Chapter 1

M-Estimators

Except for some Bayesian estimators that require integration or simulation to get the values, almost all frequentist estimators in econometrics involve steps of minimizing some criterion functions. The M-estimator is a class of estimators that each individual observation contributes to the overall criterion function in an additive, separable manner. The name M-estimator was coined by Peter J. Huber as *Maximum likelihood-type* estimator.

We are familiar with OLS. The OLS estimator boasts an explicit solution in that it can be written as a function of the data only. This is due to the fact that the loss function of OLS is quadratic and the regressors are linearly combined by the slope coefficients. Most estimators do not enjoys such simplicity. For example, the logistic regression does not have a closed-form solution but must be solved numerically.

1.1 Formulation

Let the loss function be $\rho_i(\theta) = \rho(z_i, \theta)$, where z_i is the ith observation. It can be a scalar random variable or a multivariate one. The sample criterion is the average of $\rho_i(\theta)$ over $i = 1, 2, \ldots, n$:

$$S_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \rho_i(\theta)$$

The M-estimator minimizes the sample criterion function:

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} S_n(\theta).$$

The M-estimator includes many examples as special cases. For example, OLS, nonlinear least squares (NLS), maximum likelihood estimation (MLE), and quantile regressions are all M-estimators. GMM, another familiar econometric estimator, is not an M-estimator; it belongs to the broader class of minimuM-distance estimators.

Example 1.1. The most familiar example of M-estimator is the OLS. Let $y_i = x_i'\theta + e_i$. The loss function is $\rho_i(\theta) = \frac{1}{2}(y_i - x_i'\theta)^2$ where $z_i = (y_i, x_i')'$, and the sample criterion function is $\frac{1}{2n}\sum_i (y_i - x_i'\theta)^2$.

Example 1.2. Silva and Tenreyro (2006) propose a nonlinear conditional mean model $E[y_i|x_i] = \exp(x_i'\theta)$, where the non-negative $\exp(\cdot)$ in the right-hand side is used to model the non-negative trade volumes between countries. NLS sets up the loss function $\rho_i(\theta) = (y_i - \exp(x_i'\theta))^2$.

Example 1.3. If $y \sim \text{Poisson}(\lambda)$, then $\Pr(y = k) = \lambda^k \exp(-\lambda)/k!$ for $k \in \{0, 1, 2, \ldots\}$. An alternative way to set up the criterion is to pretend that y_i is draw from a conditional Poisson model with mean $\lambda_i = \exp(x_i'\theta)$ and thus

$$\rho_i(\theta) = -\log \Pr(y_i|x_i) = -y_i \cdot x_i'\theta + \exp\left(x_i'\theta\right)$$

where we drop $y_i!$ from the above express as it is irrelevant to the parameter θ. Here we use the "minus log-likelihood" to be consistent with the minimization for M-estimators, instead of maximizing the log-likelihood. This is the loss function of so called PPML (Pseudo Poisson Maximum Likelihood) in Silva and Tenreyro (2006).

For simplicity, in this lecture we work with iid data. Let

$$S(\theta) = E[S_n(\theta)] = E[\rho_i(\theta)]$$

be the population criterion function.

In econometrics, identification has different meanings in different contexts. Lewbel (2019) is an overview of the "identification zoo". The definition here is rather mechanical and has nothing to do with economics. Let Θ be the **parameter space**. We say θ is **identified** if $\theta_0 = \arg\min_{\theta \in \Theta} S(\theta)$ is unique. More formally:

Definition 1.1 (Identification)**.** For any $\varepsilon > 0$ there exists a $\delta = \delta(\varepsilon)$ such that

$$\inf_{\theta \in \Theta \setminus N_\delta(\theta_0)} S(\theta) > S(\theta_0) + \varepsilon$$

where $N_\delta(\theta_0) := \{\theta \in \Theta : \|\theta - \theta_0\| < \delta\}$ is an open δ-neighborhood around θ_0.

Identification is a property of the underlying probabilistic model. It has nothing to do with the data or the randomness in sampling.

