta) = g_1(\mathbf{w}) - g_2(\mathbf{w})\beta$ and $\mathbf{r}(\beta) = \mathbf{R}'\beta - \mathbf{c}$, $W_n = D_n = LM_n$.
- (i) In the just-identified case, $\overline{g}_n\left(\widehat{\beta}\right) = \mathbf{0}$, so $D_n = J_n(\widetilde{\beta}) = n \cdot \overline{g}_n\left(\widetilde{\beta}\right)' \mathbf{W}_n \overline{g}_n\left(\widetilde{\beta}\right)$.

  On the other hand, given $\mathbf{G}_n\left(\widetilde{\beta}\right)$ is invertible,

$$\begin{aligned} LM_n &= n\left[\overline{g}_n\left(\widetilde{\beta}\right)' \mathbf{W}_n \mathbf{G}_n\left(\widetilde{\beta}\right)\right] \widetilde{\mathbf{V}} \left[\mathbf{G}_n\left(\widetilde{\beta}\right)' \mathbf{W}_n \overline{g}_n\left(\widetilde{\beta}\right)\right] \\ &= n \cdot \overline{g}_n\left(\widetilde{\beta}\right)' \mathbf{W}_n \mathbf{G}_n\left(\widetilde{\beta}\right) \left[\mathbf{G}_n\left(\widetilde{\beta}\right)^{-1} \mathbf{W}_n^{-1} \mathbf{G}_n\left(\widetilde{\beta}\right)'^{-1}\right] \mathbf{G}_n\left(\widetilde{\beta}\right)' \mathbf{W}_n \overline{g}_n\left(\widetilde{\beta}\right) \\ &= n \cdot \overline{g}_n\left(\widetilde{\beta}\right)' \mathbf{W}_n \overline{g}_n\left(\widetilde{\beta}\right). \end{aligned}$$

- (ii) does not include W_n because it involves the Jacobian of the constraints when $\mathbf{r}(\cdot)$ is nonlinear. (iii) is an exercise.
- The linear projection case: $LM_n = D_n$ even if the constraints are nonlinear; when the constraints are linear, all three are the same.

  - $D_n = n \cdot \overline{g}_n\left(\widetilde{\beta}\right)' \mathbf{W}_n \overline{g}_n\left(\widetilde{\beta}\right) \neq \sum_{i=1}^n (y_i - \mathbf{x}_i'\widetilde{\beta})^2 - \sum_{i=1}^n (y_i - \mathbf{x}_i'\widehat{\beta})^2$, where $\overline{g}_n\left(\widetilde{\beta}\right) = n^{-1}\sum_{i=1}^n \mathbf{x}_i(y_i - \mathbf{x}_i'\widetilde{\beta})$.

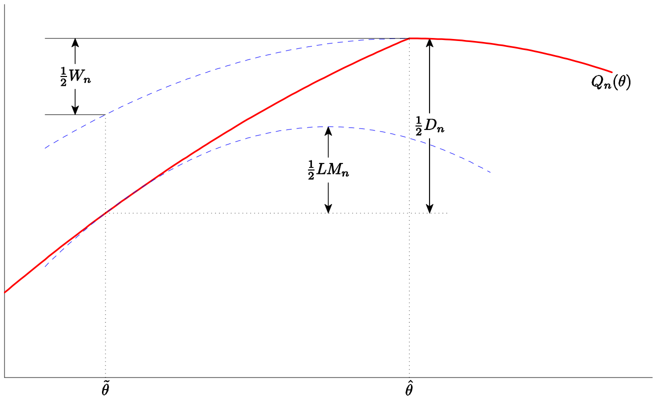

Figure: Trinity

Confidence Region - Inverting the Distance Statistic

- Use the distance statistic (rather than the Wald statistic) because of its better performance in hypothesis testing.
- Suppose we want to construct confidence region for θ_2, where $\theta = (\theta_1', \theta_2')' \in \mathbb{R}^k$ and $\theta_2 \in \mathbb{R}^{k_2}$ is a subvector of θ.
- We need to find θ_2 such that

