8. Generalized Method of Moments

- introduction
- method of moments estimator
- GMM estimator
- distribution of GMM estimators

Introduction

- method of moments (MM) estimators solves the sample moment conditions that correspond to the population moment conditions
- general methods of moments (GMM) estimators extends MM approach to accommodate the case when there are more moment conditions to solve than the number of parameters
- GMM estimator defines a class of estimators; using different population moment conditions gives different GMM estimators (just as different densities lead to different ML estimators)

GMM estimators are based on the analog principle that **population** moment conditions lead to **sample** moment conditions that can be used to estimate parameters suppose y is i.i.d. with mean μ, in population we have

$$\mathbf{E}[y - \mu] = 0$$

replacing the expectation by the average operator yields the corresponding sample moment

$$(1/N) \sum_{i=1}^{N} (y_i - \mu) = 0$$

solving for μ leads to the estimator $\hat{\mu}_{\mathrm{mm}} = (1/N) \sum_{i=1}^{N} y_i = \bar{y}$

the MM estimate of the population mean is the **sample mean**

Examples of GMM estimators

- linear regression
- nonlinear regression
- maximum likelihood
- instrumental variables regression

Linear regression as an example of MM

consider the linear regression model: $y = x^T\beta + u$ where we assume $\mathbf{E}[u|x] = 0$

using the **law of iterated expectations**

$$\mathbf{E}[xu] = \mathbf{E}[\mathbf{E}[xu|x]] = \mathbf{E}[x\mathbf{E}[u|x]] = 0$$

hence, we obtain $\mathbf{E}[xu] = \mathbf{E}[x(y - X\beta)] = 0$

replacing $\mathbf{E}$ by the average operator gives the **sample moment** condition:

$$(1/N)\sum_{i=1}^{N} x_i(y_i - x_i^T\beta) = 0$$

this yields

$$\hat{\beta}_{\mathrm{mm}} = (\sum_i x_i x_i^T)^{-1} \sum_i x_i y_i$$

LS estimator is therefore just a special case of MM estimation

Nonlinear regression as an example of MM

the nonlinear regression model with **additive error** is

$$y = g(x, \beta) + u$$

the assumption $\mathbf{E}[u|x] = 0$ implies that for any function $h(x)$ we have

$$\mathbf{E}[h(x)(y - g(x, \beta))] = 0$$

a particular choice is

$$h(x) = \nabla_\beta g(x, \beta)$$

that leads to the sample moment condition:

$$(1/N) \sum_{i=1}^{N} \nabla g(x_i, \beta)(y_i - g(x_i, \beta)) = 0$$

which is the first-order conditions for the NLS estimators

Quasi-maximum likelihood as an example of MM

the quasi MLE $\hat{\theta}_{\mathrm{mle}}$ is defined to be the estimator that maximizes a log-likelihood function that is **misspecified**, as the result of specification of the wrong density

- let $f(y|\theta)$ denoted the **assumed** joint density of $y_1, \ldots, y_N$
- let $h(y)$ denoted the **true** density
- define the **Kullback-Leibler information criterion (KLIC)**

$$\mathrm{KL} = \mathbf{E}\left[\log \frac{h(y)}{f(y|\theta)}\right]$$

  where expectation is w.r.t. $h(y)$

- KL takes a minimum of 0 when $\exists \theta^\star$ s.t. $h(y) = f(y|\theta^\star)$
- KL indicate greater ignorance about the true density

definition: the quasi-MLE minimizes KL, the distance between $h(y)$ and $f(y|\theta)$

but we can write KL as

$$\mathrm{KL} = \mathbf{E}[\log h(y)] - \mathbf{E}[\log f(y|\theta)]$$

hence, equivalently, the quasi-MLE estimate maximizes

$$\mathbf{E}[\log f(y|\theta)]$$

as $\mathbf{E}[h(y)]$ does not depend on θ

conclusion: a local maximum of KL occurs if $\mathbf{E}[\nabla \log f(y|x,\theta)] = 0$

replacing by the sample moment conditions gives an estimator that solves

$$(1/N)\sum_{i=1}^{N} \nabla \log f(y_i|x_i,\theta) = 0$$

so a quasi-MLE can be motivated as an MM estimator

IV regression as an example of MM

assume the existence of instrument z:

- $\mathbf{E}[u|z] = 0$ or that $\mathbf{E}[y - X\beta|z] = 0$
- z are correlated with x

using law of iterated expectation, the population moment conditons are

$$\mathbf{E}[z(y - x^T\beta)] = 0$$

the MM estimator solves the sample moment condition

$$\frac{1}{N}\sum_{i=1}^{N} z_i(y_i - x_i^T\beta) = 0$$

- if z has the same dimension as x then the MM estimator is

$$\hat{\beta}_{\mathrm{mm}} = \left(\sum_i z_i x_i^T\right)^{-1} \sum_i z_i y_i$$

which is the linear IV estimator $\hat{\beta}_{\mathrm{iv}} = (Z^T X)^{-1} Z^T y$

- if z has a higher dimension that x, then we choose β to minimize

$$Q(\beta) = \left[\frac{1}{N}\sum_{i=1}^{N} z_i(y_i - x_i^T\beta)\right]^T W_N \left[\frac{1}{N}\sum_{i=1}^{N} z_i(y_i - x_i^T\beta)\right]$$

where W_N is $p \times p$ if $z \in \mathbf{R}^p$

this choice is the **general method of moments estimator**

Generalized Method of Moments

GMM defines a class of estimators where different choice of moment condition and weighting matrix lead to different GMM estimators, just as different choices of distribution lead to different ML estimators

