Microeconometría I

Maestría en Econometría

Lecture 4

Agenda

1. M-Estimation
   - Introduction
   - Identification, Uniform Convergence, and Consistency
   - Asymptotic Normality
2. Two-Step M-Estimation
   - Consistency
   - Asymptotic Normality of Two-Step M-Estimators
   - Estimating the Asymptotic Variance
   - Adjustments for when we cannot ignore the first-stage estimation

Agenda

M-Estimation

- M-estimation methods include maximum likelihood, nonlinear least squares, least absolute deviations, quasi-maximum likelihood, and many other procedures used by econometricians.
- In a nonlinear regression model, we have a random variable, y, and we would like to model $E(y|x)$ as a function of the explanatory variables x, a K-vector.
- We already know how to estimate models of $E(y|x)$ when the model is linear in its parameters: OLS produces consistent, asymptotically normal estimators.
- What happens if the regression function is nonlinear in its parameters?

M-Estimation

- Generally, let $m(x;\theta)$ be a parametric model for $E(y|x)$, where m is a known function of x and θ, and θ is a $P \times 1$ parameter vector.
- This is a parametric model because $m(x;\theta)$ is assumed to be known up to a finite number of parameters.
- The dimension of the parameters, P, can be less than or greater than K. The parameter space, Θ, is a subset of $\mathbb{R}^P$
- This is the set of values of y that we are willing to consider in the regression function. Unlike in linear models, for nonlinear models the asymptotic analysis requires explicit assumptions on the parameter space

M-Estimation

- An example of a nonlinear regression function is the logistic function, $m(x;\theta) = \exp(x\theta)/[1+\exp(x\theta)]$. The logistic function is nonlinear in θ.
- We say that we have a correctly specified model for the conditional mean, $E(y|x)$, if, for some $\theta_{\mathrm{o}} \in \Theta$,

$$\mathrm{E}(y \mid \mathbf{x}) = m\left(\mathbf{x}, \boldsymbol{\theta}_{\mathrm{o}}\right) \tag{1}$$

- We introduce the subscript "o" on theta to distinguish the parameter vector appearing in $E(y|x)$ from other candidates for that vector.
- Often, the value θ_{o} is called the true value of theta.

M-Estimation

- Equation (1) is the most general way of thinking about what nonlinear least squares is intended to do: estimate models of conditional expectations.
- As a statistical matter, equation (1) is equivalent to a model with an additive, unobservable error with a zero conditional mean:

$$y = m(\mathbf{x}, \boldsymbol{\theta}_{\mathrm{o}}) + u, \quad \mathrm{E}(u \mid \mathbf{x}) = 0, \tag{2}$$

- Given equation (1), we obtain equation (2) by defining the error to be $u \equiv y - m(\mathbf{x}, \boldsymbol{\theta}_{\mathrm{o}})$.
- We formalize the first nonlinear least squares (NLS) assumption as follows: Assumption NLS.1: For some $\boldsymbol{\theta}_{\mathrm{o}} \in \boldsymbol{\Theta}, \mathrm{E}(y \mid \mathbf{x}) = m(\mathbf{x}, \boldsymbol{\theta}_{\mathrm{o}})$.

M-Estimation

- If we let $\mathbf{w} \equiv (\mathbf{x}, y)$, then $\boldsymbol{\theta}_{\mathrm{o}}$ indexes a feature of the population distribution of $\mathbf{w}$, namely, the conditional mean of y given x.
- More generally, let $\mathbf{w}$ be an M-vector of random variables with some distribution in the population.
- We let $\mathscr{W}$ denote the subset of $\mathbb{R}^M$ representing the possible values of $\mathbf{w}$.
- Let $\boldsymbol{\theta}_{\mathrm{o}}$ denote a parameter vector describing some feature of the distribution of $\mathbf{w}$ (i.e. a conditional mean).
- We assume that $\boldsymbol{\theta}_{\mathrm{o}}$ belongs to a known parameter space $\boldsymbol{\Theta} \subset \mathbb{R}^P$.
- We assume that our data come as a random sample of size N from the population; we label this random sample $\{\mathbf{w}_i : i = 1, 2, \ldots\}$, where each $\mathbf{w}_i$ is an M-vector.

