Lecture 18. Maximum Likelihood Estimation, Large Sample Properties

November 28, 2011

At the end of the previous lecture, we show that the maximum likelihood (ML) estimator is UMVU if and only if the score function can be written into certain form.

Example 1. Consider an i.i.d. sample $\{X_1, ..., X_n\}$ from a Poisson distribution with parameter λ. The pmf of X_i is $f_{X_i}(x;\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$ $x = 0, 1, 2, ...$ The pmf (or likelihood) of the sample is

$$f_{\mathbb{X}}(X_1, ..., X_n; \lambda) = \times_{i=1}^n f_{X_i}(X_i; \lambda). \tag{1}$$

Thus, the log-likelihood function is

$$\begin{aligned} L_n(\mathbb{X}; \lambda) &= \sum_{i=1}^n \ln f_{X_i}(X_i; \lambda) \\ &= \sum_{i=1}^n [\ln(e^{-\lambda}) + X_i \ln \lambda - \ln(X_i!)] \\ &= -n\lambda + n(\ln \lambda)\bar{X}_n - \sum_{i=1}^n \sum_{j=1}^{X_i} \ln j. \end{aligned} \tag{2}$$

The score function of the log-likelihood is

$$\begin{aligned} S_n(\mathbb{X}; \lambda) &= -n + n\bar{X}_n/\lambda \\ &= n\lambda^{-1}(\bar{X}_n - \lambda). \end{aligned} \tag{3}$$

Thus, the UMVU estimator (and also the ML estimator) for λ is $\bar{X}_n$.

In general, the score function cannot be written as that particular form. In that case, UMVU estimators do not exist, and the ML estimator is certainly not UMVU. However, in large samples, the ML estimator has similar properties as an UMVU estimator.

Example 2. Consider an i.i.d. sample $\{X_1, ..., X_n\}$ from an exponential distribution with parameter λ: $f_{X_i}(x, \lambda) = \lambda e^{-\lambda x}$, $(\lambda > 0)$. The likelihood of the sample is

$$f_{\mathbb{X}}(X_1, ..., X_n; \lambda) = \times_{i=1}^{n} f_{X_i}(X_i, \lambda) = \times_{i=1}^{n} (\lambda e^{-\lambda X_i}). \tag{4}$$

Thus, the log-likelihood function is

$$\begin{aligned} L_n(\mathbb{X}; \lambda) &= \sum_{i=1}^{n} [\ln \lambda - \lambda X_i] \\ &= n \ln \lambda - n\lambda \bar{X}_n. \end{aligned} \tag{5}$$

The score function of the log-likelihood function is

$$S_n(\mathbb{X}; \lambda) = n/\lambda - n\bar{X}_n. \tag{6}$$

The score function cannot be written as the shape $Q(\lambda)(T(\mathbb{X}) - \lambda)$. Thus, UMVU estimator for λ does not exists.

Nonetheless, we can find the ML estimator by setting the score function to zero:

$$\hat{\lambda}_{mle} = \bar{X}_n^{-1}. \tag{7}$$

We know that (by the Slutsky Theorem), $\hat{\lambda}_{mle}$ is consistent because $\bar{X}_n \to_p E_\lambda X_i = 1/\lambda$ and $g(y) = 1/y$ is continuous at $y = 1/\lambda \neq 0$. We also know that $\hat{\lambda}_{mle}$ is $\sqrt{n}$-asymptotically normal (i.e. $\sqrt{n}(\hat{\lambda}_{mle} - \lambda)$ converges in distribution to a normal random variable) by the delta method:

$$\begin{aligned} \sqrt{n}(\hat{\lambda}_{mle} - \lambda) &= \sqrt{n}(g(\bar{X}_n) - g(E_\lambda X_i)) \\ &\to_d g'(E_\lambda X_i) Z, \end{aligned}$$

where $Z \sim N(0, 1/\lambda^2)$ is the asymptotic distribution of $\sqrt{n}(\bar{X}_n - E_\lambda X_i)$. Thus,

$$\sqrt{n}(\hat{\lambda}_{mle} - \lambda) \to_d -\lambda^2 Z \sim N(0, \lambda^2).$$

Therefore approximately,

$$\hat{\lambda}_{mle} \sim N(\lambda, \lambda^2/n). \tag{8}$$

The Cramer-Rao lower bound in this example is

$$\frac{1}{Var_\lambda(n/\lambda - n\bar{X}_n)} = \frac{1}{n^2 Var_\lambda(\bar{X}_n)} = \frac{1}{n^2 Var_\lambda(X_i)/n} = \frac{1}{n/\lambda^2} = \lambda^2/n. \tag{9}$$

Approximately, in large samples, the ML estimator is unbiased and achieves the Cramer-Rao lower bound.

