Method of Simulated Moments©

Dean Corbae

August 5, 2024

Method of Simulated Moments©

Dean Corbae

August 5, 2024

Lecture 1 and 2 Outline

1. Micro Data and Macro Models
2. A Refresher on Generalized Method of Moments (GMM)
3. Method of Simulated Moments (MSM)
   - There is a Problem Set for this section
4. Indirect Inference (II) by Guvenen

Micro Data and Macro Models

Micro Data and Macro Models

A Couple Macro Questions:

1. Positive: How do financial frictions affect the distribution of wealth in the U.S.?

2. Normative: What are the welfare consequences of inequality in the U.S.?

One way to answer these questions is to build a dynamic structural model with idiosyncratic earnings shocks and incomplete markets which induces an endogenous cross-section of wealth.

Solving the Model: Nested Fixed Point Problem

- Start with *exogenous* earnings y_{t+1} that follows a finite state Markov Process which is iid across agents.

- Solve the model via dynamic programming for an endogenous savings decision rule $a_{t+1} = g(a_t, y_t; \theta)$ as a function of parameters θ.

- Use the decision rule and Markov process to generate an endogenous cross-sectional wealth distribution $\mu(a_t, y_t; \theta)$.

- Once you have these two functions you can use the model to simulate many different moments (e.g. Gini coefficient) to compare to data moments.

- Choose parameters θ so that the model moments match relevant data moments (MSM) or are consistent with an auxiliary model like a panel regression (II).

Refresher on GMM:

The building block of MSM and II

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

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

- 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).

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.

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 $s_{t+1} = 1$.

A GMM Example

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

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

  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

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

  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. $b = (\beta, \psi)$),
- x_{t+1} is a $k \times 1$ vector of variables observed by agents (and the econometrician) as of $t + 1$ (e.g. $\{c_n, y_n, p_n\}_{n=0}^{t+1}$)

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. $b = (\beta, \psi)$),
  - 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 where b_0 is the true parameter

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 where b_0 is the true parameter

- 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 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.

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 info set, then from (8) and the law of iterated expectations we know

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

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

GMM

- Define the $n \times 1$ moment vector

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

(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 $n \times 1$ vector

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

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 $(1 \times n)(n \times n)(n \times 1)$) is a weighted sum of squared errors
- W_T is an arbitrary weighting $(n) \times (n)$ 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)
  - $u(x, b)$ is continuous at each b
  - $E[\sup_b \|u(x, b)\|] < \infty$.
- 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 $(n \times \ell)$) 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:

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

- 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) similar to GLS above.

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

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

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.

$J_T(b)$

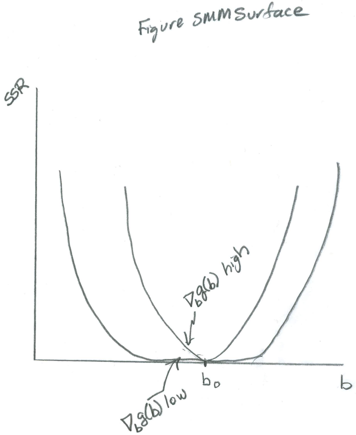


Figure SMMSurface

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.

Method of Simulated Moments

Method of Simulated Moments

- Suppose you don't have access to the data $\{x_t\}_{t=1}^{T}$ or the $n \times 1$ vector valued function $u(x_{t+1}, b)$ is unknown or extremely hard to evaluate (e.g. the foc is not well-behaved).

- Instead, we will construct $u(x_{t+1}, b)$ via simulation methods using only relevant moments $M_T(x)$ of the data (which could come from some other empirical paper) and your parameterized structural model to generate analogous moments $M_N(y(b))$.
    - Let $u(x, b) \equiv M_T(x) - M_N(y(b))$.

- MSM simply chooses b to minimize the weighted sum of squared errors between the data moments and the simulated model moments.

- The statistical foundations for MSM follow from GMM.

