M Estimators

Eduardo Rossi
University of Pavia
A basic unifying notion is that most econometric estimators are defined as the minimizers of certain functions constructed from the sample data, called criterion functions, $C_T$

Parameter Space:

$$\Theta \subseteq \mathbb{R}^p$$

The true value of the parameter:

$$\theta_0 \in \Theta$$

An estimator of $\theta_0$ based on a data set of size $T$ is given by:

$$\hat{\theta}_T = \arg \min_{\theta \in \Theta} C_T(\theta)$$

For instance, the OLS for the MLR model is an $M$-estimator where

$$C_T(\theta) = \sum_t (y_t - x_t'(\theta))^2$$
The analysis is centered around the properties of $C_T$ and its derivatives.

\[ g_T(\theta) = \frac{\partial C_T}{\partial \theta} = \begin{bmatrix} \frac{\partial C_T}{\partial \theta_1} \\ \vdots \\ \frac{\partial C_T}{\partial \theta_p} \end{bmatrix} \quad (p \times 1) \] (2)

\[ Q_T(\theta) = \frac{\partial^2 C_T}{\partial \theta \partial \theta'} = \begin{bmatrix} \frac{\partial^2 C_T}{\partial \theta_1^2} & \cdots & \frac{\partial^2 C_T}{\partial \theta_1 \partial \theta_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 C_T}{\partial \theta_p \partial \theta_1} & \cdots & \frac{\partial^2 C_T}{\partial \theta_p^2} \end{bmatrix} \quad (p \times p) \] (3)
\[ g_T(\theta) = \nabla C_T \quad (4) \]
\[ Q_T(\theta) = \nabla^2 C_T \quad (5) \]

All these quantities are random variables and functions on the parameter space.

The \textit{Criterion Function} can be thought of as having the generic form

\[ C_T(\theta) = \phi \left( \frac{1}{T} \sum_{t=1}^{T} \psi_t(\theta) \right) \]

\( \phi(\cdot) \) is a continuous function of \( r \) variables, \( \psi_t(\theta) \) is an \( r \)-vector of continuous functions. The factor \( \frac{1}{T} \) ensures that \( C_T \) is bounded in the limit as \( T \to \infty \).
**Maximum Likelihood Estimator:** $\psi_t$ is scalar, corresponds to the negative of the log-probability or log-probability density associated with each observation.

**Method of Moments Estimator:** Solution to sets of equations relating data moments and unknown parameters, the moments in question being replaced by sample averages to obtain the estimates.
Asymptotic properties

Under regularity conditions, that we have to establish in particular cases, $M$-estimators satisfy the usual demanded properties of econometric estimators.

Let $(\Omega, \mathcal{F})$ be a measurable space, and $\Theta$ a compact (closed and bounded) subset of $\mathbb{R}^p$. Let

$$C : \Omega \times \Theta \to \mathbb{R}$$

be a function, such that $C(\cdot, \omega)$ is $\mathcal{F}/\mathcal{B}$ measurable for every $\theta \in \Theta$ and $C(\omega, \cdot)$ is continuous on $\Theta$ for every $\omega \in F$ for some $F \in \mathcal{F}$. Then there exists a $\mathcal{F}/\mathcal{B}^p$ measurable function $\hat{\theta} : \Omega \to \Theta$ such that for all $\omega \in F$,

$$C(\omega, \hat{\theta}(\omega)) = \inf_{\theta \in \Theta} C(\omega, \theta)$$
The important conditions are:

1. Continuity of $C(\omega, \cdot)$ as a function of $\theta$ for given data

2. Compactness of $\Theta$
Compactness of $\Theta$

- Boundedness is an innocuous requirement, large parameter are unreasonable.

- We cannot exclude points that are closure points. The problem is that the minimum of a function on a set may not exist as a member of the set, unless the set is closed.

The asymptotic analysis is only concerned with the behavior of the estimator on a small open neighbourhood of $\theta_0$. 

7
Consistency

To analyze the consistency problem we need some new convergence concepts. Let \( \{X_t(\theta), t = 1, 2, \ldots\} \) be a random sequence whose members are functions of parameters \( \theta \), on a domain \( \Theta \).

**Definition** If \( X_t(\theta) \to X(\theta) \) for every \( \theta \in \Theta \), \( X_t \) is said to converge in probability to \( X \) pointwise on \( \Theta \).

**Definition** If \( \sup_{\theta \in \Theta} |X_t(\theta) - X(\theta)| \xrightarrow{p} 0 \), \( X_t \) is said to converge in probability to \( X \) uniformly on \( \Theta \).

