

Lecture 19: Asymptotic normality and efficiency

Consistency is a very basic requirement only:

- An inconsistent estimator should not be used.
- A consistent estimator, however, may not be always good.
- In a given problem, there may be many consistent estimators.

If variances of consistent estimators are finite, then we may compare their variances and MSEs.

For any sequence $\{T_n\}$ of consistent estimators of $g(\theta)$, let $\{k_n\}$ be a sequence of positive constants (typically, $k_n = \sqrt{n}$ or n^a for some $a > 0$) and $\sigma^2(\theta) > 0$ be a function of θ such that

$$\lim_{n \rightarrow \infty} k_n \text{Var}_{\theta}(T_n) = \tau_T^2(\theta) \quad \theta \in \Theta$$

Then $\tau_T^2(\theta)$ (or $\tau_T^2(\theta)/k_n$) is called the **limiting variance** of $k_n T_n$ (or T_n) and can be used as a performance measure for T_n .

However, in many cases variances do not exist (e.g., $\bar{X}^{-1}$) and, thus, we consider the following alternative.

Definition 10.1.9.

Let $\{T_n\}$ be a sequence of estimators of $g(\theta)$ and $\{k_n\}$ be a sequence of positive numbers such that, for every $\theta \in \Theta$, $k_n[T_n - g(\theta)]$ converges in distribution to $N(0, \sigma_T^2)$, where $\sigma_T^2 > 0$ depends on θ as well as the sequence of estimators. Then σ_T^2/k_n (or σ_T^2) is called the **asymptotic variance** of T_n (or $k_n T_n$).

Example.

For iid random variables $X_1, \dots, X_n$ with finite $\mu = E(X_1)$ and $\sigma^2 = \text{Var}(X_1)$, by the CLT, $\sqrt{n}(\bar{X} - \mu)$ converges in distribution to $N(0, \sigma^2)$.

Hence, the asymptotic variance of $\sqrt{n}\bar{X}$ is σ^2 , which is the same as the limiting variance of $\sqrt{n}\bar{X}$.

If g is a differentiable function of μ , then the delta method says that $\sqrt{n}[g(\bar{X}) - g(\mu)]$ converges in distribution to $N(0, [g'(\mu)]^2 \sigma^2)$.

Hence, the asymptotic variance of $\sqrt{n}g(\bar{X})$ is $[g'(\mu)]^2 \sigma^2$, regardless of whether $g(\bar{X})$ has a finite variance or not.

The variance of $g(\bar{X})$ is often not finite.

- The limiting variance and asymptotic variance are not uniquely defined; but the limit $\tau_T^2(\theta)$ or $\sigma_T^2(\theta)$ is uniquely defined.
- The limiting variance and the asymptotic variance, when both exist, are typically the same, but in theory they may be different.
- It is almost always true that the asymptotic variance is no larger than the limiting variance.
- For the assessment of point estimators, the asymptotic variance is more useful than the limiting variance, because the latter often is not defined.
- We may replace the normal distribution in Definition 10.1.9 by any other distribution with mean 0 and variance σ_T^2 ; but using normal distribution is enough for many statistical problems.

Is there a sequence of estimators that is optimal in terms of asymptotic variance?

In the spirit of Cramér-Rao inequality introduced in Chapter 7, we have the following definition.

Definition 10.1.11.

To estimate a parameter $g(\theta)$, a sequence of estimators $\{T_n = T_n(X_1, \dots, X_n)\}$ is asymptotically efficient (or optimal) iff $[T_n - g(\theta)]/\sqrt{v_n(\theta)}$ converges in distribution to $N(0, 1)$ and

$$v_n(\theta) = \left[\frac{\partial}{\partial \theta} g(\theta) \right]' [I_n(\theta)]^{-1} \frac{\partial}{\partial \theta} g(\theta), \quad \theta \in \Theta$$

where $I_n(\theta)$ is the Fisher information contained in $X_1, \dots, X_n$.

That is, the asymptotic variance of T_n achieves the Cramér-Rao lower bound for every $\theta \in \Theta$.

Recall that Theorem 10.1.6 states that, under some conditions, MLE's are consistent.

The next result shows that, under some conditions somewhat stronger than those in Theorem 10.1.6, MLE's are also asymptotically efficient (optimal) according to Definition 10.1.11.

This result shows why the method of MLE is so popular.

Theorem 10.1.12.

Let $X_1, \dots, X_n$ be iid with pdf or pmf $f_\theta(x)$ satisfying the regularity conditions in §10.6.2. Let $g(\theta)$ be a differentiable function of θ and $\hat{\theta}_n$ be the MLE of θ . Then, $\sqrt{n}[g(\hat{\theta}_n) - g(\theta)]$ converges in distribution to $N(0, v(\theta))$ and $v(\theta)$ achieves the Cramér-Rao lower bound, i.e.,

$$v(\theta) = \left[\frac{\partial}{\partial \theta} g(\theta) \right]' [I_1(\theta)]^{-1} \frac{\partial}{\partial \theta} g(\theta)$$

Proof.

