Generalized Method of Moment

CHUNG-MING KUAN
Department of Finance & CRETA
National Taiwan University

June 16, 2010

Lecture Outline

1. Generalized Method of Moments
   - GMM Estimation
   - Asymptotic Properties of the GMM Estimator
   - Different GMM Estimators
   - Examples
2. Large Sample Tests
   - Over-Identifying Restriction Test
   - Hausman Test
   - Wald Test
3. Conditional Moment Restrictions
   - Estimation of Conditional Moment Restrictions

Moment Conditions

- Given $y_t = \mathbf{x}_t'\boldsymbol{\beta} + e_t$, consider the moment function:

$$\mathbb{E}(\mathbf{x}_t e_t) = \mathbb{E}[\mathbf{x}_t(y_t - \mathbf{x}_t'\boldsymbol{\beta})].$$

When $\mathbf{x}_t'\boldsymbol{\beta}$ is correctly specified for the linear projection, we have the moment condition:

$$\mathbb{E}[\mathbf{x}_t(y_t - \mathbf{x}_t'\boldsymbol{\beta}_o)] = \mathbf{0}.$$

Its sample counterpart,

$$\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t(y_t - \mathbf{x}_t'\boldsymbol{\beta}) = \mathbf{0},$$

is also the FOC for OLS estimation.

- Similarly, given $y_t = f(\mathbf{x}_t; \boldsymbol{\beta}) + e_t$ and the moment function:

$$\mathbb{E}[\nabla f(\mathbf{x}_t; \boldsymbol{\beta}) e_t] = \mathbb{E}\{\nabla f(\mathbf{x}_t; \boldsymbol{\beta})[y_t - f(\mathbf{x}_t; \boldsymbol{\beta})]\}.$$

  When $f(\mathbf{x}_t; \boldsymbol{\beta})$ is correctly specified for $\mathbb{E}(y_t | \mathbf{x}_t)$, we have the moment condition

$$\mathbb{E}\{\mathbf{x}_t[y_t - f(\mathbf{x}_t; \boldsymbol{\beta}_o)]\} = \mathbf{0}.$$

  Its sample counterpart is the FOC for NLS estimation:

$$\frac{1}{T} \sum_{t=1}^{T} \nabla f(\mathbf{x}_t; \boldsymbol{\beta})[y_t - f(\mathbf{x}_t; \boldsymbol{\beta})] = \mathbf{0}.$$

- When the quasi-likelihood function $f(\mathbf{x}_t; \boldsymbol{\theta})$ is correctly specified, the moment condition $\mathbb{E}[\nabla \ln f(\mathbf{x}_t; \boldsymbol{\theta}_o)] = \mathbf{0}$ holds. The sample counterpart of this moment condition is the average of the score functions and yields the QMLE $\tilde{\boldsymbol{\theta}}$.

- Given $y_t = \mathbf{x}_t'\boldsymbol{\beta} + e_t$ and the moment function:

$$\mathbb{E}(\mathbf{z}_t e_t) = \mathbb{E}[\mathbf{z}_t(y_t - \mathbf{x}_t'\boldsymbol{\beta})].$$

When the variables $\mathbf{z}_t$ are proper instrument variables such that

$$\mathbb{E}[\mathbf{z}_t(y_t - \mathbf{x}_t'\boldsymbol{\beta}_o)] = \mathbf{0}.$$

The sample counterpart of this moment condition is the FOC for IV estimation:

$$\frac{1}{T}\sum_{t=1}^{T} \mathbf{z}_t(y_t - \mathbf{x}_t'\boldsymbol{\beta}) = \mathbf{0}.$$

- Note that we can not solve for unknown parameters if the number of moment conditions is more than the number of parameters.

GMM Estimation

Consider q moment functions $\mathbb{E}[\mathbf{m}(\mathbf{z}_t;\boldsymbol{\theta})]$, where $\boldsymbol{\theta}$ $(k \times 1)$ is the parameter vector. Suppose there exists unique $\boldsymbol{\theta}_o$ such that

$$\mathbb{E}[\mathbf{m}(\mathbf{z}_t;\boldsymbol{\theta}_o)] = \mathbf{0}.$$

These conditions are exactly identified if $q = k$ and over-identified if $q > k$.

When the conditions are exactly identified, $\boldsymbol{\theta}_o$ can be estimated by solving their sample counterpart:

$$\bar{\mathbf{m}}_T(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T}\mathbf{m}(\mathbf{z}_t;\boldsymbol{\theta}) = \mathbf{0}.$$

This is the method of moment, which is not applicable when the conditions are over-identified.

Given a $q \times q$ symmetric and p.d. weighting matrix $\mathbf{W}_o$, note that the following quadratic objective function,

$$\bar{Q}(\boldsymbol{\theta}; \mathbf{W}_o) := \mathbb{E}[\mathbf{m}(\mathbf{z}_t; \boldsymbol{\theta})]' \mathbf{W}_o \, \mathbb{E}[\mathbf{m}(\mathbf{z}_t; \boldsymbol{\theta})],$$

is minimized at $\boldsymbol{\theta} = \boldsymbol{\theta}_o$. The generalized method of moments (GMM) of Hansen (1982, *Econometrica*) suggests estimating $\boldsymbol{\theta}_o$ by minimizing the sample counterpart of $\bar{Q}(\boldsymbol{\theta}; \mathbf{W}_o)$:

$$Q_T(\boldsymbol{\theta}; \mathbf{W}_T) = \left[\bar{\mathbf{m}}_T(\boldsymbol{\theta})\right]' \mathbf{W}_T \left[\bar{\mathbf{m}}_T(\boldsymbol{\theta})\right],$$

where $\mathbf{W}_T$ is a symmetric and p.d. weighting matrix, possibly dependent on the sample, and $\mathbf{W}_T$ converges to $\mathbf{W}_o$ in probability. The GMM estimator is thus

$$\hat{\boldsymbol{\theta}}_T(\mathbf{W}_T) = \arg\min_{\boldsymbol{\theta} \in \Theta} Q_T(\boldsymbol{\theta}; \mathbf{W}_T),$$

which clearly depends on the choice of $\mathbf{W}_T$.