M-Estimation

- What allows us to estimate $\boldsymbol{\theta}_{\mathrm{o}}$ when it indexes $E(y|x)$? It is the fact that $\boldsymbol{\theta}_{\mathrm{o}}$ is the value of $\boldsymbol{\theta}$ that minimizes the expected squared error between y and $m(x;\boldsymbol{\theta})$.
- That is, $\boldsymbol{\theta}_{\mathrm{o}}$ solves the population problem

$$\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathrm{E}\left\{[y - m(\mathbf{x},\boldsymbol{\theta})]^2\right\}, \tag{3}$$

  where the expectation is over the joint distribution of $(\mathbf{x}, y)$.
- Because $\boldsymbol{\theta}_{\mathrm{o}}$ solves the population problem in expression (3), the analogy principle suggests estimating $\boldsymbol{\theta}_{\mathrm{o}}$ by solving the sample analogue.

M-Estimation

- In other words, we replace the population moment $\mathrm{E}\left\{[(y - m(\mathbf{x}, \boldsymbol{\theta})]^2\right\}$ with the sample average.
- The NLS estimator of $\boldsymbol{\theta}_{\mathrm{o}}$, $\hat{\boldsymbol{\theta}}$, solves

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} N^{-1} \sum_{i=1}^{N} \left[y_i - m\left(\mathbf{x}_i, \boldsymbol{\theta}\right)\right]^2 \tag{4}$$

  For now, we assume that a solution to this problem exists.
- The NLS objective function in expression (3) is a special case of a more general class of estimators. Let $q(\mathbf{w}, \boldsymbol{\theta})$ be a function of the random vector $\mathbf{w}$ and the parameter vector $\boldsymbol{\theta}$.

M-Estimation

- An M-estimator of $\boldsymbol{\theta}_{\mathrm{o}}$ solves the problem

$$\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} N^{-1}\sum_{i=1}^{N} q\left(\mathbf{w}_i, \boldsymbol{\theta}\right), \tag{5}$$

  assuming that a solution, call it $\hat{\boldsymbol{\theta}}$, exists. The estimator clearly depends on the sample $\{\mathbf{w}_i : i = 1, 2, \ldots\}$, but we suppress that fact in the notation.
- The parameter vector $\boldsymbol{\theta}_{\mathrm{o}}$ is assumed to uniquely solve the population problem

$$\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathrm{E}[q(\mathbf{w}, \boldsymbol{\theta})], \tag{6}$$

Agenda

M-Estimation

- How do we translate the fact that $\boldsymbol{\theta}_{\mathrm{o}}$ solves the population problem (6) into consistency of the M-estimator $\hat{\boldsymbol{\theta}}$ that solves problem (5)?
- Heuristically, the argument is as follows. Since for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}, \{q(\mathbf{w}_i, \boldsymbol{\theta}) : i = 1, 2, \ldots\}$ is just an i.i.d. sequence, the law of large numbers implies that

$$N^{-1} \sum_{i=1}^{N} q(\mathbf{w}_i, \boldsymbol{\theta}) \xrightarrow{p} \mathrm{E}[q(\mathbf{w}, \boldsymbol{\theta})], \tag{7}$$

  under very weak finite moment assumptions.
- Since $\hat{\boldsymbol{\theta}}$ minimizes the function on the left hand side of (7) and $\boldsymbol{\theta}_{\mathrm{o}}$ minimizes the function on the right, it seems plausible that $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_{\mathrm{o}}$.

M-Estimation

- There are essentially two issues to address.
- The first is identifiability of $\boldsymbol{\theta}_{\mathrm{o}}$, which is purely a population issue.
- The second is the sense in which the convergence in equation (7) happens across different values of $\boldsymbol{\theta}$ in $\boldsymbol{\Theta}$.
- For nonlinear regression, we showed how $\boldsymbol{\theta}_{\mathrm{o}}$ solves the population problem (3). However, we did not argue that $\boldsymbol{\theta}_{\mathrm{o}}$ is always the unique solution to problem (3).
- Whether or not this is the case depends on the distribution of $\mathbf{x}$ and the nature of the regression function:

  Assumption NLS.2: $\mathrm{E}\left\{\left[m\left(\mathbf{x}, \boldsymbol{\theta}_{\mathrm{o}}\right) - m(\mathbf{x}, \boldsymbol{\theta})\right]^{2}\right\} > 0, \text{ all } \boldsymbol{\theta} \in \boldsymbol{\Theta}, \boldsymbol{\theta} \neq \boldsymbol{\theta}_{\mathrm{o}}$
- Assumption NLS.2 plays the same role as the assumption of no multicollinearity in OLS.