Method of Simulated Moments

- Let $\{x_t\}_{t=1}^{T}$ be a realization of a $k \times 1$ vector valued stationary and ergodic stochastic process in the data.

- 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 where b is an $\ell \times 1$ vector of parameters. By ergodicity, we may take H simulations of length T.

- Let $M_T(x)$ be an $n \times 1$ vector of data moments 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.

MSM 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 almost everything you need to know about economics. :-)
- It says there is a link between data and theory.
- The Cowles Commission focused on linking economic theory to mathematics and statistics. It's motto is "Theory and Measurement".

MSM 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.

MSM 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, MSM is just GMM where the errors are 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 for consistency and efficiency mirror his paper.

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 $J_{TN}(b^{i+1}) \leq J_{TN}(b^i)$.

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 $J_{TN}(b^i)$.
- Use the same draw in step a1, otherwise don't know if change to objective comes from change in parameter or draw.

A Simple Example

- To illustrate the theory, here is an example based on Michaelides and Ng (2000, Journal of Econometrics).
- The true data generation process is a $k = 1$ MA(1) process

$$x_t = \varepsilon_t - b_0 \varepsilon_{t-1}, \varepsilon_t \overset{i.i.d}{\sim} N(0,1) \tag{15}$$

  with $\ell = 1$ parameter $b_0 = 0.5$ and $\varepsilon_0 = 0$.
- The model generation process is

$$y_t(b) = e_t - b e_{t-1}, \quad e_t \overset{i.i.d}{\sim} N(0,1) \tag{16}$$

  with parameter b and $e_0 = 0$.
- We do not know the true parameter value b_0 so will estimate it via simulated method of moments.

Moments

- Let m denote the mapping from some $k \times 1$ vector z_t (which could be true data (x_t) or simulated data ($y_t(b)$)) to an $n \times 1$ moment vector.
- Just for example, suppose we take $k = 1$ and consider $n = 4$ moments: mean, variance, first order autocorrelation, and second order autocorrelation given by:

$$m(z_t) = \begin{bmatrix} z_t \\ (z_t - \bar{z})^2 \\ (z_t - \bar{z})(z_{t-1} - \bar{z}) \\ (z_t - \bar{z})(z_{t-2} - \bar{z}) \end{bmatrix}. \tag{17}$$

- Thus with $n > k$ we have an overidentified model (remember I'm just trying to illustrate things).

Population Moments

- Note that we can write the population (unconditional) moment vector for the true data using m as $\mu(x) = E[m(x)]$.
- In our simple $n = 4$ example, the population data moments are:

$$\mu(x) = \begin{bmatrix} E\left[\varepsilon_t\right] - b_0 E\left[\varepsilon_{t-1}\right] \\ E\left[\left(\varepsilon_t - b_0\varepsilon_{t-1}\right)^2\right] \\ E\left[\left(\varepsilon_t - b_0\varepsilon_{t-1}\right)\left(\varepsilon_{t-1} - b_0\varepsilon_{t-2}\right)\right] \\ E\left[\left(\varepsilon_t - b_0\varepsilon_{t-1}\right)\left(\varepsilon_{t-2} - b_0\varepsilon_{t-3}\right)\right] \end{bmatrix}$$

$$= \begin{bmatrix} E\left[\varepsilon_t\right] - b_0 E\left[\varepsilon_{t-1}\right] \\ E\left[\varepsilon_t^2\right] - 2b_0 E\left[\varepsilon_t\varepsilon_{t-1}\right] + b_0^2 E\left[\varepsilon_{t-1}^2\right] \\ E\left[\varepsilon_t\varepsilon_{t-1}\right] - b_0 E\left[\varepsilon_t\varepsilon_{t-2}\right] - b_0 E\left[\varepsilon_{t-1}^2\right] + b_0^2 E\left[\varepsilon_{t-1}\varepsilon_{t-2}\right] \\ E\left[\varepsilon_t\varepsilon_{t-2}\right] - b_0 E\left[\varepsilon_t\varepsilon_{t-3}\right] - b_0 E\left[\varepsilon_{t-1}\varepsilon_{t-2}\right] + b_0^2 E\left[\varepsilon_{t-1}\varepsilon_{t-3}\right] \end{bmatrix} \tag{18}$$

Population Moments -cont.

