Econometric Methods (Econometrics I)

Lecture 8: Generalized Method of Moments

Prof. Dr. Kai Carstensen

Kiel University

Winter Term 2023/2024

Outline of this lecture

1. The GMM estimator
2. Asymptotic Properties
3. Efficient GMM estimation
4. Inference
5. Estimation under Orthogonality Restrictions

Reference: Wooldridge, Chapter 14.

Example: Estimation of Euler Equations

A standard result in macro ($\rightarrow$ intertemporal consumption decision) and finance ($\rightarrow$ CCAPM) is the Euler equation for consumption,

$$\mathrm{E}\left[\beta \frac{u'(C_{t+1})}{u'(C_t)}(1+\rho_{t+1})\middle|\mathcal{I}_t\right] = 1,$$

where ρ_t is the real rate of return, C_t is consumption, $\mathcal{I}_t$ is the information set available in period t, β is a discount factor and total utility in period t is given by

$$U_t = u(C_t) + \beta u(C_{t+1}) + \beta^2 u(C_{t+2}) + \ldots.$$

A typical specification of instantaneous utility is

$$u(C_t) = C_t^{1-\gamma}/(1-\gamma),$$

which yields the first derivative $u'(C_t) = C_t^{-\gamma}$. Substituting yields the Euler equation

$$\mathrm{E}\left[\beta \left(\frac{C_t}{C_{t+1}}\right)^{\gamma}(1+\rho_{t+1})\middle|\mathcal{I}_t\right] = 1.$$

How can we estimate the preference parameters β and γ?

Suppose we have a sample of N homogenous households for which we observe (non-durable) consumption levels $C_{i,t}$ and $C_{i,t+1}$, the real rate of their portfolios $\rho_{i,t+1}$, and some variables $\mathbf{z}_i$ of their information set $\mathcal{I}_t$.

Under rational expectations, the conditional expectation implies that the forecast error, i.e., the deviation from 1,

$$r(\mathbf{w}_i, \boldsymbol{\theta}_o) \equiv \beta \left(\frac{C_t}{C_{t+1}} \right)^{\gamma} (1 + \rho_{t+1}) - 1,$$

where $\mathbf{w}_i = (C_t, C_{t+1}, \rho_{t+1})'$ and $\boldsymbol{\theta}_o = (\beta_o, \gamma_o)'$, is unpredictable by, and thus uncorrelated with, the information variables $\mathbf{z}_i$.

Hence, the Euler equation together with rational expectations imply

$$\mathrm{E}\left[\mathbf{z}_i' r(\mathbf{w}_i, \boldsymbol{\theta}_o)\right] = \mathrm{E}\left[\mathbf{z}_i' \left\{ \beta \left(\frac{C_{i,t}}{C_{i,t+1}} \right)^{\gamma} (1 + \rho_{i,t+1}) - 1 \right\}\right] = 0.$$

In general, the generalized method of moments (GMM) estimator uses moment conditions of the type

$$\mathrm{E}\left[\mathbf{z}_i' r(\mathbf{w}_i, \boldsymbol{\theta}_o)\right] = 0$$

or, even more general, using the notation of the textbook,

$$\mathrm{E}\left[\mathbf{g}(\mathbf{w}_i, \boldsymbol{\theta}_o)\right] = 0$$

to estimate the unknown parameters.

Here is some useful literature for this example:

- Hall, R.E. (1978) "The Stochastic Implications of the Life Cycle–Permanent Income Hypothesis: Theory and Evidence," Journal of Political Economy, 86, 971-87. (Seminal paper on the PIH)
- Hansen, L. P. and K.J. Singleton (1982) "Generalized instrumental variables estimation of nonlinear rational expectation models," Econometrica, 50, 1269-86. (One of the seminal papers on GMM for which Lars Peter Hansen was awarded the Nobel prize 2013.)
- Hansen, L.P. and K.J. Singleton (1983) "Stochastic Consumption, Risk Aversion and the Temporal Behavior of Asset Returns," Journal of Political Economy, 91, 249-65. (One of the seminal papers on GMM for which Lars Peter Hansen was awarded the Nobel prize 2013.)
- Attanasio, O.P. (1999) "Consumption," in: Handbook of Macroeconomics, Vol. 1, 741-811. (Good overview over how to estimate Euler equations and which problems arise when using cross sectional data.)

