Generalized Method of Moments

Ethan Ligon

April 17, 2024

Generalized Method of Moments

We start with a set of ℓ equations implied by a parametric model of behavior which is meant to hold for observations (e.g., people, firms, households) indexed by j, say

$$\mathbb{E} g_j(\beta) = \mathbf{0}_\ell. \tag{1}$$

Here β is a k-vector of parameters, and where $g_j(b)$ can be computed for each observation i (i.e., g_i is a known function) for all $b \in B \subset \mathbb{R}^k$, where B is a compact set with $\beta \in B$. Each function $g_j : B \to \mathbb{R}^\ell$.

$k = \ell$ Just identified;

$k > \ell$ Under identified;

$k < \ell$ Over (Über) identified.

Well-specified models

A well-specified model will have a set (2) which uniquely determine β; i.e., for which there's a unique solution.

Production Example

We talked earlier about the problem facing a price-taking firm; that problem led to a collection of first order conditions of the form

$$\mathbb{E}(pF_i(\boldsymbol{x})|\boldsymbol{Z}) - w_i = 0 \qquad i = 1, \ldots, m,$$

where F_i is the partial derivative of the production function w.r.t. the ith input. If we choose some parametric form for F (with say k parameters) and some moments implied by the CEF then we'd obtain a problem which takes the method of moments form.

Causality?

These moment conditions aren't obviously interpretable as regressions, may lack some of the implicit causal structure we saw in structural equations, Cowles' commission style.

Analogy Principle

If our model has $\mathbb{E}g_j(\beta) = 0$ and we have a independent sample of observations $j = 1, \ldots, N$ we construct the obvious sample analog:

$$\boldsymbol{g}_N(b) = \frac{1}{N}\sum_{j=1}^{N} g_j(b)$$

which has the property that $\mathbb{E}\boldsymbol{g}_N(\beta) = \mathbb{E}g_j(\beta) = 0$.

Identification & the Analogy Principle

One way of thinking about the identification of a model with k unknown parameters is that each parameter can be written as a function of moments of observable variables. Once conceived of this way the analogy principle makes passage to actual estimation rather immediate.

Examples

- Producer
- OLS
- Linear IV
- Central moments
- Non-linear Least Squares

Overidentification

If we have ℓ moment restrictions and $k < \ell$ parameters, then even though we may have $\mathbb{E}g_j(\beta) = 0$ in theory, we'll basically *never* have a b which solves the overidentified $\boldsymbol{g}_N(b) = 0$, if only because of sampling variation.

What to do?

Criterion

It may be better to consider formulating the problem of solving a system of moment equations as a minimization problem.

Least Squares Criterion?

One Idea: Treat each sample moment as though it was an observation. And in a sense it is! But even though each sample moment is a random variable, if we can apply a Central Limit Theorem then we know something about its limiting distribution that usually we don't know in the standard regression setting.

$$\min_{b\in B}\frac{1}{\ell}\sum_{i=1}^{\ell}\boldsymbol{g}_{iN}(b)^2=\frac{1}{\ell}\boldsymbol{g}_N(b)^\top\boldsymbol{g}_N(b).$$

Under standard conditions this defines an estimator which is consistent and asymptotically normal.

Weighting

In general we can do better! The least squares criterion above is like OLS, in that it assigns the same weight to each moment restriction. But maybe different weights should be assigned to different moment conditions?

Different Weight

Choose $\boldsymbol{A}$ positive definite, and solve:

$$\min_{b \in B} N \boldsymbol{g}_N(b)^\top \boldsymbol{A} \boldsymbol{g}_N(b).$$

Extremum Estimators

Posed in the form above, we can regard GMM as an *extremum* estimator (or M-estimator). We have some excellent general results on the asymptotic behavior of such estimators, including standard results on consistency and central limit theorems (Newey & McFadden, 1994).