- Given the i.i.d. assumption, the population data moments are then:

$$\mu(x) = \begin{bmatrix} 0 \\ 1 + b_0^2 \\ -b_0 \\ 0 \end{bmatrix}. \tag{19}$$

  and the population model moments

$$\mu(y(b)) = \begin{bmatrix} 0 \\ 1 + b^2 \\ -b \\ 0 \end{bmatrix}. \tag{20}$$

Moment conditions

- The n moment conditions we are going to use in MSM is given by

$$u(x_t, b) = m(x_t) - \frac{1}{H}\sum_{h=1}^{H} m(y_t^h(b)) \tag{21}$$

where H is the number of simulations of the model.

- Then define an $n \times 1$ vector (where $n = 4$) as in the GMM section

$$\begin{aligned} g(b) \equiv E[u(x_t, b)] = E\left[m(x_t)\right] - \frac{1}{H}\sum_{h=1}^{H} E\left[m(y_t^h(b))\right] \\ = \mu(x) - \mu(y(b)) \end{aligned}$$

Global Identification

- Note that **Global Identification** requires $g(b) = 0 \iff b = b_0$.
- In this example, the Global Identification condition is that there is a unique solution $b = b_0$ to the following equation

$$\begin{aligned} & g(b) = \mu(x) - \mu(y(b)) = 0 \\ \iff & \mu(x) = \begin{bmatrix} 0 \\ 1 + b_0^2 \\ -b_0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 + b^2 \\ -b \\ 0 \end{bmatrix} = \mu(y(b)). \end{aligned} \tag{22}$$

- In particular, in (22) the mean and second autorcorrelation do not identify b, the variance does not uniquely identify it since $b = \pm b_0$, while the autocorrelation uniquely identifies $b = b_0$.

Optimal Weighting Matrix

- We next obtain an analytic expression for the variance-covariance matrix S and hence the optimal weighting matrix $W^* = S^{-1}$.
- Let S denote the $n \times n$ asymptotic var-covar matrix of the n moment conditions u at the true parameter value $b = b_0$:

$$S = \sum_{j=-\infty}^{\infty} E[u(x_t, b_0)u(x_{t-j}, b_0)'] \tag{23}$$

- Under a set of assumptions (see bigger slides), from (21) we know $S = \left(1 + \frac{1}{H}\right) S_x$, where S_x is the asymptotic var-covar matrix of $m(x_t)$:

$$S_x \quad \equiv \quad \sum_{j=-\infty}^{\infty} E[\{m(x_t) - \mu(x)\}\{m(x_{t-j}) - \mu(x)\}'] \equiv S_y$$

Optimal Weighting Matrix - cont.

- Letting $\Gamma_j \equiv E[\{m(x_t) - \mu(x)]\}\{m(x_{t-j}) - \mu(x)]\}']$ denote the j-th autocovariance of m, the asymptotic var-covar matrix of $m(x_t)$ is given by

$$S_x = \Gamma_0 + \sum_{j=1}^{\infty} (\Gamma_j + \Gamma_j') \tag{24}$$

- Since we know the true DGP, can compute Γ_js analytically:

$$\Gamma_0 = \begin{bmatrix} \sigma_x^2 & 0 & 0 & 0 \\ 0 & 2\sigma_x^4 & -2b_0\sigma_x^2 & 0 \\ 0 & -2b_0\sigma_x^2 & \sigma_x^4 + b_0^2 & -b_0\sigma_x^2 \\ 0 & 0 & -b_0\sigma_x^2 & \sigma_x^4 \end{bmatrix}$$

$$\Gamma_1 = \begin{bmatrix} -b_0 & 0 & 0 & 0 \\ 0 & 2b_0^2 & 0 & 0 \\ 0 & -2b_0\sigma_x^2 & b_0^2 & 0 \\ 0 & 2b_0^2 & -b_0\sigma_x^2 & b_0^2 \end{bmatrix}$$

with Γ_j zero $\forall j \geq 2$ and where $\sigma_x^2 = 1 + b_0^2$ is the variance of x_t.