1.2 Consistency

To establish the consistency of the M-estimator, identification is a necessary condition. In the previous lecture we have covered the consistency of a single sequence of random variables. In M-estimation the model is index by $\theta \in \Theta$. Pointwise consistency $S_n(\theta) \stackrel{p}{\to} S(\theta)$ is insufficient to guarantee the consistency of $\widehat{\theta}$ to the true parameter θ_0. We need to strength it to the **uniform consistency** of the sample mean.

Definition 1.2 (Uniform law of large numbers)**.** For any $\eta, \varepsilon > 0$, there exists an $N = N(\varepsilon, \eta)$ such that

$$\Pr\left\{\sup_{\theta \in \Theta} |S_n(\theta) - S(\theta)| \geq \varepsilon\right\} \leq \eta$$

for all $n > N$. More concisely, we can write $\sup_{\theta \in \Theta} |S_n(\theta) - S(\theta)| \stackrel{p}{\to} 0$ as $n \to \infty$.

Notice that uniform consistency is stronger than **pointwise consistency**: $S_n(\theta) \stackrel{p}{\to} S(\theta)$ for every $\theta \in \Theta$.

Remark 1.1. Recall the definition of pointwise continuity and uniform continuity in undergraduate calculus. A function $f(x)$ is **pointwisely continuous** in an open set A, if given any $\varepsilon > 0$, for every $x \in A$ there exists a $\delta = \delta(\varepsilon, x) > 0$ such that

$$\left|x' - x\right| \leq \delta \Rightarrow \left|f\left(x'\right) - f(x)\right| < \varepsilon.$$

The function $f(x)$ is **uniformly continuous** if δ does not depend on x. In other words, there exists a finite bound L such that

$$\sup_{x,x'\in A,\, x\neq x'} \frac{|f(x') - f(x)|}{\|x - x'\|} \leq L$$

For example, $\sin(x)$ is uniform continuous in $\mathbb{R}$; $1/x$ in $(0,\infty)$ is pointwisely continuous but not uniformly continuous, and the same applies to x^2 in $(-\infty,\infty)$.

Theorem 1.1. *If (i) ULLN:* $\sup_{\theta\in\Theta} |S_n(\theta) - S(\theta)| \xrightarrow{p} 0$*; (ii)* θ_0 *is identified, then* $\hat{\theta} \xrightarrow{p} \theta_0$ *as* $n \to \infty$.

Proof. We start from the condition of identification.

$$\begin{aligned}
\Pr\left(|\hat{\theta} - \theta| > \delta\right) &\leq \Pr\left(S(\hat{\theta}) - S(\theta_0) > \varepsilon\right) \\
&= \Pr\left(S(\hat{\theta}) - S_n(\hat{\theta}) + S_n(\hat{\theta}) - S_n(\theta_0) + S_n(\theta_0) - S(\theta_0) > \varepsilon\right) \\
&\leq \Pr\left(S(\hat{\theta}) - S_n(\hat{\theta}) + S_n(\theta_0) - S(\theta_0) > \varepsilon\right) \\
&\leq \Pr\left(\left|S_n(\hat{\theta}) - S(\hat{\theta})\right| + |S(\theta_0) - S_n(\theta_0)| > \varepsilon\right) \\
&\leq \Pr\left(\sup_{\theta\in\Theta} |S_n(\theta) - S(\theta)| \geq \frac{\varepsilon}{2}\right) \to 0
\end{aligned}$$

where the second inequality follows from the definition of the M-estimator that $S_n(\hat{\theta}) \leq S_n(\theta_0)$. □

We have established consistent of the M-estimator $\widehat{\theta}$ to the true parameter θ_0.

1.3 Asymptotic Normality

To further characterize the uncertainty of the estimator, we seek its asymptotic distribution.