Extremum Estimators (Identification)

If there exists a L such that

$$\beta = \arg\max_{b \in B} L(b) \qquad \beta \text{ unique},$$

then an extremum estimator solves sample analog problem

$$b_N = \arg\max_{b \in B} L_N(b)$$

Consistency

We need:

1. $L(b)$ continuous on B;
2. B compact;
3. $\beta \in B$ unique maximizer of $L(b)$.
4. $L_N(b) \to L(b)$ *uniformly*

Uniform Law of Large Numbers

If the above conditions hold, then $b_N \xrightarrow{p} \beta$.

Efficient GMM

Back to our weighting problem. What is the right weighting matrix? Standard answer draws on GLS logic: the matrix that minimizes the variance of a consistent estimator b_N.

Efficient GMM as a GLS Estimator

We can use a Gauss-Markov sort of argument to derive the optimal weighting matrix. Let

$$\Omega = \mathbb{E}g_j(\beta)g_j(\beta)^\top - \mathbb{E}g_j(\beta)\mathbb{E}g_j(\beta)^\top.$$

This is positive definite; let $\boldsymbol{S}\boldsymbol{S}^\top = \Omega^{-1}$ be the Cholesky decomposition of Ω^{-1}.

Then $\sqrt{N}\boldsymbol{S}^\top\boldsymbol{g}_N(\beta) \xrightarrow{d} \mathcal{N}(\mathbf{0}, I_\ell)$.

Overidentification Tests

We're familiar with Wald tests for testing restrictions in a linear regression. A similar logic allows us to construct tests of the moment restrictions in GMM. In particular, for the efficient GMM estimator our criterion function is

$$J_N = \min_{b \in B} N(\boldsymbol{g}_N(b)^\top \boldsymbol{S})(\boldsymbol{S}^\top \boldsymbol{g}_N(b)),$$

which is asymptotically the sum of ℓ products of standard normal random variables, which any statistics textbook will remind us has a $\chi^2_{\ell-k}$ distribution.

Feasible Optimally Weighted GMM

Problem: In general the matrix Ω is unknown, and the analog estimator depends on b. This problem is analogous to the typical problem with GLS, which is that we don't know the covariance matrix of disturbances. This suggests analogous solutions.
We start with a set of ℓ equations implied by a parametric model of behavior which is meant to hold for observations (e.g., people, firms, households) indexed by j, say

$$\mathbb{E}g_j(\beta) = \mathbf{0}_\ell. \tag{2}$$

Here β is a k-vector of parameters, and where $g_j(b)$ can be computed for each observation i (i.e., g_i is a known function) for all $b \in B \subset \mathbb{R}^k$, where B is a compact set with $\beta \in B$. Each function $g_j : B \to \mathbb{R}^\ell$.

Solving the GMM Estimation Problem

Consider the GMM Criterion associated with a weighting matrix $\boldsymbol{A}$ (so not necessarily efficient!):

$$b_N^A = \operatorname*{arg\,min}_{b \in B} N \boldsymbol{g}_N(b)^\top \boldsymbol{A} \boldsymbol{g}_N(b).$$

Provided $g_j(b)$ is continuously differentiable the GMM estimator will satisfy the first order conditions:

$$\frac{1}{N} \sum_{j=1}^{N} \frac{\partial g_j(b)}{\partial b^\top} \boldsymbol{A} \boldsymbol{g}_N(b) = \mathbf{0}_k.$$

Limiting Distribution

Let $\boldsymbol{D}_N(b) = \frac{\partial g_j(b)}{\partial b^\top}$, an $\ell \times k$ matrix. Then provided β is in the interior of B we'll have

$$\mathbb{E}\boldsymbol{D}_N(\beta)^\top \boldsymbol{A}\boldsymbol{g}_N(\beta) = \mathbf{0}_k.$$

Let $\mathbb{E}\boldsymbol{D}_N(\beta) = \boldsymbol{D}$. What's the limiting distribution of the GMM estimator?