We only prove the case where $\theta \in \mathcal{R}$ and $g(\theta) = \theta$.

Let $l_n(\theta) = n^{-1} \sum_{i=1}^n \log f_\theta(X_i)$.

By Taylor's expansion,

$$l'_n(\hat{\theta}_n) = l'_n(\theta) + l''_n(\theta)(\hat{\theta}_n - \theta) + R_n$$

where R_n is the remainder term satisfying that $\sqrt{n}R_n$ converges in probability to 0.

Since $\hat{\theta}_n$ is the MLE, $l'(\hat{\theta}_n) = 0$ and, hence,

$$\sqrt{n}(\hat{\theta}_n - \theta) = \frac{-\sqrt{n}l'_n(\theta)}{l''_n(\theta)} + \frac{\sqrt{n}R_n}{l''_n(\theta)}$$

By the WLLN and Lemma 7.3.11,

$$l_n''(\theta) = \frac{1}{n} \sum_{i=1}^n \frac{\partial^2}{\partial \theta^2} \log f_{\theta}(X_i)$$

converges in probability to

$$E_{\theta} \left[\frac{\partial^2}{\partial \theta^2} \log f_{\theta}(X_i) \right] = -I_1(\theta) = -E_{\theta} \left[\frac{\partial}{\partial \theta} \log f_{\theta}(X_1) \right]^2$$

Thus, the term $\sqrt{n}R_n/l_n''(\theta)$ converges in probability to 0.

By the CLT,

$$\sqrt{n}l_n'(\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\partial}{\partial \theta} \log f_{\theta}(X_i)$$

converges in distribution to

$$N \left(0, E_{\theta} \left[\frac{\partial}{\partial \theta} \log f_{\theta}(X_1) \right]^2 \right) = N(0, I_1(\theta))$$

By Slutsky's theorem, the term $-\sqrt{n}l_n'(\theta)/l_n''(\theta)$ converges in distribution to $N(0, I_1(\theta))$ divided by $I_1(\theta)$, which is $N(0, [I_1(\theta)]^{-1})$.

By Slutsky's theorem again, $\sqrt{n}(\hat{\theta}_n - \theta)$ converges in distribution to $N(0, [I_1(\theta)]^{-1})$.

- The property of asymptotic efficiency gives us a benchmark for what we can hope to attain in asymptotic variance.
- There exist other estimators that are asymptotically efficient but are simpler than the MLE's.
- The conditions in §10.6.2 may not be satisfied in some problems and, hence, the MLE's may not be asymptotically optimal and there may not exist any asymptotically optimal estimators.
- We can still use the asymptotic variance to compare different estimators even if no asymptotically optimal estimator exists.

Definition 10.1.16.

If T_n and U_n are estimators of $g(\theta)$ and satisfy

$$\sqrt{n}[T_n - g(\theta)] \text{ converges in distribution to } N(0, \sigma_T^2)$$

$$\sqrt{n}[U_n - g(\theta)] \text{ converges in distribution to } N(0, \sigma_U^2)$$

then the asymptotic relative efficiency (ARE) of U_n with respect to T_n is

$$\text{ARE}(U_n, T_n) = \sigma_T^2 / \sigma_U^2$$

U_n is asymptotically more efficient than T_n if $\text{ARE}(U_n, T_n) > 1$.

Example.

Let $X_1, \dots, X_n$ be iid random variables with $EX_i = \mu \neq 0$, $\text{Var}(X_i) = 1$, and $EX_i^4 < \infty$.

Consider the following two estimators of μ^2 :

$$T_{1n} = n^{-1} \sum_{i=1}^n X_i^2 - 1 \quad \text{and} \quad T_{2n} = \bar{X}^2 - n^{-1}$$

Since $EX_1^2 = \text{Var}(X_1) + \mu^2 = 1 + \mu^2$, by applying the CLT, we obtain that

$$\sqrt{n}(T_{1n} - \mu^2) = \sqrt{n} \left[\frac{1}{n} \sum_{i=1}^n X_i^2 - (1 + \mu^2) \right]$$

converges in distribution to $N(0, \text{Var}(X_1^2))$.

Also, by the CLT, the delta-method, and Slutsky's theorem,

$$\sqrt{n}(T_{2n} - \mu^2) = \sqrt{n}(\bar{X}^2 - \mu^2) - \frac{1}{\sqrt{n}}$$

converges in distribution to $N(0, 4\mu^2)$.