Optimal Weighting Matrix - cont.

- Then since the asymptotic var-covar matrix $S_x = \Gamma_0 + \Gamma_1 + \Gamma_1'$, we can evaluate it at $b_0 = 0.5$ use it to construct the numerical optimal weighting matrix for our hypothetical MA(1):

$$W^* = \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 1.1115 & 1.5705 & 0.6823 \\ 0 & 1.5705 & 2.8621 & 1.3539 \\ 0 & 0.6823 & 1.3539 & 1.1400 \end{bmatrix} \tag{25}$$

Estimation

- To estimate the parameter vector, the population analogue of the SMM objective function as

$$J(b) = g(b)' W^* g(b)$$

- The first order condition is

$$\nabla_b \left(g(b)' W^* g(b) \right) = 0 \iff \nabla_b g(b)' W^* g(b) = 0$$

- The derivative of $g(b)$ (an $n \times \ell$ matrix) defined in (22) is:

$$\nabla_b g(b) = -\nabla_b \mu(y(b)) = - \begin{bmatrix} 0 \\ 2b \\ -1 \\ 0 \end{bmatrix} \qquad (26)$$

Estimation - cont.

- In that case we can write $\nabla_b g(b)' W^* g(b) = 0$ as

$$-\begin{bmatrix} 0 & 2b & -1 & 0 \end{bmatrix} \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 1.1115 & 1.5705 & 0.6823 \\ 0 & 1.5705 & 2.8621 & 1.3539 \\ 0 & 0.6823 & 1.3539 & 1.1400 \end{bmatrix} g(b) = 0$$

- If we evaluate $\nabla_b g(b)$ at $b = b_0$ and compute $\nabla_b g(b_0)' W^*$, then we obtain

$$\left[\begin{array}{cccc} 0 & 0.4590 & 1.2916 & 0.6715 \end{array}\right] g(b_0) = 0 \qquad (27)$$

which is one equation in one unknown b_0.

Estimation - cont.

- What does $\begin{bmatrix} 0 & 0.4590 & 1.2916 & 0.6715 \end{bmatrix} g(b) = 0$ imply?
- Recall this is an overidentified system. Further, recall from (22) that the mean (1) and 2nd order autocovariance (4) do not identify b, the variance (2) does not uniquely identify b, while the 1st order autocovariance (3) uniquely identifies b.
- The most weight is placed on (3) as it should.
- Zero weight is placed on (1) because it is not useful at all for the estimation of b.
- Then why does (4) receive positive weight?
  - Even though the 2nd autocovariance doesn't depend on b, it is correlated with the variance and 1st autocovariance, which is useful for the estimation of b.
  - That is, if we want to make the estimator efficient, we should take the information in the 2nd autocovariance into account.

Computing Standard Errors

- Recall, from Theorem 3.2 of Hansen to compute standard errors we need more than just the optimal weighting matrix.
- Specifically:

$$\sqrt{T}(b_T - b_0) \to N(0, \left[\nabla_b g(b_0)' W^* \nabla_b g(b_0)\right]^{-1})$$

- Hence, we need the derivative of g (an $n \times \ell$ matrix) defined in (22):

$$\begin{aligned} \nabla_b g(b_0) &= -\nabla_b \mu(y(b_0)) \\ &= -\begin{bmatrix} 0 \\ 2b_0 \\ -1 \\ 0 \end{bmatrix} \end{aligned} \tag{28}$$

