University of Illinois | Department of Economics
Spring 2017 | Econ 574 | Roger Koenker

Lecture 9
"Consistency and Asymptotic Efficiency of the MLE"

Ref: Wald (1949), Lehmann §6.2.

We will begin with a very simple special case which illustrates the main line of argument. Let $Z_1, \dots, Z_n$ be iid from $\{P_\theta, f(z|\theta)\}$, where $P_\theta(A) = \int_A f(z|\theta)dz$. Assume

A1. The elements of P_θ are distinct,
A2. The elements of P_θ have common support
A3. The parameter space $\Theta \in \Re$ contains an open interval Θ_0 containing θ_0 the true parameter.

Lemma: Under A1-3 for any fixed $\theta \neq \theta_0$,

$$P_{\theta_0}\{\prod_{i=1}^n f(Z_i|\theta_0) > \prod_{i=1}^n f(Z_i|\theta)\} \to 1 \qquad \text{as } n \to \infty$$

Proof: The event in { } is $\Leftrightarrow$ to

$$\frac{1}{n}\sum \log f(Z_i|\theta)/f(Z_i|\theta_0) < 0$$

By the WLLN the lhs converges to $E_{\theta_0}\log(f/f_0)$. Since $-\log(x)$ is *strictly convex*

$$\begin{aligned} E_{\theta_0}\log(f/f_0) &< \log(E_{\theta_0}(f/f_0)) \\ &= \log(\int f dy) \\ &= 0 \qquad \Box \end{aligned}$$

This is the essence of Wald's argument. If the parameter space Θ is finite, then the Lemma implies directly that $\hat{\theta}$ is consistent, since it shows that, eventually, the likelihood is larger at θ_0 than at any other θ.

Theorem: Under A1-3, if Θ is finite, then the mle $\hat{\theta}_n$ exists, is unique with probability tending to 1 and is consistent, i.e., $\hat{\theta} \to \theta_0$.

Proof: Let $\Theta = \{\theta_0, \theta_1, \dots, \theta_k\}$ and E_{in} be the event $\sum \log(f_i/f_0) < 0$. Then since

$$P\{E_{in}\} \to 1 \quad i = 1, \dots, k \quad \Rightarrow P(E_{1n} \cap \dots \cap E_{kn}\} \to 1$$

the result follows. This implication is immediate from Bonferroni's inequality

$$P(\cap E_{in}) \geq 1 - \sum P(E_{in}^c)$$

Bonferroni's Digression Recall that (De Morgan's law) $\cap A_i = (\cup A_i^c)^c$ so that

$$P(\cap A_i) = P(\cup A_i^c)^c = 1 - P(\cup A_i^c) \geq 1 - \sum P(A_i^c).$$

This is usually used to adjust critical values for confidence interval computations: if you have g contrasts and want to do simultaneous confidence intervals then you can use c_α^* where $\alpha^* = \alpha/(2g)$, and α is the desired overall confidence level. There are several variants and strengthenings of the basic Bonferroni inequality. For example one can show that,

$$\sum P(A_i) \leq P(\cup A_i) + \sum\sum P(A_i \cap A_j)$$

Without further conditions on f one can't go further, even the uncountable Θ is fraught with danger. Possible escapes

(Wald)	(i)	*ad hoc* assumptions about $\lim f(z\|\theta)$
(Cramer)	(ii)	differentiability assumptions of f.

we will try to illustrate the latter approach.

Theorem: Under A1-3, if $f(z|\theta)$ is differentiable wrt to θ in Θ_0 with derivative $f'(z|\theta)$, then $wp \to 1$

$$\sum_{i=1}^{n} \frac{f'(z|\theta)}{f(z_i|\theta)} = 0$$

has a root $\hat{\theta}_n$ such that $\hat{\theta}_n \to \theta_0$.