The above properties (approximately UMVU) turns out to hold generally for ML estimators. For a general distribution family $f_X(X, \theta)$. Suppose that we observe a random sample $\{X_1, ..., X_n\}$ from $f_X(X, \theta_0)$ for some true value θ_0. (From now on, we designate "θ_0" for the true value, i.e., $f_X(X, \theta_0)$ is the true distribution from which the random sample is generated. The symbol θ without subscript denotes a generic value of the parameter.) Under regularity conditions in Theorem 6.2.2. in HMC, it can be shown that the ML estimator for θ_0 is $\sqrt{n}-$consistent[1] and asymptotically normal with asymptotic variance $I(\theta_0)^{-1}$, where $I(\theta_0)$ is the information matrix defined in the previous lecture:

$$I(\theta) := E_\theta \left[\frac{\partial \ln f_X(X, \theta)}{\partial \theta} \frac{\partial \ln f_X(X, \theta)}{\partial \theta'} \right] = -E_\theta \left(\frac{\partial^2 \ln f_X(X, \theta)}{\partial \theta \partial \theta'} \right).$$

(Recall that the second equality is due to the information identity.)

Remark about ML estimator. (1) Because the ML estimator is $\sqrt{n}$-consistent and asymptotically normal with asymptotic variance equal to the inverse of the information matrix, we call the ML estimator "asymptotically efficient". A $\sqrt{n}$-consistent and asymptotically normal estimator is called "asymptotic efficient" iff its asymptotic variance equals the inverse of the information matrix.

(2) ML estimators are attractive partly because they are asymptotically efficient estimators under fairly general conditions (ML-regularity conditions).

(3) Another reason that ML estimator is attractive is that it is invariant to nonlinear transformations. Reparameterizing the density will not affect the ML estimator for a parameter of interest. For any function g of θ, we have $\widehat{g(\theta)}_{mle} = g(\hat{\theta}_{mle})$. (Question: in Example 2 above, what is the ML estimator of $1/\lambda$? Is it a UMVU estimator?)

(4) ML estimators in general are not unbiased, and in certain cases can be quite biased, especially in small samples.

(5) If the ML-regularity conditions are not satisfied, ML estimators might not be good estimators.

[1] An estimator $\hat{\theta}_n$ of θ_0 is "$\sqrt{n}$-consistent" if $\sqrt{n}(\hat{\theta}_n - \theta_0) = O_p(1)$.

(6) There are certain models in which ML estimators are not well defined. ($L_n(\mathbb{X};\theta)$ might have no maximizer.)

Now I give some heuristics for why $\hat{\theta}_{mle}$ is consistent. By definition, $\hat{\theta}_{mle}$ maximizes $n^{-1}L_n(X_1,...,X_n;\theta)$. On the other hand,$\theta=\theta_0$ maximizes $E_{\theta_0}(\ln f_X(X,\theta))$. This is because for any θ,

$$\begin{aligned} E_{\theta_0}\left(\ln\frac{f_X(X,\theta)}{f_X(X,\theta_0)}\right) &\leq \ln\left[E_{\theta_0}\left(\frac{f_X(X,\theta)}{f_X(X,\theta_0)}\right)\right] \quad \text{by Jensen's} \\ &= \ln\left[\int_{\mathcal{X}}\frac{f_X(x,\theta)}{f_X(x,\theta_0)}f_X(x,\theta_0)dx\right] \quad \mathcal{X} \text{ is support of } X \\ &= \ln\left[\int_{\mathcal{X}} f_X(x,\theta)dx\right] \\ &= \ln(1)=0. \end{aligned} \tag{10}$$

That is:

$$\max_{\theta} E_{\theta_0}(\ln f_X(X,\theta)) \leq E_{\theta_0}(\ln f_X(X,\theta_0)).$$

Thus, $\theta=\theta_0$ maximizes $E_{\theta_0}(\ln f_X(X,\theta))$. The law of large number implies that $n^{-1}L_n(X_1,...,X_n;\theta)\equiv n^{-1}\sum_{i=1}^n \ln f_X(X_i,\theta) \to_p E_{\theta_0}\ln f_X(X,\theta)$ for all θ. Thus, the maximizer of $n^{-1}L_n(X_1,...,X_n;\theta)$ should converge to the maximizer of $E_{\theta_0}\ln f_X(X,\theta)$.

