Empirical Methods in Finance MFE230E
Week 4: GMM

Martin Lettau

Spring 2019

Haas School of Business

General Method of Moments (GMM)

A very versatile estimation method

Agenda

- Review of classical Method of Moments (MM) estimator
- Motivation of General Method of Moments (GMM) estimator
- Asymptotic properties of GMM estimator
- Small sample properties of GMM
- Examples

Pre-reading: Lecture notes Pre-Program Stats section on Method of Moments, DGS section 7.6

Outline

Example 1: χ^2 distribution

If $Z_1, \ldots, Z_k$ are independent, standard normal RVS, then

$$X = \sum_{i=1}^{k} Z_i^2 \sim \chi^2(k).$$

Suppose we have an iid sample $(x_1, \ldots, x_T)$.

Question: How can we estimate k?

Alternative to MLE: Method of Moments

The first 2 moments are

$$m_1(k) \equiv \mathrm{E}(X) = k$$

$$m_2(k) \equiv \mathrm{E}(X^2) = k(k+2)$$

The sample estimate of the first moments is

$$\widehat{m}_1 \equiv \frac{1}{T}\sum_{t=1}^{T} x_t$$

So, the MM estimator of k using the first moment is

$$\widehat{m}_1 \equiv \frac{1}{T}\sum_{t=1}^{T} x_t = \widehat{k}.$$

MM NOTATION

We will write the MM estimator as follows:

The **moment condition** is

$$m_1(k) \equiv \mathrm{E}(X) = k.$$

The sample equivalent of the moment condition is

$$\widehat{m}_1 \equiv \frac{1}{T}\sum_{t=1}^{T} x_t = \widehat{k}.$$

Write the MM estimator so that the moment condition is equal to zero:

Population MC: $m_1(k) \equiv \mathrm{E}(X) - k = 0$

Sample MC: $\widehat{m}_1(\widehat{k}) \equiv \frac{1}{T}\sum_{t=1}^{T} x_t - \widehat{k} = 0$

Note that we could also have used the second moment to estimate k:

$$m_2(k) \equiv \mathrm{E}(X^2) = k(k+2)$$

The sample equivalent of the second moments is

$$\widehat{m}_2 \equiv \frac{1}{T}\sum_{t=1}^{T} x_t^2$$

So, the MM estimator of k using the second moment $\widehat{k}$ solves

$$\widehat{m}_2 \equiv \frac{1}{T}\sum_{t=1}^{T} x_t^2 - \widehat{k}(\widehat{k}+2) = 0.$$

$\hat{k}$ would solve this equation
different moments lead to $\neq$ estimators.

Key insight: The MM can be applied to *any* type of moment condition.

In the χ^2 example, we exploited the fact that we can derive the moments in closed-from.

But a moment condition can take various forms.

Important example: Univariate OLS regression

$$y_i = x_i\beta + e_i$$

OLS condition: x_i is orthogonal to e_i

$$\mathrm{E}[x_i e_i] = 0$$

This is also a moment condition!!

Example: OLS as GMM estimator

Sample analog:

$$\frac{1}{N}\sum_i x_i \widehat{e}_i = 0$$

$$\frac{1}{N}\sum_i x_i (y_i - x_i \widehat{\beta}_{MM}) = 0$$

$$\widehat{\beta}_{mm} = \left(\sum_i x_i x_i\right)^{-1} \left(\sum_i x_i y_i\right)$$

The MM estimator is identical to the OLS estimator!

Or, more precisely: OLS is an example of a moment estimator!

OLS IS A SPECIAL CASE OF MM!

It turns out that OLS is a special case of MM!

Moment estimators are a much more general class of estimators than OLS.

General formulation: Let $\mathbf{z}_t$ be a vector of (stationary) data.

- In the χ^2 example: $\mathbf{z}_t = (x_t)$
- in the OLS example; $\mathbf{z}_t = (x_t, y_t)'$ data matrix

Let $\boldsymbol{\theta}$ be the vector of parameters to be estimated.

The moment condition(s) can be written as parameter we want to estimate

moment condition

$$E[\mathbf{g}(\boldsymbol{\theta}, \mathbf{z}_t)] = 0,$$

data

- χ_k^2 example with first moment: $\mathbf{g}(k, x_t) = x_t - k$
- χ_k^2 example with second moment: $\mathbf{g}(k, x_t) = x_t^2 - k(k+2)$
- Univariate OLS: $\mathbf{g}(\beta, (x_i, y_i)) = x_i e_i = x_i(y_i - x_i\beta)$

The MM estimator sets the sample mean of the moment conditions to 0:

$$\frac{1}{T}\sum_t \mathbf{g}(\widehat{\boldsymbol{\theta}}, \mathbf{z}_t) = 0$$

Note: So far, one moment condition and one parameter. → only have the mean or variance for now

- Multivariate OLS where $\mathbf{x}_t$ and $\boldsymbol{\beta}$ are $k \times 1$ vectors

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + e_i$$

- Now we have k moment conditions:

$$\mathrm{E}[g(\boldsymbol{\beta}; (\mathbf{x}_i, y_i))] = \mathrm{E}[\mathbf{x}_i e_i] = 0$$

x is a vector

- $\mathbf{x}_i$ is a $\underline{k \times 1}$ vector $\rightarrow$ k moment conditions for k parameter.

system of k moment conditions

- Sample analog:

$$\frac{1}{N} \sum_i \mathbf{x}_i' (y_i - \mathbf{x}_i' \widehat{\boldsymbol{\beta}}) = 0$$

k linear equation

k parameters

$$\widehat{\boldsymbol{\beta}}_{mm} = \left(\sum_i \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \left(\sum_i \mathbf{x}_i y_i \right)$$

- MM: Same number of moment conditions as we have parameters.

Recall: If the assumption $E[\mathbf{x}_i e_i] = 0$ is violated, OLS is inconsistent.

Instrument $\mathbf{z}_i$ that is correlated with $\mathbf{x}_i$ but uncorrelated with e_i.

The MM moment condition is

$$E(\mathbf{z}_i e_i) = 0.$$

Sample analog:

$$\forall i, \quad \frac{1}{N}\sum_i \mathbf{z}_i'(y_i - \mathbf{x}_i'\widehat{\beta}) = 0$$

$$\widehat{\beta}_{mm} = \left(\sum_i \mathbf{z}_i \mathbf{x}_i'\right)^{-1} \left(\sum_i \mathbf{z}_i y_i\right),$$

which is equal to the IV regressor.

MLE as GMM

Log-likelihood function:

$$\log L = \frac{1}{n}\sum_{i=1}^{n} \log f(y_i|x_i;\boldsymbol{\theta})$$

Population expectation of FOC:

$$\mathrm{E}\left[\frac{\partial \log L}{\partial \theta_k}\right] = 0$$

GMM sample equivalent

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \log f(y_i|x_i;\boldsymbol{\theta})}{\partial \theta_k} = 0$$

K nonlinear equations with K unknowns

MM TWO-STAGE ESTIMATION WITH MEASUREMENT ERROR

- Later we will consider estimation of factor models, e.g. CAPM.
- Some factor models need to be estimated in 2 steps
- Suppose we have returns of N different stocks
  1. Estimation of the "beta" of each stock i:

$$R_{i,t} = \beta_i R_{M,t} + e_{i,t}$$

  2. Estimation of the "factor risk premium" with $\widehat{\beta}_i$ as an explanatory variable.

$$\mu_i = \underset{\substack{\uparrow \\ \text{market risk} \\ \text{premium}}}{\lambda} \widehat{\beta}_i + u_i, \quad i = 1, \ldots N,$$

  where $\mu_i = \mathrm{E}[R_i]$
- Problem: Stage 1 estimation of $\widehat{\beta}_i$ has estimation error that needs to be taken into account in the stage 2 estimation

- Solution: Stack both equations together and estimate $\theta = (\beta, \lambda)$ simultaneously

$$\mathbf{g}(\theta) = \mathrm{E}\begin{bmatrix} e_t R_{M,t} \\ u_i \beta_i \end{bmatrix} = \mathrm{E}\begin{bmatrix} (R_{i,t} - \beta R_{M,t}) R_{M,t} \\ (\mu_i - \beta_i \lambda)\beta_i \end{bmatrix} = 0$$

- Details later but this GMM system takes the effect of estimation uncertainty of $\widehat{\beta}_i$ on the standard error of $\widehat{\lambda}$ correctly into account.
- GMM is useful for many multi-step estimation method

Theorem 1 (Properties of Method of Moments estimator).