1. The GMM estimator

OLS

To estimate the linear model with P regressors

$$y = \mathbf{x}\boldsymbol{\beta} + u$$

by OLS we assume $\mathrm{E}(u|\mathbf{x}) = 0$ or at least

$$\mathrm{E}(\mathbf{x}'u) = \mathbf{0}.$$

Using the sample equivalent, we generate P normal equations

$$\sum_{i=1}^{N} \mathbf{x}_i'\hat{u}_i = \sum_{i=1}^{N} \mathbf{x}_i'(y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{N} \mathbf{x}_i'y_i\right) - \left(\sum_{i=1}^{N} \mathbf{x}_i'\mathbf{x}_i\right)\hat{\boldsymbol{\beta}} = \mathbf{0},$$

which are solved analytically for the OLS estimator.

Note: here we have as many normal equations as we have unknown parameters, so $\hat{\boldsymbol{\beta}}$ is chosen such that the normal equations are exactly zero in sample.

2SLS

The 2SLS estimator uses the $L \geq P$ orthogonality conditions

$$\mathrm{E}(\mathbf{z}'u) = \mathbf{0}.$$

In the case of overidentification, $L > P$, the L normal equations

$$\sum_{i=1}^{N} \mathbf{z}_i' \hat{u}_i = \sum_{i=1}^{N} \mathbf{z}_i'(y_i - \mathbf{x}_i \hat{\boldsymbol{\beta}})$$

cannot all be exactly zero. Hence, the 2SLS estimator is chosen such that the weighted squared deviations from zero,

$$\left(\sum_{i=1}^{N} \mathbf{z}_i' \hat{u}_i \right)' \mathbf{W} \left(\sum_{i=1}^{N} \mathbf{z}_i' \hat{u}_i \right),$$

are minimized. Recall that the 2SLS uses $\mathbf{W} = \sum_{i=1}^{N} \mathbf{z}_i' \mathbf{z}_i$.

Due to linearity a closed form solution is feasible.

GMM

In general, let $\{\mathbf{w}_i \in \mathbb{R}^M \colon i = 1, 2, \ldots\}$ denote a set of independent, identically distributed random vectors, where some feature of the distribution of $\mathbf{w}_i$ is indexed by the $P \times 1$ parameter vector $\boldsymbol{\theta}$.

Now assume that for some function $\mathbf{g}(\mathbf{w}_i, \boldsymbol{\theta}) \in \mathbb{R}^L$, the parameter $\boldsymbol{\theta}_o \in \boldsymbol{\Theta} \subset \mathbb{R}^P$ satisfies

$$\mathrm{E}\left[\mathbf{g}\left(\mathbf{w}_i, \boldsymbol{\theta}_o\right)\right] = \mathbf{0}.$$

If $L = P$, then the analogy principle suggests estimating $\boldsymbol{\theta}_o$ by setting to zero the sample counterpart,

$$N^{-1} \sum_{i=1}^{N} \mathbf{g}(\mathbf{w}_i, \hat{\boldsymbol{\theta}}) \stackrel{!}{=} \mathbf{0}.$$

If $L > P$, we can choose $\hat{\boldsymbol{\theta}}$ to make the sample average close to zero in an appropriate metric.

A **generalized method of moments (GMM)** estimator, $\hat{\boldsymbol{\theta}}$, minimizes a quadratic form in $\sum_{i=1}^{N} \mathbf{g}(\mathbf{w}_i, \boldsymbol{\theta})$:

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left[\sum_{i=1}^{N} \mathbf{g}\left(\mathbf{w}_i, \boldsymbol{\theta}\right) \right]' \hat{\boldsymbol{\Xi}} \left[\sum_{i=1}^{N} \mathbf{g}\left(\mathbf{w}_i, \boldsymbol{\theta}\right) \right],$$

where $\hat{\boldsymbol{\Xi}}$ is an $L \times L$ symmetric, positive semidefinite weighting matrix.

2. Asymptotic Properties

Consistency

Recall that the M-estimator $\hat{\boldsymbol{\theta}}$ solves the minimization problem

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} N^{-1} \sum_{i=1}^{N} q(\mathbf{w}_i, \boldsymbol{\theta}).$$

The aim is to estimate the population parameter $\boldsymbol{\theta}_o$ that minimizes the population function

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathrm{E}\left[q(\mathbf{w}, \boldsymbol{\theta})\right].$$

Consistency requires that the sample moment converges uniformly in probability towards the population moment. Then the estimator $\hat{\boldsymbol{\theta}}$ that minimizes the sample moment converges in probability towards the population parameter $\boldsymbol{\theta}_o$ that minimizes the population moment.

