Advanced Econometrics

Chapter 11: M-Estimation

In Choi

Sogang University

References

- Newey, W. and D. McFadden (1994). "Large sample estimation and hypothesis testing," Chapter 36 in R. Engle and D. McFadden (eds.) *Handbook of Econometrics*, Vol. IV, North-Holland.
- Van der Vaart, A. W. (2000). *Asymptotic statistics* (Cambridge series in statistical and probabilistic mathematics).
- Wooldridge, J. (2002). *Econometric Analysis of Cross Section and Panel Data*, The MIT Press.

Introduction

- In nonlinear regressions, $E(y \mid x)$ is modelled as a nonlinear function of x and $\theta_0 \in \Theta \subset R^P$. Let $m(x, \theta)$ be a known function of x and θ. Then, a typical nonlinear regression model is written as

$$y_i = m(x, \theta_0) + u, \;\; E(u \mid x) = 0.$$

Here, θ_0 denotes "the true value of θ."

- Why is it possible to estimate θ_0? Because θ_0 solves the population problem

$$\min_{\theta \in \Theta} E[y - m(x, \theta)]^2, \tag{1}$$

where the expectation is taken over the joint distribution of (x, y).

- Write

$$[y - m(x,\theta)]^2 = [y - m(x,\theta_0)]^2 + 2[m(x,\theta_0) - m(x,\theta)]u + [m(x,\theta_0) - m(x,\theta)]^2.$$

Because $E[m(x,\theta_0) - m(x,\theta)]u$ due to the given assumption $E(u \mid x) = 0$ and because $[m(x,\theta_0) - m(x,\theta)]^2 \geq 0$ with probability 1, we have

$$E[y - m(x,\theta)]^2 \geq E[y - m(x,\theta_0)]^2, \ \theta \in \Theta.$$

- The nonlinear least squares (NLS) estimator of θ_0 is obtained by solving the sample analogue of the population problem (1). That is, the NLS estimator of θ_0, $\hat{\theta}$, solves the problem

$$\min_{\theta \in \Theta} N^{-1} \sum_{i=1}^{N} [y_i - m(x_i,\theta)]^2.$$

Introduction

- The M-estimator generalizes the NLS estimator. Let $w_i = (x_i, y_i)$. An M-estimator of θ_0, $\hat{\theta}$, solves the problem

$$\min_{\theta \in \Theta} N^{-1} \sum_{i=1}^{N} q(w_i, \theta). \tag{2}$$

  Obviously, the NLS estimator is a special case of the M-estimator.
- The parameter vector θ_0 is assumed to uniquely solve the population problem

$$\min_{\theta \in \Theta} Eq(w, \theta).$$

Example

Suppose that $x_1, \ldots, x_n$ are independent samples having the common density $p(\cdot \mid \theta)$. The MLE maximizes the log likelihood function

$$\theta \longmapsto \sum_{i=1}^{n} \log p(x_i \mid \theta).$$

Introduction

- Intuition for the consistency of the M-estimator: Since

$$N^{-1} \sum_{i=1}^{N} q(w_i, \theta) \xrightarrow{p} Eq(w, \theta), \tag{3}$$

  it should follow that $\hat{\theta} \xrightarrow{p} \theta$.

- Two issues:
  (i) Identifiability of θ_0 (a population issue).
  (ii) Convergence in equation (3) across different values of $\theta \in \Theta$.

- It is required that θ_0 is the unique solution for the inequality

$$Eq(w, \theta_0) < Eq(w, \theta), \;\; \theta \in \Theta, \;\; \theta \neq \theta_0. \tag{4}$$

- We need the uniform weak law of large numbers for consistency of the M-estimator

$$\max_{\theta\in\Theta}\left|N^{-1}\sum_{i=1}^{N}q(w_i,\theta)-Eq(w,\theta)\right|\xrightarrow{p}0. \tag{5}$$

Theorem 1 Let w be a random vector taking values in $\mathcal{W}\subset\mathbb{R}^M$, let Θ be a subset of $\mathbb{R}^P$, and let $q:\mathcal{W}\times\times\rightarrow\mathbb{R}$ be a real-valued function. Assume
(a) Θ is compact;
(b) for each $\theta\in\Theta$, $q(\cdot,\theta)$ is Borel measurable on $\mathcal{W}$;
(c) for each $w\in\mathcal{W}$, $q(w,\cdot)$ is continuous on Θ;
(d) $|q(w,\theta)|\leq b(w)$ for all $\theta\in\Theta$, where b is a nonnegative function on $\mathcal{W}$ such that $Eb(w)<\infty$.
Then, the uniform weak law of large numbers (5) holds.

