Lecture 3: Asymptotic Properties of MLE

Econ 205A: Econometric Methods I

Ruoyao Shi

Fall 2018

Contents

1 Asymptotic Properties of An Estimator

Consistency

- In the last lecture, we learned some methods of evaluating: unbiasedness, loss function (MSE) optimality and efficiency. They were all finite-sample criteria.
- In this lecture, we will introduce asymptotic properties of an estimator, properties describing the behaviors of an estimator as the sample size approaches infinity.
- Asymptotic criteria can be applied to a variety of models, for most of which the finite sample criteria will be hard to compute, if not impossible.
- In the following, we will first introduce the asymptotic properties of an estimator that we are usually interested in. Then we will use the MLE as an example to illustrate these properties.

- Asymptotic properties concern a sequence of estimators rather than a single estimator. If we observe i.i.d. $X_1, X_2, \ldots$ from a population $f(x|\theta)$, we can construct a sequence of estimators $\hat{\theta}_n \equiv \theta_n(X_1, \ldots, X_n)$ by performing the same estimation procedure for each sample size n.

- **Definition 1** ***(Consistency)*** *A sequence of estimators $\hat{\theta}_n \equiv \theta_n(X_1, \ldots, X_n)$ is a* ***consistent (sequence of) estimator(s)*** *of the parameter θ if, for any $\epsilon > 0$ and every $\theta \in \Theta$,*

$$\lim_{n\to\infty} P_\theta(\|W_n - \theta\| < \epsilon) = 1. \tag{1}$$

- Consistency means that $\hat{\theta}_n \xrightarrow{p.} \theta$; that is, as the sample size becomes larger and larger, the estimator $\hat{\theta}_n$ becomes arbitrarily close to the true parameter value θ. Recall the notation introduce in Lecture 1, consistency means $\hat{\theta}_n = \theta + o_p(1)$.

- Note that the consistency of an estimator (as the outcome of an estimation procedure) requires the estimate to be consistent *regardless of* the true parameter value. If equation (1) holds for some values $\theta \in \Theta$ but not for others, then the estimator is *not* consistent.

- From now on, we will refer to W_n as "an estimator" instead of "a sequence of estimators".

- **Theorem 1** *If W_n is an estimator of a parameter θ satisfying:*

  1. $\lim_{n\to\infty} var_\theta(\hat{\theta}_n) = 0$*, and*
  2. $\lim_{n\to\infty} bais_\theta(\hat{\theta}_n) = 0$

  for every $\theta \in \Theta$, then $\hat{\theta}_n$ is a consistent estimator of θ.

- A consistent estimator could be biased for any finite sample. But a consistent estimator must be **asymptotically unbiased**.

- **Definition 2** *If $MSE(\hat{\theta}_n) \to 0$ as $n \to \infty$, the $\hat{\theta}_n$ is called* ***squared error consistent****, denoted as $\hat{\theta}_n \xrightarrow{m.s.} \theta$.*

- Recall that

$$MSE(\hat{\theta}_n) \equiv \mathbb{E}_\theta[(\hat{\theta}_n - \theta)'(\hat{\theta}_n - \theta)] = [bias_\theta(\hat{\theta}_n)]' bias_\theta(\hat{\theta}_n) + var_\theta(\hat{\theta}_n),$$

  then by Theorem 1, $\hat{\theta}_n \xrightarrow{m.s.} \theta$ implies $\hat{\theta}_n \xrightarrow{p.} \theta$. The converse is in general not true.

Asymptotic Normality

- **Definition 3** *Let Θ be the set of all the values that θ can take. An estimator $\hat{\theta}_n$ of θ is **asymptotically normal** if there exists a sequence $\{c_n\}$ such that $c_n > 0$ for all n, $c_n \to \infty$ as $n \to \infty$, and*

$$c_n(\hat{\theta}_n - \theta) \xrightarrow{d.} \mathcal{N}(0, \Sigma),$$

  for some positive semi-definite matrix Σ.[1]

- c_n is called the **rate of convergence** of $\hat{\theta}_n$, and Σ is called the **asymptotic variance** of $\hat{\theta}_n$ (or asymptotic variance-covariance matrix if θ is a vector).