The FOC of GMM estimation contains k equations in k unknowns:

$$\mathbf{G}_T(\boldsymbol{\theta})'\mathbf{W}_T\bar{\mathbf{m}}_T(\boldsymbol{\theta}) = \mathbf{0},$$

where $\mathbf{G}_T(\boldsymbol{\theta}) = T^{-1}\sum_{t=1}^{T}\nabla\mathbf{m}(\mathbf{z}_t;\boldsymbol{\theta})$. The GMM estimator is then solved using a nonlinear optimization algorithm.

For example, in the linear regression case,

$$\bar{\mathbf{m}}_T(\boldsymbol{\beta}) = \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t(y_t - \mathbf{x}_t'\boldsymbol{\beta}).$$

When $\mathbf{W}_T = \mathbf{I}_k$, the FOC is

$$\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'\right)\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t(y_t - \mathbf{x}_t'\boldsymbol{\beta})\right) = \mathbf{0}.$$

Clearly, the resulting GMM estimator is the OLS estimator.

Consistency

By invoking a suitable ULLN, $Q_T(\boldsymbol{\theta}; \mathbf{W}_T)$ is close to $\bar{Q}(\boldsymbol{\theta}; \mathbf{W}_o)$ uniformly in $\boldsymbol{\theta}$ when T becomes large. Hence, the GMM estimator $\hat{\boldsymbol{\theta}}_T(\mathbf{W}_T)$ ought to be close to $\boldsymbol{\theta}_o$, the minimizer of $\bar{Q}(\boldsymbol{\theta}; \mathbf{W}_o)$, for sufficiently large T. This is the underlying idea of establishing GMM consistency. This approach is analogous to that for NLS and QMLE consistency. Note that consistency does not depend on the weighting matrix. For example, $\hat{\boldsymbol{\theta}}_T(\mathbf{I}_q)$ is consistent for $\boldsymbol{\theta}_o$, which is also the minimizer of $\bar{Q}(\boldsymbol{\theta}; \mathbf{I}_q)$.

Consider the mean value expansion:

$$\sqrt{T}\bar{\mathbf{m}}_T(\hat{\boldsymbol{\theta}}_T(\mathbf{W}_T)) = \sqrt{T}\bar{\mathbf{m}}_T(\boldsymbol{\theta}_o) + \mathbf{G}_T(\boldsymbol{\theta}_T^{\dagger})\sqrt{T}(\hat{\boldsymbol{\theta}}_T(\mathbf{W}_T) - \boldsymbol{\theta}_o),$$

where $\boldsymbol{\theta}_T^{\dagger}$ lies between $\hat{\boldsymbol{\theta}}_T(\mathbf{W}_T)$ and $\boldsymbol{\theta}_o$.

Asymptotic Distribution

Using the FOC of GMM estimation,

$$\begin{aligned}\mathbf{0} &= \mathbf{G}_T\big(\hat{\boldsymbol{\theta}}_T(\mathbf{W}_T)\big)'\mathbf{W}_T\sqrt{T}\bar{\mathbf{m}}_T\big(\hat{\boldsymbol{\theta}}_T(\mathbf{W}_T)\big)\\ &= \mathbf{G}_T\big(\hat{\boldsymbol{\theta}}_T(\mathbf{W}_T)\big)'\mathbf{W}_T\sqrt{T}\bar{\mathbf{m}}_T(\boldsymbol{\theta}_o)\\ &\qquad + \mathbf{G}_T\big(\hat{\boldsymbol{\theta}}_T(\mathbf{W}_T)\big)'\mathbf{W}_T\mathbf{G}_T(\boldsymbol{\theta}_T^\dagger)\sqrt{T}\big(\hat{\boldsymbol{\theta}}_T(\mathbf{W}_T)-\boldsymbol{\theta}_o\big).\end{aligned}$$

When $\nabla\mathbf{m}(\mathbf{z}_t;\boldsymbol{\theta})$ obey a ULLN, so that $\mathbf{G}_T\big(\hat{\boldsymbol{\theta}}_T(\mathbf{W}_T)\big)$ converges to $\mathbf{G}_o = \mathbb{E}[\nabla\mathbf{m}(\mathbf{z}_t;\boldsymbol{\theta}_o)]$. Then,

$$\begin{aligned}\sqrt{T}\big(\hat{\boldsymbol{\theta}}_T(\mathbf{W}_T)-\boldsymbol{\theta}_o\big) &= -\big\{\mathbb{E}[\nabla\mathbf{m}(\mathbf{z}_t;\boldsymbol{\theta}_o)]'\mathbf{W}_o\,\mathbb{E}[\nabla\mathbf{m}(\mathbf{z}_t;\boldsymbol{\theta}_o)]\big\}^{-1}\\ &\qquad \mathbb{E}[\nabla\mathbf{m}(\mathbf{z}_t;\boldsymbol{\theta}_o)]'\mathbf{W}_o[\sqrt{T}\bar{\mathbf{m}}_T(\boldsymbol{\theta}_o)] + o_{\mathbb{P}}(1)\\ &= -\big(\mathbf{G}_o'\mathbf{W}_o\mathbf{G}_o\big)^{-1}\mathbf{G}_o'\mathbf{W}_o\sqrt{T}\bar{\mathbf{m}}_T(\boldsymbol{\theta}_o) + o_{\mathbb{P}}(1).\end{aligned}$$