- Under the assumptions of a compact parameter space and continuity, both the population and sample optimization problems have solutions.

Theorem 2 Suppose that (4) and assumptions in Theorem 1 hold. Then, the M-estimator solves the problem (2) and is consistent for θ_0.

Lemma Suppose that $\hat{\theta} \xrightarrow{p} \theta_0$ and assume that $r(w, \theta)$ satisfies the same assumptions on $q(w, \theta)$ in Theorem 1. Then,

$$N^{-1} \sum_{i=1}^{N} r(w_i, \hat{\theta}) \xrightarrow{p} Er(w, \theta_0).$$

Asymptotic normality

- If $q(w,\cdot)$ is continuously differentiable on the interior of Θ, we have

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

where $s(w_i,\theta) = \frac{\partial q(w_i,\theta)}{\partial \theta}$.

- If $q(w,\cdot)$ is twice continuously differentiable on the interior of Θ, we have by the mean-value theorem

$$\sum_{i=1}^{N} s(w_i,\hat{\theta}) = \sum_{i=1}^{N} s(w_i,\theta_0) + \left(\sum_{i=1}^{N} H_i^*\right)(\hat{\theta}-\theta_0),$$

where $H_i^* = \frac{\partial^2 q(w_i,\theta)}{\partial\theta\partial\theta'}_{|\theta=\theta^*}$ and θ^* is one line connecting $\hat{\theta}$ and θ_0.

Asymptotic normality

- Thus, we have

$$\sqrt{N}(\hat{\theta} - \theta_0) = \left(N^{-1} \sum_{i=1}^{N} H_i^* \right) \left[-N^{-1/2} \sum_{i=1}^{N} s(w_i, \theta_0) \right].$$

- The law of large numbers works for $N^{-1} \sum_{i=1}^{N} H_i^*$, and CLT for $N^{-1/2} \sum_{i=1}^{N} s(w_i, \theta_0)$. Thus, $\sqrt{N}-$asymptotic normality follows.

Asymptotic normality

Theorem 3 Suppose that assumptions in Theorem 2 hold. In addition, assume

(a) θ_0 is in the interior of Θ;

(b) $q(w, \cdot)$ is twice continuously differentiable on the interior of Θ for all $w \in \mathcal{W}$;

(c) Each element of $H(w, \theta)$ is bounded in absolute value by a function $b(w)$ with $Eb(w) < \infty$;

(d) $A_0 = EH(w, \theta_0)$ is positive definite;

(e) $Es(w, \theta_0) = 0$;

(f) each element of $s(w, \theta_0)$ has finite second moment.

Then, letting $B_0 = Es(w, \theta_0)s(w, \theta_0)'$, we have

$$\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, A_0^{-1} B_0 A_0^{-1}).$$

Asymptotic normality

- Theorem 3 does not cover the LAD estimator because $q(w_i, \theta) = |y_i - m(x_i, \theta)|$ is not twice continuously differentiable and because the Hessian is a zero matrix.

Asymptotic normality

- In case of the NLS estimator, let

$$q(w,\theta) = [y - m(x,\theta)]^2/2.$$

The score vector is written as

$$s(w,\theta) = -\nabla_\theta m(x,\theta)[y - m(x,\theta)],$$

where $\nabla_\theta m(x,\theta) = \partial m(x,\theta)/\partial\theta$. Then,

$$E[s(w,\theta_0) \mid x] = -\nabla_\theta m(x,\theta)[E(y \mid x) - m(x,\theta)] = 0$$

and

$$B_0 = E[s(w,\theta_0)s(w,\theta_0)'] = E[u^2 \nabla_\theta m(x,\theta)\nabla_\theta m(x,\theta)'].$$

Asymptotic normality

- The Hessian of $q(w, \theta)$ is

$$H(w, \theta) = \nabla_\theta m(x, \theta) \nabla_\theta m(x, \theta)' - \nabla_\theta^2 m(x, \theta)[y - m(x, \theta)],$$

giving

$$E[H(w, \theta_0) \mid x] = \nabla_\theta m(x, \theta_0) \nabla_\theta m(x, \theta_0)'$$

and

$$A_0 = E[\nabla_\theta m(x, \theta_0) \nabla_\theta m(x, \theta_0)'].$$

Estimating the asymptotic variance

- A_0 can be estimated by

$$\hat{A}_0 = N^{-1} \sum_{i=1}^{N} H(w_i, \hat{\theta}).$$

  This may not be positive-definite.

- Alterantively, define $A(x, \theta_0) = E[H(w, \theta) \mid x]$. Since $E[A(x, \theta_0)] = E[H(w, \theta)] = A_0$, it follows that

$$\hat{A}_0 = N^{-1} \sum_{i=1}^{N} A(x_i, \hat{\theta}) \xrightarrow{p} A_0.$$

  This estimator is convinient to use when $E[H(w, \theta) \mid x]$ can be obtained in closed form.