Let $\mathbf{g}(\boldsymbol{\theta}, \mathbf{z}_t)$ be a $k \times 1$ vector of moment conditions, where $\boldsymbol{\theta}$ is a $k \times 1$ vector of parameters and $\mathbf{z}_t$ is a sequence of stationary data.

The MM estimator $\widehat{\boldsymbol{\theta}}$ sets the sample moments conditions to zero:

$$\frac{1}{T}\sum_t \mathbf{g}(\widehat{\boldsymbol{\theta}}, \mathbf{z}_t) = 0$$

The mean of conditions

Under some mild technical conditions, the MM estimator has the following properties:

moments exist

1. *The estimator is consistent: $\widehat{\boldsymbol{\theta}}_n \xrightarrow{P} \boldsymbol{\theta}$.*
2. *The estimator is asymptotically normal:*

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \rightarrow N(0, \mathbf{V})$$

Outline

From MM to GMM

So far: # of moment conditions = # of parameters
$\Rightarrow$ MM is **exactly identified**

If there are more parameters than moment conditions
$\Rightarrow$ MM is **under-identified** and the parameters cannot be estimated

What if # of moment conditions $>$ # of parameters
$\Rightarrow$ MM is **over-identified** $\Rightarrow$ GMM

From MM to GMM

- GMM is a versatile estimation method
- Consistent and asymptotically normal under mild assumptions: Stationarity
- Trick: Convert a given problem into moment conditions
- Once an estimator is written as a moment condition, we know it is consistent and asymptotically normal
- Limitation in practice: GMM can have poor small sample properties
- Cochrane's *Asset Pricing* textbook frames most of asset pricing into the GMM framework!
- Hayashi's *Econometrics* textbook includes a very good, if somewhat advanced, treatment of GMM

stochastic discount factor

$$E_t[M_{t+1}R_{t+1}] = 1$$

→ moment condition

Recall the χ^2_k example: $X \sim \chi^2_k$

The first 2 moments are

$$m_1(k) \equiv \mathrm{E}(X) = k$$
$$m_2(k) \equiv \mathrm{E}(X^2) = k(k+2)$$

In MM, we used **either the first or second moment** to estimate k.

EXAMPLE 1, CONT'D

- **Generalized method of moments (GMM)**: Combine moment conditions to estimate k.
- In MM, we have as many moment conditions as parameters. Hence, we can pick the parameters to set the moment conditions to exactly zero.
- If we have more moment conditions than parameters, not all moment conditions can be exactly satisfied.
- **GMM: Pick the parameters that minimize a *weighted average of the moment conditions*.**

In the χ^2_k example, the moment conditions were:

$$m_1(k) \equiv \mathrm{E}(X) - k = 0$$
$$m_2(k) \equiv \mathrm{E}(X^2) - k(k+2) = 0$$

Let $\mathbf{m}(k) = (m_1(k), m_2(k))'$ be the vector of moment conditions and let $\mathbf{W}$ be a symmetric positive semidefinite matrix

$$\mathbf{W} = \begin{pmatrix} w_{11} & w_{12} \\ w_{12} & w_{22} \end{pmatrix}$$

GMM SETUP

The GMM estimator of k minimizes the quadratic form

$$\widehat{k}_{GMM} = \underset{k}{\operatorname{argmin}}\, \mathbf{m}(k)'\, \mathbf{W}\, \mathbf{m}(k)$$

$$\mathbf{m}(k)'\, \mathbf{W}\, \mathbf{m}(k) = w_{11} m_1(k)^2 + 2 w_{12}\, m_1(k)\, m_2(k) + w_{22} m_2(k)^2$$

m_i moment conditions, not moment

Note that the GMM estimator is the same for $\mathbf{W}$ and $\alpha\mathbf{W}$.

Example 1, cont'd.

Pop. moments:

$$m_1(k) \equiv \mathrm{E}(X) - k = 0$$

$$m_2(k) \equiv \mathrm{E}(X^2) - k(k+2) = 0$$

Sample equivalents:

$$\widehat{m}_1(k) \equiv \frac{1}{T}\sum_t x_t - k = 0$$

$$\widehat{m}_2(k) \equiv \frac{1}{T}\sum_t x_t^2 - k(k+2) = 0$$

Given sample

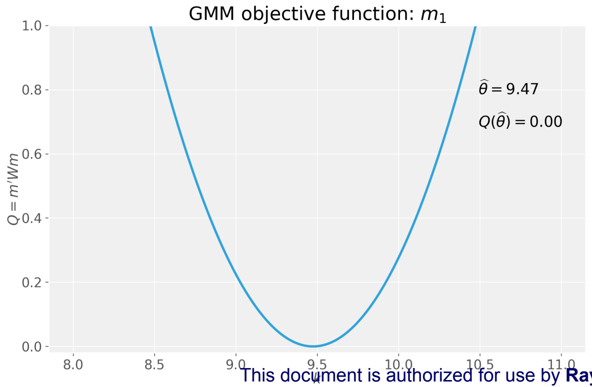
$$\widehat{m}_1(k) \equiv \frac{1}{T}\sum_t x_t = 9.47$$

$$\widehat{m}_2(k) \equiv \frac{1}{T}\sum_t x_t^2 = 104.18$$

MM estimator for $\mathrm{E}(X) - k = 0 \Rightarrow \widehat{k} = 9.47$. Special case of GMM with

$\tilde{m}(k)$

$$\mathbf{W} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$

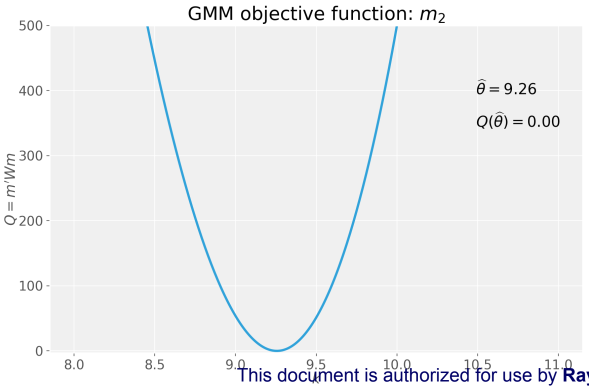


Example 1, cont'd.

MM estimator for $\mathrm{E}(X^2) - k(k+2) = 0 \Rightarrow \widehat{k} = 9.26$. Special case of GMM with

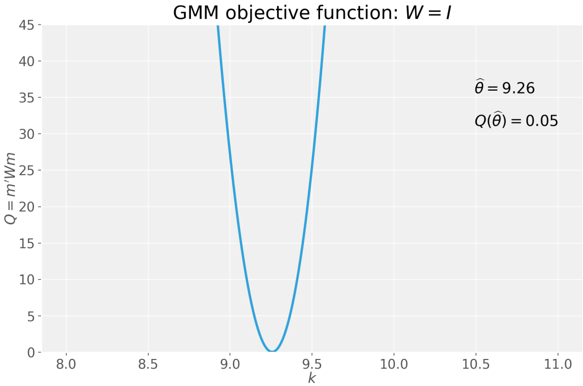
$\hat{m}_2(k)$

$$\mathbf{W} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$$



Example 1, cont'd.