We go with a heuristic argument. Define $\psi_i(\theta) = \frac{\partial}{\partial\theta}\rho_i(\theta)$ and $\bar{\psi}(\theta) = \frac{\partial}{\partial\theta} S_n(\theta)$. A Taylor expansion of $\bar{\psi}(\hat{\theta})$ around θ_0 gives

$$0 = \bar{\psi}(\hat{\theta}) = \bar{\psi}(\theta_0) + \frac{\partial^2}{\partial\theta\partial\theta'} S_n(\dot{\theta})\left(\hat{\theta} - \theta_0\right)$$

where $\dot{\theta}$ lies in between $\widehat{\theta}$ and θ_0. If $\frac{\partial^2}{\partial\theta\partial\theta'} S_n(\dot{\theta})$ is invertible, we can rearrange the above inequality,

$$\sqrt{n}\left(\hat{\theta} - \theta\right) = -\left[\frac{\partial^2}{\partial\theta\partial\theta'} S_n(\dot{\theta})\right]^{-1} \sqrt{n}\bar{\psi}(\theta_0).$$

Since $\hat{\theta} \xrightarrow{p} \theta_0$, we also have $\dot{\theta} \xrightarrow{p} \theta_0$. By the continuous mapping theorem:

$$\frac{\partial^2}{\partial\theta\partial\theta'} S\left(\dot{\theta}\right) \xrightarrow{p} \frac{\partial^2}{\partial\theta\partial\theta'} S(\theta_0) = Q$$

if $\frac{\partial^2}{\partial\theta\partial\theta'} S(\cdot)$ is continuous. (Ultimately we want to show $\frac{\partial^2}{\partial\theta\partial\theta'} S_n(\dot{\theta}) \xrightarrow{p} Q$ but our heuristic argument here has a gap, because $\dot{\theta}$ is moving as $n \to \infty$. The textbook provides a rigorous proof invoking the empirical process theory.) In the population, $E\left[\bar{\psi}(\theta_0)\right] = E\left[\psi(\theta_0)\right] = 0$, and

$$\sqrt{n}\bar{\psi}(\theta_0) \xrightarrow{d} N(0, \Omega)$$

where $\Omega = E\left[\psi_i(\theta_0)\,\psi_i'(\theta_0)\right]$. As a result,

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right) \xrightarrow{d} N\left(0, Q^{-1}\Omega Q^{-1}\right)$$

where the asymptotic variance follows a sandwich form.

1.4 Examples

1.4.1 OLS

For OLS, the population criterion function is

$$\begin{aligned}S(\theta) &= E\left[\left(y_i - x_k'\theta\right)^2\right] = E\left[\left(y - x_i'\theta_0 + x'\theta_0 - x_i'\theta\right)^2\right] = E\left[\left(e_i + x_i'(\theta_0-\theta)\right)^2\right] \\ &= E\left[e_i^2\right] + E\left[\left(x_i'(\theta_0-\theta)\right)^2\right] = E\left[e_i^2\right] + (\theta_0-\theta)\,E\left[x_i x_i'\right](\theta_0-\theta).\end{aligned}$$

Exercise 1.1. Verify that θ_0 is identified if $E\left[x_i x_i'\right]$ is of full rank.

Given the sample, we have $\psi_i(\theta) = -x_i(y_i - x_i'\theta)$ and $\frac{\partial^2}{\partial\theta\partial\theta'}\rho_i(\theta) = x_i x_i'$. Evaluated at $\theta = \theta_0$, we have

$$\psi_i(\theta_0) = -x_i\left(y_i - x_i'\theta_0\right) = -x_i e_i,$$

and obviously the assumption that x_i and e_i are orthogonal implies $E\left[\psi_i(\theta_0)\right] = 0$. When $E\left[x_i x_i' e_i^2\right] < \infty$ is finite, the Lindeberg-Levy CLT gives

$$\sqrt{n}\bar{\psi}(\theta_0) \xrightarrow{d} N\left(0, E\left[x_i x_i' e_i^2\right]\right).$$

Moreover $\Omega = E\left[x_i x_i'\right]$. Therefore we conclude

$$\sqrt{n}\left(\widehat{\theta}-\theta_0\right) \xrightarrow{d} N\left(0, \left(E\left[x_i x_i'\right]\right)^{-1} E\left[x_i x_i' e_i^2\right]\left(E\left[x_i x_i'\right]\right)^{-1}\right).$$

This is the asymptotic distribution of the OLS estimator under conditional heteroskedasticity.