Estimating the asymptotic variance

- B_0 can be estimated by

$$\hat{B}_0 = N^{-1}\sum_{i=1}^{N} s(w_i,\hat{\theta})s(w_i,\hat{\theta})'.$$

- The estimator of asymptotic variance is

$$\hat{V} = \hat{A}_0^{-1}\hat{B}_0\hat{A}_0^{-1}.$$

Estimating the asymptotic variance

- In the case of nonlinear least squares,

$$\hat{A}_0 = N^{-1} \sum_{i=1}^{N} \nabla_\theta m(x, \hat{\theta}) \nabla_\theta m(x, \hat{\theta})';$$

$$\hat{B}_0 = N^{-1} \sum_{i=1}^{N} \hat{u}_i^2 \nabla_\theta m(x, \hat{\theta}) \nabla_\theta m(x, \hat{\theta})',$$

where $\hat{u}_i = y_i - m(x_i, \hat{\theta})$.

- The null hypothesis is

$$H_0 : c(\theta_0) = 0,$$

where $c(\theta_0)$ is a $Q \times 1$ vector.

- The Wald statistic for this null hypothesis is

$$W = c(\hat{\theta})'[\hat{C}\hat{V}\hat{C}']^{-1}c(\hat{\theta}),$$

where $\hat{C} = \nabla_\theta c(\hat{\theta})$ and $\hat{V}$ is the estimator of asymptotic variance.

- Under H_0,

$$W \xrightarrow{d} \chi^2(Q).$$

- The null hypothesis is

$$H_0 : c(\theta_0) = 0,$$

  where $c(\theta_0)$ is a $Q \times 1$ vector.
- When the unrestricted model is diffult to estimate but the restricted model is relatively easy to do so, it is convenient to use the score test.
- Assume we can write $\theta_0 = d(\lambda_0)$, where $d : R^{P-Q} \rightarrow R^P$.
- Assume λ_0 is interior of its parameter space Λ under the null hypothesis.

- Let $\tilde{\lambda}$ be the solution to the constrained minimization problem

$$\min_{\lambda \in \Lambda} N^{-1} \sum_{i=1}^{N} q(w_i, d(\lambda)),$$

which gives the constrained estimator of θ_0, $\tilde{\theta} = d(\tilde{\lambda})$.

Example

Suppose that

$$m(x, \theta) = \exp\left[x'\beta + \delta_1 (x'\beta)^2 + \delta_2 (x'\beta)^3\right].$$

If $H_0 = \delta_1 = \delta_2 = 0$, the model with the restrictions imposed is just an exponential regression function $m(x, \beta) = \exp(x'\beta)$.

- A mean value expansion gives

$$N^{-1/2}\sum_{i=1}^{N} s_i(\tilde{\theta}) = N^{-1/2}\sum_{i=1}^{N} s_i(\theta_0) + A_0 N^{-1/2}(\tilde{\theta}-\theta_0) + o_p(1).$$

Because

$$0 = N^{-1/2}c(\tilde{\theta}) = N^{-1/2}c(\theta_0) + \ddot{C}N^{-1/2}(\tilde{\theta}-\theta_0),$$

where $\ddot{C}$ is the first derivative of $c(\theta)$ evaluated at θ oin the line joining $\tilde{\theta}$ and θ_0 converging to C_0, and because $c(\theta_0)=0$ under the null,

$$\ddot{C}N^{-1/2}(\tilde{\theta}-\theta_0) = o_p(1).$$

- Thus,

$$C_0 A_0^{-1} N^{-1/2}\sum_{i=1}^{N} s_i(\tilde{\theta}) = C_0 A_0^{-1} N^{-1/2}\sum_{i=1}^{N} s_i(\theta_0) + o_p(1)$$

- By CLT,

$$C_0 A_0^{-1} N^{-1/2} \sum_{i=1}^{N} s_i(\theta_0) \xrightarrow{d} N(0, C_0 A_0^{-1} B_0 A_0^{-1} C_0').$$

- The LM test statistic is defined by

$$LM = \left(\sum_{i=1}^{N} \tilde{s}_i \right)' \tilde{A}^{-1} \tilde{C}' \left(\tilde{C} \tilde{A}^{-1} B \tilde{A}^{-1} \tilde{C}' \right)^{-1} \tilde{C} \tilde{A}^{-1} \left(\sum_{i=1}^{N} \tilde{s}_i \right)$$

and its limiting distribution is $\chi^2(Q)$.