- **Example 1** *Let $X_1, \ldots, X_n$ be a random sample from a population with mean μ and variance σ^2. Let $\bar{X} \equiv n^{-1} \sum_{i=1}^n X_i$, then by the CLT*

$$\sqrt{n}(\bar{X} - \mu) \xrightarrow{d.} \mathcal{N}(0, \sigma^2).$$

  *Here, the rate of convergence is $c_n = \sqrt{n}$, often referred to as **root-n-consistency**), or denoted as $\hat{\theta}_n - \theta = O_p(n^{-1/2})$.*

- The rate of convergence of many estimators in parametric models is root-n. If $n^{1/2}/c_n = o_p(1)$, then we say that $\hat{\theta}_n$ converges **faster then root-n rate** (or **super-consistent**); if $c_n/n^{1/2} = o_p(1)$, then we say that $\hat{\theta}_n$ converges **slower than root-n rate**. In Theorem 19 of Lecture 1 (second-order delta method), the estimator $g(Y_n)$ converges at n rate (to a χ^2 distribution), which is faster than root-n rate. Most estimators in nonparametric models converge slower than root-n rate.

- **Theorem 2** *Asymptotic normality implies consistency. (Note that we assume $\mathbb{E}(\hat{\theta}_n) = \theta$.)*

[1] Here we only focus on the case where the center of the asymptotic distribution is the true parameter value θ. There exist cases where it is not but the estimator is still asymptotically normal; that is

$$c_n(\hat{\theta}_n - \mathbb{E}(\hat{\theta}_n)) \xrightarrow{d.} \mathcal{N}(0, \Sigma)$$

where $\mathbb{E}(\hat{\theta}_n) \neq \theta$.

- *Proof.* For simplicity, we assume that θ is a scalar. For any $\epsilon > 0$,

$$\begin{aligned} P_\theta(|\hat{\theta}_n - \theta| < \epsilon) &= P_\theta\left(\left|\frac{c_n(\hat{\theta}_n - \theta)}{\sigma}\right| < \frac{c_n\epsilon}{\sigma}\right) \\ &\to \Phi\left(\frac{c_n\epsilon}{\sigma}\right) - \Phi\left(-\frac{c_n\epsilon}{\sigma}\right) \\ &= 2\Phi\left(\frac{c_n\epsilon}{\sigma}\right) - 1 \\ &\to 1, \end{aligned}$$

  since $c_n \to \infty$. And this implies the result of the theorem.

Asymptotic Efficiency

- **Definition 4** *For an estimator $\hat{\theta}_n$, suppose that $c_n(\hat{\theta}_n - \theta) \xrightarrow{d.} \mathcal{N}(0, \sigma^2)$. Then σ^2 is called the **asymptotic variance** or **variance of the limit distribution** of $\hat{\theta}_n$.*[2]

- In Example 1, σ^2 is the asymptotic variance of $\bar{X}$.

- An estimator $\hat{\theta}_n$ is **asymptotically efficient** for a parameter θ if

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d.} \mathcal{N}\left(0, [I(\theta_0)]^{-1}\right),$$

  where $S(\theta|\mathbf{x}) \equiv \partial l(\theta|\mathbf{x})/\partial\theta$ is the score function, i.e. the partial derivative of the log likelihood function. That is, $\hat{\theta}_n$ is asymptotically efficient for the parameter θ if the asymptotic variance of $\hat{\theta}_n$ achieves the CRLB.

- For estimators with the same rate of convergence, the one with smaller asymptotic variance is more accurate. The asymptotically efficient estimators are (asymptotically) the most accurate estimators among the root-n-consistent estimators.

[2] A related concept is the **limiting variance** or **limit of the variances**. For an estimator $\hat{\theta}_n$, if $\lim_{n\to\infty} c_n var(\hat{\theta}_n) = \tau^2 < \infty$, where $\{c_n\}$ is a sequence of constants, then τ^2 is called the **limiting variance** or **limit of the variances**. Limiting variance and asymptotic variances sometimes are the same, but not always.