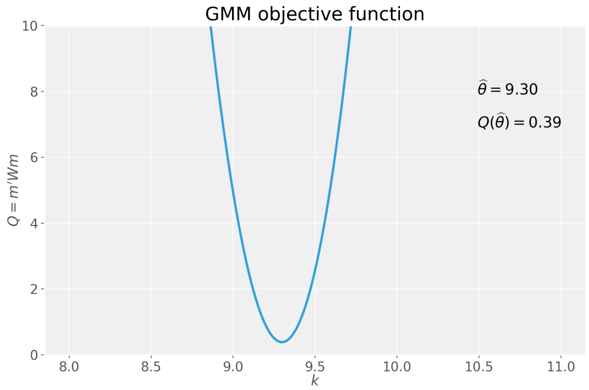
$$\mathbf{W} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$



EXAMPLE 1, CONT'D.

$$\mathbf{W} = \begin{pmatrix} 10 & 0 \\ 0 & 1/10 \end{pmatrix}$$

more weight on 1st moment



Example 1, cont'd.

Question: How to pick the weighting matrix **W**?

Intuition: Put more weight on moments that are "more informative" about the true $\boldsymbol{\theta}$.

As in MLE, "informativeness" is linked to the curvatures of moment conditions. variance of moments

BUT: Curvature cannot be assessed by looking at the plots above since the moments have different "units".

The first moment is in terms of k while the second moment is in terms of k^2.

GMM

- GMM is a very powerful way of looking at an estimation problem.
- All we need is a moment condition that holds.
- The problem does not have to be linear.
- No distributional assumptions are needed.
- We can use GMM to estimate
  - the non-linearized version of the Consumption CAPM.
  - nonlinear processes, such as ARCH, GARCH, etc.
  - interest rate models

GMM

- Practical Considerations:
  - We need at least as many conditions as parameters (just-identified case) wha
  - If there are more moments, they can be used to test the model (J-test). test the null of $g = 0$.
  - Too many moments are not desirable in practice.
  - The conditioning information matters (what variables are included in the moments–as with other estimators).
  - How to pick the “best” weighting matrix **W**?

Next: The general GMM methodology following Hansen (1982)

Note: The idea is rather simple, so don’t get confused by the sometimes complex notation!

Outline

THE GMM METHODOLOGY (USING HANSEN'S NOTATION)

- Suppose you have an economic model which implies a set of r moment conditions that take the form

g?

$$\mathrm{E}\left[\boldsymbol{h}(\boldsymbol{\theta}, \mathbf{z}_t)\right] = 0,$$

where $\mathbf{z}_t$ is an $l \times 1$ vector of variables known at time t, and $\boldsymbol{\theta}$ is an $k \times 1$ vector of coefficients we seek to estimate.

- This is a population mean. The sample equivalent is

$$\boldsymbol{g}(\boldsymbol{\theta}, \mathbf{Z}_T) \equiv \frac{1}{T}\sum_{t=1}^{T}\mathbf{h}(\boldsymbol{\theta}, \mathbf{z}_t),$$

where $\mathbf{z}_T \equiv \left(\mathbf{z}_1', \mathbf{z}_2', \ldots \mathbf{z}_T'\right)'$.

The GMM estimator of $\boldsymbol{\theta}$ is the value of $\boldsymbol{\theta}$ that minimizes the scalar

$$Q(\boldsymbol{\theta}; \mathbf{z}_T) = \left[\mathbf{g}(\boldsymbol{\theta}; \mathbf{z}_T)\right]' \mathbf{W} \left[\mathbf{g}(\boldsymbol{\theta}; \mathbf{z}_T)\right] \qquad (*)$$

where $\mathbf{W}$ is a $r \times r$ positive definite weighting matrix.

g are the moment conditions

$\begin{pmatrix} m_1(k) - k \\ m_2(k) - E(k^2) \ldots \end{pmatrix}$

Classical Method of Moments:

If $r = k$ (the number of parameters to be estimated is equal to the number of moment conditions), then typically the objective function $(*)$ will be minimized by setting

$$\mathbf{g}(\widehat{\boldsymbol{\theta}}^{MM}, \mathbf{Z}_T) = \mathbf{0}$$

General Method of Moments:

If $r > k$ we cannot set all moment conditions to exactly zero.

Instead:

$$\widehat{\boldsymbol{\theta}}^{gmm} = \operatorname{argmin} \mathbf{g}(\boldsymbol{\theta}; \mathbf{Z}_T)' \, \mathbf{W} \, \mathbf{g}(\boldsymbol{\theta}; \mathbf{Z}_T)$$

The quadratic form can be minimized with respect to $\boldsymbol{\theta}$ using analytic or numerical methods.

Theorem 2 (Hansen (1982)).

If $\mathbf{z}_t$ are strictly stationary:

$$\mathbf{g}(\boldsymbol{\theta}; \mathbf{Z}_T) \xrightarrow{p} E\left[\mathbf{h}(\boldsymbol{\theta}; \mathbf{z}_t)\right].$$

If $\mathbf{h}(\theta, \mathbf{z}_t)$ is continuous in θ, then the GMM estimator is consistent:

$$\widehat{\boldsymbol{\theta}}^{gmm} \xrightarrow{p} \boldsymbol{\theta}_0$$

- GMM is consistent for any positive semidefinite weighting matrix!
- How should we choose the weighting matrix?
- What is the asymptotic distribution of the GMM estimator?

Example: Consumption CAPM I

In the MFE Investments class, you derived the stochastic discount factor M_{t+1} that prices an asset j with payoff $X_{j,t+1}$:

$$P_{j,t} = \mathrm{E}_t\left[M_{t+1}X_{j,t+1}\right]$$

$$R_{j,t+1} = \frac{X_{j,t+1}}{P_{j,t}} = \frac{P_{j,t+1} + D_{j,t+1}}{P_{j,t}}$$

$$\Rightarrow \ \mathrm{E}_t\left[M_{t+1}R_{j,t+1}\right] = 1$$

Consumption CAPM: M_{t+1} depends consumption C_t and preferences $U(C_t)$:

$$\mathrm{E}_t\left[\beta \frac{U'(C_{t+1})}{U'(C_t)} R_{j,t+1}\right] = 1$$

Example: Consumption CAPM II

Assumption: Utility function is of **constant relative risks aversion** form

$$U(C_t) = \beta^t \frac{C_t^{1-\gamma}}{1-\gamma}$$

$$U'(C_t) = \beta^t C_t^{-\gamma}$$

$$\mathrm{E}_t \left[\beta \left(\frac{C_{t+1}}{C_t} \right)^{-\gamma} R_{j,t+1} - 1 \right] = 0$$

⇒ This is a **moment condition**!

Example: Consumption CAPM III

We can test whether this equation holds in the data:

- Returns $R_{j,t+1} = (P_{j,t+1} + D_{j,t+1})/P_{j,t}$ of $j = 1, \ldots, J$ assets
- Data on aggregate consumption C_t

$$\mathrm{E}\left[\mathbf{h}\left(\boldsymbol{\theta}, \mathbf{z}_t\right)\right] = 0$$

$$\mathbf{h}\left(\theta, \mathbf{z}_t\right) = \left(h_1(\boldsymbol{\theta}, z_t), \ldots, h_J(\boldsymbol{\theta}, z_t)\right)'$$

$$\Rightarrow h_j(\boldsymbol{\theta}, \mathbf{z}_t) = \beta\left(\frac{C_{t+1}}{C_t}\right)^{-\gamma} R_{j,t+1} - 1$$

moment condition!

$$\boldsymbol{\theta} = (\beta, \gamma)'$$

$$\mathbf{z}_t = \left(R_{1,t}, \ldots, R_{J,t}, C_t\right)'$$

- There are J moment conditions (one for each asset) and 2 parameters.

