Simulated Method of Moments©

Dean Corbae

October 20, 2021

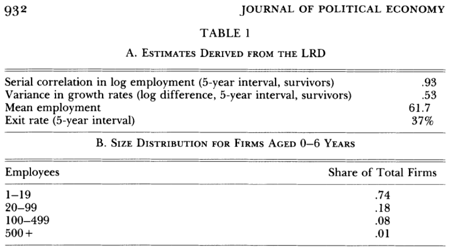
Simulated Method of Moments©

Dean Corbae

October 20, 2021

An Example

Suppose you want to estimate the parameters of a heterogeneous firm model as in Hopenhayn and Rogerson (1993) but you do not have access to Census (LRD) data, only the following table:



932 JOURNAL OF POLITICAL ECONOMY

TABLE 1

A. Estimates Derived from the LRD

Serial correlation in log employment (5-year interval, survivors)	.93
Variance in growth rates (log difference, 5-year interval, survivors)	.53
Mean employment	61.7
Exit rate (5-year interval)	37%

B. Size Distribution for Firms Aged 0–6 Years

Employees	Share of Total Firms
1–19	.74
20–99	.18
100–499	.08
500 +	.01

How can you do it? Choose the parameters of your model to miminize differences between model moments and the above data moments. Just an application of Method of Moments.

OLS as Method of Moments

- The first model we usually see in econometrics is a linear one where the true model is assumed to be $y_t = \beta x_t + u_t$ with $E[x_t u_t] = 0$, $E[u_t] = 0$, and demeaned data.
- We try to estimate the parameter vector β that maps the model βx_t to the data of interest y_t.
- An implication of $E[x_t u_t] = 0$ is that

$$E[x_t \left(y_t - \beta x_t \right)] = 0. \tag{1}$$

- The sample analogue of the moment condition (1) is

$$\frac{1}{T}\sum_{t=1}^{T} x_t \left(y_t - \widehat{\beta}_T^{MM} x_t \right) = 0 \tag{2}$$

yielding

$$\widehat{\beta}_T^{MM} = \frac{\sum_{t=1}^{T} x_t y_t}{\sum_{t=1}^{T} x_t x_t}. \tag{3}$$

OLS as Method of Moments -cont.

- An alternative is to choose β that minimizes the sum of squared deviations of the data y_t from the model βx_t or

  $$\widehat{\beta}_T^{OLS} = \arg\min_{\beta} \sum_{t=1}^{T} (y_t - \beta x_t)^2. \tag{4}$$

  The first order condition is

  $$-2 \sum_{t=1}^{T} (y_t - \widehat{\beta}_T^{OLS} x_t) x_t = 0. \tag{5}$$

  But this foc is identical to the moment condition in (2).
- Generalized Least Squares is simply a more general moment condition than (1) given by

  $$E[x_t \left(y_t - \beta x_t\right) / \sigma^2(x_t)] = 0. \tag{6}$$

  Instead of equal weights as in OLS, GLS upweights moments inversely related to variation in x_t (info from perturbing low variation variables provides more info than noisy vars.)

A GMM Example

- Consider Lucas' (1978) representative agent asset pricing model (see Hansen and Singleton (1982)):

$$\begin{aligned} \max_{\{c_t, s_{t+1}\}_{t=0}^{\infty}} \quad & E_0 \sum_{t=0}^{\infty} \beta^t U(c_t) \\ s.t.\ c_t + p_t s_{t+1} \quad &= \quad (y_t + p_t) s_t \end{aligned}$$

  with market clearing conditions $c_t = y_t$ and $s_{t+1} = 1$.

- After parameterizing preferences as $U(c_t) = \frac{c_t^{1-\psi} - 1}{1-\psi}$, the first order necessary condition is given by

$$\begin{aligned} p_t c_t^{-\psi} \quad &= \quad E_t \beta c_{t+1}^{-\psi} \left(p_{t+1} + y_{t+1}\right) \qquad (7) \\ \iff \quad & E_t \left[\beta \left(\frac{c_t}{c_{t+1}} \right)^{\psi} \left(\frac{p_{t+1} + y_{t+1}}{p_t} \right) - 1 \right] = 0 \end{aligned}$$

  which is a moment condition.