Computing Standard Errors

- This derivative is also useful to see if the parameter is **Locally Identified**. To see this, take the first order approximation of $g(b)$ around b_0:

$$\begin{aligned} g(b) &\approx g(b_0) + \nabla_b g(b_0)(b - b_0) \\ &= \nabla_b g(b_0)(b - b_0) \end{aligned}$$

  since $g(b_0) = 0$.
- For $b = b_0$ to be the unique solution to $\nabla_b g(b_0)(b - b_0) = 0$, it must be true that

$$\text{rank}(\nabla_b g(b_0)) = \ell$$

- From (28), we can see that $\text{rank}(\nabla_b g(b_0)) = 1 = \ell$, so the parameter is locally identified.
- In contrast, if $\nabla_b g(b_0)$ in (28) was the zero vector (which has rank $0 < \ell$), then the objective would not respond to changes in the parameter.

Small Samples

- In general we never actually have analytical expressions for $\mu(x)$ and $\mu(y(b))$ so cannot obtain an estimate as above.

- That's why we will use SMM to estimate b in a finite sample.

- With a finite sample of size T data we must construct the $n \times 1$ vector of data moments $M_T(x)$.

- Given the finite sample, in general $M_T(x) \neq \mu(x)$, but $M_T(x) \overset{a.s.}{\to} \mu(x)$ as $T \to \infty$.

True Data Sample

- We first generate a series of random sample $\{\varepsilon\}_{t=1}^{T}$ from $N(0,1)$ and then construct a series of $\{x_t\}_{t=1}^{T}$ using the true DGP in (15) or $x_t = \varepsilon_t - b_0\varepsilon_{t-1}$ with $\varepsilon_0 = 0, b_0 = 0.5$, and $T = 200$.

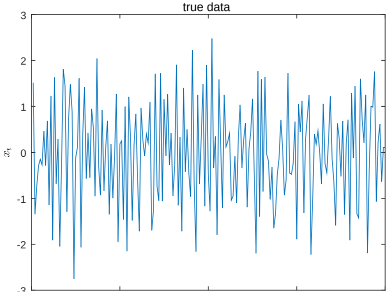


Figure: Simulated 'true' data

Sample moments for the true data

- Using this data, we can compute the $m \times 1$ data moment vector by

$$M_T(x) = \frac{1}{T}\sum_{t=1}^{T} m(x_t). \tag{29}$$

- For our case, the $m = 4$ data moment vector obtained from this simulation is

$$M_T(x) = \begin{bmatrix} -0.0153 \\ 1.1874 \\ -0.4269 \\ -0.0868 \end{bmatrix}. \tag{30}$$

- Recall the population moment matrix is

$$M_T(x) = \begin{bmatrix} 0 \\ 1.25 \\ -0.5 \\ 0 \end{bmatrix}.$$

Sample Var-Covar for the true data

- This data is used to estimate the sample var-covar matrix.
- Unlike the general SMM slides where it was assumed that there was no autocorrelation in $u(x_t, b)$, here we have autocorrelation so apply Newey-West correction in (24).
- Let

$$\widehat{\Gamma}_{T,j} \equiv \frac{1}{T} \sum_{t=j+1}^{T} \left[m(x_t) - M_T(x)\right]\left[m(x_{t-j}) - M_T(x)\right]'$$

  denote the j-th autocovariance of m.
- Then the estimated sample var-covar matrix of $m(x_t)$ is

$$\widehat{S}_{x,T} = \widehat{\Gamma}_{T,0} + \sum_{j=1}^{\infty} \left(1 - \frac{j}{i(T)+1}\right)(\widehat{\Gamma}_{T,j} + \widehat{\Gamma}'_{T,j})$$

  where $i(T)$ is the key to the Newey-West correction (here taken to be 4).

Sample Var-Covar for the true data - cont.