- **Definition 5** ***(Asymptotic Relative Efficiency)*** *Given two consistent and asymptotically normal estimators $\hat{\theta}_n$ and $\tilde{\theta}_n$ of θ such that*

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d.} \mathcal{N}(0, \Sigma_1), \text{ and}$$
$$\sqrt{n}(\tilde{\theta}_n - \theta) \xrightarrow{d.} \mathcal{N}(0, \Sigma_2).$$

  If $\Sigma_2 \geq \Sigma_1$, then $\hat{\theta}_n$ is said to be ***asymptotically efficienty relative to*** $\tilde{\theta}_n$.[3]

- The following theorem follows in light of the delta method.

- **Theorem 3** *Suppose $\tau(\cdot)$ is a function defined in a neighborhood of θ and that $\tau'(\cdot)$ exists and is non-zero. Then an estimator $\hat{\tau}_n$ is asymptotcally efficient for the parameter $\tau(\theta)$ if*

$$\sqrt{N}(\hat{\tau}_n - \tau(\theta)) \xrightarrow{d.} \mathcal{N}\left(0, [I(\theta_0)]^{-1}\right).$$

- An immediate corollary is that if $\hat{\theta}_n$ achieves the CRLB of θ, then $\tau(\hat{\theta})$ achieves the CRLB of $\tau(\theta)$, for function $\tau(\cdot)$ satisfying the conditions in the above theorem.

- **Example 2** ***(Bernoulli MLE)*** *Recall Example 9 of Lecture 2, where $X_1, \ldots, X_n$ is a random sample from Bernoulli population with pmf*

$$f(x|p) = p^x(1-p)^{1-x},$$

  where $0 \leq p \leq 1$. In that example we showed that for any *sample size n, the MLE estimator $\bar{X}_n$ of p achieves the CRLB $p(1-p)/n$ (efficient in finite samples). This should also hold asymptotically. Formally, by the CLT, we have*

$$\sqrt{n}(\bar{X}_n - p) \xrightarrow{d.} \mathcal{N}(0, p(1-p)).$$

  Recall the calculation in EXample 9 of Lecture 2, we have $var[S(p|\mathbf{X})] = \frac{n}{p(1-p)}$. This implies that the CRLB in the asymptotic analysis is $n\left(var[S(p|\mathbf{X})]\right)^{-1} = p(1-p)$, which is exactly the asymptotic variance of $\bar{X}_n$. Therefore, we have shown that the MLE estimator is also asymptotically *efficient.*

[3] For matrices, $\Sigma_2 \geq \Sigma_1$ means that the matrix $\Sigma_2 - \Sigma_1$ is positive semi-definite.

- **Example 3** ***(Bernoulli MLE cont')*** *Now in the same model, suppose we are interested in the* odds ratio*, i.e.* $\frac{p}{1-p}$*. Let the function* $\tau(t) \equiv \frac{t}{1-t}$*, then we have* $\tau'(t) = \frac{1}{(1-t)^2}$*. By the delta method, it is easy to verify that we have*

$$\sqrt{n}\left(\frac{\bar{X}_n}{1-\bar{X}_n} - \frac{p}{1-p}\right) \xrightarrow{d.} \mathcal{N}\left(0, \frac{p}{(1-p)^3}\right).$$

In addition, we have

$$n[\tau'(\theta)]^2 \left(var[S(\theta|\mathbf{X})]\right)^{-1} = n\left(\frac{1}{(1-p)^2}\right)^2 \frac{p(1-p)}{n} = \frac{p}{(1-p)^3}.$$

So the estimator $\frac{\bar{X}_n}{1-\bar{X}_n}$ *is also* asymptotically *efficient for the odds ratio.*[4]

2 Asymptotic Properties of MLE

Consistency

- In this section, we study the asymptotic properties of general MLE. Suppose $X_1, \ldots, X_n$ is a random sample with pdf of pmf $f(\mathbf{x}|\theta)$, which is not necessarily normal or Bernoulli, etc. But we know the functional form of f, and it is determined only by a finite dimensional parameter $\theta \in \Theta$. The MLE is obtained by maximizing the (log) likelihood function; that is

$$\hat{\theta}_n \equiv \arg\max_{\theta \in \Theta} n^{-1} \sum_{i=1}^{n} \log f(X_i|\theta).$$

- Under suitable regularity conditions, we can show that the MLE are consistent, asymptotically normal and asymptotically efficient.