A GMM Example -cont.

- We can rewrite (7) in terms of errors

$$u_{t+1}(x_{t+1}, b) \equiv \beta \left(\frac{c_t}{c_{t+1}} \right)^{\psi} \left(\frac{p_{t+1} + y_{t+1}}{p_t} \right) - 1$$

  where

  - $u_{t+1}(x_{t+1}, b)$ is an $n \times 1$ vector of errors (with finite second moments (stationarity)).
  - b is an $\ell \times 1$ vector of parameters (e.g. (β, ψ)),
  - x_{t+1} is a $k \times 1$ vector of variables observed by agents (and the econometrician) as of $t+1$ (e.g. $\{c_n, y_n, p_n\}_{n=0}^{t+1}$)

- We then estimate the true $\ell = 2$ parameters b_0 to solve the $n = 1$ moment condition:

$$E_t \left[u_{t+1}(x_{t+1}, b_0) \right] = 0.$$

Order Conditions

- Suppose there are n necessary conditions of the model:

$$E_t\left[u_{t+1}(x_{t+1}, b_0)\right] = 0 \tag{8}$$

  where
  - u_{t+1} is an $n \times 1$ vector of "errors"(e.g. foc in the asset pricing model; differences between model and data moments in the SMM case).
  - x_{t+1} is a $k \times 1$ vector of data
  - b is an $\ell \times 1$ vector of parameters where b_0 stands for the true parameter vector
- Recall the following order conditions necessary (but not sufficient) for identification:
  - If $\ell < n$, the model is said to be overidentified.
  - If $\ell = n$, the model is said to be just identified.
  - If $\ell > n$, the model is said to be underidentified.

Order Conditions - cont.

- In the asset pricing example above, we have $\ell = 2$ (i.e. (β, ψ)) and $n = 1$ (i.e. the foc wrt s_{t+1}) so we are in the underidentified case (very bad).
- Fix it by adding more "equations". If z_t is a $q \times 1$ vector of variables in the econometrician's (and agent's) info set, then from (8) and the law of iterated expectations we know

$$E_t \left[u_{t+1}(x_{t+1}, b_0) \otimes z_t \right] = 0 \otimes z_t = 0 \Longrightarrow E \left[u_{t+1}(x_{t+1}, b_0) \otimes z_t \right] = 0 \tag{9}$$

is an $nq \times 1$ vector where $\otimes$ is the Kroenecker product.

- Loosely speaking, one way to interpret the z_t is as a vector of instrumental variables.
- For example, in the Hansen and Singleton (1982) paper, they include past consumption growth in z_t (i.e. $z_t = [1 \ c_t/c_{t-1}]'$). This is similar to using a lagged dependent variable as an instrument provided the true errors are not autocorrelated.

Order Conditions - cont.

- Letting $f(x_{t+1}, z_t, b) \equiv u_{t+1}(x_{t+1}, b) \otimes z_t$, define the $nq \times 1$ moment vector

$$g(b) \equiv E[f(x_{t+1}, z_t, b)] \tag{10}$$

(i.e. the unconditional average error). By (9), $g(b_0) = 0$. This is the analogue of the OLS condition (1).

- The sample analogue of (10) is the $nq \times 1$ vector

$$g_T(b) \equiv \frac{1}{T} \sum_{t=1}^{T} f(x_{t+1}, z_t, b) \tag{11}$$

The basic idea of GMM is that as $T \to \infty$, (9) implies $g_T(b_0) = 0$.This is the analogue of the OLS condition (2).

GMM Estimation

Assuming that $g_T(b)$ is continuous in b, the GMM estimate of b solves

$$b_T = \arg\min_b J_T(b) \tag{12}$$

where

- $J_T(b) \equiv g'_T(b) W_T g_T(b)$ (which is $(1 \times nq)(nq \times nq)(nq \times 1)$) is a weighted sum of squared errors
- W_T is an arbitrary weighting $(nq) \times (nq)$ matrix that can depend on the data.
- (12) is the analogue of the OLS condition (4).
- In the just identified case, the weighting matrix does not matter provided the Jacobian of J_T wrt b is invertible.