The asymptotic distribution of $\sqrt{T}\big(\hat{\boldsymbol{\theta}}_T(\mathbf{W}_T) - \boldsymbol{\theta}_o\big)$ is thus determined by

$$\sqrt{T}\bar{\mathbf{m}}_T(\boldsymbol{\theta}_o) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \mathbf{m}(\mathbf{z}_t; \boldsymbol{\theta}_o).$$

When $\mathbf{m}(\mathbf{z}_t; \boldsymbol{\theta}_o)$ obey a central limit theorem:

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \mathbf{m}(\mathbf{z}_t; \boldsymbol{\theta}_o) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_o),$$

we have $\sqrt{T}\big(\hat{\boldsymbol{\theta}}_T(\mathbf{W}_T) - \boldsymbol{\theta}_o\big) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}_o)$, where

$$\boldsymbol{\Omega}_o = \big(\mathbf{G}_o'\mathbf{W}_o\mathbf{G}_o\big)^{-1}\mathbf{G}_o'\mathbf{W}_o\boldsymbol{\Sigma}_o\mathbf{W}_o\mathbf{G}_o\big(\mathbf{G}_o'\mathbf{W}_o\mathbf{G}_o\big)^{-1}.$$

When $\mathbf{W}_o = \boldsymbol{\Sigma}_o^{-1}$, it is easy to see that $\boldsymbol{\Omega}_o$ simplifies to

$$\big(\mathbf{G}_o'\boldsymbol{\Sigma}_o^{-1}\mathbf{G}_o\big)^{-1}\mathbf{G}_o'\boldsymbol{\Sigma}_o^{-1}\mathbf{G}_o\big(\mathbf{G}_o'\boldsymbol{\Sigma}_o^{-1}\mathbf{G}_o\big)^{-1} = \big(\mathbf{G}_o'\boldsymbol{\Sigma}_o^{-1}\mathbf{G}_o\big)^{-1} =: \boldsymbol{\Omega}_o^*.$$

Asymptotic Efficiency

To compare $\boldsymbol{\Omega}_o$ and $\boldsymbol{\Omega}_o^*$, note that

$$\begin{aligned}
&(\boldsymbol{\Omega}_o^*)^{-1} - \boldsymbol{\Omega}_o^{-1} \\
&\quad = \mathbf{G}_o'\boldsymbol{\Sigma}_o^{-1}\mathbf{G}_o - \mathbf{G}_o'\mathbf{W}_o\mathbf{G}_o\big(\mathbf{G}_o'\mathbf{W}_o\boldsymbol{\Sigma}_o\mathbf{W}_o\mathbf{G}_o\big)^{-1}\mathbf{G}_o'\mathbf{W}_o\mathbf{G}_o \\
&\quad = \mathbf{G}_o'\boldsymbol{\Sigma}_o^{-1/2} \\
&\qquad\quad \big[\mathbf{I} - \boldsymbol{\Sigma}_o^{1/2}\mathbf{W}_o\mathbf{G}_o\big(\mathbf{G}_o'\mathbf{W}_o\boldsymbol{\Sigma}_o^{1/2}\boldsymbol{\Sigma}_o^{1/2}\mathbf{W}_o\mathbf{G}_o\big)^{-1}\mathbf{G}_o'\mathbf{W}_o\boldsymbol{\Sigma}_o^{1/2}\big] \\
&\qquad\quad \boldsymbol{\Sigma}_o^{-1/2}\mathbf{G}_o,
\end{aligned}$$

which is p.s.d. because the matrix in the square bracket is symmetric and idempotent. Thus, $\boldsymbol{\Omega}_o - \boldsymbol{\Omega}_o^*$ is p.s.d. This shows that, given the weighting matrix whose limit is $\boldsymbol{\Sigma}_o^{-1}$, the resulting GMM estimator is asymptotically efficient. $\boldsymbol{\Sigma}_o^{-1}$ is thus known as the optimal (limiting) weighting matrix.

Two-Step Optimal GMM Estimator

Hansen and Singleton (1982, *Econometrica*):

1. Compute a preliminary, consistent estimator $\hat{\boldsymbol{\theta}}_{1,T}$ based on the pre-specified weighting matrix $\mathbf{W}_{0,T}$:

$$\hat{\boldsymbol{\theta}}_{1,T} := \arg\min_{\boldsymbol{\theta} \in \Theta} \left[\bar{\mathbf{m}}_T(\boldsymbol{\theta})\right]' \mathbf{W}_{0,T} \left[\bar{\mathbf{m}}_T(\boldsymbol{\theta})\right].$$

   For example, we may set $\mathbf{W}_{0,T} = \mathbf{I}_q$.

2. Compute a consistent estimator for $\boldsymbol{\Sigma}_o$ based on $\hat{\boldsymbol{\theta}}_{1,T}$ and use its inverse as the optimal weighting matrix, i.e., $\mathbf{W}_T(\hat{\boldsymbol{\theta}}_{1,T}) = \widehat{\boldsymbol{\Sigma}}_T^{-1}$.

3. The two-step optimal GMM estimator is obtained as

$$\hat{\boldsymbol{\theta}}_{2,T} := \arg\min_{\boldsymbol{\theta} \in \Theta} \left[\bar{\mathbf{m}}_T(\boldsymbol{\theta})\right]' \widehat{\boldsymbol{\Sigma}}_T^{-1} \left[\bar{\mathbf{m}}_T(\boldsymbol{\theta})\right],$$

   which is asymptotically efficient.