The sample var-covar matrix is given by

$$\hat{S}_{x,T} = \begin{bmatrix} 0.4147 & 0.0058 & -0.0895 & -0.0244 \\ 0.0058 & 1.8946 & -0.8869 & -0.1872 \\ -0.0895 & -0.8869 & 1.2988 & -0.6078 \\ -0.0244 & -0.1872 & -0.6078 & 1.5729 \end{bmatrix} \tag{31}$$

while recall that the population var-covar matrix is given by

$$S_x = \begin{bmatrix} 0.25 & 0 & 0 & 0 \\ 0 & 4.125 & -2.5 & 0.5 \\ 0 & -2.5 & 2.3125 & -1.25 \\ 0 & 0.5 & -1.25 & 2.0625 \end{bmatrix}$$

SMM Estimation

- Next we consider the model moments.
- We first draw a series of random sample $\{\{e_t^h\}_{t=1}^T\}_{h=1}^H$.
- We will use the same draw in the whole estimation process.
- Given parameter value b, we can compute $\{\{y_t^i(b)\}_{t=1}^T\}_{h=1}^H$ in (16) or

  $$y_t^h = e_t^h - be_{t-1}^h$$

  where $e_0^h = 0,\ T = 200$, and $H = 10$.
- Then given b, we can compute the simulated model moment

$$M_{TH}(y(b)) = \frac{1}{H}\sum_{i=1}^{H}\frac{1}{T}\sum_{t=1}^{T} m(y_t^h(b)). \tag{32}$$

SMM Estimation - cont.

- Our objective is to choose b so that the weighted sum of squared residuals between the model moments $M_{TH}(y(b))$ in (32) and data moments $M_T(x)$ in (29) is minimized.
- The estimation procedure depends on whether we have the data to form the optimal weighting matrix $\hat{S}_{x,T}^{-1}$ as in (31) or we only have data moments so we cannot directly estimate the var-covar matrix from the data.
- The consistent estimate of the parameter b solves

$$\hat{b}_{TH} = \arg\min_{b} J_{TH}(b) \tag{33}$$

where

$$J_{TH}(b) \equiv [M_T(x) - M_{TH}(y(b))]' W [M_T(x) - M_{TH}(y(b))].$$

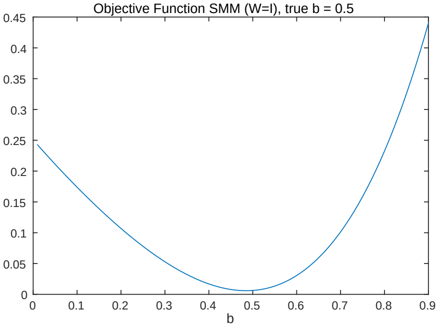
SMM Estimation - cont.

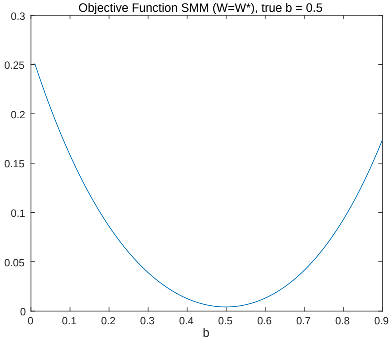
The SMM estimate that solves (33) depends on W:

- In the former case where $W^* = \hat{S}_{x,T}^{-1}$, the consistent and efficient estimate of b is $\hat{b}_T^* = 0.4993$.
- In the latter case where we don't have data to estimate $\hat{S}_{x,T}^{-1}$, we can use a two-step procedure.
  - In the first stage where we use $W = I$ second stage, we obtain a consistent estimate of $\hat{b}_{TH}^1 = 0.4850$.
  - We can then use $\hat{b}_{TH}^1$ to generate a model equivalent of $\hat{S}_{y,TH}^{-1}$ to obtain a consistent, efficient estimate $\hat{b}_{y,TH}^2 = 0.4970$.