In general, pointwise convergence \( \to \) uniform convergence.
Theorem

1. \( \Theta \) is compact

2. \( C_T(\theta) \xrightarrow{p} \overline{C}(\theta) \) (a non-stochastic function of \( \theta \)) uniformly in \( \Theta \)

3. \( \theta_0 \in \text{int}(\Theta) \) is the unique minimum of \( \overline{C}(\theta) \)

then \( \hat{\theta}_T \xrightarrow{p} \theta_0 \).

Condition 2 can also be stated in the form

\[
\sup_{\theta \in \Theta} |C_T(\theta) - \overline{C}(\theta)| \xrightarrow{p} 0
\]  

(8)
It depends essentially on the vectors $\psi_t(\theta)$, for if their average converges uniformly, this property will carry over to $C_T$ by Slutsky’s Theorem, given that $\phi(\cdot)$ is continuous. A sufficient condition on the sequences $\{\psi_{it}(\theta), t = 1, 2, \ldots\}$ for $i = 1, \ldots, r$ is stochastic uniform equicontinuity ($\psi_{it}(\theta)$ uniformly continuous, not merely for every finite $t$, but also in the limit).
Identification

Condition 3 is called *identification condition*, ensuring the problem is properly specified. If condition 3 is satisfied for a given $\theta_0$, this defines the probability limit of the estimator by construction.

The basic requirement for a consistent estimator is that the structure can be distinguished from alternative possibilities, given sufficient data. This requirement fails when there is *observational equivalence*. 
**Definition** Structures $\theta_1$ and $\theta_2$ are said to be *observationally equivalent* with respect to $C_T$ if for every $\varepsilon > 0$ there exists $h_\varepsilon \geq 1$ such that for all $T \geq T_\varepsilon$

$$P(|C_T(\theta_1) - C_T(\theta_2)| < \varepsilon) > 1 - \varepsilon$$

(9)

Thus the criterion changes by only an arbitrary amount between points $\theta_1$ and $\theta_2$ of $\Theta$, with probability arbitrary close to 1.

The criterion changes by only an arbitrarily small amount between point $\theta_1$ and $\theta_2$ of $\Theta$, with probability close to 1. In almost every sample of size exceeding $T_\varepsilon$, it is not possible to distinguish between the two structures using $C_T$. 
**Definition** The true structure $\theta_0$ is said to be *globally (locally) identified* by $C_T$ if no other point in $\Theta$ (in an open neighborhood of $\theta_0$) is observationally equivalent to $\theta_0$ with respect to $C_T$.

Identification by $C_T$ is equivalent to condition 3 in the following sense:

**Theorem** Let $C_T(\theta) \xrightarrow{p} \overline{C}(\theta)$ uniformly in $\Theta$ and let $\theta_0$ be a *global (local) minimum* of $\overline{C}(\theta)$. The minimum is unique (unique in an open neighborhood of $\theta_0$) if and only if $\theta_0$ is asymptotically globally (locally) identified by $C_T$. 

13
**Assumption** $C_T$ is twice differentiable with respect to $\theta$, at all interior points of $\Theta$, with probability 1. The expectations $E(C_T)$, $E(g_T)$, and $E(Q_T)$ exist, with

$$E(C_T) \to \overline{C}$$
$$E(g_T) \to \overline{g}$$
$$E(Q_T) \to \overline{Q}$$

uniformly in $\Theta$.

**Theorem** Given the above assumptions and $\theta_0 \in \text{int}(\Theta)$, if $\theta_0$ is the unique minimum of $\overline{C}$ and $\overline{g}(\theta_0) = 0$ and $\overline{Q}(\theta_0)$ is positive definite.

**Proof** If $\theta$ is an interior point of $\Theta$,

$$\nabla E[C_T(\theta)] = E[g_T(\theta)] \quad \text{ (10)}$$
$$\nabla^2 E[C_T(\theta)] = E[Q_T(\theta)] \quad \text{ (11)}$$
In fact, by the following theorem

**Theorem** If:

for fixed real $\theta$, $f(\theta)$ is an integrable random variable;

for almost all $\omega$ in the sample space, $f(\theta, \omega)$ is a function of $\theta$, differentiable at the point $\theta_0$, with

$$\frac{[f(\theta_0 + h, \omega) - f(\theta_0, \omega)]}{h} \leq d(\omega)$$

for all $h$ in an open neighborhood of 0 not depending on $\omega$, where $d$ is an integrable random variable;
then

\[
E \left( \frac{df}{d\theta} \bigg|_{\theta=\theta_0} \right) = \frac{dE(f)}{d\theta} \bigg|_{\theta=\theta_0}. 
\]  

This is called differentiation under the integral sign.