Drawbacks of Two-Step Estimators

- The finite-sample performance of the two-step estimator depends on the preliminary GMM estimator $\hat{\boldsymbol{\theta}}_{1,T}$ and hence the initial weighting matrix $\mathbf{W}_{0,T}$.
- Note that $\widehat{\boldsymbol{\Sigma}}_T$ is also determined by $\mathbf{m}(\mathbf{z}_t;\boldsymbol{\theta})$. Hence, the correlation between $\bar{\mathbf{m}}_T$ and $\widehat{\boldsymbol{\Sigma}}_T$ results in finite-sample bias in $\hat{\boldsymbol{\theta}}_{2,T}$.
- Computing a consistent estimator of $\boldsymbol{\Sigma}_o$ may not be straightforward, especially when the data are serially correlated and heterogeneously distributed.

Iterative GMM Estimator

1. At the j th iteration, compute the j th iterative GMM estimator using the weighting matrix $\mathbf{W}_T(\hat{\boldsymbol{\theta}}_{j-1,T})$:

$$\hat{\boldsymbol{\theta}}_{j,T} := \arg\min_{\boldsymbol{\theta}\in\Theta} \left[\bar{\mathbf{m}}_T(\boldsymbol{\theta})\right]' \mathbf{W}_T(\hat{\boldsymbol{\theta}}_{j-1,T}) \left[\bar{\mathbf{m}}_T(\boldsymbol{\theta})\right].$$

   At the first iteration, we can set the initial weighting matrix $\mathbf{W}_{0,T}$, as in the two-step estimator.

2. Use $\hat{\boldsymbol{\theta}}_{j,T}$ to construct the optimal weighting matrix $\mathbf{W}_T(\hat{\boldsymbol{\theta}}_{j,T})$ and set $j = j + 1$.

3. Iterate the procedure above till one of the following convergence criteria is satisfied: For some pre-specified ε,

$$\|Q_T(\hat{\boldsymbol{\theta}}_{j,T}) - Q_T(\hat{\boldsymbol{\theta}}_{j-1,T})\|, \leq \varepsilon \quad \textit{or}$$
$$\|\hat{\boldsymbol{\theta}}_{j,T} - \hat{\boldsymbol{\theta}}_{j-1,T}\| \leq \varepsilon.$$

Continuous Updating GMM Estimator

Hansen, Heaton and Yaron (1996, *JBES*): Continuous updating (CU) estimator is based on one-time optimization:

$$\hat{\boldsymbol{\theta}}_T := \arg\min_{\boldsymbol{\theta}\in\Theta} \left[\bar{\mathbf{m}}_T(\boldsymbol{\theta})\right]' \mathbf{W}_T(\boldsymbol{\theta}) \left[\bar{\mathbf{m}}_T(\boldsymbol{\theta})\right].$$

The objective function clearly changes when $\mathbf{W}_T$ depends explicitly on $\boldsymbol{\theta}$. This does not affect the limiting distribution of the resulting estimator, however; see Pakes and Pollard (1989, *Econometrica*).

Remark: The CU estimator is invariant when the moment conditions are re-scaled, even when the scale factor is parameter dependent; the two-stage or iterative GMM estimator is sensitive to such transformation, however.

Independently Weighted Optimal Estimator

Altonji and Segal (1996, *JBES*): The independently weighted optimal estimator avoids possible correlation between $\bar{\mathbf{m}}_T$ and $\widehat{\mathbf{\Sigma}}_T$ by splitting the sample and computing $\bar{\mathbf{m}}_T$ and $\widehat{\mathbf{\Sigma}}_T$ with different sub-samples.

- Split the sample into ℓ groups, and let $\bar{\mathbf{m}}_{T_j}(\boldsymbol{\theta})$ be the sample average of $\mathbf{m}(\mathbf{z}_t, \boldsymbol{\theta})$ for t in the j th group with T_j observations.
- Also let $\widehat{\mathbf{\Sigma}}_{T_j^{\perp}}^{-1}$ be the optimal weighting matrix based on the observations not in the j th group.
- The resulting GMM estimator is obtained by solving:

$$\hat{\boldsymbol{\theta}}_T := \arg\min_{\boldsymbol{\theta}\in\Theta} \sum_{j=1}^{\ell} \left[\bar{\mathbf{m}}_{T_j}(\boldsymbol{\theta})\right]' \widehat{\mathbf{\Sigma}}_{T_j^{\perp}}^{-1} \left[\bar{\mathbf{m}}_{T_j}(\boldsymbol{\theta})\right].$$

  A common choice of ℓ is 2.

Example: Regression with Symmetric Error

Given the specification $y_t = \mathbf{x}_t'\boldsymbol{\beta} + e_t$, let

$$\mathbf{m}(y_t, \mathbf{x}_t; \boldsymbol{\beta}) = \begin{bmatrix} \mathbf{x}_t(y_t - \mathbf{x}_t'\boldsymbol{\beta}) \\ \mathbf{x}_t(y_t - \mathbf{x}_t'\boldsymbol{\beta})^3 \end{bmatrix}.$$

The moment condition $\mathbb{E}[\mathbf{m}(y_t, \mathbf{x}_t; \boldsymbol{\beta}_o)] = \mathbf{0}$ suggests estimating $\boldsymbol{\beta}_o$ while taking into account symmetry of the error term. The gradient vector of $\mathbf{m}$ is:

$$\nabla\mathbf{m}(y_t, \mathbf{x}_t; \boldsymbol{\beta}) = \begin{bmatrix} -\mathbf{x}_t\mathbf{x}_t' \\ -3\mathbf{x}_t\mathbf{x}_t'(y_t - \mathbf{x}_t'\boldsymbol{\beta})^2 \end{bmatrix}.$$