Proof: Choose a such that $(\theta_0 \pm a) \subset \Theta_0$ and set

$$S_n = \{z | l_n(\theta_0) > l_n(\theta_0 - a) \text{ and } l_n(\theta_0) > l_n(\theta_0 + a)\}$$

where $l_n(\theta) = \sum_{i=1}^n \log f(z_i|\theta)$. By the previous Theorem, $P_{\theta_0}\{S_n\} \to 1$.

For any $z \in S_n$, there exists $\hat{\theta}_n$ such that $\hat{\theta}_n \in (\theta_0 \pm a)$ at which $l(\theta, z)$ has a local max, and therefore $l'(\theta) = 0$ Hence, for any $a > 0$, but sufficiently small, *there exists* a sequence $\{\hat{\theta}_n\} = \{\hat{\theta}_n(a)\}$ of roots such that

$$* \quad P_{\theta_0}\{|\hat{\theta}_n - \theta_0| < a\} \to 1$$

It remains to show that sequence doesn't depend on a. For this, let θ_n^* be root closest to θ_0 (which exists because the limit of sequence of roots is a root by continuity of l). Now θ_n^* satisfies $(*)$ but is independent of a. $\square$

Remarks: In some problems the likelihood isn't concave so we can't guarantee a unique maximum, and in this case it is sometimes difficult to choose the right root. Often we will see that it is possible to find a root near an initial consistent estimator – this helps. Cauchy likelihood is an interesting example.

Asymptotic Normality of the MLE

Theorem: Let $Z_1, \dots, Z_n$ be iid from $\{P_\theta, f(z|\theta)\}$ and assume:

(i) Θ is an open interval not necessarily finite.

(ii) P_θ are distinct and have common support

(iii) $f(z|\theta)$ is thrice differentiable wrt to θ and f'' is continuous wrt θ.

(iv) $\int f(z|\theta)dz$ is 3 times differentiable under $\int$.

(v) $I(\theta) = V(\partial \log f/\partial\theta)$ satisfies $0 < I(\theta) < \infty$.

(vi) For any $\theta_1 \in \Theta$, there exists $c > 0$ and $M(z)$ such that

$$\partial^3 \log f(z|\theta)/\partial\theta^3 \leq M(z) \text{ for all } z \in Z \quad \text{and} \quad \theta \in (\theta_0 \pm c)$$

and $E_{\theta_0} M(Z) < \infty$,

Then for any consistent sequence, $\hat{\theta}_n \to \theta_0$, of roots to the likelihood satisfies,

$$(*) \quad \sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow \mathcal{N}(0, I(\theta_0)^{-1}),$$

Proof: Let $l(\theta) = \sum \log f(z_i|\theta)$ as above and expand $l'(\hat{\theta}_n)$ about θ_0 for any fixed z,

$$l'(\hat{\theta}_n) = l'(\theta_0) + (\hat{\theta}_n - \theta_0)l''(\theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)^2 l'''(\theta_n^*)$$

where $|\theta_n^* - \theta_0| < |\hat{\theta}_n - \theta_0|$. By hypothesis $l'(\hat{\theta}_n) = 0$ so that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{-n^{-1/2} l'(\theta_0)}{n^{-1}l''(\theta_0) + \frac{1}{2}n^{-1}(\hat{\theta}_n - \theta_0)l'''(\theta_n^*)}$$

Consider these terms in turn:

(1) $n^{-1/2}l'(\theta_0) \rightsquigarrow \mathcal{N}(0, I(\theta_0))$

$$\begin{aligned} n^{-1/2}l'(\theta_0) &= \sqrt{n}[n^{-1}\sum_{i=1}^{n} l_i' - El_i'] \\ &\equiv \sqrt{n}[n^{-1}\sum(X_i - \mu)] \end{aligned}$$

where $X_i = \partial \log f(Z_i|\theta_0)/\partial\theta,\ E(X_i) = 0,$ and $V(X_i) = I(\theta_0)$ so $n^{-1/2}l'(\theta_0) \rightsquigarrow \mathcal{N}(0, I(\theta))$ by the simplest form of the CLT for iid r.v.'s.