The GMM estimator $\hat{\boldsymbol{\theta}}$ solves the minimization problem

$$Q_N(\boldsymbol{\theta}) \equiv \left[N^{-1}\sum_{i=1}^{N}\mathbf{g}\left(\mathbf{w}_i,\boldsymbol{\theta}\right)\right]' \hat{\boldsymbol{\Xi}} \left[N^{-1}\sum_{i=1}^{N}\mathbf{g}\left(\mathbf{w}_i,\boldsymbol{\theta}\right)\right].$$

- Under standard moment conditions, $N^{-1}\sum_{i=1}^{N}\mathbf{g}(\mathbf{w}_i,\boldsymbol{\theta})$ satisfies the uniform law of large numbers.
- In addition, assume that $\hat{\boldsymbol{\Xi}} \xrightarrow{p} \boldsymbol{\Xi}_o$, where $\boldsymbol{\Xi}_o$ is an $L \times L$ positive definite matrix.

Thus the random function $Q_N(\boldsymbol{\theta})$ converges uniformly in probability to the population function

$$Q(\boldsymbol{\theta}) \equiv \mathrm{E}\left[\mathbf{g}\left(\mathbf{w}_i,\boldsymbol{\theta}\right)\right]' \boldsymbol{\Xi}_o \,\mathrm{E}\left[\mathbf{g}\left(\mathbf{w}_i,\boldsymbol{\theta}\right)\right].$$

Because $\boldsymbol{\Xi}_o$ is positive definite, $\boldsymbol{\theta}_o$ uniquely minimizes $Q(\boldsymbol{\theta})$.

Now consistency of GMM follows from the same reasoning used for the M-estimator:

- The GMM estimator $\hat{\boldsymbol{\theta}}$ minimizes the sample function $Q_N(\boldsymbol{\theta})$.
- The true parameter $\boldsymbol{\theta}_o$ minimizes the population function $Q(\boldsymbol{\theta})$.
- The sample function $Q_N(\boldsymbol{\theta})$ converges uniformly in probability towards the population function $Q(\boldsymbol{\theta})$.

Thus, the GMM estimator $\hat{\boldsymbol{\theta}}$ converges in probability towards the true parameter $\boldsymbol{\theta}_o$,

$$\hat{\boldsymbol{\theta}} \xrightarrow{\text{p}} \boldsymbol{\theta}_o.$$

Asymptotic normality

Start by noting that by assumption

$$\mathrm{E}\left[\mathbf{g}(\mathbf{w}_i, \boldsymbol{\theta}_o)\right] = \mathbf{0}.$$

Hence, for an iid random sample of size N, the CLT implies

$$N^{-1/2} \sum_{i=1}^{N} \mathbf{g}(\mathbf{w}_i, \boldsymbol{\theta}_o) \xrightarrow{\mathrm{d}} \mathrm{Normal}(\mathbf{0}, \boldsymbol{\Lambda}_o),$$

where

$$\boldsymbol{\Lambda}_o \equiv \mathrm{E}\left[\mathbf{g}(\mathbf{w}_i, \boldsymbol{\theta}_o)\mathbf{g}(\mathbf{w}_i, \boldsymbol{\theta}_o)'\right] = \mathrm{Var}\left[\mathbf{g}(\mathbf{w}_i, \boldsymbol{\theta}_o)\right].$$

For later use consider the first-order conditions for $\hat{\boldsymbol{\theta}}$,

$$\left[\sum_{i=1}^{N} \nabla_{\boldsymbol{\theta}} \mathbf{g}(\mathbf{w}_i, \hat{\boldsymbol{\theta}})\right]' \hat{\boldsymbol{\Xi}} \left[\sum_{i=1}^{N} \mathbf{g}(\mathbf{w}_i, \hat{\boldsymbol{\theta}})\right] \equiv \mathbf{0}.$$

Recall: the first derivative of the scalar quadratic function $\mathbf{f}(\boldsymbol{\theta})' \boldsymbol{\Sigma} \mathbf{f}(\boldsymbol{\theta})$, where $\mathbf{f}(\cdot)$ is a $L \times 1$ vector function of $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$ is a symmetric matrix of constants, is

$$\frac{\partial \, \mathbf{f}(\boldsymbol{\theta})' \boldsymbol{\Sigma} \, \mathbf{f}(\boldsymbol{\theta})}{\partial \, \boldsymbol{\theta}} = \nabla_{\boldsymbol{\theta}}' \left[\mathbf{f}(\boldsymbol{\theta})' \boldsymbol{\Sigma} \, \mathbf{f}(\boldsymbol{\theta})\right] = 2 \left[\nabla_{\boldsymbol{\theta}} \mathbf{f}(\boldsymbol{\theta})\right]' \boldsymbol{\Sigma} \, \mathbf{f}(\boldsymbol{\theta}).$$