M-Estimation

- For the general M-estimation case, we assume that $q(\mathbf{w}, \boldsymbol{\theta})$ has been chosen so that $\boldsymbol{\theta}_{\mathrm{o}}$ is a solution to problem (6).
- Identification requires that $\boldsymbol{\theta}_{\mathrm{o}}$ be the unique solution:

$$\mathrm{E}\left[q\left(\mathbf{w}, \boldsymbol{\theta}_{\mathrm{o}}\right)\right] < \mathrm{E}[q(\mathbf{w}, \boldsymbol{\theta})], \quad \text{all } \boldsymbol{\theta} \in \boldsymbol{\Theta}, \quad \boldsymbol{\theta} \neq \boldsymbol{\theta}_{\mathrm{o}}, \tag{8}$$

- The second component for consistency of the M-estimator is convergence of the sample average $N^{-1} \sum_{i=1}^{N} q\left(\mathbf{w}_{i}, \boldsymbol{\theta}\right)$ to its expected value.
- It is not enough to simply invoke the usual weak law of large numbers at each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

M-Estimation

- Instead, uniform convergence in probability is sufficient. Mathematically,

$$\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left| N^{-1} \sum_{i=1}^{N} q\left(\mathbf{w}_i, \boldsymbol{\theta}\right) - \mathrm{E}[q(\mathbf{w}, \boldsymbol{\theta})] \right| \xrightarrow{p} 0, \tag{9}$$

- Uniform convergence clearly implies pointwise convergence, but the converse is not true: it is possible for equation (7) to hold but equation (9) to fail.
- To state a formal result concerning uniform convergence, we need to be more careful in stating assumptions about the function $q(\cdot, \cdot)$ and the parameter space $\boldsymbol{\Theta}$.
- Technically, we should assume that $q(\cdot, \boldsymbol{\theta})$ is a Borel measurable function on $\mathscr{W}$ for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

M-Estimation

- The next assumption concerning q is practically more important.
- We assume that, for each$\mathbf{w} \in \mathscr{W}$, $q(\mathbf{w}, \cdot)$ is a continuous function over the parameter space $\mathbf{\Theta}$.
- We can now state a theorem concerning uniform convergence appropriate for the random sampling environment. This result, known as the uniform weak law of large numbers (UWLLN), dates back to LeCam (1953).

  Theorem 1 (Uniform Weak Law of Large Numbers): Let $\mathbf{w}$ be a random vector taking values in $\mathscr{W} \subset \mathbb{R}^M$, let $\mathbf{\Theta}$ be a subset of $\mathbb{R}^P$ and let $q : \mathscr{W} \times \mathbf{\Theta} \to \mathbb{R}$ be a real valued function. Assume that (a) $\mathbf{\Theta}$ is compact; (b) for each $\boldsymbol{\theta} \in \mathbf{\Theta}$, $q(\cdot, \boldsymbol{\theta})$ is Borel measurable on $\mathscr{W}$; (c) for each $\mathbf{w} \in \mathscr{W}$, $q(\mathbf{w}, \cdot)$ is continuous on $\mathbf{\Theta}$; and (d) $|q(\mathbf{w}, \boldsymbol{\theta})| \leq b(\mathbf{w})$ for all $\boldsymbol{\theta} \in \mathbf{\Theta}$, where b is a nonnegative function on $\mathscr{W}$ such that $\mathrm{E}[b(\mathbf{w})] < \infty$. Then equation (9) holds.

M-Estimation

- Theorem 2 (Consistency of M-Estimators): Under the assumptions of Theorem 1, assume that the identification assumption (8) holds. Then a random vector, $\hat{\boldsymbol{\theta}}$, solves problem (5), and $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_{\mathrm{o}}$.
- Lemma 1: Suppose that $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_{\mathrm{o}}$, and assume that $r(\mathbf{w}, \boldsymbol{\theta})$ satisfies the same assumptions on $q(\mathbf{w}, \boldsymbol{\theta})$ in Theorem 2. Then

$$N^{-1} \sum_{i=1}^{N} r\left(\mathbf{w}_i, \hat{\boldsymbol{\theta}}\right) \xrightarrow{p} \mathrm{E}\left[r\left(\mathbf{w}, \boldsymbol{\theta}_{\mathrm{o}}\right)\right], \tag{10}$$

That is $N^{-1} \sum_{i=1}^{N} r\left(\mathbf{w}_i, \hat{\boldsymbol{\theta}}\right)$ is a consistent estimator of $\mathrm{E}\left[r\left(\mathbf{w}, \boldsymbol{\theta}_{\mathrm{o}}\right)\right]$.