If the data are independent over t,

$$\boldsymbol{\Sigma}_o = \mathbb{E}\begin{bmatrix} \epsilon_t^2\mathbf{x}_t\mathbf{x}_t' & \epsilon_t^4\mathbf{x}_t\mathbf{x}_t' \\ \epsilon_t^4\mathbf{x}_t\mathbf{x}_t' & \epsilon_t^6\mathbf{x}_t\mathbf{x}_t' \end{bmatrix}.$$

Let $\hat{e}_t = y_t - \mathbf{x}_t' \hat{\boldsymbol{\beta}}_{1,T}$, where $\hat{\boldsymbol{\beta}}_{1,t}$ is a first-step GMM estimator based on a preliminary weighting matrix. Then, $\boldsymbol{\Sigma}_o$ may be estimated by the sample counterpart:

$$\widehat{\boldsymbol{\Sigma}}_T(\hat{\boldsymbol{\beta}}_{1,T}) = \frac{1}{T} \sum_{t=1}^{T} \begin{bmatrix} \hat{e}_t^2 \mathbf{x}_t \mathbf{x}_t' & \hat{e}_t^4 \mathbf{x}_t \mathbf{x}_t' \\ \hat{e}_t^4 \mathbf{x}_t \mathbf{x}_t' & \hat{e}_t^6 \mathbf{x}_t \mathbf{x}_t' \end{bmatrix}.$$

Note that $\hat{\boldsymbol{\beta}}_{1,T}$ here may be the OLS estimator (in fact, we only need a consistent estimator for $\boldsymbol{\beta}_o$). The two-step GMM estimator is computed with $\big[\widehat{\boldsymbol{\Sigma}}_T(\hat{\boldsymbol{\beta}}_{1,T})\big]^{-1}$ the weighting matrix; that is,

$$\hat{\boldsymbol{\theta}}_{2,T} := \arg\min_{\boldsymbol{\theta} \in \Theta} \bar{\mathbf{m}}_T(\boldsymbol{\theta}) \big[\widehat{\boldsymbol{\Sigma}}_T(\hat{\boldsymbol{\beta}}_{1,T})\big]^{-1} \bar{\mathbf{m}}_T(\boldsymbol{\theta}).$$

Example: Generalized Instrumental Variables Estimator

For the specification $y_t = \mathbf{x}_t'\boldsymbol{\beta} + e_t$, consider the moment condition:

$$\mathbb{E}[\mathbf{m}_t(\boldsymbol{\beta}_o)] = \mathbb{E}[\mathbf{z}_t(y_t - \mathbf{x}_t'\boldsymbol{\beta}_o)] = \mathbf{0},$$

where $\mathbf{z}_t$ contains $q > k$ instrumental variables. The GMM estimator minimizes

$$\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{z}_t(y_t - \mathbf{x}_t'\boldsymbol{\beta})\right)'\mathbf{W}_T\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{z}_t(y_t - \mathbf{x}_t'\boldsymbol{\beta})\right),$$

and solves

$$\left(\sum_{t=1}^{T}\mathbf{x}_t\mathbf{z}_t'\right)\mathbf{W}_T\left(\sum_{t=1}^{T}\mathbf{z}_t(y_t - \mathbf{x}_t'\boldsymbol{\beta})\right) = (\mathbf{X}'\mathbf{Z})\mathbf{W}_T[\mathbf{Z}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})] = \mathbf{0}.$$

where $\mathbf{Z}$ $(T \times q)$ is the matrix of instrumental variables and $\mathbf{X}$ $(T \times k)$ is the matrix of regressors.

The GMM estimator here is also the generalized instrumental variables estimator:

$$\hat{\beta}(\mathbf{W}_T) = (\mathbf{X}'\mathbf{Z}\mathbf{W}_T\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\mathbf{W}_T\mathbf{Z}'\mathbf{y}.$$

When the data are independent over t and there is no condition heteroskedasticity,

$$\boldsymbol{\Sigma}_o = \text{var}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\mathbf{z}_t(y_t - \mathbf{x}_t'\boldsymbol{\beta}_o)\right) = \frac{\sigma_o^2}{T}\sum_{t=1}^{T}\mathbb{E}(\mathbf{z}_t\mathbf{z}_t').$$

Ignoring σ_o^2 in $\boldsymbol{\Sigma}_o$, $T^{-1}\sum_{t=1}^{T}\mathbb{E}(\mathbf{z}_t\mathbf{z}_t')$ can be estimated by $\mathbf{Z}'\mathbf{Z}/T$. The two-step GMM estimator now is:

$$\hat{\boldsymbol{\beta}}_{2,T} = [\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y},$$

which is also known as the two-stage least squares (2SLS) estimator.

The 2SLS estimator can be expressed as

$$\hat{\boldsymbol{\theta}}_{2,T} = [\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}]^{-1}\widetilde{\mathbf{X}}'\mathbf{y},$$

where $\widetilde{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$ is the matrix of fitted values from the OLS regression of $\mathbf{X}$ on $\mathbf{Z}$. Hence, $\hat{\boldsymbol{\theta}}_{2,T}$ is also the OLS estimator of regressing y on $\widetilde{\mathbf{X}}$ (Theil, 1953).

When the errors are not conditionally homoskedastic, the 2SLS estimator remains consistent (why?), but it is no longer the most efficient. A two-step GMM estimator with a properly estimated weighting matrix $[\widehat{\boldsymbol{\Sigma}}_T(\hat{\boldsymbol{\beta}}_{1,T})]^{-1}$ is more efficient. For example,

$$\widehat{\boldsymbol{\Sigma}}_T(\hat{\boldsymbol{\beta}}_{1,T}) = \frac{1}{T}\sum_{t=1}^{T} \hat{e}_t^2 \mathbf{z}_t\mathbf{z}_t',$$

where $\hat{e}_t = y_t - \mathbf{x}_t'\hat{\boldsymbol{\beta}}_{1,T}$ are the residuals from the first-step GMM estimation.