In the simple case $\mathbf{f}(\boldsymbol{\theta}) = \boldsymbol{\theta}$ this yields the familiar result

$$\frac{\partial \, \boldsymbol{\theta}' \boldsymbol{\Sigma} \, \boldsymbol{\theta}}{\partial \, \boldsymbol{\theta}} = \nabla_{\boldsymbol{\theta}}' \left[\boldsymbol{\theta}' \boldsymbol{\Sigma} \, \boldsymbol{\theta}\right] = 2 \, \boldsymbol{\Sigma} \, \boldsymbol{\theta}.$$

Define the $L \times P$ matrix

$$\mathbf{G}_o \equiv \mathrm{E}\left[\nabla_\theta \mathbf{g}\left(\mathbf{w}_i, \boldsymbol{\theta}_o\right)\right]$$

which we assume to have full rank P.

Then, by a weak law of large numbers,

$$N^{-1} \sum_{i=1}^{N} \nabla_{\boldsymbol{\theta}} \mathbf{g}(\mathbf{w}_i, \boldsymbol{\theta}_o) \xrightarrow{\mathrm{p}} \mathbf{G}_o.$$

Now apply the familiar mean value expansion (and use the short form $\mathbf{g}_i(\boldsymbol{\theta}) = \mathbf{g}(\mathbf{w}_i, \boldsymbol{\theta})$)

$$\sum_{i=1}^{N} \mathbf{g}_i(\hat{\boldsymbol{\theta}}) = \sum_{i=1}^{N} \mathbf{g}_i(\boldsymbol{\theta}_o) + \left(\sum_{i=1}^{N} \ddot{\mathbf{G}}_i \right) (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o),$$

where $\ddot{\mathbf{G}}_i$ is the $L \times P$ matrix of first derivatives, $\nabla_{\boldsymbol{\theta}} \mathbf{g}_i(\boldsymbol{\theta})$, evaluated row-wise at (unknown) vectors that satisfy the mean value theorem.

Asymptotically, as the unknown vectors are trapped between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_o$,

$$N^{-1} \sum_{i=1}^{N} \ddot{\mathbf{G}}_i \xrightarrow{\text{p}} \mathbf{G}_o.$$

However, $\sum_{i=1}^{N} \mathbf{g}_i(\hat{\boldsymbol{\theta}})$ is not generally zero because there might be more moment conditions L than parameters P. But we need to set the lhs of the mean value expansion to zero as in previous proofs. What can we do?

Multiply the mean value expansion from the left by $\left[\sum_{i=1}^{N}\nabla_{\boldsymbol{\theta}}\mathbf{g}_i(\hat{\boldsymbol{\theta}})\right]'\hat{\boldsymbol{\Xi}}$, which yields

$$\left[\sum_{i=1}^{N}\nabla_{\boldsymbol{\theta}}\mathbf{g}_i(\hat{\boldsymbol{\theta}})\right]'\hat{\boldsymbol{\Xi}}\left[\sum_{i=1}^{N}\mathbf{g}_i(\hat{\boldsymbol{\theta}})\right] = \left[\sum_{i=1}^{N}\nabla_{\boldsymbol{\theta}}\mathbf{g}_i(\hat{\boldsymbol{\theta}})\right]'\hat{\boldsymbol{\Xi}}\left[\sum_{i=1}^{N}\mathbf{g}_i(\boldsymbol{\theta}_o)\right] + \left[\sum_{i=1}^{N}\nabla_{\boldsymbol{\theta}}\mathbf{g}_i(\hat{\boldsymbol{\theta}})\right]'\hat{\boldsymbol{\Xi}}\left(\sum_{i=1}^{N}\ddot{\mathbf{G}}_i\right)(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_o).$$