The role of the weighting matrix

One possibility: $\mathbf{W} = \mathbf{I}$, in other words, all moments have the same weight

Hansen (1982): Optimal weighting matrix

Recall GLS: Observations weighted according to their variance

Same idea here: Put more weight on moments whose variance is smaller

If the data is iid:

$$\mathbf{W} = \mathbf{S}^{-1}$$

$$\mathbf{S} \equiv \mathrm{E}\left[\mathbf{h}(\widehat{\boldsymbol{\theta}}, \mathbf{z}_t)\mathbf{h}(\widehat{\boldsymbol{\theta}}, \mathbf{z}_t)'\right]$$

$$= \mathrm{Var}\left[\mathbf{h}(\mathbf{z}_t; \widehat{\boldsymbol{\theta}})\right].$$

The sample equivalent is

$$\widehat{\mathbf{W}} = \widehat{\mathbf{S}}^{-1}$$

$$\boxed{\widehat{\mathbf{S}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{h}(\widehat{\boldsymbol{\theta}}, \mathbf{z}_t)\, \mathbf{h}(\widehat{\boldsymbol{\theta}}, \mathbf{z}_t)'}$$

EXAMPLE: χ^2 DISTRIBUTION I

If $X_1, \ldots, X_k$ are independent and standard normal RVS, then

$$Z = \sum_{i=1}^{k} X_i^2 \sim \chi^2(k).$$

The first 4 moments are

$$m_1(k) = k$$
$$m_2(k) = k(k+2)$$
$$m_3(k) = k(k+2)(k+4)$$
$$m_4(k) = k(k+2)(k+4)(k+6)$$

Suppose we have a sample of $\mathbf{z} = (z_1, \ldots, z_T)'$ observations.

The moment conditions for the first two moments are

$$g_1(k, \mathbf{z}) \equiv m_1(k) - k = 0,$$

$$g_2(k, \mathbf{z}) \equiv m_2(k) - k(k+2) = 0$$

$m_1(k)$ = psudo moments

The GMM estimator using the first 2 moments and $\mathbf{W} = \mathbf{I}$ minimizes

$$\begin{bmatrix} \widehat{m}_1(\mathbf{z}) - k \\ \widehat{m}_2(\mathbf{z}) - k(k+2) \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \widehat{m}_1(\mathbf{z}) - k \\ \widehat{m}_2(\mathbf{z}) - k(k+2) \end{bmatrix} \; 2$$

$$\left(m_1(z) - k\right)^2 + \left[m_2(z) - k(k+2)\right]^2$$

$z_t^2 - m_2(k))$

Next, let's compute GMM using the optimal $\mathbf{W} = \mathbf{S}^{-1}$. $(z_t - m_1(k)$

$$\mathbf{S} \equiv \begin{bmatrix} S_{11} & S_{12} \\ S_{12} & S_{22} \end{bmatrix} = \mathrm{E}\begin{bmatrix} z_t - m_1(k) \\ z_t^2 - m_2(k) \end{bmatrix}\begin{bmatrix} z_t - m_1(k) \\ z_t^2 - m_2(k) \end{bmatrix}^T$$

$z_t^2 - 2z_t m_1(k) + m_1^2(k)$

$$\begin{aligned} S_{11} &= Ez_t^2 - 2m_1^2 + m_1^2 &&= k(k+2) - 2k^2 + k^2 &&= 2k \\ S_{12} &= m_3(k) - m_1(k)m_2(k) &&= k(k+2)(k+4) - k^2(k+2) &&= 4k(k+2) \\ S_{22} &= m_4(k) - [m_2(k)]^2 &&= k(k+2)(k+4)(k+6) - k^2(k+2)^2 &&= 8k(k+2)(k+3) \end{aligned}$$

$(z_t - m_1)(z_t^2 - m_2) = m_3 - \cancel{m_2} z + \cancel{m_1} z^2 - m_1 m_2$

$m_2 = E[X^2] E[X]$

Recall earlier example

$$\widehat{m}_1(k) \equiv \frac{1}{T}\sum_t x_t = 9.47$$

$$\widehat{m}_2(k) \equiv \frac{1}{T}\sum_t x_t^2 = 104.18$$

For $\mathbf{W} = \mathbf{I} : \widehat{k} = 9.26$

Compute $\mathbf{S}$ for $k = 9.26$:

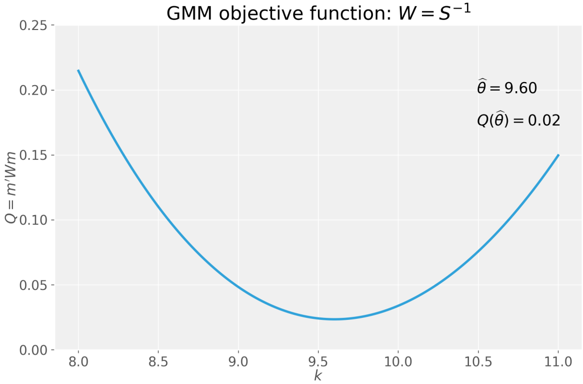
$$\mathbf{S} = \begin{pmatrix} 18.51 & 416.78 \\ 416.78 & 10216.43 \end{pmatrix}$$

$$\mathbf{W} = \mathbf{S}^{-1} = \begin{pmatrix} 0.66 & -0.03 \\ -0.03 & 0.01 \end{pmatrix}$$

$\mathbf{W} = \mathbf{I}$ and $\mathbf{W} = \mathbf{S}^{-1}$ are very different. Optimal GMM puts almost all the weight on the first moment.

Example 1: Optimal W I

$$\mathbf{W} = \begin{pmatrix} 0.66 & -0.03 \\ -0.03 & 0.01 \end{pmatrix}$$

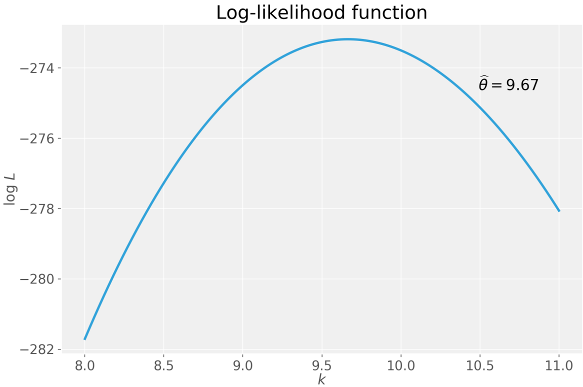


Note: The optimal weighting matrix depends on k, which we don't know in practice!

Example: Use $\mathbf{W} = \mathbf{I}$ first, and then use $\widehat{k}$ to compute $\mathbf{W} = \mathbf{S}^{-1}$ and compute new $\widehat{k}$, and so on:

	k	Q
0	9.26	0.05
1	9.60	0.02
2	9.65	0.02
3	9.65	0.02
4	9.65	0.02

Comparison: MLE



The data was from a $\chi^2(10)$ with sample size of 50.

- In practice, **S** can usually not be computed analytically and has to be estimated.
- Many different estimators for **S** have been proposed. The most popular one is the **Newey-West** (1987) estimator:

$$\widehat{\mathbf{S}} = \sum_{j=-q}^{q} \left(\frac{q - |j|}{q} \right) \frac{1}{T} \sum_{t=q+1}^{T-q} \left[\mathbf{h}\left(\widehat{\theta}, \mathbf{z}_t\right) \right] \left[\mathbf{h}\left(\widehat{\theta}, \mathbf{z}_{t-j}\right) \right]'.$$

- Note:
  - Down-weights higher-order autocorrelations
  - Only autocorrelations up to lag q are used
  - q must be chosen ex ante
  - $\widehat{\mathbf{S}}$ depends on $\widehat{\theta}$ which depends on $\widehat{\mathbf{S}}$

default value of q should be 0 (white covariance)

ESTIMATING S ITERATIVELY

1. Obtain an initial estimate of θ, $\widehat{\theta}^{(1)}$, by minimizing $Q(\theta; \mathbf{Z}_T)$ for a given weighting matrix, usually $\mathbf{W} = \mathbf{I}$.
2. Use the initial estimate $\widehat{\theta}^{(1)}$ to produce an initial estimate of $\widehat{\mathbf{S}}^{(1)}$.
3. Re-minimize $Q(\theta; \mathbf{Z_T})$ using this initial estimate $\widehat{\mathbf{S}}^{(1)}$ to arrive at a new estimate $\widehat{\theta}^{(2)}$.
4. One can continue iterating in this manner until estimates at successive iterations converge.