Over-Identifying Restrictions Test

To test whether the model for $\mathbb{E}[\mathbf{m}(\mathbf{z}_t, \theta_0)] = \mathbf{0}$ is correctly specified, it is natural to check if $\bar{\mathbf{m}}_T(\hat{\boldsymbol{\theta}}_T)$ is sufficiently close to zero. For example, when the moment conditions are the Euler equations in different capital asset pricing models, this amounts to checking if the "pricing errors" are zero.

The over-identifying restrictions (OIR) test of Hansen (1982), also known as the $\mathcal{J}$ test, is based on the value of the GMM objective function:

$$\mathcal{J}_T := T\left[\bar{\mathbf{m}}_T(\hat{\boldsymbol{\theta}}_T(\mathbf{W}_T))\right]' \mathbf{W}_T \left[\bar{\mathbf{m}}_T(\hat{\boldsymbol{\theta}}_T(\mathbf{W}_T))\right].$$

As the test statistic involves the GMM estimator obtained from the same objective function, the limiting distribution is $\chi^2(q-k)$.

To derive its limiting distribution, note that

$$\begin{aligned}
&\mathbf{W}_T^{1/2}\sqrt{T}\bar{\mathbf{m}}_T\big(\hat{\boldsymbol{\theta}}_T(\mathbf{W}_T)\big) \\
&\qquad = \mathbf{W}_T^{1/2}\sqrt{T}\bar{\mathbf{m}}_T(\boldsymbol{\theta}_o) + \mathbf{W}_T^{1/2}\mathbf{G}_T(\boldsymbol{\theta}_T^{\dagger})\sqrt{T}\big(\hat{\boldsymbol{\theta}}_T(\mathbf{W}_T) - \boldsymbol{\theta}_o\big).
\end{aligned}$$

As $\sqrt{T}\big(\hat{\boldsymbol{\theta}}_T(\mathbf{W}_T) - \boldsymbol{\theta}_o\big) = -\big(\mathbf{G}_o'\mathbf{W}_o\mathbf{G}_o\big)^{-1}\mathbf{G}_o'\mathbf{W}_o\sqrt{T}\bar{\mathbf{m}}_T(\boldsymbol{\theta}_o) + o_{\mathbf{P}}(1)$,

$$\mathbf{W}_T^{1/2}\sqrt{T}\bar{\mathbf{m}}_T\big(\hat{\boldsymbol{\theta}}_T(\mathbf{W}_T)\big) = \mathbf{P}_o\mathbf{W}_o^{1/2}\sqrt{T}\bar{\mathbf{m}}_T(\boldsymbol{\theta}_o) + o_{\mathbf{P}}(1)$$

where $\mathbf{P}_o = \mathbf{I} - \mathbf{W}_o^{1/2}\mathbf{G}_o\big(\mathbf{G}_o'\mathbf{W}_o\mathbf{G}_o\big)^{-1}\mathbf{G}_o'\mathbf{W}_o^{1/2}$ which is symmetric and idempotent with rank $q-k$ (why?), and

$$\mathbf{W}_o^{1/2}\sqrt{T}\bar{\mathbf{m}}_T(\boldsymbol{\theta}_o) \xrightarrow{D} \mathcal{N}(\mathbf{0},\, \mathbf{W}_o^{1/2}\boldsymbol{\Sigma}_o\mathbf{W}_o^{1/2}).$$

When $\mathbf{W}_o = \mathbf{\Sigma}_o^{-1}$, the limiting distribution simplifies:

$$\mathbf{W}_o^{1/2}\sqrt{T}\bar{\mathbf{m}}_T(\boldsymbol{\theta}_o) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \mathbf{I}_q).$$

Consequently,

$$\mathcal{J}_T = T\,\bar{\mathbf{m}}_T(\boldsymbol{\theta}_o)'\mathbf{W}_o^{1/2}\mathbf{P}_o\mathbf{P}_o\mathbf{W}_o^{1/2}\bar{\mathbf{m}}_T(\boldsymbol{\theta}_o) \xrightarrow{D} \chi^2(q-k).$$

Remarks:

- The weighting matrix in the $\mathcal{J}$ statistic and the weighting matrix for the GMM estimator in $\bar{\mathbf{m}}_T$ must be the same. And the weighting matrix $\mathbf{W}_T$ for the GMM estimator must converge to $\mathbf{W}_o = \mathbf{\Sigma}_o^{-1}$. That is, $\mathcal{J}$ test requires the optimal GMM estimator.
- Lee and Kuan (2010) propose an OIR test that does not requires the weighting matrix to converge to $\mathbf{\Sigma}_o$ and hence avoids the optimal GMM estimation.

Hausman Test

Given two estimators $\hat{\boldsymbol{\theta}}_T$ and $\check{\boldsymbol{\theta}}_T$ of the parameter $\boldsymbol{\theta}_o$, suppose that both are consistent under the null hypothesis but only one, say $\check{\boldsymbol{\theta}}_T$, is consistent under the alternative. The Hausman test suggests testing the null hypothesis by comparing these two estimators.

The Hausman test is particularly useful for testing model specification, which may not be expressed as parameter restrictions. For example, consider testing the null hypothesis that regressors are exogenous against the alternative of endogenous regressors. Under "classical" conditions, the OLS estimator $\hat{\boldsymbol{\theta}}_T$ and the 2SLS estimator $\check{\boldsymbol{\theta}}_T$ are consistent under the null, but only the 2SLS estimator is consistent under the alternative. Thus, we can test for endogeneity by checking if $\hat{\boldsymbol{\theta}}_T - \check{\boldsymbol{\theta}}_T$ is sufficiently large.