Agenda

M-Estimation

- The simplest asymptotic normality proof proceeds as follows.
- Assume that $\boldsymbol{\theta}_{\mathrm{o}}$ is in the interior of $\boldsymbol{\Theta}$, which means that $\boldsymbol{\Theta}$ must have nonempty interior (this assumption is true in most applications). Then, since $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_{\mathrm{o}}$, $\hat{\boldsymbol{\theta}}$ is in the interior of $\boldsymbol{\Theta}$ with probability approaching one.
- If $q(\mathbf{w}, \cdot)$ is continuously differentiable on the interior of $\boldsymbol{\Theta}$, then (with probability approaching one) $\hat{\boldsymbol{\theta}}$ solves the first-order condition

$$\sum_{i=1}^{N} \mathbf{s}\left(\mathbf{w}_i, \hat{\boldsymbol{\theta}}\right) = \mathbf{0}, \tag{11}$$

where $\mathbf{s}(\mathbf{w}, \boldsymbol{\theta})$ is the $P \times 1$ vector of partial derivatives of $q(\mathbf{w}, \boldsymbol{\theta}) : \mathbf{s}(\mathbf{w}, \boldsymbol{\theta})' = \nabla_{\theta} q(\mathbf{w}, \boldsymbol{\theta}) \equiv [\partial q(\mathbf{w}, \boldsymbol{\theta})/\partial\theta_1, \partial q(\mathbf{w}, \boldsymbol{\theta})/\partial\theta_2, \ldots, \partial q(\mathbf{w}, \boldsymbol{\theta})/\partial\theta_P]$. (That is, $\mathbf{s}(\mathbf{w}, \boldsymbol{\theta})$ is the transpose of the gradient of $q(\mathbf{w}, \boldsymbol{\theta})$).
- We call $\mathbf{s}(\mathbf{w}, \boldsymbol{\theta})$ **the score of the objective function** $q(\mathbf{w}, \boldsymbol{\theta})$.

M-Estimation

- If $q(\mathbf{w}, \boldsymbol{\theta})$ is twice continuously differentiable, then each row of the left-hand side of equation (11) can be expanded about $\boldsymbol{\theta}_{\mathrm{o}}$ in a mean-value expansion:

$$\sum_{i=1}^{N} \mathbf{s}\left(\mathbf{w}_i, \hat{\theta}\right) = \sum_{i=1}^{N} \mathbf{s}\left(\mathbf{w}_i, \boldsymbol{\theta}_{\mathrm{o}}\right) + \left(\sum_{i=1}^{N} \ddot{\mathbf{H}}_i\right)\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\mathrm{o}}\right), \tag{12}$$

- The notation $\ddot{\mathbf{H}}_i$ denotes the $P \times P$ Hessian of the objective function, $q\left(\mathbf{w}_i, \boldsymbol{\theta}\right)$, with respect to $\boldsymbol{\theta}$, but with each row of $\mathbf{H}\left(\mathbf{w}_i, \boldsymbol{\theta}\right) \equiv \partial^2 q\left(\mathbf{w}_i, \boldsymbol{\theta}\right) / \partial \boldsymbol{\theta} \partial \hat{\boldsymbol{\theta}}' \equiv \nabla^2_{\theta} q\left(\mathbf{w}_i, \boldsymbol{\theta}\right)$ evaluated at a different mean value.

M-Estimation

- Combining equations (11) and (12) and multiplying through by $1/\sqrt{N}$ gives

$$\mathbf{0} = N^{-1/2}\sum_{i=1}^{N}\mathbf{s}\left(\mathbf{w}_i, \boldsymbol{\theta}_{\mathrm{o}}\right) + \left(N^{-1}\sum_{i=1}^{N}\ddot{\mathbf{H}}_i\right)\sqrt{N}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\mathrm{o}}\right), \tag{13}$$