The above heuristics is not exactly correct, though it does outline the consistency proof. For example, the pointwise convergence $n^{-1}L_n(X_1,...,X_n;\theta) \to_p E_{\theta_0}\ln f_X(X,\theta)$ (i.e. convergence for every θ) is not sufficient to guarantee that the maximizer of the left-hand-side converges to the right-hand-side. Uniform convergence is needed. More details of this will be covered in Econ 715.

The outline of the asymptotic normality proof is also given now. Suppose that $L_n(X_1,...,X_n;\theta)$ is twice-continuous differentiable, then $\hat{\theta}_{mle}$ satisfies the first order condition:

$$n^{-1}\sum_{i=1}^n \frac{\partial \ln f_X(X_i,\hat{\theta}_{mle})}{\partial\theta}=0. \tag{11}$$

Take a taylor expansion of $n^{-1}\sum_{i=1}^n \frac{\partial \ln f_X(X_i,\hat{\theta}_{mle})}{\partial\theta}$ around θ_0, we have

$$0=n^{-1}\sum_{i=1}^n \frac{\partial \ln f_X(X_i,\theta_0)}{\partial\theta}+n^{-1}\sum_{i=1}^n \frac{\partial^2 \ln f_X(X_i,\theta_0)}{\partial\theta\partial\theta'}(\hat{\theta}_{mle}-\theta_0)+o_p(||\hat{\theta}_{mle}-\theta_0||),$$

where $\bar{\theta}_n$ lies between $\hat{\theta}_{mle}$ and θ_0. Suppose that $n^{-1}\sum_{i=1}^n \frac{\partial^2 \ln f_X(X_i,\bar{\theta}_n)}{\partial\theta\partial\theta'}$ is invertible, then

$$\begin{aligned}\sqrt{n}(\hat{\theta}_{mle}-\theta_0) &= \left(n^{-1}\sum_{i=1}^n \frac{\partial^2 \ln f_X(X_i,\theta_0)}{\partial\theta\partial\theta'}\right)^{-1} n^{-1/2}\sum_{i=1}^n \frac{\partial \ln f_X(X_i,\theta_0)}{\partial\theta} \\ &\quad +o_p(\sqrt{n}||\hat{\theta}_{mle}-\theta_0||). \end{aligned} \tag{12}$$

By the law of large numbers, $n^{-1}\sum_{i=1}^n \frac{\partial^2 \ln f_X(X_i,\theta_0)}{\partial\theta\partial\theta'} \to_p -I(\theta_0)$ (suppose that the information identity holds). By the CLT, $n^{-1/2}\sum_{i=1}^n \frac{\partial \ln f_X(X_i,\theta_0)}{\partial\theta} \to_d N(0, I(\theta_0))$. Suppose that $I(\theta_0)$ is invertible, then by the continuous mapping theorem, the first summand on the right-hand-side of the above display converges in distribution to

$$(-I(\theta_0))^{-1}N(0,I(\theta_0)) =_d N(0, I(\theta_0)^{-1}). \tag{13}$$

If the second summand is $o_p(1)$, then we immediately have

$$\sqrt{n}(\hat{\theta}_{mle}-\theta_0) \to_d N(0, I(\theta_0)^{-1}). \tag{14}$$

We have yet to show that the second summand in the RHS of (12) is $o_p(1)$. To show this, let Y_n denote $||\left(n^{-1}\sum_{i=1}^n \frac{\partial^2 \ln f_X(X_i,\theta_0)}{\partial\theta\partial\theta'}\right)^{-1} n^{-1/2}\sum_{i=1}^n \frac{\partial \ln f_X(X_i,\theta_0)}{\partial\theta}||$. Then$Y_n \to_d ||N(0, I(\theta_0)^{-1}||$ and thus $Y_n = O_p(1)$. Take norm on both sides and divide both sides by $\sqrt{n}||\hat{\theta}_{mle}-\theta_0||$, we have

$$1 = \frac{Y_n}{\sqrt{n}||\hat{\theta}_{mle}-\theta_0||} + o_p(1).$$

This implies that $\frac{Y_n}{\sqrt{n}||\hat{\theta}_{mle}-\theta_0||} \to_p 1$. By the Slutsky theorem, $\frac{\sqrt{n}||\hat{\theta}_{mle}-\theta_0||}{Y_n} \to_p 1$. Thus, $\sqrt{n}||\hat{\theta}_{mle}-\theta_0|| = O_p(Y_n)) = O_p(O_p(1)) = O_p(1)$. This shows that the second summand in the RHS of (12) is $o_p(O_p(1)) = o_p(1)$.