(2) Now consider the first term in the denominator. Set $X_i = \frac{\partial^2}{\partial\theta^2}\log f(Z_i|\theta_0)$ and recall that $EX_i = -I(\theta_0)$, so

$$\begin{aligned} n^{-1}l''(\theta_0) &= n^{-1}\sum_{i=1}^{n}\left(\frac{f_i'}{f_i}\right)^2 - \frac{f_i''}{f_i} \\ &= n^{-1}\sum X_i \\ &\to -I(\theta_0) \end{aligned}$$

(3) Finally, consider, $n^{-1}(\hat{\theta}_n - \theta_0)^2 l'''(\theta_n^*) \to 0$

$$\begin{aligned}
|n^{-1}l'''(\theta)| &= \left| \frac{1}{n} \sum \frac{\partial^3}{\partial \theta^3} \log f(Z_i|\theta_0) \right| \\
&\le \frac{1}{n}[M(Z_1) + \ldots + M(Z_n)] \\
&\to E_{\theta_0} M(Z_1) \qquad (\text{ by } (vi))
\end{aligned}$$

But, by Slutsky, since $\hat{\theta}_n \to \theta_0$ the whole term tends to zero.

Then, putting the pieces back together using Slutsky again we have the result.

Example 1: $1pxf$'s

Finding the mle for the natural parameter η in a $1pxf$ involves solving

$$(*) \qquad \sum T(z_i) + nd_0'(\eta) = 0$$

Checking the second order conditions we have

$$\frac{\partial^2 l}{\partial \eta^2} = nd_0''(\eta)$$

but recall that $V(T(z)) = -d_0''(\eta)$ so $d_0''(\eta) < 0$ so the $1pxf$ likelihood must be globally concave. Note that the 3rd derivative conditions are trivially satisfied since all higher derivatives are independent of z_i. Thus, $wp \to 1$, (*) has a unique root $\hat{\eta}$ which is consistent and asymptotically efficient

$$\sqrt{n}(\hat{\eta} - \eta) \rightsquigarrow \mathcal{N}(0, I(\eta)^{-1}).$$

Example 2: *Location Model*

Suppose $Z_1, \ldots, Z_n$ come from $f(z - \theta)$ where f is differentiable and $f(z) > 0$ for all z. Then the likelihood equation is

$$\sum_{i=1}^{n} \frac{f'(z_i - \theta)}{f(z_i - \theta)} = 0$$

If f is strongly unimodal, i.e., f'/f strictly decreasing, i.e., $\log f$ is strictly concave, then the objective function, i.e., log likelihood is strictly concave and therefore has a unique root. The Laplace distribution, or double exponential, is a borderline case. since $f'/f = 1/2sgn(\cdot)$ which is "just barely monotone."

Example 3: $Z_i \ iid \ U[0, \theta]$

Here none of the theorems apply. What about the mle? Recall that the MLE is $\hat{\theta}_n = Z_{(n)}$. Suppose for convenience $\theta_0 = 1$.

$$P(Z_{(n)} < z) = \begin{cases} z^n & \text{for } z \in [0, 1] \\ 0 & z < 0 \\ 1 & z > 1 \end{cases}$$

Now consider transformed $Z_{(n)}$, with $Y_n = (1 - Z_{(n)})/b_n$ so $Z_{(n)} = 1 - b_n Y_n$ so

$$\begin{aligned} P(Y_n < y) &= P((1 - Z_{(n)})/b_n < y) \\ &= P(1 - b_n y < Z_{(n)}) \\ &= \begin{cases} 1 - (1 - b_n y)^n & y \in (0, 1/b_n) \\ 0 & y < 0 \\ 1 & y > 1/b_n \end{cases} \end{aligned}$$