Consistency

- Under certain conditions, Hansen 1982 (Theorem 2.1) proves that this estimator b_T exists and converges in probability to b_0.
- It is essential for consistency that the limit $J_\infty(b)$ have a unique maximum at the true parameter value b_0.
- This condition is related to identification; the distribution of the data at b_0 is different than that at any other possible parameter value.
- Further, Hansen 1982 (Theorem 3.1) establishes asymptotic normality of the estimator.

Consistency - cont.

- The consistency conditions are:
  - $W_T \to W$ in probability, where W is a positive semi-definite matrix
  - $g(b) = 0$ (an $nq \times 1$ vector) only for $b = b_0$.
  - $b_0 \in B$ (a compact set)
  - $f(x, z, b)$ is continuous at each b
  - $E[\sup_b \|f(x, z, b)\|] < \infty$.
- The second condition (known as **Global Identification**) is hard to verify.

Local Identification

A simpler necessary but not sufficient condition is known as **Local Identification**.

- If $g(b)$ is continuously differentiable in a neighborhood of b_0, then the Jacobian matrix $\nabla_b g(b)$ (which is $(nq \times \ell)$) must have full column rank (i.e. there are ℓ linearly independent columns).
- If $\nabla_{b_i} g(b) = 0$ (i.e. the parameter b_i does not have any impact on the objective of lowering the error variance), then the Jacobian matrix in (12) does not have full column rank since it has a column of zeros. This implies b_i is not well-identified.

Efficiency

- While the above result shows that the GMM estimator is consistent for arbitrary weighting matrices (e.g. $W = I$), it is not necessarily efficient.
- Hansen (1982, Theorem 3.2) shows that the statistically optimal weighting matrix is given by $W^* = S^{-1}$ where the asymptotic variance covariance matrix is:

$$S = \sum_{j=-\infty}^{\infty} E\left[f(x_t, z_{t-1}, b_0) f(x_{t-j}, z_{t-j-1}, b_0)'\right] \tag{13}$$

- Why does this weighting matrix make sense? Some moments will have more variance than others. This downweights errors from high variance moments (i.e. those with low signal to noise).

Efficiency - cont.

- The problem is that we do not know S^{-1} nor $g(b)$.
- If the errors are serially uncorrelated, then a consistent estimate of the asymptotic var-covar matrix S is given by

$$S_T = \frac{1}{T} \sum_{t=1}^{T} f(x_{t+1}, z_t, b_T) f(x_{t+1}, z_t, b_T)'$$

  where b_T is a consistent estimate of b_0.

- In this case, the distribution of the estimator is given by

$$\sqrt{T}(b_T - b_0) \rightarrow N(0, \left[\nabla_b g_T(b_T)' S_T^{-1} \nabla_b g_T(b_T)\right]^{-1}) \quad (14)$$

Efficiency - cont.

- Notice that the precision of the estimates in (14) is related to $\nabla_b g_T(b_T)$.

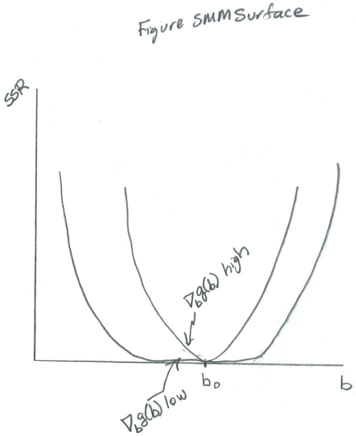
- If the objective is very sensitive to changes in the parameters (i.e. $\nabla_b g_T(b_T)$ is high), then there will be a low variance of the estimate (since $\nabla_b g_T(b_T)' S_T^{-1} \nabla_b g_T(b_T)$ is inverted).

- If the objective is not very sensitive to changes in the parameters (i.e. $\nabla_b g_T(b_T)$ is low), it will produce a high variance for the estimates. See Figure SMMsurface.