Linear GMM

The linear case is particularly simple to compute, because estimates of the matrix $\boldsymbol{D}$ don't depend on parameters. We'll focus on the IV problem (but this generalizes straight-forwardly to any linear regression problem):

$$y = X\beta + u \qquad \mathbb{E}(Z^\top u) = 0.$$

Here $g_j(b) = Z_j(y_j - X_j b)$ and $\boldsymbol{g}_N(b) = \boldsymbol{Z}^\top(\boldsymbol{y} - \boldsymbol{X}b)/N$, the sample average of the $g_j(b)$s. And note that

$$\boldsymbol{D}_N(b) = \partial\boldsymbol{g}_N(b)/\partial b^\top = \boldsymbol{Z}^\top\boldsymbol{X}.$$

Finite Sample Performance of GMM Estimator

See notebook exploring finite sample performance of GMM.

Efficient GMM

We've seen that an "efficient" GMM estimator solves

$$b_N = \operatorname*{arg\,min}_{b \in B} \boldsymbol{g}_N(b)^\top \Omega^{-1} \boldsymbol{g}_N(b)),$$

where

$$\Omega = \mathbb{E} g_j(\beta) g_j(\beta)^\top - \mathbb{E} g_j(\beta) \mathbb{E} g_j(\beta)^\top.$$

Now we want to think about what "efficient" means.

Maximum Likelihood & Efficiency

What's an efficient consistent estimator? The one with the smallest covariance matrix. (Note that this ordering of estimators may be different in finite samples than it is in the limit.)

Maximum Likelihood

Suppose we have a continuous random variable X with density $f(X|\beta)$; we obtain realizations $\boldsymbol{X} = (X_1, \dots, X_N)$. Then the maximum likelihood estimator for β is

$$b_{MLE} = \arg\max_{b \in B} \frac{1}{N} \sum_{j=1}^{N} \log f(X_j|b).$$

Provided f is continuously differentiable, this implies first order conditions ("Scores"):

$$\frac{1}{N} \sum_{j} \frac{\partial f(X_j|b)/\partial b^\top}{f(X_j|b)} = 0$$

Efficient Scores

If β is in the interior of B and the support of X isn't a function of β, then we have

$$\mathbb{E}\frac{\partial f(X_j|\beta)/\partial b^\top}{f(X_j|\beta)} = 0;$$

In this case we say the "scores are efficient". But note this is a set of moment conditions! So "efficient scores" implies that we can pose maximum likelihood problem as a GMM problem.

Information Matrix

Let

$$g_j(b) = \frac{\partial f(X_j|b)/\partial b^\top}{f(X_j|b)};$$

If scores are efficient

$$\mathbb{E}g_j(\beta) = 0.$$

We've already seen that the optimal weighting matrix for GMM is the inverse of

$$\Omega = \mathbb{E}g_j(\beta)g_j(\beta)^\top - \mathbb{E}g_j(\beta)\mathbb{E}g_j(\beta)^\top.$$

But in this ML context Ω has another interpretation: this is the information matrix $\mathcal{I}$.

Interpretation

We can interpret the information matrix as measuring the information X conveys about the vector of parameters β.

Cramér-Rao Lower Bound

Recall: Under the same conditions for the score to be efficient, then the variance of any unbiased estimator V_b satisfies

$$V_b \geq \frac{1}{N}\mathcal{I}^{-1}$$

So: GMM exploiting efficient scores asymptotically achieves the Cramér-Rao bound, in the sense that

$$\sqrt{N}(b_N - \beta) \xrightarrow{d} Z \qquad \text{with } \mathbb{E}Z = 0 \text{ and } \mathrm{Var}(Z) = \mathcal{I}^{-1}.$$

Maximum Likelihood Efficiency

GMM exploiting scores as moment conditions achieves the same efficiency as ML using the same information. Note that this is a *just-identified* estimator by construction. Some questions:

- We've assumed parametric density functions. How does this affect our interpretation of efficiency?
- Why is it not possible to *improve* efficiency by exploiting additional moment conditions? (Or is it?)

Interpreting Efficiency Bounds

Maximum likelihood asymptotically achieves the Cramér-Rao lower bound; in this sense it makes efficient use of the information provided in the estimation. What is this information?