Now the lhs equals the FOC and thus is identically zero. Hence, we obtain

$$\mathbf{0} = \left[\sum_{i=1}^{N}\nabla_{\boldsymbol{\theta}}\mathbf{g}_i(\hat{\boldsymbol{\theta}})\right]'\hat{\boldsymbol{\Xi}}\left[\sum_{i=1}^{N}\mathbf{g}_i(\boldsymbol{\theta}_o)\right] + \left[\sum_{i=1}^{N}\nabla_{\boldsymbol{\theta}}\mathbf{g}_i(\hat{\boldsymbol{\theta}})\right]'\hat{\boldsymbol{\Xi}}\left(\sum_{i=1}^{N}\ddot{\mathbf{G}}_i\right)(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_o).$$

Dividing the equation by $N^{3/2}$ yields

$$\mathbf{0} = \left[\frac{1}{N}\sum_{i=1}^{N}\nabla_{\boldsymbol{\theta}}\mathbf{g}_i(\hat{\boldsymbol{\theta}})\right]'\hat{\boldsymbol{\Xi}}\left[\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{g}_i(\boldsymbol{\theta}_o)\right]+\left[\frac{1}{N}\sum_{i=1}^{N}\nabla_{\boldsymbol{\theta}}\mathbf{g}_i(\hat{\boldsymbol{\theta}})\right]'\hat{\boldsymbol{\Xi}}\left(\frac{1}{N}\sum_{i=1}^{N}\ddot{\mathbf{G}}_i\right)\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_o).$$

Replacing the sample moments with their population moments is asymptotically equivalent because the difference is $o_P(1)$:

$$\mathbf{0} = \mathbf{G}_o'\boldsymbol{\Xi}_o\left[N^{-1/2}\sum_{i=1}^{N}\mathbf{g}_i(\boldsymbol{\theta}_o)\right] + \mathbf{G}_o'\boldsymbol{\Xi}_o\mathbf{G}_o\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_o) + o_P(1).$$

Defining $\mathbf{A}_o \equiv \mathbf{G}_o'\boldsymbol{\Xi}_o\mathbf{G}_o$ and solving for $\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_o)$ yields

$$\sqrt{N}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_o\right) = -\mathbf{A}_o^{-1}\mathbf{G}_o'\boldsymbol{\Xi}_o N^{-1/2}\sum_{i=1}^{N}\mathbf{g}_i(\boldsymbol{\theta}_o) + o_p(1).$$

Finally, from

$$N^{-1/2}\sum_{i=1}^{N}\mathbf{g}_i(\boldsymbol{\theta}_o)\xrightarrow{\mathrm{d}}\text{Normal}(0,\boldsymbol{\Lambda}_o),$$

we infer

$$\sqrt{N}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_o\right)=-\mathbf{A}_o^{-1}\mathbf{G}_o'\boldsymbol{\Xi}_o N^{-1/2}\sum_{i=1}^{N}\mathbf{g}_i(\boldsymbol{\theta}_o)+o_p(1)\xrightarrow{\mathrm{d}}\text{Normal}\left(\mathbf{0},\mathbf{A}_o^{-1}\mathbf{B}_o\mathbf{A}_o^{-1}\right)$$

where

$$\mathbf{B}_o\equiv\mathbf{G}_o'\boldsymbol{\Xi}_o\boldsymbol{\Lambda}_o\boldsymbol{\Xi}_o\mathbf{G}_o.$$

Estimation of the variance

Estimating the asymptotic variance of the GMM estimator is easy once $\hat{\boldsymbol{\theta}}$ has been obtained. A consistent estimator of $\boldsymbol{\Lambda}_o$ is given by

$$\hat{\boldsymbol{\Lambda}} \equiv N^{-1} \sum_{i=1}^{N} \mathbf{g}_i(\hat{\boldsymbol{\theta}})\, \mathbf{g}_i(\hat{\boldsymbol{\theta}})'$$

and $\mathrm{Avar}(\hat{\boldsymbol{\theta}})$ is estimated as $\hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{A}}^{-1}/N$, where

$$\hat{\mathbf{A}} \equiv \hat{\mathbf{G}}'\hat{\boldsymbol{\Xi}}\hat{\mathbf{G}}, \qquad \hat{\mathbf{B}} \equiv \hat{\mathbf{G}}'\hat{\boldsymbol{\Xi}}\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Xi}}\hat{\mathbf{G}}$$

and

$$\hat{\mathbf{G}} \equiv N^{-1} \sum_{i=1}^{N} \nabla_{\boldsymbol{\theta}} \mathbf{g}_i(\hat{\boldsymbol{\theta}}).$$

3. Efficient GMM estimation

Optimal weighting matrix

In general, the asymptotic variance of the GMM estimator is based on

$$\mathbf{A}_o^{-1}\mathbf{B}_o\mathbf{A}_o^{-1} = \left(\mathbf{G}_o'\boldsymbol{\Xi}_o\mathbf{G}_o\right)^{-1}\left(\mathbf{G}_o'\boldsymbol{\Xi}_o\boldsymbol{\Lambda}_o\boldsymbol{\Xi}_o\mathbf{G}_o\right)\left(\mathbf{G}_o'\boldsymbol{\Xi}_o\mathbf{G}_o\right)^{-1}.$$

As in the linear case, the optimal weighting matrix minimizes this variance.