Therefore, the ARE of T_{1n} with respect to T_{2n} is equal to

$$\text{ARE}(T_{1n}, T_{2n}) = \frac{4\mu^2}{\text{Var}(X_1^2)}$$

If the distribution of $X_1 - \mu$ is symmetric about 0, then $E(X_1 - \mu)^3 = 0$ and, thus,

$$\begin{aligned} \text{Var}(X_1^2) &= EX_1^4 - (EX_1^2)^2 \\ &= E[(X_1 - \mu) + \mu]^4 - (1 + \mu^2)^2 \\ &= E(X_1 - \mu)^4 + 4\mu E(X_1 - \mu)^3 + 6\mu^2 E(X_1 - \mu)^2 \\ &\quad + 4\mu^3 E(X_1 - \mu) + \mu^4 - (1 + 2\mu^2 + \mu^4) \\ &= E(X_1 - \mu)^4 + 4\mu^2 - 1 \\ &> 4\mu^2, \end{aligned}$$

where the inequality follows from the Jensen's inequality

$$E(X_1 - \mu)^4 \geq [E(X_1 - \mu)^2]^2 = 1$$

and the fact that X_1 is not a constant.

Therefore,

$$\text{ARE}(T_{1n}, T_{2n}) < 1$$

This means that T_{2n} is asymptotically more efficient if $E(X_1 - \mu)^3 = 0$.

Suppose now that the common distribution of X_i is the distribution of $Y/\sqrt{p(1-p)}$, where Y is a Bernoulli random variable with $P(Y = 1) = p$.

Then $EX_i = \sqrt{p/(1-p)} = \mu$, $\text{Var}(X_1) = 1$, and $EX_1^4 < \infty$.

Note that

$$\begin{aligned}\text{Var}(X_1^2) &= \text{Var}(Y^2)/[p^2(1-p)^2] \\ &= \text{Var}(Y)/[p^2(1-p)^2] \\ &= [p(1-p)]^{-1}\end{aligned}$$

Then

$$\text{ARE}(T_{1n}, T_{2n}) = \frac{4\mu^2}{\text{Var}(X_1^2)} = 4p^2$$

which is larger than 1 iff $p \in (1/2, 1)$.

Thus, T_{1n} is asymptotically more efficient when $p > 1/2$ and T_{2n} is asymptotically more efficient when $p < 1/2$.

Asymptotic results in linear models

In a linear model without normality assumption, we usually apply asymptotic results.

Theorem (asymptotic normality of the LSE)

Assume the general linear model $Y = X\beta + \mathcal{E}$ and that

$$\lim_{n \rightarrow \infty} \max_{1 \leq i \leq n} x_i'(X'X)^{-1}x_i = 0$$

- (i) If components of $\mathcal{E}$ are iid with mean 0 and finite variance σ^2 , then for any $l \in \mathcal{R}^p$,

$$l'(\hat{\beta} - \beta) / \sqrt{\text{Var}(l'\hat{\beta})} \text{ converges in distribution to } N(0, 1)$$

- (ii) The same result holds if components of $\mathcal{E}$ are independent, $\inf_i E(\varepsilon_i^2) > 0$ and $\sup_i E|\varepsilon_i|^{2+\delta} < \infty$ for some $\delta > 0$.

Proof.

For $l \in \mathcal{R}^p$, defining $c_{ni} = l'(X'X)^{-1}x_i$, $i = 1, \dots, n$, we have

$$l'(\hat{\beta} - \beta) = l'(X'X)^{-1}X'\mathcal{E} = \sum_{i=1}^n c_{ni}\varepsilon_i$$

Since

$$\sum_{i=1}^n c_{ni}^2 = I'(X'X)^{-1}X'X(X'X)^{-1}I = I'(X'X)^{-1}I$$

$$\max_{1 \leq i \leq n} c_{ni}^2 = \max_{1 \leq i \leq n} [I'(X'X)^{-1}x_i]^2 \leq I'(X'X)^{-1}I \max_{1 \leq i \leq n} x_i'(X'X)^{-1}x_i$$

where the last inequality follows from the Cauchy-Schwartz inequality,

$$\lim_{n \rightarrow \infty} \left(\max_{1 \leq i \leq n} c_{ni}^2 / \sum_{i=1}^n c_{ni}^2 \right) \leq \lim_{n \rightarrow \infty} \max_{1 \leq i \leq n} x_i'(X'X)^{-1}x_i = 0$$

which is the given condition.

Then, the result in (i) follows from the CLT for a linear combination of iid random variables (stat 609, lecture 31).