- Simply put, this suggests that if you find big standard errors, it is because the objective is not very sensitive to changes in the parameters so it is hard to find the true unique maximum.

- This is how **Local Identification** is linked to standard errors.

$J_T(b)$



Implementation

To implement this, can use a two step procedure (not necessary if you already have the optimal weighting matrix):

1. the first stage estimate of b minimizes a quadratic form of the sample mean of errors for $W = I$, which is consistent;
2. estimate a var-covar matrix of the residuals S_T from the first stage to form $W_T = S_T^{-1}$ in a second stage minimization of $g_T'(b)S_T^{-1}g_T(b)$.

This is like the two step procedure in Generalized Least Squares.

Simulated Method of Moments

One way to think about SMM is that it is the statistical approach to do calibration.

- Let $\{x_t\}_{t=1}^{T}$ be a realization of a $k \times 1$ vector valued stationary and ergodic stochastic process generating the observed data (e.g. detrended GDP).
- Let $\{y_t(b)\}_{t=1}^{T}$ be a realization of a $k \times 1$ vector valued stationary stochastic and ergodic process generating the simulated data (e.g. GDP generated by the model) where b is an $\ell \times 1$ vector of parameters. In general we may take H simulations of length T.
- Let $M_T(x)$ be an $n \times 1$ vector of data moments (e.g. standard deviation of detrended GDP) and $M_N(y(b))$ be a $n \times 1$ vector of model moments of the simulated data where $N = H \cdot T$.
- Assume that $M_T(x) \overset{a.s.}{\to} \mu(x)$ as $T \to \infty$ and that $M_N(y(b)) \overset{a.s.}{\to} \mu(y(b))$ as $N \to \infty$ where $\mu(x)$ and $\mu(y(b))$ are the population moments.

SMM key idea

- Under the null that the model is correct at the true parameter vector b_0, then $\mu(x) = \mu(y(b_0))$. If you understand this equality you understand everything you need to know about economics. It says there is a link between data and theory.

- In summary, x_t is observed data (which we may not even have), y_t is simulated data, $M_T(x)$ is observed moments (which we will assume we have), $M_N(y(b))$ is simulated moments, and the reason we can use the model to say something about the data we don't have is that if the model is the true data generation process, then the asymptotic moments have to be equal at the true parameter values (i.e. $\mu(x) = \mu(y(b_0))$).

SMM analogue of $J_T(b)$

- Given a symmetric $n \times n$ weighting matrix W_T (which may depend on the data - hence the subscript T), Lee and Ingram show that under certain conditions the simulation estimator $\widehat{b}_{TN}$ which minimizes the weighted sum of squared errors of the model moments from the data moments:

  $$\widehat{b}_{TN} = \arg\min_b [M_T(x) - M_N(y(b))]' W_T [M_T(x) - M_N(y(b))]$$

  - is a consistent and asymptotically normal estimator of b_0.
- Basically, SMM is just GMM where the errors are just the difference between the data moment and the model moment $g_{TN} = M_T - M_N(y(b))$.
- Since the solution to this problem is essentially a special case of Hansen's (1982) GMM estimator, the conditions mirror his paper: (i) x and $y(b)$ are independent; (ii) the model must be identified; and (iii) $M_N(y(b))$ must be continuous in the mean.

Estimation

Estimation of parameters conducted in two steps (function calls):

1. For any given value of b, say b^i,
   a. simulate artificial data from the model
      a1. H draws of $\{\varepsilon_t\}_{t=1}^{T}$ (**You must use the same random draw throughout each simulation**)
      a2. induce technology shocks and via decision rules, which depend on parameters b^i, induce a realization of real output $y(b^i)$)
   b. compute a moment based on those (i.e. $M_N(y(b^i))$), and evaluate the objective function $J_{TN}(b^i) = [M_T(x) - M_N(y(b^i))]' W [M_T(x) - M_N(y(b^i))]$; and
2. choose a new value for the parameters, say b^{i+1}, for which $J_{TN}(b^{i+1}) \leq J_{TN}(b^i)$.