▸ Small Sample Weighting Matrices

Comparing Data and Model Weighted Sum of Squared Errors





Standard Errors

- Once we have computed the point estimate, we want to compute the standard errors of the estimate.
- In the case of the simulated weighted matrix, we have

$$\sqrt{T}(\widehat{b}^2_{TH} - b_0) \to N(0, (1+1/H)\left[\nabla_b g_T(\widehat{b}^2_{TH})' \widehat{S}^{-1}_{y,TH} \nabla_b g_T(\widehat{b}^2_{TH})\right]^{-1})$$

where

$$\begin{aligned} g_T(b) &\equiv \frac{1}{T}\sum_{t=1}^{T} u(x_t, b) = \frac{1}{T}\sum_{t=1}^{T} m(x_t) - \frac{1}{H}\sum_{h=1}^{H}\frac{1}{T}\sum_{t=1}^{T} m(y^h_t(b)) \\ &= M_T(x) - M_{TH}(y(b)) \end{aligned}$$

- The derivative of g_T is given by

$$\begin{aligned} \nabla_b g_T(\hat{b}^2_{TH}) &= -\nabla_b M_{TH}(y(\hat{b}^2_{TH})) \\ &= -\frac{1}{TH}\sum_{h=1}^{H}\sum_{t=1}^{H}\frac{\partial m(y^h_t(\hat{b}^2_{TH}))}{\partial b} \end{aligned}$$

Standard Errors - cont.

- In general we don't have an analytical formula for this derivative, so we will use the numerical derivative.
- Once can compute $M_{TH}(y(\hat{b}^2_{TH}))$, then compute $M_{TH}(y(\hat{b}^2_{TH} - s))$, take the difference, and divide by the step size s.
- The result is

$$\frac{\Delta M_{TH}(\hat{b}^2_{TH})}{\Delta b} = \begin{bmatrix} -0.0104 \\ 0.9342 \\ -0.9330 \\ -0.0234 \end{bmatrix}$$

- Since there is small sample error, this numerical derivative is broadly consistent with the theoretical result computed in (28) evaluated at $b_0 = 0.5$ given by $[0\ 1\ \ -1\ 0]'$.

Standard Errors - cont.

- The standard error of the $\hat{S}_{y,TH}^{-1}$ weighted estimator is then

$$\text{Std}(\hat{b}_{TH}^2) = \sqrt{\frac{1}{T}\left[\nabla_b g_T(\hat{b}_{TH}^2)' \left\{\left(1+\frac{1}{H}\right)\hat{S}_{y,TH}\right\}^{-1} \nabla_b g_T(\hat{b}_{TH}^2)\right]^{-1}} = 0.089.$$

- The standard error of the $\hat{S}_{x,T}^{-1}$ weighted estimator is

$$\text{Std}(\hat{b}_T^*) = \sqrt{\frac{1}{T}\left[\nabla_b g_T(\hat{b}_T^*)' \left\{\hat{S}_{x,T}\right\}^{-1} \nabla_b g_T(\hat{b}_T^*)\right]^{-1}} = 0.075.$$

J-Test

- Once we have estimated the parameter, we can also test if the moment condition is true or not.

$$T\frac{H}{1+H}\times[M_T(x)-M_{TH}(y(\hat{b}^2_{TH}))]'W^*_{TH}[M_T(x)-M_{TH}(y(\hat{b}^2_{TH}))] = 0.7588.$$

- The asymptotic distribution of this test statistics is $\chi(n-k)$, where n is the number of moments $(=4)$ and k is the number of parameters $(=1)$.

- The p value is 0.14, so we cannot reject the hypothesis that the model is true.

Bootstrap

- In order to see the finite sample distribution of the estimators, we can use the bootstrap method. The algorithm is as follows.

  1. Draw ε_t and e_t^h from $N(0,1)$ for $t = 1, 2, \ldots, T$ and $h = 1, 2, \ldots, H$. Compute $(\hat{b}_{TH}^1, \hat{b}_{TH,data}^2, \hat{b}_{TH,sim}^2)$ as described.
  2. Repeat 1 using another seed.