- Using Lemma 1 we get $N^{-1}\sum_{i=1}^{N}\ddot{\mathbf{H}}_i \xrightarrow{p} \mathrm{E}\left[\mathbf{H}\left(\mathbf{w}, \boldsymbol{\theta}_{\mathrm{o}}\right)\right]$.
- If $\mathbf{A}_{\mathrm{o}} \equiv \mathrm{E}\left[\mathbf{H}\left(\mathbf{w}, \boldsymbol{\theta}_{\mathrm{o}}\right)\right]$ is nonsingular, then $N^{-1}\sum_{i=1}^{N}\ddot{\mathbf{H}}_i$ is nonsingular w.p.a. 1 and $\left(N^{-1}\sum_{i=1}^{N}\ddot{\mathbf{H}}_i\right)^{-1} \xrightarrow{p} \mathbf{A}_{\mathrm{o}}^{-1}$.
- Therefore, we can write

$$\sqrt{N}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\mathrm{o}}\right) = \left(N^{-1}\sum_{i=1}^{N}\ddot{\mathbf{H}}_i\right)^{-1}\left[-N^{-1/2}\sum_{i=1}^{N}\mathbf{s}_i\left(\boldsymbol{\theta}_{\mathrm{o}}\right)\right], \tag{14}$$

where $\mathbf{s}_i\left(\boldsymbol{\theta}_{\mathrm{o}}\right) \equiv \mathbf{s}\left(\mathbf{w}_i, \boldsymbol{\theta}_{\mathrm{o}}\right)$.

M-Estimation

- Since $o_p(1) \cdot O_p(1) = o_p(1)$ we have,

$$\sqrt{N}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\text{o}}\right) = \mathbf{A}_{\text{o}}^{-1}\left[-N^{-1/2}\sum_{i=1}^{N}\mathbf{s}_i\left(\boldsymbol{\theta}_{\text{o}}\right)\right] + o_p(1), \tag{15}$$

- This is an important equation. It shows that $\sqrt{N}\left(\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{o}}\right)$ inherits its limiting distribution from the average of the scores, evaluated at $\boldsymbol{\theta}_{\text{o}}$. The matrix $\mathbf{A}_0^{-1}$ simply acts as linear transformation.
- Absorbing this linear transformation into $\mathbf{s}_i\left(\boldsymbol{\theta}_{\text{o}}\right)$, we can write

$$\sqrt{N}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\text{o}}\right) = N^{-1/2}\sum_{i=1}^{N}\mathbf{e}_i\left(\boldsymbol{\theta}_{\text{o}}\right) + o_p(1), \tag{16}$$

where $\mathbf{e}_i\left(\boldsymbol{\theta}_{\text{o}}\right) \equiv -\mathbf{A}_{\text{o}}^{-1}\mathbf{s}_i\left(\boldsymbol{\theta}_{\text{o}}\right)$; this is sometimes called the **influence function representation** of $\boldsymbol{\theta}$, where $\mathbf{e}(\mathbf{w}, \boldsymbol{\theta})$ is the influence function.

M-Estimation

- THEOREM 3 (Asymptotic Normality of M-Estimators): In addition to the assumptions in Theorem 2, assume (a) $\boldsymbol{\theta}_{\mathrm{o}}$ is in the interior of $\boldsymbol{\Theta}$; (b)$\mathbf{s}(\mathbf{w},\cdot)$ is continuously differentiable on the interior of $\boldsymbol{\Theta}$ for all $\mathbf{w} \in \mathscr{W}$; (c) Each element of $\mathbf{H}(\mathbf{w},\boldsymbol{\theta})$ is bounded in absolute value by a function $b(\mathbf{w})$, where $\mathrm{E}[b(\mathbf{w})] < \infty$; (d) $\mathbf{A}_0 \equiv \mathrm{E}\left[\mathbf{H}\left(\mathbf{w},\boldsymbol{\theta}_{\mathbf{o}}\right)\right]$ is positive definite; (e) $\mathrm{E}\left[\mathbf{s}\left(\mathbf{w},\boldsymbol{\theta}_{\mathbf{o}}\right)\right] = \mathbf{0}$; and (f) each element of $\mathbf{s}\left(\mathbf{w},\boldsymbol{\theta}_{\mathbf{o}}\right)$ has finite second moment. Then

$$\sqrt{N}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\mathrm{o}}\right) \xrightarrow{d} \text{Normal}\left(0, \mathbf{A}_{\mathrm{o}}^{-1}\mathbf{B}_0\mathbf{A}_{\mathrm{o}}^{-1}\right), \tag{17}$$