- A standard minimization routine can construct this sequence of increasingly smaller $J_{TN}(b^i)$.
- If you don't use the same draw in step a1. you wouldn't know whether the change in the objective comes from a change in the parameter or a change in the draw.

Optimal Weighting Matrix

- If you don't have data, to obtain the optimal weighting matrix, you can use a two stage procedure (see Section 4.2.3 of Gourieroux and Monfort (1996) for justification).
    - In the first stage, minimize $J^1_{TN}(b)$ constructed using $W = I$. Given consistency, you now have the true data generation process $y(b_0)$.
    - The second stage is different from standard GMM since there is not a vector of residuals to construct the variance-covariance matrix. Here, you don't have enough "data" to get a var-covar matrix.
    - Since the resulting estimate $\widehat{b}^1_{TN}$ of b_0 is consistent, generate H repetitions of model moments (from T length simulated samples) analogous to the data moments in order to construct an estimate of the var-covar matrix $\widehat{S}_T$ of the "data moments".
    - Use $W_T = \widehat{S}_T^{-1}$ to construct the second stage J^2_{TN} and obtain the corrected estimate $\widehat{b}^2_{TN}$.
- Once you have the optimal weighting matrix, you can generate standard errors according to Hansen (1982).

Standard Errors -cont.

- Alternatively, you can run a Monte Carlo experiment to compute standard errors of the estimates.
- Recall that each estimate of $\widehat{b}_{TN}$ is derived for a given HT draw of the shocks ε_t to the underlying data generation process.
- Different draws will generate different estimates of $\widehat{b}_{TN}$.
- You can generate a histogram and summary statistics (mean and sd of $\widehat{b}_{TN}$) which are interpretable as the point estimate and standard error of b.

Model Parameter Sensitivity

Based on Andrews, Gentzkow, and Shapiro (2017) “Measuring the Sensitivity of Parameter Estimates to Estimation Moments”, *Quarterly Journal of Economics*, p.1553-1592.

- Recall that SMM employs Hansen’s GMM methodology.
- Recall from (14) that the distribution of the GMM estimator is given by

$$\sqrt{T}(b_T - b_0) \rightarrow N\left(0, \left[\nabla_b g_T(b_T)' S_T^{-1} \nabla_b g_T(b_T)\right]^{-1}\right) \quad (15)$$

  where the $(nq \times \ell)$ Jacobian matrix $\nabla_b g(b)$ must have full column rank (i.e. no occurrences of $\nabla_{b_i} g(b) = 0$ where the parameter b_i does not have any impact on the objective of lowering the WSSE.)
- But if $\nabla_b g_T(b_T)$ is very low, then there will be a high variance of the estimate (since $\nabla_b g_T(b_T)' S_T^{-1} \nabla_b g_T(b_T)$ is inverted).
- AGS use the Jacobian matrix $\nabla_b g(b)$ to construct a measure of parameter sensitivity to the data.

Some Quotes from AGS p. 1553-1554

- "We propose a local measure of the relationship between parameter estimates and the moments of the data they depend on."
- "Our measure can be computed at negligible cost even for complex structural models." This is because it depends on the above Jacobian.
- "We argue that reporting this measure can increase the transparency of structural estimates, making it easier for readers to predict the way violations of identifying assumptions would affect the results."

Sensitivity Matrix Λ

Definition 1. (p.1562) The sensitivity of $\hat{b}$ (a vector of estimated parameters) to $\hat{g}(b_0)$ (model moments) is:

$$\Lambda = -(G'WG)^{-1}G'W$$

where

- G is the Jacobian matrix $\nabla_b g(b_0)$
- W is a weight matrix

Hence Λ acts like the inverse of the Jacobian matrix.

- In the case of minimum distance estimators like SSM (section IV.A.) where $g(b) = M_T(x) - M_N(y(b))$, it measures how variation in model moments impact parameter estimates.
- A big number means the parameter is very sensitive to a given data moment. We will illustrate this in next a specific case.