- Every time you do step 1, the seed needs to change (which is done automatically by matlab if you don't specify it). Otherwise you will keep getting the same estimators.

Bootstrap - cont.

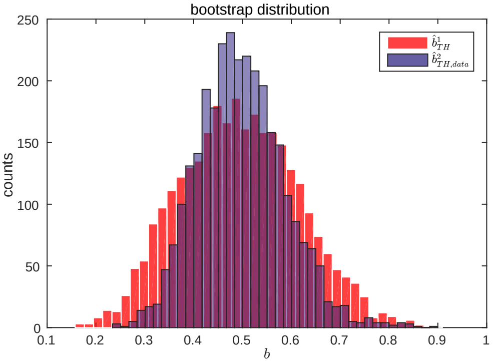


Figure: Bootstrap distributions: histogram

- The histogram of the estimator is plotted in figure 4.
- As theory predicts, $\hat{b}^*_T$, which is the efficient estimator, has a smaller variance than $\hat{b}^1_{y,TH}$.

Bootstrap - cont.

- To make it easier for us to compare the distributions, figure 5 plots the density function of the estimators, obtained by Kernel density estimation

$$f(x) = \frac{1}{\sqrt{2\pi}} \frac{1}{cI} \sum_{i=1}^{I} \exp\left[-\frac{1}{2}\left(\frac{x - x_i}{c}\right)^2\right]$$

  where I is the number of data and c is the bandwidth.
- We can see that the distribution of $\hat{b}^*_{x,T}$ looks very similar to that of $\hat{b}^2_{y,TH}$.
- This is because the model nests the true DGP (in the sense that it is the true DGP at b_0), so even if we use the simulated data to estimate the variance-covariance matrix, we can obtain the efficient estimator.

Bootstrap - cont.

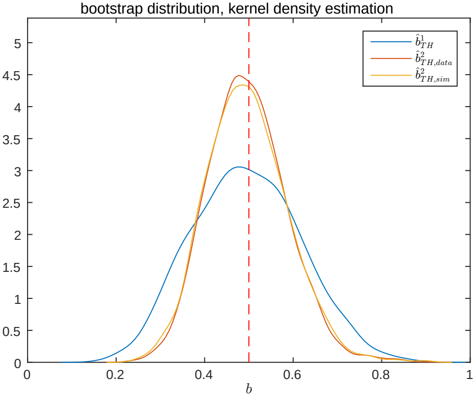


Figure: Bootstrap distributions, approximated by the kernel density estimation.

Summary

- If you know GMM, you know MSM.
- MSM is super useful if
  - you only have access to empirical studies providing sample moments (e.g. hard to get special sworn status into the Census data)
  - it is difficult to derive or evaluate the moment conditions (e.g. Euler equation does not hold for those who are borrowing constrained) or full likelihood function
- Indirect Inference also useful if you have an auxiliary model (e.g. a reduced form regression) to match and has many other advantages as will be explained by Fatih Guvenen via zoom at a future date.

Comparing Data and Model Simulated Weighting Matrices

$$\hat{S}^{-1}_{x,T} = \begin{bmatrix} 2.5146 & 0.2155 & 0.4282 & 0.2301 \\ 0.2155 & 1.0110 & 0.9316 & 0.4836 \\ 0.4282 & 0.9316 & 1.8197 & 0.8206 \\ 0.2301 & 0.4836 & 0.8206 & 1.0140 \end{bmatrix}$$

$$\hat{S}^{-1}_{y,TH} = \begin{bmatrix} 2.2441 & 0.0369 & 0.0023 & 0.0395 \\ 0.0369 & 0.8928 & 1.1272 & 0.4225 \\ 0.0023 & 1.1272 & 2.1335 & 0.9009 \\ 0.0395 & 0.4225 & 0.9009 & 0.9688 \end{bmatrix}.$$

▸ Back