To prove (ii), we apply Liapounov's CLT, and need to show that

$$\lim_{n \rightarrow \infty} \frac{1}{v_n^{1+\delta/2}} \sum_{i=1}^n E|c_{ni}\varepsilon_i|^{2+\delta} = 0$$

where δ is given in the condition and

$$v_n = \text{Var}(I'\hat{\beta}) = \sum_{i=1}^n c_{ni}^2 \text{Var}(\varepsilon_i) \geq \inf_i E(\varepsilon_i^2) \sum_{i=1}^n c_{ni}^2$$

Then

$$\begin{aligned}\frac{1}{v_n^{1+\delta/2}} \sum_{i=1}^n E|c_{ni}\varepsilon_i|^{2+\delta} &= \frac{1}{v_n^{1+\delta/2}} \sum_{i=1}^n (c_{ni}^2)^{1+\delta/2} E|\varepsilon_i|^{2+\delta} \\ &\leq \frac{\sup_i E|\varepsilon_i|^{2+\delta}}{[\inf_i E(\varepsilon_i^2)]^{1+\delta/2}} \sum_{i=1}^n (c_{ni}^2)^{1+\delta/2} \bigg/ \left(\sum_{i=1}^n c_{ni}^2 \right)^{1+\delta/2} \\ &\leq \frac{\sup_i E|\varepsilon_i|^{2+\delta}}{[\inf_i E(\varepsilon_i^2)]^{1+\delta/2}} \left(\max_{1 \leq i \leq n} c_{ni}^2 \right)^{\delta/2} \bigg/ \left(\sum_{i=1}^n c_{ni}^2 \right)^{\delta/2}\end{aligned}$$

which tends to 0 as $n \rightarrow \infty$.

We often need to estimate $L\beta$ by $L\hat{\beta}$, where L is a fixed $s \times p$ matrix of rank $s \leq p$.

By the asymptotic normality of $\hat{\beta} - \beta$ and the transformation theorem,

$$(LV_{\hat{\beta}}L')^{-1/2}(L\hat{\beta} - L\beta) \text{ converges in distribution to } N(0, I_s)$$

where $V_{\hat{\beta}} = \text{Var}(\hat{\beta})$.

This result is useful for large sample inference on $L\beta$.

The weighted LSE

When $V = \text{Var}(\mathcal{E})$ is not $\sigma^2 I_n$ and $\hat{V}$ is an estimator of V , the WLSE is of the form

$$\hat{\beta}_W = (X'WX)^{-1}X'WY, \quad W = \hat{V}^{-1}$$

Asymptotic properties of the weighted LSE depend on the asymptotic behavior of $\hat{V}$.

Theorem

Let $\hat{\beta}_{V^{-1}}$ be the WLSE with $W = V^{-1}$ and $\hat{\beta}_W$ be the WLSE with $W = \hat{V}^{-1}$. If $\hat{V}$ is consistent for V , then the asymptotic distributions of $\hat{\beta}_W$ and $\hat{\beta}_{V^{-1}}$ are the same and, for any $l \in \mathcal{R}^p$, $l \neq 0$,

$$l'(\hat{\beta}_W - \beta)/a_n \text{ converges in distribution to } N(0, 1)$$

where $l \in \mathcal{R}^p$, $l \neq 0$, and $a_n^2 = \text{Var}(l'\hat{\beta}_{V^{-1}}) = l'(X'V^{-1}X)^{-1}l$.

Using the same argument as in the proof for the LSE, we show that $l'(\hat{\beta}_{V^{-1}} - \beta)/a_n$ converges in distribution to $N(0, 1)$.

The proof of the asymptotic equivalence of $\hat{\beta}_W$ and $\hat{\beta}_{V^{-1}}$ is omitted.

- Note that the consistency of $\hat{V}$ means

$$\|\hat{V}V^{-1} - I_p\|_{\max} \rightarrow 0$$

- As long as $\hat{V}$ is consistent, the WLSE $\hat{\beta}_W$ is asymptotically as efficient as $\hat{\beta}_{V^{-1}}$, which is the BLUE if V is known.
- From the theorems, the asymptotic relative efficiency of the LSE $l'\hat{\beta}$ with respect to the WLSE $l'\hat{\beta}_W$ is

$$\frac{l'(X'V^{-1}X)^{-1}l}{l'(X'X)^{-1}X'VX(X'X)^{-1}l}$$

which is always less than 1 and equals 1 if $l'\hat{\beta} = l'\hat{\beta}_{V^{-1}}$.

- Finding a consistent $\hat{V}$ is possible when V has some structure. An example is that $Y' = (Y'_1, \dots, Y'_k)$, where each Y_i is q -dimensional with dependent components and $\text{Var}(Y_i) = \Sigma$, $Y_1, \dots, Y_k$ are independent, q is small and k is large. This type of data are called **panel data**, where each Y_i is a panel, or **longitudinal data**, where components of Y_i are repeated measurements.