The Hausman test reads:

$$\mathcal{H}_T = T(\hat{\boldsymbol{\theta}}_T - \check{\boldsymbol{\theta}}_T)' \widehat{\mathbf{V}}_T^{-1} (\hat{\boldsymbol{\theta}}_T - \check{\boldsymbol{\theta}}_T) \xrightarrow{D} \chi^2(k),$$

where $\widehat{\mathbf{V}}_T$ is a consistent estimator for the asymptotic covariance matrix of $\sqrt{T}(\hat{\boldsymbol{\theta}}_T - \check{\boldsymbol{\theta}}_T)$:

$$\mathbf{V}(\hat{\boldsymbol{\theta}}_T - \check{\boldsymbol{\theta}}_T) = \mathbf{V}(\hat{\boldsymbol{\theta}}_T) + \mathbf{V}(\check{\boldsymbol{\theta}}_T) - 2\,\mathrm{cov}(\hat{\boldsymbol{\theta}}_T, \check{\boldsymbol{\theta}}_T).$$

This asymptotic covariance matrix is simplified when $\hat{\boldsymbol{\theta}}_T$ is also asymptotically efficient under the null. In particular, we can show

$$\mathbf{V}_{1,2} := \mathrm{cov}(\hat{\boldsymbol{\theta}}_T, \check{\boldsymbol{\theta}}_T) = \mathbf{V}(\hat{\boldsymbol{\theta}}_T),$$

so that $\mathbf{V}(\hat{\boldsymbol{\theta}}_T - \check{\boldsymbol{\theta}}_T)$ depends only on the respective asymptotic covariance matrices of these two estimators:

$$\mathbf{V}(\hat{\boldsymbol{\theta}}_T - \check{\boldsymbol{\theta}}_T) = \mathbf{V}(\check{\boldsymbol{\theta}}_T) - \mathbf{V}(\hat{\boldsymbol{\theta}}_T).$$

To see this, consider the following estimator which is a linear combination of $\hat{\boldsymbol{\theta}}_T$ and $\check{\boldsymbol{\theta}}_T$:

$$\hat{\boldsymbol{\theta}}_T^{\dagger} = \hat{\boldsymbol{\theta}}_T + \big[\mathbf{V}(\hat{\boldsymbol{\theta}}_T) - \mathbf{V}_{1,2}\big]\mathbf{V}(\hat{\boldsymbol{\theta}}_T - \check{\boldsymbol{\theta}}_T)^{-1}(\check{\boldsymbol{\theta}}_T - \hat{\boldsymbol{\theta}}_T).$$

It is easy to verify that

$$\mathbf{V}(\hat{\boldsymbol{\theta}}_T^{\dagger}) = \mathbf{V}(\hat{\boldsymbol{\theta}}_T) - \big[\mathbf{V}(\hat{\boldsymbol{\theta}}_T) - \mathbf{V}_{1,2}\big]\mathbf{V}(\hat{\boldsymbol{\theta}}_T - \check{\boldsymbol{\theta}}_T)^{-1}\big[\mathbf{V}(\hat{\boldsymbol{\theta}}_T) - \mathbf{V}_{1,2}\big]'.$$

If $\mathbf{V}(\hat{\boldsymbol{\theta}}_T)$ is not the same as $\mathbf{V}_{1,2}$, the second term on the RHS is p.d. It follows that $\hat{\boldsymbol{\theta}}_T^{\dagger}$ is asymptotically more efficient than $\hat{\boldsymbol{\theta}}_T$. But this contradicts the assumption that $\hat{\boldsymbol{\theta}}_T$ is asymptotically more efficient. This suggests that $\mathbf{V}(\hat{\boldsymbol{\theta}}_T) = \mathbf{V}_{1,2}$.

Wald Test

Hypothesis: $\mathbf{R}(\boldsymbol{\theta}) = \mathbf{0}$, where $\mathbf{R} : \mathbb{R}^K \mapsto \mathbb{R}^r$. A mean-value expansion yields

$$\mathbf{R}(\hat{\boldsymbol{\theta}}_T) = \mathbf{R}(\boldsymbol{\theta}_0) + \nabla \mathbf{R}(\boldsymbol{\theta}^\dagger)(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0),$$

where $\sqrt{T}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}_0)$. Therefore,

$$\sqrt{T}\big[\mathbf{R}(\hat{\boldsymbol{\theta}}_T) - \mathbf{R}(\boldsymbol{\theta}_0)\big] \xrightarrow{D} \mathcal{N}\big(\mathbf{0},\, [\nabla \mathbf{R}(\boldsymbol{\theta}_o)]\boldsymbol{\Omega}_0[\nabla \mathbf{R}(\boldsymbol{\theta}_o)]'\big).$$

It follows that, under the null hypothesis,

$$\mathcal{W}_T := T\mathbf{R}(\hat{\boldsymbol{\theta}}_T)' \Big(\big[\nabla \mathbf{R}(\hat{\boldsymbol{\theta}}_T)\big] \widehat{\boldsymbol{\Omega}}_T \big[\nabla \mathbf{R}(\hat{\boldsymbol{\theta}}_T)\big]' \Big)^{-1} \mathbf{R}(\hat{\boldsymbol{\theta}}_T) \xrightarrow{D} \chi^2(r).$$

where $\widehat{\boldsymbol{\Omega}}_T$ is a consistent estimator of $\boldsymbol{\Omega}_o$. For the linear hypothesis $\mathbf{R}\boldsymbol{\theta}_0 = \mathbf{r}$ with $\mathbf{R}$ a $r \times k$ matrix,

$$\mathcal{W}_T = T(\mathbf{R}\hat{\boldsymbol{\theta}}_T - \mathbf{r})'(\mathbf{R}\widehat{\boldsymbol{\Omega}}_T\mathbf{R}')^{-1}(\mathbf{R}\hat{\boldsymbol{\theta}}_T - \mathbf{r}) \xrightarrow{D} \chi^2(r).$$

Estimation of Conditional Moment Restrictions

Consider the conditional moment restrictions:

$$\mathbb{E}[\mathbf{h}(\boldsymbol{\eta}_t;\boldsymbol{\theta}_o)|\mathcal{F}^t] = 0,$$

where $\mathbf{h}$ is $r \times 1$ and $\mathcal{F}^t$ is the information set up to time t.