- **Theorem 4** ***(Consistency of MLE)*** *Under suitable regularity conditions,*[5] $\hat{\theta}_n \xrightarrow{p.} \theta_0$*.*

[4] Unlike Example 2, the efficiency of the odds ratio estimator only holds *asymptotically* but not in finite samples, since the delta method only works asymptotically.

[5] A set of suffieint regularity conditions are:

- Consistency of the MLE is a special case of the consistency of the more general **M-estimators**. The proof requires some knowledge beyond the scope of this course and thus is skipped.

- The idea, however, can be explained. First we need to show that $n^{-1}\sum_{i=1}^{n}\log f(X_i|\theta)$ converges $\mathbb{E}_\theta[\log f(X|\theta)]$ in probability regardless of the value of θ (i.e. uniform convergence in probability). That is, $n^{-1}\sum_{i=1}^{n}\log f(X_i|\theta)$ should be very close to $\mathbb{E}_\theta[\log f(X|\theta)]$ with probability approaching one in large samples. If θ_0 is the unique maximizer of $\mathbb{E}_\theta[\log f(X|\theta)]$, then by the continuity of $f(x|\theta)$ and the compactness of Θ (these are all taken care of by the regularity conditions), we can show that $\hat{\theta}_n$ as the maximizer of $n^{-1}\sum_{i=1}^{n}\log f(X_i|\theta)$ should also be very close to θ_0 with probability approaching one in large samples.

Asymptotic Normality of MLE

- **Theorem 5** ***(Asymptotic Normality of MLE)*** *Under suitable regularity conditions,*[6] *and suppose $\theta_0 \in \Theta^\circ$. Then*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d.} \mathcal{N}\left(0, [I(\theta_0)]^{-1}\right).$$

(i) The parameter is *identified*; that is, if $\theta \neq \theta'$, then $f(x|\theta) \neq f(x|\theta')$.

(ii) The densisites $f(x|\theta)$ have common support (for different values of θ), and $f(x|\theta)$ is differentiable in θ.

(iii) The parameter set Θ contains an open set $\mathcal{O}$ of which the true parameter value θ_0 is an interior point.

[6] A set of sufficient regularity conditions include those for the consistency and:

(iv) For every $x \in \mathcal{X}$, the density $f(x|\theta)$ is three times differentiable with respect to θ, the third derivative is continuous in θ, and $\int f(x|\theta)dx$ can be differentiated three times under the integral sign.

(v) For any $\theta_0 \in \Theta$, there exists a positive number c and a function $M(x)$ (both of which may depend on θ_0) such that

$$\left|\frac{\partial^3}{\partial\theta^3} f(x|\theta)\right| \leq M(x) \text{ for all } x \in \mathcal{X},\ \theta_0 - c < \theta < \theta_0 + c,$$

with $\mathbb{E}[M(X)] < \infty$.

- This theorem states that the MLE is both asymptotically normal and asymptotically efficient, since its asymptotic variance attains the CRLB.

- Recall the information equality, it is easy to verify that the asymptotic variance can alternatively be written as

$$n\left(var[S(\theta_0|\mathbf{X})]\right)^{-1} = -n\left[\mathbb{E}\left(\frac{\partial^2 l(\theta_0|\mathbf{X})}{\partial\theta\partial\theta'}\right)\right]^{-1} = [I(\theta_0)]^{-1} = -[H(\theta_0)]^{-1},$$

where $H(\theta_0) = \mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X,\theta)\right]$.

- Again, the rigorous proof belongs to a more advanced course. But here it helps to outline the idea heuristically. By the definition of the MLE, we have the FOC

$$n^{-1}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(X_i|\hat{\theta}_n) = 0.$$