$$J_n\left(\widetilde{\theta}_1(\theta_2), \theta_2\right) - J_n\left(\widehat{\theta}\right) \le \chi^2_{k_2,\alpha},$$

  where $\widetilde{\theta}_1(\theta_2) = \arg\min_{\theta_1} J_n(\theta_1, \theta_2)$ for a given θ_2, the df of the χ^2 limiting distribution is k_2 because the df of $J_n\left(\widetilde{\theta}_1(\theta_2), \theta_2\right)$ is $l - k_1$ and the df of $J_n\left(\widehat{\theta}\right)$ is $l - k$ so the difference is $(l - k_1) - (l - k) = k - k_1 = k_2$.
- We can also construct confidence region for θ_2 by collecting θ_2's such that $J_n\left(\widetilde{\theta}_1(\theta_2), \theta_2\right) \le \chi^2_{l-k_1,\alpha}$ directly.
- However, $J_n\left(\widetilde{\theta}_1(\theta_2), \theta_2\right) = \left[J_n\left(\widetilde{\theta}_1(\theta_2), \theta_2\right) - J_n\left(\widehat{\theta}\right)\right] + J_n\left(\widehat{\theta}\right)$, so this confidence region is based on the joint test of overidentification and $\theta_2 = \theta_{20}$.
  - If the model is misspecified so that the overidentifying conditions are invalid, this confidence region can be null.

Conditional Moment Restrictions

Conditional Moment Restrictions

- In many cases, the model may imply conditional moment restrictions

$$E\left[u\left(\mathbf{w},\beta_0\right)|\mathbf{x}\right]=\mathbf{0},$$

where $u\left(\mathbf{w},\beta\right)$ is some $s\times 1$ function of the observation and the parameters.

- For example, in linear regression, $u\left(\mathbf{w},\beta\right)=y-\mathbf{x}'\beta$, $\mathbf{w}=\left(y,\mathbf{x}'\right)'$, and $s=1$; in a joint model of conditional mean and variance,

$$u\left(\mathbf{w},\beta\right)=\begin{pmatrix} y-\mathbf{x}'\beta \\ \left(y-\mathbf{x}'\beta\right)^2-f\left(\mathbf{x}\right)'\gamma \end{pmatrix}$$

for a specification $Var\left(y|\mathbf{x}\right)=f\left(\mathbf{x}\right)'\gamma$, so $s=2$.

- Conditional moment restrictions imply infinite unconditional moment conditions, since for any function of $\mathbf{x}$, say $\phi\left(\mathbf{x}\right)$, $E\left[\phi(\mathbf{x})u_i\left(\mathbf{w},\beta_0\right)\right]=\mathbf{0}$.
- So a natural question is which instruments are optimal, or what is the semiparametric efficiency bound for β_0.
- Chamberlain (1987) derived this bound by approximating the CDF $F\left(\mathbf{x}\right)$ and the conditional CDF $F\left(\mathbf{w}|\mathbf{x}\right)$ with multinomial distributions.

Semiparametric Efficiency Bound

- It turns out that the optimal instruments are
$$\mathbf{A}(\mathbf{x}) = \mathbf{G}(\mathbf{x})'\Omega(\mathbf{x})^{-1},$$
where $\mathbf{G}(\mathbf{x}) = E\left[\partial u(\mathbf{w},\beta_0)/\partial\beta' \middle| \mathbf{x}\right]$, and $\Omega(\mathbf{x}) = E\left[u(\mathbf{w},\beta_0)u(\mathbf{w},\beta_0)' \middle| \mathbf{x}\right]$.
- $\mathbf{A}(\mathbf{x})$ is similar to the optimal linear combination $\mathbf{B}$ in the unconditional moment case, but now we condition every random variable on $\mathbf{x}$.
- Using the optimal instruments, the unconditional moment conditions are
$$E\left[\mathbf{A}(\mathbf{x})u(\mathbf{w},\beta_0)\right] = \mathbf{0}.$$
- Applying the formula of the asymptotic variance for the MoM estimator, we have the semiparametric efficiency bound for β_0
$$\begin{aligned} & E\left[\mathbf{A}(\mathbf{x})\partial u(\mathbf{w},\beta_0)/\partial\beta'\right]^{-1} \cdot E\left[\mathbf{A}(\mathbf{x})u(\mathbf{w},\beta_0)u(\mathbf{w},\beta_0)'\mathbf{A}(\mathbf{x})'\right]\cdots \\ = \; & E\left[\mathbf{G}(\mathbf{x})'\Omega(\mathbf{x})^{-1}\mathbf{G}(\mathbf{x})'\right]^{-1}. \end{aligned}$$
- In the linear regression case, $\mathbf{G}(\mathbf{x}) = \mathbf{x}'$, and $\Omega(\mathbf{x}) = \sigma^2(\mathbf{x})$, so the optimal instrument is $\mathbf{x}/\sigma^2(\mathbf{x})$, which corresponds to the generalized least squares estimator, and the semiparametric efficiency bound for β_0 is $E\left[\mathbf{x}\mathbf{x}'/\sigma^2(\mathbf{x})\right]$.