To this end, choose $\hat{\boldsymbol{\Xi}}$ such that it converges to $\boldsymbol{\Lambda}_o^{-1}$. Then, substituting $\boldsymbol{\Xi}_o = \boldsymbol{\Lambda}_o^{-1}$ in the above expression yields

$$\mathbf{A}_o^{-1}\mathbf{B}_o\mathbf{A}_o^{-1} = \left(\mathbf{G}_o'\boldsymbol{\Lambda}_o^{-1}\mathbf{G}_o\right)^{-1}.$$

It can be shown that this variance is smaller (or equal) to the one above.

Since $\boldsymbol{\Lambda}_o$ is unknown it has to be estimated, leading to the following estimation procedure.

Efficient GMM estimation procedure

a. Let $\hat{\hat{\boldsymbol{\theta}}}$ be an initial consistent estimator of $\boldsymbol{\theta}$ using some initial weighting matrix $\hat{\hat{\boldsymbol{\Xi}}}$.

b. Obtain the $L \times 1$ vectors

$$\mathbf{g}_i(\hat{\hat{\boldsymbol{\theta}}}), \qquad i = 1, 2, \ldots, N.$$

c. As a consistent estimator of $\boldsymbol{\Lambda}_o$, compute $\hat{\boldsymbol{\Lambda}} = N^{-1} \sum_{i=1}^{N} \mathbf{g}_i(\hat{\hat{\boldsymbol{\theta}}}) \mathbf{g}_i(\hat{\hat{\boldsymbol{\theta}}})'$.

d. Use the weighting matrix

$$\hat{\boldsymbol{\Xi}} = \hat{\boldsymbol{\Lambda}}^{-1} = \left[N^{-1} \sum_{i=1}^{N} \mathbf{g}_i(\hat{\hat{\boldsymbol{\theta}}}) \mathbf{g}_i(\hat{\hat{\boldsymbol{\theta}}})' \right]^{-1}$$

to obtain the asymptotically efficient GMM estimator which solves

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left[\sum_{i=1}^{N} \mathbf{g}_i(\boldsymbol{\theta}) \right]' \hat{\boldsymbol{\Lambda}}^{-1} \left[\sum_{i=1}^{N} \mathbf{g}_i(\boldsymbol{\theta}) \right].$$

Estimation of the variance

The asymptotic distribution of the efficient GMM estimator is

$$\sqrt{N}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o\right) \xrightarrow{\text{d}} \text{Normal}\left(\mathbf{0}, \left[\mathbf{G}_o' \boldsymbol{\Lambda}_o^{-1} \mathbf{G}_o\right]^{-1}\right).$$

The variance of $\boldsymbol{\theta}$ can be estimated as

$$\widehat{\text{Avar}}(\hat{\boldsymbol{\theta}}) = \left(\hat{\mathbf{G}}' \hat{\boldsymbol{\Lambda}}^{-1} \hat{\mathbf{G}}\right)^{-1} / N,$$

where the estimator

$$\hat{\mathbf{G}} \equiv N^{-1} \sum_{i=1}^{N} \nabla_{\boldsymbol{\theta}} \mathbf{g}_i(\hat{\boldsymbol{\theta}})$$

is based on the efficient GMM estimator (and not the initial estimator $\hat{\hat{\boldsymbol{\theta}}}$).

4. Inference

Testing the validity of the moment conditions

We started GMM by assuming that the L population moment conditions

$$\mathrm{E}\left[\mathbf{g}(\mathbf{w}_i, \boldsymbol{\theta}_o)\right] = \mathbf{0}$$

are satisfied.

Whenever there are more moment conditions L than parameters P, the sample counterpart

$$N^{-1} \sum_{i=1}^{N} \mathbf{g}(\mathbf{w}_i, \hat{\boldsymbol{\theta}})$$

is not exactly zero.