1.4.2 Logistic Regression

The classical econometric *random utility model* is

$$y_i^* = x_i'\theta + \varepsilon_i$$

where y_i^* is a latent response variable ("latent" means unobservable by the econometrician). What is observable is $y_i = 1\{y_i^* \geq 0\}$. That is, if the latent utility is greater than a threshold (set as 0, without loss of generality), then we observe $y_i = 1$; otherwise $y_i = 0$. While y_i^* is a continuous random variable, y_i is a binary random variable.

The conditional probability of observing $y_i = 1$ is

$$\Pr(y_i = 1|x_i) = \Pr(y_i^* \geq 0|x_i) = \Pr\left(x_i'\theta + \varepsilon_i \geq 0|X\right) = \Pr\left(-\varepsilon_i \leq x_i'\theta|x_i\right).$$

Assume ε_i is independent of x_i and its PDF symmetric around 0, then $\Pr(-\varepsilon_i \leq x_i'\theta|x_i) = F_\varepsilon(x_i'\theta)$, where $F_\varepsilon(\cdot)$ is the CDF of ε. When $\varepsilon \sim$ Logistic, we call it the *Logit regression* or *Logistic regression*; if $\varepsilon \sim N(0,1)$, we call it the *Probit regression*.

Exercise 1.2. Let $\Lambda(z) = \frac{1}{1+\exp(-z)}$. Verify

$$\frac{d}{dz}\Lambda(z) = \Lambda(z)(1-(z)).$$

This is a useful property for the Logistic regression.

Logistic regression. Given x_i, the conditional probability

$$\Pr(y_i = 1|x_i) = \Lambda\left(x_i'\theta\right).$$

For simplicity we denote $\Lambda_i = \Lambda(x_i'\theta)$. The probability of observing y_i conditional on x_i is $f(y_i|x_i) = \Lambda_i^{y_i}(1-\Lambda_i)^{(1-y_i)}$, and thus the *negative conditional* log-likelihood for (y_i, x_i) is

$$\rho_i(\theta) = -y_i \log \Lambda_i - (1-y_i)\log(1-\Lambda_i)$$

and

$$\begin{aligned} E[\rho_i(\theta)|x_i] &= -E[y_i|x_i]\log\Lambda_i - (1-E[y_i|x_i])\log(1-\Lambda_i) \\ &= -\Lambda_{i0}\log\Lambda_i - (1-\Lambda_{i0})\log(1-\Lambda_i) \end{aligned}$$

where and $\Lambda_{i0} = \Lambda(x_i'\theta_0)$.

The score function is

$$\psi_i(\theta) = \frac{\partial}{\partial\theta}\rho_i(\theta) = x_i\left[\frac{y_i}{\Lambda_i}\Lambda_i(1-\Lambda_i) - \frac{1-y_i}{1-\Lambda_i}\Lambda_i(1-\Lambda_i)\right] = x_i\left[y_i(1-\Lambda_i) - (1-y_i)\Lambda_i\right].$$

The conditional expectation of the score is

$$E[\psi_i(\theta)|x_i] = x_i\left[\Lambda_0(1-\Lambda_i) - (1-\Lambda_0)\Lambda_i\right]$$

and therefore $E[\psi_i(\theta_0)|x_i] = 0$. The Hessian is

$$H_i(\theta) = \frac{\partial}{\partial\theta'}\psi_i(\theta) = x_i x_i'\left[y_i\Lambda_i(1-\Lambda_i) - (1-y_i)\Lambda_i(1-\Lambda_i)\right] = x_i x_i'\Lambda_i(1-\Lambda_i)$$

is irrelevant to y_i and it is positive definite for any bounded θ, under which $\Lambda_i \in (0,1)$.

`Zhentao Shi. January 24, 2024`

Bibliography

Lewbel, A. (2019). The identification zoo: Meanings of identification in econometrics. *Journal of Economic Literature 57*(4), 835–903. 1.1

Silva, J. S. and S. Tenreyro (2006). The log of gravity. *The Review of Economics and Statistics 88*(4), 641–658. 1.2, 1.3