(13)
\[ E(g_T) = 0 \]
\[ E(Q_T) \quad p.d.m. \]

are necessary and sufficient for identification (unique local minimum of \( E(C_T) \)). \( \square \)
Asymptotic Normality

The form of the criterion function, with assumptions on the data series, must be such as to lead to the result

\[ \sqrt{T} g_T(\theta_0) \xrightarrow{L} N(0, A_0) \]

\[ A_0 = \lim_{T \to \infty} TE\left[ g_T(\theta_0) g_T(\theta_0)' \right] \]

while, under the assumption of consistency

\[ g_T(\theta_0) \xrightarrow{p} 0 \quad (14) \]

To obtain this CLT the vector in question must have the following generic form

\[ g_T(\theta_0) = J_T \frac{1}{\sqrt{T}} \sum_{t=1}^{T} v_t \quad (15) \]
where $J_T$, $(p \times q)$, is a random matrix converging in probability to a finite limit $\overline{J}$, and $v_t$ $(q \times 1)$ is a zero mean vector. The essential step is a quadratic approximation argument based on an application of the mean value theorem.

**Theorem** If

$$\sqrt{T}g_T(\theta_0) \xrightarrow{L} N(0, A_0)$$

and

$$Q_T(\theta) \xrightarrow{p} Q(\theta)$$

uniformly in a neighborhood of $\theta_0$, then

$$\sqrt{T}(\hat{\theta}_T - \theta_0) \xrightarrow{asy} -Q_0^{-1}\sqrt{T}g_T(\theta_0) \xrightarrow{L} N(0, V_0)$$

where $V_0 = Q_0^{-1}A_0Q_0^{-1}$. 

19
To prove the theorem we need the following

**Lemma.** If $Y_T(\theta)$ is a stochastic function of $\theta$,

$$Y_T(\theta) \xrightarrow{Pr} Y(\theta)$$

uniformly in an open neighbourhood of $\theta_0$, and $\hat{\theta}_T \xrightarrow{p} \theta_0$, then

$$Y_T(\hat{\theta}_T) \xrightarrow{p} Y(\theta_0)$$

this can be thought of as an extension of Slutsky’s Theorem. In this case there is a sequence of stochastic functions (i.e., functions depending on random variables other than $\hat{\theta}_T$) converging to a limit.
**Slutsky’s Theorem** Let \( \{X_T\} \) be a random sequence converging in probability to a constant \( a \), and let \( g(\cdot) \) be a function continuous at \( a \). Then

\[
g(X_T) \xrightarrow{p} g(a)
\]

**Cramer’s Theorem** Let \( \{Y_T\} \) and \( \{X_T\} \) be random sequences, with \( Y_T \xrightarrow{d} Y \) and \( X_T \xrightarrow{p} a \) (a constant). Then

1. \( X_T + Y_T \xrightarrow{d} a + Y \)

2. \( X_T Y_T \xrightarrow{d} aY \)

3. \( \frac{Y_T}{X_T} \xrightarrow{d} \frac{Y}{a} \) when \( a \neq 0 \).
Cramér-Wold device If \( \{X_T\} \) is a sequence of random vectors, \( X_T \xrightarrow{D} X \) if and only if for all conformable fixed vectors \( \lambda \), \( \lambda^\prime X_T \xrightarrow{D} \lambda^\prime X \).
Proof Letting $g_{Ti}$ denote the $i$-th element of $g_T$, and

$$g_{Ti}(\hat{\theta}_T) = 0$$

identically by definition of $\hat{\theta}_T$, the mean value theorem yields

$$g_{Ti}(\theta_0) + (\nabla g_{Ti}|_{\theta=\theta^*_{Ti}})'(\hat{\theta}_T - \theta_0) = 0 \quad i = 1, \ldots, p$$

where

$$\theta^*_{Ti} = \theta_0 + \lambda_i(\hat{\theta}_T - \theta_0) \quad for \quad 0 \leq \lambda \leq 1$$

Stacking these equations up to form a column vector

$$g_T(\theta_0) + Q^*_{T}(\hat{\theta}_T - \theta_0) = 0 \quad (p \times 1)$$

where $Q^*_{T}$ denotes $Q_T$ with row $i$ evaluated at the point $\theta^*_{Ti}$ for each $i$. 

23
After multiplying by $\sqrt{T}$

$$\sqrt{T} \left[ g_T(\theta_0) + Q^*_T(\hat{\theta}_T - \theta_0) \right] = 0$$

$$\sqrt{T}(\hat{\theta}_T - \theta_0) = -(Q^*_T)^{-1}\sqrt{T}g_T(\theta_0)$$

Since

$$\hat{\theta}_T \overset{p}{\to} \theta_0$$

$$\theta^*_{T_i} \overset{p}{\to} \theta_0 \quad \forall i$$

If follows from $Q_T(\theta) \overset{p}{\to} \overline{Q}(\theta)$ and the Lemma that

$$Q^*_T \overset{p}{\to} \overline{Q}(\theta_0)$$

Finally, from the Cramer’s Theorem follows the theorem, noting that $\overline{Q}_0$ is symmetric.