In practice, usually stop at $\widehat{\theta}^{(2)}$.

Practical considerations

- $\widehat{\mathbf{S}}$ is often close to being singular.
- Research has shown that computing $\widehat{\mathbf{S}}^{-1}$ is often numerically unstable.
- Reason: Inverting large matrices is computationally difficult if they are close to being singular
- Since we only need the optimal weighing matrix S for efficiency (smallest variance), is it possible to find a matrix that, although not yielding efficient estimates, yields robust estimates?

- In practice: If the units of all moments are comparable, then the most robust results are obtained with $\mathbf{W} = \mathbf{I}$
- If the units are different, redefine the moments (in χ^2 example?)
- Empirical rule of thumb: Try the identity matrix first. Then, try the optimal weighing matrix, $\mathbf{W} = \widehat{\mathbf{S}}^{-1}$.
  $\Rightarrow$ If the results are substantially different, figure out why: $\widehat{\mathbf{S}}^{-1}$ poorly behaved? OLS very inefficient?
- Same tradeoff as with OLS vs. GLS

Outline

Asymptotic distribution of GMM

Since $\mathbf{g}(\boldsymbol{\theta}; \mathbf{Z}_T)$ is just a sample mean of a process with population mean of zero and with finite, positive definite covariance matrix, the ergodic theorem and CLT apply...

... **as long as the data are ergodic and stationary!!**

Theorem 3 (Hansen (1982)).

Assume that the stochastic process that generates the data $\mathbf{Z}_T$ is ergodic and stationary. Under certain regularity conditions, the GMM estimator is asymptotically normal:

$$\sqrt{T}\,\mathbf{g}(\boldsymbol{\theta};\mathbf{Z}_T) \xrightarrow{d} N(\mathbf{0},\mathbf{S})$$

$$\sqrt{T}\left(\widehat{\theta}-\theta\right) \xrightarrow{d} N\left(\mathbf{0},(\mathbf{D}\mathbf{S}^{-1}\mathbf{D}')^{-1}\right)$$

$$\text{where}\quad \mathbf{D}' = \begin{bmatrix} \frac{\partial g_1\left(z_t;\widehat{\theta}\right)}{\partial\theta'} \\ \vdots \\ \frac{\partial g_r\left(z_t;\widehat{\theta}\right)}{\partial\theta'} \end{bmatrix}.$$

Asymptotic distribution of GMM

- The asymptotic normality of the GMM estimator is an important result!
- If an estimator can be written as a moment condition, it is generally consistent and asymptotically normal.
- Important assumption: The data has to be stationary.
- Many “standard” problems can be written in GMM form.
- The real power of GMM is that one framework can handle a lot of interesting problems.
- Usually the moment conditions are directly implied by the definition of the estimator.

EXAMPLE: OLS AS GMM ESTIMATOR

Regression model with predetermined regressors, i.i.d. data and homoskedasticity:

$$y_t = \mathbf{x}_t'\beta + e_t$$

Moment conditions:

$$E[\mathbf{x}_t e_t] = 0$$

$$E[\mathbf{x}_t(y_t - \mathbf{x}_t'\beta)] = 0 \quad = E[x_t y_t] - E(x_t x_t')\beta$$

$$E[g(\beta; \mathbf{z}_t)] = 0, \quad \mathbf{z}_i = (\mathbf{x}_t', y_t)'$$

Note: $\mathbf{x}_i$ is a $k \times 1$ vector, hence there are k moment conditions for k parameter. Hence GMM is exactly identified.

Sample analog:

$$\frac{1}{N}\sum_t \mathbf{x}_t'(y_t - \mathbf{x}_t'\widehat{\beta}) = 0$$

$$\widehat{\beta}_{gmm} = \left(\frac{1}{T}\sum_t \mathbf{x}_t\mathbf{x}_t'\right)^{-1}\left(\frac{1}{T}\sum_t \mathbf{x}_t y_t\right)$$

The GMM estimator is identical to the OLS estimator.

The GMM estimator is also asymptotically normal

$$\sqrt{T}\left(\widehat{\theta}_T - \theta\right) \xrightarrow{d} N(\mathbf{0}, (\mathbf{D}\mathbf{S}^{-1}\mathbf{D}')^{-1})$$

How about the variance-covariance matrix $(\mathbf{D}\mathbf{S}^{-1}\mathbf{D}')^{-1}$?

k×1

$$\mathbf{D}' = \frac{\partial g}{\partial \beta'} = -\frac{1}{T}\left[\sum_t \mathbf{x}_t \mathbf{x}_t'\right] = -\mathrm{E}\left[\mathbf{x}_t \mathbf{x}_t'\right] \quad \in (k \times k)$$

$$\mathrm{E}[\mathbf{x}_t e_t] = 0$$

For simplicity, assume $\mathbf{x}_t$ is a scalar, x_t.

$$\mathrm{E}[\mathbf{x}_t(y_t - \mathbf{x}_t'\beta)] = 0 \quad = E[x_t y_t] - E(x_t x_t')\beta$$

$$\mathrm{E}[g(\beta; \mathbf{z}_t)] = 0, \quad \mathbf{z}_i = (\mathbf{x}_t', y_t)'$$

Homoskedasticity and no serial correlation:

$$\mathrm{E}\left(e_t | x_t, x_{t-1}, \ldots, e_{t-1}, e_{t-2}, \ldots\right) = 0$$

$$\mathrm{E}\left(e_t^2 | x_t, x_{t-1}, \ldots, e_{t-1}, e_{t-2}, \ldots\right) = \sigma_e^2.$$

$$\mathbf{S} = \sum_{j=-\infty}^{\infty} \mathrm{E}\left[\left(x_t e_t\right)\left(x_{t-j} e_{t-j}\right)\right] = \sum_{j=-\infty}^{\infty} \mathrm{E}\left[e_t e_{t-j} x_t x_{t-j}\right]$$

no serial correlation unless j=0

$$= \ldots + \mathrm{E}\left[e_t e_{t-1} x_t x_{t-1}\right] + \underbrace{\mathrm{E}\left[e_t^2 x_t^2\right]} + \mathrm{E}\left[e_t e_{t+1} x_t x_{t+1}\right] + \ldots$$

$$= \ldots + \mathrm{E}\left[\mathrm{E}\left(e_t | e_{t-1}, x_t, x_{t-1}\right) e_{t-1} x_t x_{t-1}\right] + \mathrm{E}\left[\mathrm{E}\left(e_t^2 | x_t\right) x_t^2\right]$$

$$+ \mathrm{E}\left[\mathrm{E}\left(e_t | e_{t+1}, x_t, x_{t+1}\right) e_{t+1} x_t x_{t+1}\right] + \ldots$$

S in OLS imposing the OLS assumptions

- Remember the moment condition $\mathrm{E}[x_t e_t] = 0$.
- Under the assumption of independent errors (**no serial correlation in e_t**):

  $\longrightarrow$ All terms $j \neq 0$ are 0.

$$\mathbf{S} = \sum_{j=-\infty}^{\infty} \mathrm{E}\left[(x_t e_t)\left(x_{t-j} e_{t-j}\right)\right]$$

$$= \mathrm{E}\left[e_t^2 x_t^2\right]$$

$$= \mathrm{E}\left[\mathrm{E}\left(e_t^2 | x_t\right) x_t^2\right]$$

$$= \sigma_e^2 \,\mathrm{E}\left[x_t^2\right]$$

OLS as GMM

Recall that

$$\sqrt{T}(\widehat{\theta}_T - \theta) \xrightarrow{d} N(\mathbf{0}, (\mathbf{D}\mathbf{S}^{-1}\mathbf{D}')^{-1})$$