Now choose b_n to stabilize $P(Y_n < y)$. Note if $b_n = b_0$, a constant $(1 - b_0 y)^n \to 0$ If $b_n = n^{-2}$, then $(1 - y/n^2)^n \to 1$ However, $b_n = n^{-1}$ we have

$$(1 - y/n)^n \to e^{-y}$$

so, as baby bear says, this rate is "just right" and the normalized version of the MLE converges to the standard exponential distribution,

$$P(n(1 - Z_{(n)}) < z) \to e^{-z}$$

or

$$n(1 - Z_{(n)}) \rightsquigarrow E(0, 1)$$

More generally we have if $Z \in U[0, \theta_0]$. then

$$n(\theta_0 - Z_{(n)}) \rightsquigarrow E(0, \theta_0)$$

It is interesting to compare the MLE $\hat{\theta}_n$ with the estimator based on the sample mean. If $Z \sim U[0, \theta]$, then $EZ = \theta_0/2$ and $VZ = \theta^2/12$ so $2\bar{Z} \to \theta_0$ and therefore

$$\sqrt{n}(2\bar{Z} - \theta_0) \to \mathcal{N}(0, \theta_0^2/3).$$

Thus $\tilde{\theta}_n = 2\bar{Z}$ is a consistent estimator of the parameter θ, but it converges only at rate $1/\sqrt{n}$, while the MLE converges at the rate $1/n$, so the mean-based estimator has zero asymptotic efficiency in this case.

We now turn to the problem posed by multiple roots of the likelihood. The first result gives a simple "solution" to this problem if we have a consistent estimator available.

Theorem: (One-Steps) Given the assumptions of the previous theorem, suppose that $\tilde{\theta}_n$ is any $\sqrt{n}$ consistent estimator of θ_0, i.e., for any ε there exists M such that

$$P(\sqrt{n}|\tilde{\theta}_n - \theta_0| \geq M] < \varepsilon.$$

Then $\hat{\theta}_n = \tilde{\theta}_n - l'(\tilde{\theta}_n)/l''(\tilde{\theta}_n)$ is asymptotically efficient, i.e., $\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow \mathcal{N}(0, I(\theta_0)^{-1})$.

Proof: (Heuristics) The name comes from the fact that $\hat{\theta}_n$ is one Gauss-Newton step toward the mle from $\hat{\theta}_n$. Suppose l were quadratic, then

$$l(\theta) = l(\tilde{\theta}) + (\theta - \tilde{\theta})l'(\tilde{\theta}) + \frac{1}{2}(\theta - \tilde{\theta})^2 l''(\tilde{\theta})$$

would hold *exactly.* Then if we wanted to maximize $l(\theta)$, we'd let $l'(\theta) = 0$ or

$$\begin{aligned} l'(\tilde{\theta}) &= -(\theta - \tilde{\theta})l''(\tilde{\theta}) \\ \text{or} \qquad \hat{\theta} &= \tilde{\theta} - l'(\tilde{\theta})/l''(\tilde{\theta}) \end{aligned}$$

so, in effect, we are behaving as if the quadratic approximator is valid near $\hat{\theta}_n$. More formally, expand as in the main theorem and substitute in definition of $\hat{\theta}_n$,

$$\begin{aligned} l'(\tilde{\theta}_n) &= l'(\theta_0) + (\tilde{\theta}_n - \theta_0)l''(\theta_0) + \frac{1}{2}(\theta_n - \theta_0)^2 l'''(\theta_n^*) \\ \sqrt{n}(\hat{\theta}_n - \theta_0) &= -\frac{n^{-1/2}l'(\theta_0)}{n^{-1}l''(\tilde{\theta}_n)} + \sqrt{n}(\tilde{\theta}_n - \theta_0)\left[1 - \frac{l''(\theta_0)}{l''(\tilde{\theta}_n)} - \frac{1}{2}(\tilde{\theta}_n - \theta_0)\frac{l'''(\theta_n^*)}{l''(\tilde{\theta}_n)}\right] \end{aligned}$$

first term as above second term has leading term $O_p(1)$ and

$$\frac{l''(\theta_0)}{l''(\tilde{\theta}_n)} \to 1$$

and last term $\to 0$ as in the proof of the main result.