- method of moments estimator
- definition of GMM estimator
- distribution of GMM estimator
- optimal GMM

General form of MM estimators

assume there are m moment conditions for n parameters:

$$\mathbf{E}[h(w, \theta^\star)] = 0$$

- $\theta \in \mathbf{R}^n$ and $\theta^\star \in \mathbf{R}^n$ is the value of θ in the dgp
- h is an $m \times 1$ vector-valued function
- w includes all observables (y, x or instrument z)

some examples of $h(w) = h(y, x, z, \theta)$

moment function $h(\cdot)$	**estimation method**
$y - \mu$	method of moments for population mean
$x(y - x^T\beta)$	ordinary least-squares regression
$z(y - x^T\beta)$	instrumental variables regression
$\partial \log f(y\|x, \theta)/\partial\theta$	maximum likelihood estimation

Definition of MM estimator

if $m = n$ then method of moments can be applied

replace the population moment by the sample moment

the **method of moments estimator** $\hat{\theta}_{\rm mm}$ is defined to the solution of

$$\frac{1}{N}\sum_{i=1}^{N} h(w_i, \hat{\theta}) = 0$$

this is the zero gradient condition of the minimization:

$$Q(\theta) = \left[\frac{1}{N}\sum_{i=1}^{N} h(w_i, \theta)\right]^T \left[\frac{1}{N}\sum_{i=1}^{N} h(w_i, \theta)\right]$$

Definition of GMM estimators

the GMM estimator is based on m conditions with n parameters to be estimated

- if $m = n$ the model is said to be **just-identified** and MM estimator is used
- if $m > n$ the model is said to be **overidentified** and MM cannot be applied

instead $\hat{\theta}$ is chosen so that $(1/N)\sum_i h(w_i, \hat{\theta})$ is as close to zero as possible

the **GMM estimators** $\hat{\theta}_{\mathrm{gmm}}$ is defined to be the problem of minimizing

$$Q(\theta) = \left[\frac{1}{N}\sum_{i=1}^{N} h(w_i, \theta)\right]^T W_N \left[\frac{1}{N}\sum_{i=1}^{N} h(w_i, \theta)\right]$$

where $W \succ 0$, possibly stochastic but does not depend on θ

First-order condition for GMM estimators

differentiating Q w.r.t. θ yields the first-order conditions:

$$\left[\frac{1}{N}\sum_{i=1}^{N}\frac{\partial h(w_i,\hat{\theta})}{\partial\theta}\right]^T W \left[\frac{1}{N}\sum_{i=1}^{N} h(w_i,\hat{\theta})\right] = 0$$

- the conditions are generally nonlinear in θ; use numerical method to solve it
- diffferent choices of W lead to different estimators with different variances
- the optimal choice of W is provided

Distribution of GMM estimator

assumptions:

1. the dgp imposes the moment condition: $\mathbf{E}[h(w, \theta^\star)] = 0$
2. $h(\cdot)$ satisfies $h(w, \beta) = h(w, \theta)$ iff $\beta = \theta$
3. the following $m \times n$ matrix exists and is finite with rank n:

$$A = \text{plim}(1/N) \sum_{i=1}^{N} \frac{\partial h(w_i, \theta^\star)}{\partial \theta}$$

4. $W_N \xrightarrow{p} W$ where W is finite positive definite
5. $(1/\sqrt{N}) \sum_{i=1}^{N} h(w_i, \theta^\star) \xrightarrow{d} \mathcal{N}(0, B)$ where

$$B = \text{plim}(1/N) \sum_{i=1}^{N} \sum_{j=1}^{N} h(w_i, \theta^\star) h(w_j, \theta^\star)^T$$

then the **GMM estimator** $\hat{\theta}_{\mathrm{gmm}}$, defined to be the root of

$$\nabla_\theta Q(\theta) = 0$$

is consistent for $\theta^\star$ and

$$\sqrt{N}(\hat{\theta}_{\mathrm{gmm}} - \theta^\star) \xrightarrow{d} \mathcal{N}(0, (A^T W A)^{-1}(A^T W B W A)(A^T W A)^{-1})$$

special case:

- if the data is independent over i then B is simplified to

$$B = \mathrm{plim}\, \frac{1}{N} \sum_{i=1}^{N} h(w_i, \theta^\star) h(w_i, \theta^\star)^T$$

- in just-identified case $(m = n)$, the matrices A, W, S are square and invertible, the result on MM becomes

$$\sqrt{N}(\hat{\theta}_{\mathrm{mm}} - \theta^\star) \xrightarrow{d} \mathcal{N}(0, A^{-1} B A^{-T})$$

Estimated asymptotic covariance

we use consistent estimates of A, B:

- estimate of A: replace $\theta^\star$ by $\hat{\theta}$

$$\hat{A} = \frac{1}{N}\sum_{i=1}^{N} \frac{\partial h(w_i, \hat{\theta})}{\partial \theta}$$

- estimate of B: consider when data is independent over i

$$B = \frac{1}{N}\sum_{i=1}^{N} h(w_i, \hat{\theta}) h(w_i, \hat{\theta})^T$$

GMM estimator is asymptotically normally distributed with mean $\theta^\star$ and estimated covariance is

$$\widehat{\mathbf{Avar}}(\hat{\theta}_{\mathrm{gmm}}) = (1/N)(\hat{A}^T W_N \hat{A})^{-1} \hat{A}^T W_N \hat{B} W_N \hat{A} (\hat{A} W_N \hat{A})^{-1}$$