Testing the validity of the moment conditions now means testing whether these deviations from zero can be explained by sampling uncertainty or whether it is systematic (from which we would conclude that the moment conditions—and thus the GMM estimator—are invalid).

To find a test statistic, start from

$$N^{-1/2}\sum_{i=1}^{N}\mathbf{g}_i(\boldsymbol{\theta}_o)\xrightarrow{\text{d}}\text{Normal}(\mathbf{0},\boldsymbol{\Lambda}_o),$$

which holds under the null hypothesis that the moment conditions are valid (because the CLT used $\text{E}\left[\mathbf{g}(\mathbf{w}_i,\boldsymbol{\theta}_o)\right]=\mathbf{0}$). Then

$$\left[N^{-1/2}\sum_{i=1}^{N}\mathbf{g}_i(\boldsymbol{\theta}_o)\right]'\hat{\boldsymbol{\Lambda}}^{-1}\left[N^{-1/2}\sum_{i=1}^{N}\mathbf{g}_i(\boldsymbol{\theta}_o)\right]\xrightarrow{\text{d}}\chi^2_L.$$

Since we do not know $\boldsymbol{\theta}_o$, we replace it by $\hat{\boldsymbol{\theta}}$ and obtain

$$\left[N^{-1/2}\sum_{i=1}^{N}\mathbf{g}_i(\hat{\boldsymbol{\theta}})\right]'\hat{\boldsymbol{\Lambda}}^{-1}\left[N^{-1/2}\sum_{i=1}^{N}\mathbf{g}_i(\hat{\boldsymbol{\theta}})\right]\xrightarrow{\text{d}}\chi^2_{L-P}.$$

The degrees of freedom reduction happens because P degrees of freedom have been used up by estimating $\hat{\boldsymbol{\theta}}$.

The test statistic is often called Hansen's J statistic or criterion statistic. It equals N times the criterion function of the efficient GMM estimator:

$$J = N\, Q_N(\hat{\boldsymbol{\theta}}) = N \left[N^{-1} \sum_{i=1}^{N} \mathbf{g}_i(\hat{\boldsymbol{\theta}}) \right]' \hat{\boldsymbol{\Lambda}}^{-1} \left[\sum_{i=1}^{N} N^{-1} \mathbf{g}_i(\hat{\boldsymbol{\theta}}) \right].$$

To test the null hypothesis that the moment conditions are valid, you would thus proceed as follows:

- Estimate the model using an optimal weighting matrix.
- This yields the efficient GMM estimator $\hat{\boldsymbol{\theta}}$.
- Substitute this estimator into the criterion function and compute $J = N\, Q_N(\hat{\boldsymbol{\theta}})$.
- Compare the J statistic with the appropriate quantile of the χ^2_{L-P} distribution.

Comments:

- Obviously, only overidentifying restrictions can be tested.
- You should use this test whenever you impose overidentifying restrictions, not only in the nonlinear GMM case. In principle, this (or a similar) test can be applied to test overidentifying restrictions in the case of 2SLS or linear system estimators like 3SLS.
- In Stata, use the postestimation command `estat overid`.

Wald statistic

Given the asymptotic normal distribution of the GMM estimator, Wald tests of linear hypotheses

$$H_0 : \mathbf{R}\boldsymbol{\theta} = \mathbf{r} \quad \text{against} \quad H_1 : \mathbf{R}\theta \neq \mathbf{r}$$

or nonlinear hypotheses

$$H_0 : \mathbf{c}(\boldsymbol{\theta}) = \mathbf{0} \quad \text{against} \quad H_1 : \mathbf{c}(\boldsymbol{\theta}) \neq \mathbf{0}$$

can be performed as discussed for the M-estimator.

GMM distance statistic

The GMM analogue to the likelihood ratio statistic is the GMM distance statistic based on the efficient GMM estimator:

$$\left\{ \left[\sum_{i=1}^{N} \mathbf{g}_i(\tilde{\boldsymbol{\theta}}) \right]' \hat{\Lambda}^{-1} \left[\sum_{i=1}^{N} \mathbf{g}_i(\tilde{\boldsymbol{\theta}}) \right] - \left[\sum_{i=1}^{N} \mathbf{g}_i(\hat{\boldsymbol{\theta}}) \right]' \hat{\boldsymbol{\Lambda}}^{-1} \left[\sum_{i=1}^{N} \mathbf{g}_i(\hat{\boldsymbol{\theta}}) \right] \right\} / N \xrightarrow{\mathrm{d}} \chi^2_Q,$$

where $\tilde{\boldsymbol{\theta}}$ is the restricted estimator (estimated under H_0) and $\hat{\boldsymbol{\theta}}$ is the unrestricted estimator (estimated under H_1), Q is the number of restrictions, and $\hat{\boldsymbol{\Lambda}}$ is obtained from an initial unrestricted estimator.