Alternative Inference Procedures and Extensions (*)

Underestimation of the Sample Variation and Solutions

- Monte Carlo studies have shown that estimated asymptotic standard errors of the efficient two-step GMM estimator can be severely downward biased in small samples.
- A key observation for the source of this bias is that the weight matrix used in the calculation of the efficient two-step GMM estimator is based on initial consistent parameter estimates whose variation is not embodied in the asymptotic covariance matrix estimation.
- Solutions:
  - nonlinear procedures: the generalized empirical likelihood (GEL) method.
  - linear procedures: incorporate the variation in the first-stage estimator explicitly.
  - bootstrap procedures: refine the inferences based on the two-step GMM estimator.

A Special GEL Estimator - Continuously-Updated Estimator (CUE)

- **Idea**: let the weight matrix be considered as a function of θ.
- The criterion function becomes

$$J_n(\theta) = n \cdot \overline{g}_n(\theta)' \left(\frac{1}{n} \sum_{i=1}^{n} g_i^*(\theta) g_i^*(\theta)' \right)^{-1} \overline{g}_n(\theta),$$

  where

$$g_i^*(\theta) = g_i(\theta) - \overline{g}_n(\theta).$$

- The $\widehat{\theta}$ which minimizes this function is called the CUE of GMM, and was introduced by Hansen et al. (1996).
- The CUE has some better properties (e.g., smaller bias) than traditional GMM, but can be numerically tricky to obtain in some cases.

Extensions

- $\mathbf{w}_i$, $i = 1, \cdots, n$, is a random sample. If $\mathbf{w}_i$, $i = 1, \cdots, n$, are time series $\mathbf{w}_t$, $t = 1, \cdots, T$, such that $g(\mathbf{w}_t, \theta)$ are correlated, then the optimal

$$\begin{aligned} \Omega &= TE\left[\overline{g}_T(\theta_0)\overline{g}_T(\theta_0)'\right] \\ &= \sum_{v=-\infty}^{\infty} E\left[g(\mathbf{w}_t, \theta_0)g(\mathbf{w}_{t-v}, \theta_0)'\right] \equiv \sum_{v=-\infty}^{\infty} \Omega_v. \end{aligned}$$

  A consistent estimator of Ω is often called the *heteroskedasticity and autocorrelation consistent* (HAC) estimator.
- $g(\mathbf{w}, \theta)$ is smooth in θ. When g is nondifferentiable and/or discontinuous in θ (e.g., the moment conditions in quantile regression), $\mathbf{G}$ is not well defined.
- $\mathbf{G}$ is full column rank. When $\mathbf{G} \approx \mathbf{C}n^{-1/2}$, the instruments are weak, and θ cannot be consistently estimated.
- l is fixed. When l can go to infinity, there are many moment conditions which will increase the bias of the GMM estimator and deteriorates the estimation of Ω.
- k is fixed. When k can go to infinity, there are nonparametric parameters in the moment conditions. For identification, we need infinite moment conditions.
- There are only moment equalities. If there are moment inequalities, θ can only be partially identified.