The GMM variance-covariance matrix $(\mathbf{D}\mathbf{S}^{-1}\mathbf{D}')^{-1}$ is

$$\mathbf{D}' = \frac{\partial g}{\partial \beta'} = -\mathrm{E}\left[\sum_t \mathbf{x}_t \mathbf{x}_t'\right] = -E\left[\mathbf{x}_t \mathbf{x}_t'\right]$$

$$\mathbf{S} = \sigma_u^2 E\left[\mathbf{x}_t \mathbf{x}_t'\right]$$

$$\Rightarrow (\mathbf{D}\mathbf{S}^{-1}\mathbf{D}')^{-1} = \sigma_u^2 \left(E\left[\mathbf{x}_t \mathbf{x}_t'\right]\right)^{-1}$$

Putting everything together:

$$\sqrt{T}\left(\widehat{\theta}-\theta\right) \xrightarrow{d} N(\mathbf{0}, (\mathbf{D}\mathbf{S}^{-1}\mathbf{D}')^{-1}) = N(\mathbf{0}, \sigma_u^2(E\left[\mathbf{x}_t\mathbf{x}_t'\right])^{-1})$$

which is equal to the standard OLS asymptotic variance-covariance matrix.

Note: In finance, the assumption of homoskedasticity is often violated. GMM allows for a simple correction for heteroskedasticity.

White-variance covariance matrix:

$$\mathbf{S} = \sum_{j=-\infty}^{\infty} E\left[(\mathbf{x}_t e_t)\left(\mathbf{x}_{t-j} e_{t-j}\right)'\right] = E\left[e_t^2 \mathbf{x}_t \mathbf{x}_t'\right].$$

so that

$$\sqrt{T}\left(\widehat{\theta}-\theta\right) \xrightarrow{d} N(\mathbf{0}, (E\left[\mathbf{x}_t\mathbf{x}_t'\right])^{-1} E\left[e_t^2\mathbf{x}_t\mathbf{x}_t'\right] E\left[\mathbf{x}_t\mathbf{x}_t'\right])^{-1})$$

white correction! special case of GMM!

- So far we assumed that errors are serially uncorrelated
- GMM gives us an easy way to correct the standard errors if the errors are serially correlated.
- The spectral density matrix is

$$\mathbf{S} = \sum_{j=-\infty}^{\infty} E\left[e_t e_{t-j} \mathbf{x}_t \mathbf{x}'_{t-j}\right]$$

- If errors are serially uncorrelated, then all the $j \neq 0$ terms are zeros
- If errors are serially correlated, all we need to do is to adjust $\mathbf{S}$.

- Most popular estimator: Newey-West (1987)

$$\widehat{\mathbf{S}}_{NW} = \sum_{j=-q}^{q} \left(\frac{q-|j|}{k} \right) E\left[e_t e_{t-j} \mathbf{x}_t \mathbf{x}'_{t-j} \right]$$

S matrix depends on the assumptions.
Specify what assumptions to put in.

- $\widehat{\mathbf{S}}_{NW}$ is an example of a Heteroskedasticity and Autocorrelation Consistent (HAC) standard error
- **One of the powers of GMM is that it allows for an easy way to compute adjustments to standard errors!**
- Figure out in what way errors deviate from i.i.d. assumption and compute corresponding **S**.

An MA(1) process

$$y_t = \epsilon_t + \theta \epsilon_{t-1}$$

cannot be estimated by OLS. How about GMM?

What are the moment conditions? at least 2 moment conditions

θ and σ^2

We started out with the model

$$\mathrm{E}[\mathbf{h}\left(\theta, \mathbf{z}_t\right)] = 0,$$

The value of the minimized objective function

$$\widehat{Q} = \left[\mathbf{g}(\widehat{\boldsymbol{\theta}}; \mathbf{Z}_T)\right]' \widehat{\mathbf{S}}^{-1} \left[\mathbf{g}(\widehat{\boldsymbol{\theta}}; \mathbf{Z}_T)\right]$$

gives us an idea whether the model is “true” or not.

If the model is “true”, $\widehat{Q}$ should be close to 0.

Thus, test whether $\widehat{Q} = 0$. If reject, then we “reject the model”.

Hansen's J-test:

$$J = T\widehat{Q} \xrightarrow{d} \chi^2(r-k)$$

- The J-test is extensively used in finance and economics
- Often used as criterion to evaluate models
- Depends on choice of moments
- Depends on estimate of $\mathbf{S}$, which is very difficult to estimate accurately
- "Imprecise" models harder to reject than "precise" models

Outline

$$\mathrm{E}\left[\mathbf{h}\left(\theta, \mathbf{z}_t\right)\right] = 0$$

$$\mathbf{h}\left(\theta, \mathbf{z}_t\right) = \left(h_1(\theta, z_t), \ldots, h_J(\theta, z_t)\right)$$

$$h_j(\theta, z_t) = \beta \left(\frac{C_{t+1}}{C_t}\right)^{-\gamma} \frac{X_{j,t+1}}{P_{j,t}} - 1$$

$$\theta = (\beta, \gamma)$$

The GMM estimator picks $\theta = (\beta, \gamma)$, to make

$$\widehat{Q} = \left[\mathbf{g}(\widehat{\boldsymbol{\theta}}; z_t)\right]' \widehat{\mathbf{S}}^{-1} \left[\mathbf{g}(\widehat{\boldsymbol{\theta}}; z_t)\right]$$

as small as possible.

Question: How “small” is “small”? Can we reject $H_0 : \widehat{Q} = 0$?

The vector $\mathbf{g}(\widehat{\boldsymbol{\theta}}; z_t) = (g_1(\widehat{\boldsymbol{\theta}}; z_t), \ldots, g_J(\widehat{\boldsymbol{\theta}}; z_t))$ tells us by how much each moment condition deviates from 0.

$$J = T\widehat{Q} \xrightarrow{d} \chi^2(r - a)$$

tells us whether we can reject the null that all moment conditions are equal to zero.

Lettau and Ludvigson, “Euler equation errors,” *Review of Economic Dynamics*, 2009

Test the Consumption CAPM for two cases:

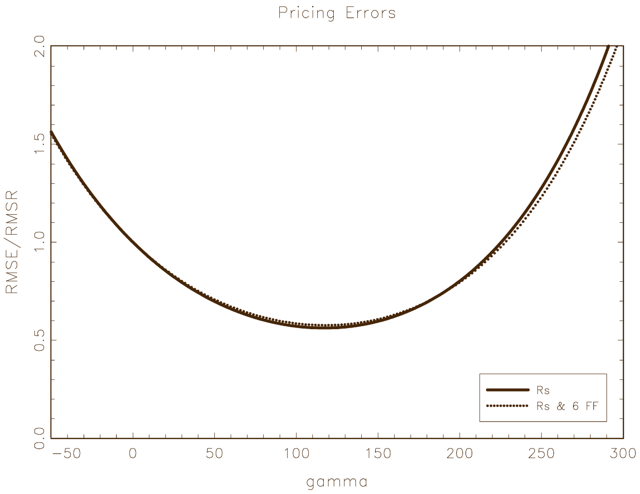
1. Use the CRSP-VW index and the 3-month T-bill rate as test assets
2. Use the CRSP-VW index, the 3-month T-bill rate as test assets plus 6 Fama-French portfolios (more on these later)

Note that Case 1 is just identified and asks what level of risk aversion and time-discount rate is required to match the moments of the CRSP-VW return and the T-bill rate

EXAMPLE: EULER EQUATION ERRORS

	Mean	St.Dev
CRSP-VW	8.31%	16.57%
T-bill	1.76%	1.06%
Equity premium	6.58%	15.59%
Sharpe-ratio	0.42	

EULER EQUATION ERRORS



Notes: The figure plots RMSE/RMSR as a function of γ for excess returns. The Euler equation errors are $e_X = E[\delta(C_{t+1}/C_t)^{\gamma}(\mathbf{R}_{t+1} - R^f_{t+1})]$. The solid line shows RMSE/RMSR for $\mathbf{R} = R^s$, the dotted line shows RMSE/RMSR for $\mathbf{R} = (R^s, 6\text{ FF})$. For each value of γ, δ is chosen to minimize the Euler equation error for the risk-free rate.