Example: *Super-Efficiency* (Hodges (1953))

Suppose $Z_1, \ldots, Z_n$ are iid $\mathcal{N}(\theta, 1)$, so $I(\theta) = 1$ and let $\hat{\theta}_n = \begin{cases} \bar{Z} & \text{if } |\bar{Z}| \geq n^{-1/4} \\ a\bar{Z} & \text{otherwise} \end{cases}$ This is a "pre-test shrinker" if $a < 1$.) Now $\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow \mathcal{N}(0, v(\theta))$ where $v(\theta) = 1$ for $\theta \pm 0$ and $v(\theta) = a^2$ when $\theta = 0$. So for $a < 1$, the CRLB is violated! By Chebyshev if $\theta = 0$, then for large n, $\hat{\theta}_n = a\bar{Z}$, if not, not.

Proof: If $\theta_0 = 0$, $P(|\bar{Z}| > \varepsilon) \leq \frac{V(\bar{Z})}{\varepsilon^2}$ so $P(|\bar{Z}|) > n^{-1/4}) \leq \frac{1/n}{1/\sqrt{n}} = \frac{1}{\sqrt{n}} \to 0$ so wp1 $\hat{\theta}_n = a\bar{Z}$.

Remark: Le Cam, Bahadur and others have shown that this has to happen on a set of Lebesgue measure zero.

Multiparameter Extensions

Theorem: Let $Z_1, \ldots, Z_n$ be iid from $\{f(z|\theta_0), P_\theta\}$.

Assume

(i) P_θ are distinct

(ii) P_θ have common support

(iii) $\exists\ \Theta_0 \subset \Theta$ s.t. $\theta_0 \in \Theta_0$ and all 3rd partials exist for $\theta \in \Theta_0$

(iv) $E_{\theta_0} \nabla \log f = 0$ and $I_{jk}(\theta_0) = E_{\theta_0}\left(\frac{\partial \log f}{\partial \theta_j}\frac{\partial \log f}{\partial \theta_k}\right) = -E_{\theta_0}\left(\frac{\partial^2}{\partial \theta_j \partial \theta_k} \log f\right)$

(v) $I(\theta)$ is positive definite for all $\theta \in \Theta_0$ and the vector of "scores" $s = \nabla \log f$ is linearly independent

(vi) There exist functions $M_{ijk}(z)$ such that, $\left|\frac{\partial}{\partial\theta_i\partial\theta_j\partial\theta_k}\log f\right| \leq M_{ijk}(z) \quad \forall\theta \in \Theta_0$. and $E_{\theta_0}M_{ijk}(z) < \infty$.

Then with probability tending to 1, there exists $\hat{\theta}_n$ solving the likelihood equations, $\nabla_\theta l(\theta|z) = 0$ such that $\| \hat{\theta}_n - \theta_0 \| \to 0$ and $\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow \mathcal{N}_p(0, I^{-1}(\theta_0))$

Proof: See Lehmann

Corollary: (One-Steps) If the previous conditions hold and $\tilde{\theta}_n$ is $\sqrt{n}$ consistent for θ_0, then

$$\hat{\theta}_n = \tilde{\theta}_n - [\nabla^2 l(\tilde{\theta}_n)]^{-1}\nabla l(\tilde{\theta}_n)$$

is asymptotically efficient.
Remark: The alternative, $\hat{\theta}_n = \tilde{\theta}_n + [I(\tilde{\theta}_n)]^{-1}\nabla l(\tilde{\theta}_n)$ will also work. This is the method-of-scoring version of the one-step.