  where $\mathbf{A}_{\mathbf{o}} \equiv \mathrm{E}\left[\mathbf{H}\left(\mathbf{w},\boldsymbol{\theta}_{\mathrm{o}}\right)\right]$ and $\mathbf{B}_{\mathrm{o}} \equiv \mathrm{E}\left[\mathbf{s}\left(\mathbf{w},\boldsymbol{\theta}_{\mathrm{o}}\right)\mathbf{s}\left(\mathbf{w},\boldsymbol{\theta}_{\mathrm{o}}\right)'\right] = \mathrm{Var}\left[\mathbf{s}\left(\mathbf{w},\boldsymbol{\theta}_{\mathrm{o}}\right)\right]$
- Thus,

$$\text{Avar}(\hat{\boldsymbol{\theta}}) = \mathbf{A}_{\mathrm{o}}^{-1}\mathbf{B}_{\mathrm{o}}\mathbf{A}_{\mathrm{o}}^{-1}/N, \tag{18}$$

Two-Step M-Estimators

- Sometimes applications of M-estimators involve a first-stage estimation (an example is OLS with generated regressors).
- Let $\hat{\gamma}$ be a preliminary estimator, usually based on the random sample $\{\mathbf{w}_i : i = 1, 2, \ldots, N\}$.
- A two-step M-estimator θ of θ_0 solves the problem

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{i=1}^{N} q\left(\mathbf{w}_i, \boldsymbol{\theta}; \hat{\gamma}\right), \tag{19}$$

where q is now defined on $\mathscr{W} \times \boldsymbol{\Theta} \times \boldsymbol{\Gamma}$, and $\boldsymbol{\Gamma}$ is a subset of $\mathbb{R}^J$.

Agenda

1. M-Estimation
   - Introduction
   - Identification, Uniform Convergence, and Consistency
   - Asymptotic Normality
2. Two-Step M-Estimation
   - Consistency
   - Asymptotic Normality of Two-Step M-Estimators
   - Estimating the Asymptotic Variance
   - Adjustments for when we cannot ignore the first-stage estimation

Two-Step M-Estimators

- For the general two-step M-estimator, when will $\hat{\boldsymbol{\theta}}$ be consistent for $\boldsymbol{\theta}_{\mathrm{o}}$?
- In practice, the important condition is the identification assumption.
- To state the identification condition, we need to know about the asymptotic behavior of $\hat{\gamma}$.
- A general assumption is that $\hat{\gamma} \xrightarrow{p} \gamma^*$, where γ^* is some element in Γ.
- The identification condition for the two-step M-estimator is

$$\mathrm{E}\left[q\left(\mathbf{w}, \boldsymbol{\theta}_{\mathrm{o}}; \boldsymbol{\gamma}^*\right)\right] < \mathrm{E}\left[q\left(\mathbf{w}, \boldsymbol{\theta}; \boldsymbol{\gamma}^*\right)\right] \text{ all } \boldsymbol{\theta} \in \boldsymbol{\Theta}, \quad \boldsymbol{\theta} \neq \boldsymbol{\theta}_{\mathrm{o}}, \tag{20}$$

Two-Step M-Estimators

- The consistency argument is essentially the same as that underlying Theorem 2. If $q(\mathbf{w}_i, \boldsymbol{\theta}; \gamma)$ satisfies the UWLLN over $\mathbf{\Theta} \times \mathbf{\Gamma}$ then expression (19) can be shown to converge to $\mathrm{E}[q(\mathbf{w}, \boldsymbol{\theta}; \gamma^*)]$ uniformly over $\mathbf{\Theta}$. Along with identification, this result can be shown to imply consistency of $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}_{\mathrm{o}}$.

Agenda

Two-Step M-Estimators

- With the two-step M-estimator, there are two cases worth distinguishing.
- The first occurs when the asymptotic variance of $\sqrt{N}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\mathrm{o}}\right)$ does not depend on the asymptotic variance of $\sqrt{N}\left(\hat{\gamma} - \gamma^{*}\right)$.
- The second occurs when the asymptotic variance of $\sqrt{N}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\mathrm{o}}\right)$ should be adjusted to account for the first-stage estimation of γ^{*}.
- We first derive conditions under which we can ignore the first-stage estimation error.

Two-Step M-Estimators

- first derive conditions under which we can ignore the first-stage estimation error.
- Using arguments similar to those used to derive the asymptotic normality of M-estimators, it can be shown that, under standard regularity conditions,

$$\sqrt{N}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\mathrm{o}}\right) = \mathbf{A}_{\mathrm{o}}^{-1}\left(-N^{-1/2}\sum_{i=1}^{N}\mathbf{s}_i\left(\boldsymbol{\theta}_{\mathrm{o}};\hat{\gamma}\right)\right) + \mathrm{o}_p(1), \tag{21}$$

  where now $\mathbf{A_o} = \mathrm{E}\left[\mathbf{H}\left(\mathbf{w},\boldsymbol{\theta}_{\mathbf{o}};\gamma^*\right)\right]$.
- In obtaining the score and the Hessian, we take derivatives only with respect to $\boldsymbol{\theta}$; γ^* simply appears as an extra argument.