Fig. 1. Euler equation errors for CRRA preferences: Excess returns.

Euler equation errors

Table 1
Euler equation errors with CRRA preferences.

Assets	$\hat{\delta}$	$\hat{\gamma}$	RMSE (in %)	RMSE/RMSR	p ($\mathbf{W}=\mathbf{I}$)	p ($\mathbf{W}=\mathbf{S}^{-1}$)
R^s, R^f	1.41	89.78	2.71	0.48	N/A	N/A
R^s, R^f, 6 FF	1.39	87.18	3.05	0.33	0.00	0.00
Excluding periods with low consumption growth						
R^s, R^f	2.55	326.11	0.73	0.13	N/A	N/A
R^s, R^f, 6 FF	2.58	356.07	1.94	0.21	0.00	0.00

Notes: This table reports the minimized annualized postwar data Euler equation errors for CRRA preferences. The preference parameters $\hat{\delta}_c$ and $\hat{\gamma}_c$ are chosen to minimize the mean square pricing error for different sets of returns: $\min_{\delta_c,\gamma_c}[\mathbf{g}(\delta_c,\gamma_c)'\mathbf{W}\mathbf{g}(\delta_c,\gamma_c)]$ where $\mathbf{g}(\delta_c,\gamma_c)=E[\delta_c(C_t/C_{t-1})^{-\gamma_c}\mathbf{R}_t-1]$. R^s is the CRSP-VW stock returns, R^f is the 3-month T-bill rate and C_t is real per capita consumption of nondurables and services excluding shoes and clothing. The table also reports results when the periods with the lowest six consumption growth rates are eliminated. The table reports estimated $\hat{\delta},\hat{\gamma}$ and the minimized value of RMSR/RMSRR where RMSE is the square root of the average squared Euler equation error and RMSR is the square root of the averaged squared returns of the assets under consideration for $\mathbf{W}=\mathbf{I}$. The last two columns report χ^2 p-values for tests for the null hypothesis that Euler equation errors are jointly zero for $\mathbf{W}=\mathbf{I}$ and $\mathbf{W}=\mathbf{S}^{-1}$ where $\mathbf{S}$ is the spectral density matrix at frequency zero. The data span the period 1951Q4 to 2002Q4.

Other weighting matrices

Messy math but easy to implement!

The inverse of the spectral density matrix is the “optimal” weighting matrix in the sense that it yields the estimator with the smallest variance.

Properties of GMM for general weighting matrices

$$\widehat{\theta} = \underset{\theta}{\operatorname{argmin}}\, \mathbf{g}_T(\boldsymbol{\theta})' \mathbf{W} \mathbf{g}_T(\boldsymbol{\theta})$$

$$\text{FOC: } \mathbf{d}_T\left(\widehat{\boldsymbol{\theta}}\right)' \mathbf{W} \mathbf{g}_T\left(\widehat{\boldsymbol{\theta}}\right) = 0$$

$$\mathbf{d}_T(\boldsymbol{\theta}_0) \equiv \frac{\partial \mathbf{g}_T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$$

$$\sqrt{T}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \xrightarrow{\mathbf{d}} N\left(0, (\mathbf{d}'\mathbf{W}\mathbf{d})^{-1} \mathbf{d}'\mathbf{W}\mathbf{S}\mathbf{W}\mathbf{d}\left(\mathbf{d}'\mathbf{W}\mathbf{d}\right)^{-1}\right)$$

(Before somebody asks: No, you don’t have to memorize this for exams...)

Discussion

- The GMM framework is rich enough that we can think of many other ways of testing the hypotheses of interest.
- For example: we can break the orthogonality restrictions into those that identify and those that over-identify the parameters

$$E\begin{pmatrix} \mathbf{h}_1(\mathbf{z}_t;\boldsymbol{\theta}_0) \\ \mathbf{h}_2(\mathbf{z}_t;\boldsymbol{\theta}_0) \end{pmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}$$

- GMM offers a simple way to derive the limiting distributions and associated variance-covariance matrices of many estimators. As long as an estimator can be written in the form of a moment condition, GMM gives us a recipe to compute its variance-covariance matrix.

- GMM has many desirable asymptotic properties
- How about finite sample properties?
- The optimal weighting matrix $\mathbf{S}^{-1}$ is a function of fourth moments of the regressors
- It takes much longer samples to estimate fourth moments reliably than to estimate first and second moments.

Small sample issues

- In addition, in many situations $\mathbf{S}$ is close to being singular, so computing $\mathbf{S}^{-1}$ can be computationally difficult.
- It is generally accepted that choosing the inefficient $\mathbf{W} = \mathbf{I}$ is superior in small samples to the efficient $\mathbf{W} = \mathbf{S}^{-1}$.
- In small samples, the chi-squared over-identifying test tends to over-reject.

Let $X_1, \ldots X_n \sim Unif[0, \theta]$.

The pdf of a uniform distribution is

$$f(x;\theta) = \begin{cases} \frac{1}{\theta} & 0 \leq x \leq \theta \\ 0 & otherwise. \end{cases}$$

Let's compute various GMM estimators and compare them to MLE.

Moments: MGF for U is

$$\psi(t) = \mathrm{E}(e^{tx})$$
$$= \int_0^\theta \frac{e^{xt}}{\theta} dx = \left[\frac{e^{xt}}{t\theta}\right]_0^\theta$$
$$= \begin{cases} \frac{e^{t\theta}-1}{\theta} & t \neq 0 \\ 1 & t = 0 \end{cases}$$

$$\psi^k(0) = \frac{d^k\psi}{dt^k}(0) = \mathrm{E}(X^k)$$

Technical issue: $\psi(t)$ is not differentiable at 0, so we need to take limits

$$m_1(\theta) = \theta/2$$

$$m_2(\theta) = \theta^2/3$$

$$m_3(\theta) = \theta^3/4$$

$$m_4(\theta) = \theta^4/5$$

GMM comparisons

We will compare MLE to:

1. MM first moment: $m_1(\theta)$
2. MM second moment: $m_2(\theta)$
3. GMM with $\mathbf{W} = \mathbf{I}$
4. GMM with $\mathbf{W} = \mathbf{S}^{-1}$ when $\mathbf{S}$ is computed analytically with true θ.
5. GMM with $\mathbf{W} = \mathbf{S}^{-1}$ when $\mathbf{S}$ is computed analytically with estimated θ from GMM with $\mathbf{W} = \mathbf{I}$.
6. GMM with $\mathbf{W} = \mathbf{S}^{-1}$ when $\mathbf{S}$ is estimated, 2nd stage
7. GMM with $\mathbf{W} = \mathbf{S}^{-1}$ when $\mathbf{S}$ is estimated, 3rd stage

Analytical **S**:

$$\mathbf{S} \equiv \begin{bmatrix} S_{11} & S_{12} \\ S_{12} & S_{22} \end{bmatrix}$$

$$= \mathrm{E}\begin{bmatrix} y - m_1(\theta) \\ y^2 - m_2(\theta) \end{bmatrix}\begin{bmatrix} y - m_1(\theta) \\ y^2 - m_2(\theta) \end{bmatrix}^\top = \mathrm{Cov}((y, y^2)'|\theta)$$