5. Estimation under Orthogonality Restrictions

The orthogonality restrictions

As in our initial example (estimation of an Euler equation) the moment condition

$$\mathrm{E}\left[\mathbf{g}(\mathbf{w}_i, \boldsymbol{\theta}_o)\right] = \mathbf{0}$$

can often be written as an orthogonality restriction,

$$\mathrm{E}\left[\mathbf{Z}_i' \mathbf{r}(\mathbf{w}_i, \boldsymbol{\theta}_o)\right] = \mathbf{0},$$

where

- the $G \times 1$ vector $\mathbf{r}_i(\boldsymbol{\theta}) = \mathbf{r}(\mathbf{w}_i, \boldsymbol{\theta})$ can be thought of as a **generalized residual function** and
- the $G \times L$ matrix $\mathbf{Z}_i$ is usually called the **matrix of instruments**.

Example: linear models

For the single-equation linear model

$$y_i = \mathbf{x}_i \boldsymbol{\theta} + u_i$$

with instrument vector $\mathbf{z}_i$ it becomes

$$\mathrm{E}\left[\mathbf{z}_i' u_i\right] = \mathbf{0}.$$

For the multiple-equation linear model

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\theta} + \mathbf{u}_i$$

with instrument matrix $\mathbf{Z}_i$ it becomes

$$\mathrm{E}\left[\mathbf{Z}_i' \mathbf{u}_i\right] = \mathbf{0}.$$

Properties

Using orthogonality restrictions

$$\mathrm{E}\left[\mathbf{Z}_i'\mathbf{r}(\mathbf{w}_i, \boldsymbol{\theta}_o)\right] = \mathbf{0}$$

yields a GMM estimator with

$$\mathbf{g}(\mathbf{w}_i, \boldsymbol{\theta}_o) = \mathbf{Z}_i'\mathbf{r}(\mathbf{w}_i, \boldsymbol{\theta}_o).$$

Hence, all the properties of the GMM estimator studied so far carry over. Just replace $\mathbf{g}(\mathbf{w}_i, \boldsymbol{\theta}_o)$ by $\mathbf{Z}_i'\mathbf{r}(\mathbf{w}_i, \boldsymbol{\theta}_o)$.

The nice thing about this specification is that it is straightforward to find an initial GMM estimator needed to compute the efficient (two-step) GMM estimator: just use nonlinear 2SLS.

Nonlinear 2SLS

Nonlinear 2SLS can be defined as an analogue to the linear 2SLS estimator that uses the same weighting matrix

$$\hat{\mathbf{\Xi}} \equiv \left(N^{-1} \sum_{i=1}^{N} \mathbf{Z}_i' \mathbf{Z}_i \right)^{-1}.$$

Then the nonlinear 2SLS estimator $\hat{\hat{\boldsymbol{\theta}}}$ solves

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left[\sum_{i=1}^{N} \mathbf{Z}_i' \mathbf{r}_i(\boldsymbol{\theta}) \right]' \left[N^{-1} \sum_{i=1}^{N} \mathbf{Z}_i' \mathbf{Z}_i \right]^{-1} \left[\sum_{i=1}^{N} \mathbf{Z}_i' \mathbf{r}_i(\boldsymbol{\theta}) \right].$$

Efficient GMM

In the second step, we can compute an estimator of the optimal weighting matrix,

$$\hat{\mathbf{\Lambda}} = N^{-1} \sum_{i=1}^{N} \mathbf{Z}_i' \mathbf{r}_i(\hat{\hat{\boldsymbol{\theta}}}) \mathbf{r}_i(\hat{\hat{\boldsymbol{\theta}}})' \mathbf{Z}_i,$$

and use it to find the efficient GMM estimator $\hat{\boldsymbol{\theta}}$ as the solution to

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left[\sum_{i=1}^{N} \mathbf{Z}_i' \mathbf{r}_i(\boldsymbol{\theta}) \right]' \hat{\mathbf{\Lambda}}^{-1} \left[\sum_{i=1}^{N} \mathbf{Z}_i' \mathbf{r}_i(\boldsymbol{\theta}) \right].$$