Two-Step M-Estimators

- Now if,

$$N^{-1/2}\sum_{i=1}^{N}\mathbf{s}_i\left(\boldsymbol{\theta}_{\mathrm{o}};\hat{\gamma}\right)=N^{-1/2}\sum_{i=1}^{N}\mathbf{s}_i\left(\boldsymbol{\theta}_{\mathrm{o}};\gamma^{*}\right)+\mathrm{o}_p(1), \tag{22}$$

- Then $\sqrt{N}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_{\mathrm{o}}\right)$ behaves the same asymptotically whether we used $\hat{\gamma}$ or its plim in defining the M-estimator.
- When does equation (22) hold?

Two-Step M-Estimators

- Assuming that $\sqrt{N}\left(\hat{\gamma}-\gamma^{*}\right)=\mathrm{O}_{p}(1)$ (which is standard).
- A mean value expansion similar to (12) gives

$$N^{-1/2}\sum_{i=1}^{N}\mathbf{s}_{i}\left(\boldsymbol{\theta}_{\mathrm{o}};\hat{\gamma}\right)=N^{-1/2}\sum_{i=1}^{N}\mathbf{s}_{i}\left(\boldsymbol{\theta}_{0};\gamma^{*}\right)+\mathbf{F}_{\mathrm{o}}\sqrt{N}\left(\hat{\gamma}-\gamma^{*}\right)+\mathrm{o}_{p}(1), \quad (23)$$

  where $\mathbf{F}_0$ is the $P\times J$ matrix $\mathbf{F}_{\mathrm{o}}\equiv\mathrm{E}\left[\nabla_{\gamma}\mathbf{s}\left(\mathbf{w},\boldsymbol{\theta}_{\mathrm{o}};\boldsymbol{\gamma}^{*}\right)\right]$.
- Therefore if

$$\mathrm{E}\left[\nabla_{\gamma}\mathbf{s}\left(\mathbf{w},\boldsymbol{\theta}_{\mathrm{o}};\boldsymbol{\gamma}^{*}\right)\right]=\mathbf{0}, \quad (24)$$

  then equation (22) holds.
- And the asymptotic variance of the two-step M-estimator is the same as if $\boldsymbol{\gamma}^{*}$ were plugged in.

Two-Step M-Estimators

- There are many problems for which assumption (24) does not hold.
- These problems include some the methods for correcting for endogeneity in Probit and Tobit models.
- In such cases we need to make an adjustment to the asymptotic variance of $\sqrt{N}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\text{o}}\right)$.
- Assume

$$\sqrt{N}\left(\hat{\gamma} - \gamma^{*}\right) = N^{-1/2}\sum_{i=1}^{N}\mathbf{r}_i\left(\gamma^{*}\right) + \text{o}_p(1), \tag{25}$$

where $\mathbf{r}_i\left(\gamma^{*}\right)$ is a $J \times 1$ vector with $\mathrm{E}\left[\mathbf{r}_i\left(\gamma^{*}\right)\right] = \mathbf{0}$.

Two-Step M-Estimators

- Now using equation (23) we can write

$$\sqrt{N}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\text{o}}\right) = \mathbf{A}_{\text{o}}^{-1} N^{-1/2} \sum_{i=1}^{N} \left[-\mathbf{g}_i\left(\boldsymbol{\theta}_{\text{o}}; \gamma^*\right)\right] + o_p(1), \tag{26}$$

  where $\mathbf{g}_i\left(\boldsymbol{\theta}_{\text{o}}; \gamma^*\right) \equiv \mathbf{s}_i\left(\boldsymbol{\theta}_{\text{o}}; \gamma^*\right) + \mathbf{F}_{\text{o}} \mathbf{r}_i\left(\gamma^*\right)$.
- Since $\mathbf{g}_i\left(\boldsymbol{\theta}_{\text{o}}; \gamma^*\right)$ has zero mean, the standardized partial sum in equation (26) can be assumed to satisfy the central limit theorem.
- Define the $P \times P$ matrix

$$\mathbf{D}_{\text{o}} \equiv \mathrm{E}\left[\mathbf{g}_i\left(\boldsymbol{\theta}_{\text{o}}; \gamma^*\right) \mathbf{g}_i\left(\boldsymbol{\theta}_{\text{o}}; \gamma^*\right)'\right] = \text{Var}\left[\mathbf{g}_i\left(\boldsymbol{\theta}_{\text{o}}; \gamma^*\right)\right], \tag{27}$$