In the $U(0,\theta)$ case we have

$$S_{11} = var(y|\theta) = \frac{\theta^2}{12}$$

$$\begin{aligned} S_{12} &= m_3 - m_1 m_2 - m_2 m_1 + m_1 m_2 \\ &= m_3(\theta) - m_1(\theta) m_2(\theta) \\ &= \frac{\theta^3}{4} - \frac{\theta}{2} \times \frac{\theta^2}{3} = \frac{\theta^3}{12} \end{aligned}$$

$$\begin{aligned} S_{22} &= m_4(\theta) - [m_2(\theta)]^2 \\ &= \frac{\theta^4}{5} - \left(\frac{\theta^2}{3}\right)^2 = \frac{4\theta^4}{45} \end{aligned}$$

For $\theta = 10$:

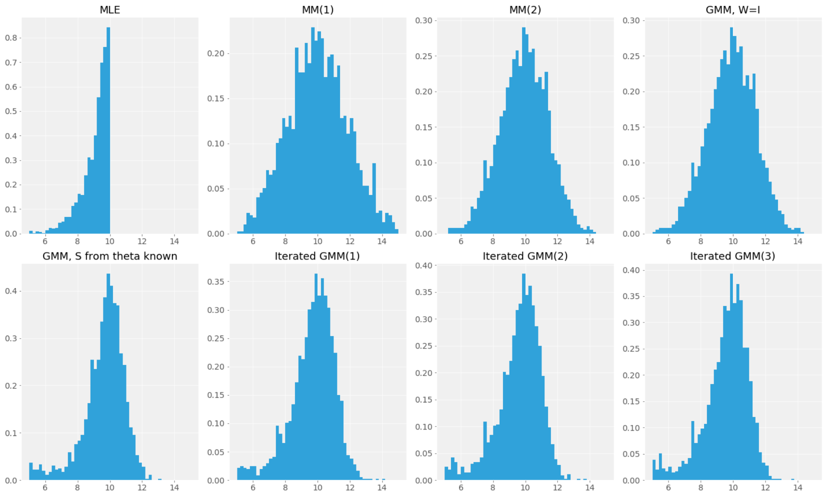
$$\mathbf{S} = \begin{bmatrix} 8.333 & 83.333 \\ 83.333 & 888.889 \end{bmatrix}$$

and

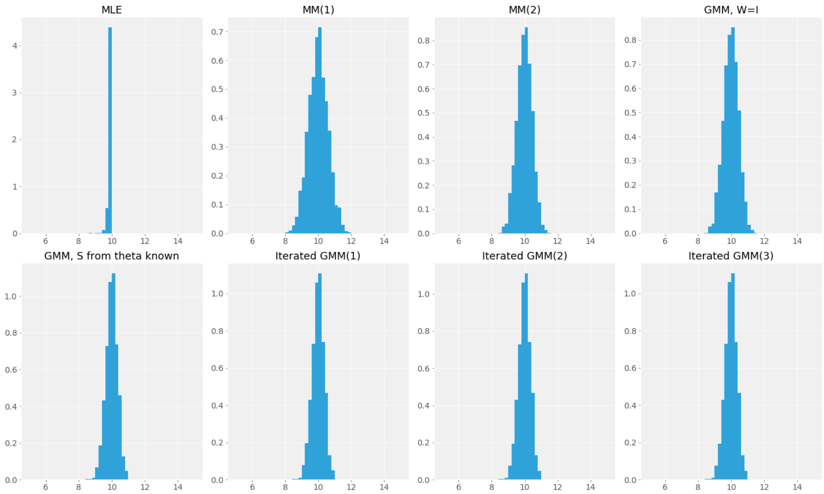
$$\mathbf{W} = \mathbf{S}^{-1} = \begin{bmatrix} 1.920 & -0.180 \\ -0.180 & 0.018 \end{bmatrix}$$

Simulate 5,000 samples of size $T = 10, 100, 1000$

N=10

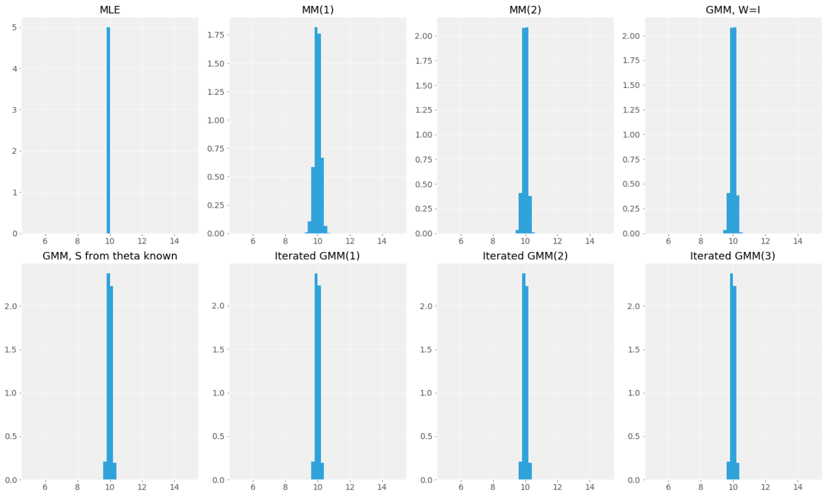

MLE
MM(1)
MM(2)
GMM, W=I
GMM, S from theta known
Iterated GMM(1)
Iterated GMM(2)
Iterated GMM(3)

N=100



MLE
MM(1)
MM(2)
GMM, W=I
GMM, S from theta known
Iterated GMM(1)
Iterated GMM(2)
Iterated GMM(3)

N=1,000



MLE
MM(1)
MM(2)
GMM, W=I
GMM, S from theta known
Iterated GMM(1)
Iterated GMM(2)
Iterated GMM(3)

Small sample properties of GMM vs. MLE

True $\theta = 10$

	$E[\widehat{\theta}]$			St.Dev.$(\widehat{\theta})$		
	10	100	1000	10	100	1000
MLE	9.08	9.90	9.99	0.83	0.09	0.01
MM, $m_1(\theta)$	9.99	10.00	9.99	1.82	0.58	0.18
MM, $m_2(\theta)$	9.86	9.99	9.99	1.64	0.45	0.14
GMM, $W = I$	10.03	10.00	9.99	1.44	0.45	0.14
GMM, $W = S^{-1}$, true θ	9.30	10.02	10.00	2.27	0.36	0.11
GMM, $W = S^{-1}$, estimated θ	9.27	10.02	10.00	2.52	0.36	0.11
GMM, $W = S^{-1}$, 2nd stage	8.98	10.02	10.00	2.64	0.36	0.11
GMM, $W = S^{-1}$, 3rd stage	8.75	10.03	10.00	2.76	0.37	0.11

GMM vs. OLS vs. MLE

- Note that both OLS and MLE can be regarded as special cases of GMM.
- OLS: Let the moments be the sample mean of $E\left[X_t \varepsilon_t\right]$
- MLE: Set moments to be the MLE first order conditions
- The GMM asymptotic distribution can be used to compute limiting behavior of other estimators
- In these cases the model is exactly identified (the number of moments equals the number of parameters to be estimated).
- The choice of weighting matrix therefore does not matter (every moment is set to zero, regardless)
- The J-test of overidentifying restrictions does not apply.

MLE vs. GMM

Robustness vs. efficiency

- MLE is (asymptotically) efficient, so why bother with GMM?
- GMM is based on a limited set of moment conditions
- MLE requires strong assumption on the exact distribution of the errors
- GMM is robust to distributional misspecification
- Tradeoff between efficiency and robustness

Sometimes, the MLE is not feasible because we cannot write down the exact likelihood function, GMM is applicable in a wider range of problems.

Estimation techniques: Summary

1. OLS
   - Under strong assumption (errors are exogenous), OLS estimator is BLUE
   - Under weaker assumptions (errors are predetermined), OLS estimator is consistent
2. MLE
   - Specific distributional assumptions required
   - If assumptions are satisfied, ML estimator is the best possible estimator
3. GMM
   - Weaker assumptions are required
   - Applicable to a wide range of problems
   - GMM estimator is consistent under weak assumption
   - Computation of spectral density matrix is challenging
   - Small sample properties may be an issue