Taking the mean value expansion of the FOC around θ_0 gives us

$$\begin{aligned}0 &= n^{-1}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(X_i|\hat{\theta}_n)\\ &= n^{-1}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(X_i|\theta_0) + n^{-1}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}\log f(X_i|\tilde{\theta}_n)(\hat{\theta}_n-\theta_0).\end{aligned}$$

where $\tilde{\theta}_n$ is some value between $\hat{\theta}_n$ and θ_0. Rearranging it, we have

$$\sqrt{n}(\hat{\theta}_n-\theta_0) = \underbrace{\left[-n^{-1}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}\log f(X_i|\tilde{\theta}_n)\right]^{-1}}_{\substack{\xrightarrow{p.} -[H(\theta_0)]^{-1} \equiv -\left\{\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X|\theta_0)\right]\right\}^{-1} \\ \text{by ULLN, CMT and that } \hat{\theta}_n \xrightarrow{p.} \theta_0}} \underbrace{\cdot n^{-1/2}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(X_i|\theta_0).}_{\xrightarrow{d.}\mathcal{N}(0,[I(\theta_0)]^{-1}) \text{ by CLT}}$$

And the result of the theorem is straightforward to verify using the information equality and the Slutsky's theorem.

3 Exercises

1. Let $X_1, \dots, X_n$ be a random sample from a population with mean μ and variance $\sigma^2 = 1$. By CLT we have

$$\sqrt{n}(\bar{X} - \mu) \xrightarrow{d.} \mathcal{N}(0, 1).$$

   Prove that $\bar{X} \xrightarrow{p.} \mu$ without invoking Theorem 2.

2. Now suppose we have

$$c_n(\hat{\theta}_n - \mathbb{E}(\hat{\theta}_n)) \xrightarrow{d.} \mathcal{N}(0, \sigma^2)$$

   where $\mathbb{E}(\hat{\theta}_n) \neq \theta$. Explain why $\hat{\theta}_n$ is not consistent.

3. Let $X_1, \dots, X_n$ be a random sample from a $\mathcal{N}(\mu, \sigma^2)$ population.

   (a) Find the MLE of μ and σ^2, respectively.

   (b) Show that the MLE of μ is unbiased, but that of σ^2 is biased.

   (c) Verify that the Hessian matrix is

$$\begin{bmatrix} \frac{\partial^2 l(\theta|\mathbf{x})}{\partial \mu^2} & \frac{\partial^2 l(\theta|\mathbf{x})}{\partial \mu \partial \sigma^2} \\ \frac{\partial^2 l(\theta|\mathbf{x})}{\partial \sigma^2 \partial \mu} & \frac{\partial^2 l(\theta|\mathbf{x})}{\partial (\sigma^2)^2} \end{bmatrix} = \begin{bmatrix} -\frac{n}{\sigma^2} & -\frac{1}{\sigma^4}\sum_{i=1}^n (x_i - \mu) \\ -\frac{1}{\sigma^4}\sum_{i=1}^n (x_i - \mu) & \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^n (x_i - \mu)^2 \end{bmatrix}.$$

   (d) Verify that

$$-\mathbb{E}\begin{bmatrix} \frac{\partial^2 l(\theta|\mathbf{x})}{\partial \mu^2} & \frac{\partial^2 l(\theta|\mathbf{x})}{\partial \mu \partial \sigma^2} \\ \frac{\partial^2 l(\theta|\mathbf{x})}{\partial \sigma^2 \partial \mu} & \frac{\partial^2 l(\theta|\mathbf{x})}{\partial (\sigma^2)^2} \end{bmatrix} = \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}.$$

   (e) Verify that the MLE of μ achieves the CRLB.

   (f) Show that $var(S_n^2) = \frac{2\sigma^4}{n-1} > \frac{2\sigma^4}{n}$, so the sample variance does not achieve the CRLB (efficient) in finite samples.

   (g) What is the finite sample variance of the MLE of σ^2? Does it violate the Cramér-Rao theorem? Why or why not?

   (h) Argue that the MLE of μ and σ^2 are both consistent.

(i) Given that[7]

$$\sqrt{n}(S_n^2 - \sigma^2) \xrightarrow{d.} \mathcal{N}(0, \mu_4 - \sigma^4),$$

where $\mu_4 \equiv \mathbb{E}[(X - \mu)^4]$ and for normal distributions $\mu_4 = 3\sigma^4$. Argue that the sample variance is *asymptotically* efficient.

(j) Show that the MLE of σ^2 is also *asymptotically* efficient.

[7] Recall that the finite sample distribution of S_n^2 is χ^2. However, by some algebra, CLT and LLN, we can show that the large sample distribution of S_n^2 can be approximated by a normal distribution shown here.