- Then

$$\text{Avar}\,\sqrt{N}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\text{o}}\right) = \mathbf{A}_{\text{o}}^{-1} \mathbf{D}_{\text{o}} \mathbf{A}_{\text{o}}^{-1}, \tag{28}$$

Agenda

Two-Step M-Estimators

- We first consider estimating the asymptotic variance of $\hat{\theta}$ in the case where there are no nuisance parameters.
- Under regularity conditions that ensure uniform converge of the Hessian, the estimator

$$N^{-1}\sum_{i=1}^{N}\mathbf{H}\left(\mathbf{w}_i,\hat{\theta}\right)\equiv N^{-1}\sum_{i=1}^{N}\hat{\mathbf{H}}_i, \tag{29}$$

is consistent for $\mathbf{A}_o$, by Lemma 1.

- By Lemma 1, under standard regularity conditions we have

$$N^{-1}\sum_{i=1}^{N}\mathbf{s}\left(\mathbf{w}_i,\hat{\theta}\right)\mathbf{s}\left(\mathbf{w}_i,\hat{\theta}\right)'\equiv N^{-1}\sum_{i=1}^{N}\hat{\mathbf{s}}_i\hat{\mathbf{s}}_i'\xrightarrow{p}\mathbf{B}_o. \tag{30}$$

Two-Step M-Estimators

- Combining equations (29) and (30) we can consistently estimate $\text{Avar}\,\sqrt{N}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\text{o}}\right)$ by

$$\text{Avar}\,\sqrt{N}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\text{o}}\right) = \hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{A}}^{-1}, \tag{31}$$

- The asymptotic standard errors are obtained from the matrix

$$\hat{\mathbf{V}} \equiv \widehat{\text{Avar}(\hat{\boldsymbol{\theta}})} = \hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{A}}^{-1}/N. \tag{32}$$

- Which can be expressed as

$$\left(\sum_{i=1}^{N}\hat{\mathbf{H}}_i\right)^{-1}\left(\sum_{i=1}^{N}\hat{\mathbf{s}}_i\hat{\mathbf{s}}_i'\right)\left(\sum_{i=1}^{N}\hat{\mathbf{H}}_i\right)^{-1}. \tag{33}$$

Agenda

Two-Step M-Estimators

- When assumption (24) is violated, the asymptotic variance estimator of $\hat{\boldsymbol{\theta}}$ must account for the asymptotic variance of $\hat{\gamma}$.
- We need to estimate equation (28).
- We already know how to consistently estimate $\mathbf{A}_{\mathrm{o}}$ using equation (29).
- Estimation of $\mathbf{D}_{\mathrm{o}}$ is also straightforward.
- First we need to estimate $\mathbf{F}_{\mathrm{o}}$,

$$\hat{\mathbf{F}} = N^{-1} \sum_{i=1}^{N} \nabla_{\gamma} \mathbf{s}_i(\hat{\boldsymbol{\theta}}; \hat{\gamma}), \tag{34}$$

Two-Step M-Estimators

- Next, replace $\mathbf{r}_i(\gamma^*)$ with $\hat{\mathbf{r}}_i \equiv \mathbf{r}_i(\hat{\gamma})$.
- Then

$$\hat{\mathbf{D}} \equiv N^{-1} \sum_{i=1}^{N} \hat{\mathbf{g}}_i \hat{\mathbf{g}}_i' \xrightarrow{p} \mathbf{D}_o, \tag{35}$$

where $\hat{\mathbf{g}}_i = \hat{\mathbf{s}}_i + \hat{\mathbf{F}} \hat{\mathbf{r}}_i$.

- The asymptotic variance of the two-step M-estimator is,

$$\left(\sum_{i=1}^{N} \hat{\mathbf{H}}_i \right)^{-1} \left(\sum_{i=1}^{N} \hat{\mathbf{g}}_i \hat{\mathbf{g}}_i' \right) \left(\sum_{i=1}^{N} \hat{\mathbf{H}}_i \right)^{-1}. \tag{36}$$