Let $\mathbf{w}_t$ be a set of variables from $\mathcal{F}^t$ and may contain some variables in $\boldsymbol{\eta}_t$. The conditional moment restrictions above imply

$$\mathbb{E}[\mathbf{D}(\mathbf{w}_t)'\mathbf{h}(\boldsymbol{\eta}_t;\boldsymbol{\theta}_o)] = \mathbf{0},$$

where $\mathbf{D}(\mathbf{w}_t)$ is a $(r \times n)$ matrix of instruments. GMM estimation of $\boldsymbol{\theta}_o$ is based on the sample moments:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbf{m}_t(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T}\mathbf{D}(\mathbf{w}_t)'\mathbf{h}(\boldsymbol{\eta}_t;\boldsymbol{\theta}).$$

The optimal GMM estimator has asymptotic covariance matrix $\left(\mathbf{G}_o'\mathbf{\Sigma}_o^{-1}\mathbf{G}_o\right)^{-1}$, with

$$\mathbf{\Sigma}_o = \mathbb{E}\big[\mathbf{D}(\mathbf{w}_t)'\mathbf{h}(\boldsymbol{\eta}_t;\boldsymbol{\theta}_o)\mathbf{h}(\boldsymbol{\eta}_t;\boldsymbol{\theta}_o)'\mathbf{D}(\mathbf{w}_t)\big], \qquad (n \times n)$$

$$\mathbf{G}_o = \mathbb{E}\big[\mathbf{D}(\mathbf{w}_t)'\nabla\mathbf{h}(\boldsymbol{\eta}_t;\boldsymbol{\theta}_o)\big]. \qquad (n \times k)$$

For a smaller set of sample moments obtained by taking a linear transformation:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbf{C}'\mathbf{D}(\mathbf{w}_t)'\mathbf{h}(\boldsymbol{\eta}_t;\boldsymbol{\theta}),$$

where $\mathbf{C}$ is $n \times p$ with $p < n$, the asymptotic covariance matrix of the resulting optimal GMM estimator is $\left[\mathbf{G}_o'\mathbf{C}(\mathbf{C}'\mathbf{\Sigma}_o\mathbf{C})^{-1}\mathbf{C}'\mathbf{G}_o\right]^{-1}$. It is readily seen that

$$\begin{aligned}\mathbf{G}_o'\mathbf{\Sigma}_o^{-1}\mathbf{G}_o &- \mathbf{G}_o'\mathbf{C}(\mathbf{C}'\mathbf{\Sigma}_o\mathbf{C})^{-1}\mathbf{C}'\mathbf{G}_o \\ &= \mathbf{G}_o'\mathbf{\Sigma}_o^{-1/2}\big[\mathbf{I}_n - \mathbf{\Sigma}_o^{1/2}\mathbf{C}(\mathbf{C}'\mathbf{\Sigma}_o^{1/2}\mathbf{\Sigma}_o^{1/2}\mathbf{C})^{-1}\mathbf{C}'\mathbf{\Sigma}_o^{1/2}\big]\mathbf{\Sigma}_o^{-1/2}\mathbf{G}_o,\end{aligned}$$

which is p.s.d. matrix.

Number of moment conditions: The optimal GMM estimator based on a larger set of moment conditions is asymptotically more efficient, yet it may have a larger bias in finite samples.

Optimal instruments:

$$\mathbf{D}^*(\mathbf{w}_t;\boldsymbol{\theta}) = \big[\mathrm{var}(\mathbf{h}(\boldsymbol{\eta}_t;\boldsymbol{\theta})|\mathcal{F}^t)\big]^{-1}\,\mathbb{E}\big[\nabla\mathbf{h}(\boldsymbol{\eta}_t;\boldsymbol{\theta})|\mathcal{F}^t\big] =: \mathbf{V}_t^{-1}\mathbf{J}_t, \qquad (r\times k)$$

which contains k instruments. In this case,

$$\begin{aligned}
\boldsymbol{\Sigma}_o &= \mathbb{E}\big(\mathbf{J}_t'\mathbf{V}_t^{-1}\mathbf{V}_t\mathbf{V}_t^{-1}\mathbf{J}_t\big) = \mathbb{E}\big(\mathbf{J}_t'\mathbf{V}_t^{-1}\mathbf{J}_t\big),\\
\mathbf{G}_o &= \mathbb{E}\big[\mathbf{J}_t'\mathbf{V}_t^{-1}\nabla\mathbf{h}(\boldsymbol{\eta}_t;\boldsymbol{\theta}_o)\big] = \mathbb{E}\big(\mathbf{J}_t'\mathbf{V}_t^{-1}\mathbf{J}_t\big).
\end{aligned}$$

The optimal GMM estimator thus has the asymptotic covariance matrix

$$\big(\mathbf{G}_o'\boldsymbol{\Sigma}_o^{-1}\mathbf{G}_o\big)^{-1} = \mathbb{E}\big(\mathbf{J}_t'\mathbf{V}_t^{-1}\mathbf{J}_t\big)^{-1}.$$

As $\mathbf{J}_t$ and $\mathbf{V}_t$ are conditional expectations and depend on unknown parameters, it is not easy to estimate the optimal instruments in practice.