

Lecture Note 15: Large Sample Properties of Maximum Likelihood Estimators, CB 10.1.1-10.1.3

Previously, we showed that if there is an Minimum Variance Unbiased Estimator with variance equal to the Cramer–Rao bound, then the MVUE is equal to the Maximum Likelihood Estimator (MLE). However, the conditions were fairly restrictive. In models which do not satisfy those conditions, the MLE is not unbiased in general, so it cannot be the MVUE. Nevertheless, it turns out that in a certain *approximate* sense, the maximum likelihood estimator is unbiased and minimum variance. The approximations hold when the sample size is large, and we refer to these as *asymptotic* or *large sample* approximations.

Example

Let X_1, \dots, X_n be a random sample from an exponential distribution with arrival rate λ^* : $f_X(x; \lambda^*) = \lambda^* \exp(-x\lambda^*)$. The Cramér-Rao bound for the variance is λ^{*2}/N . The log likelihood function is

$$L(\lambda) = \sum_{i=1}^N \ln \lambda - x_i \lambda,$$

and the maximum likelihood estimator is $\hat{\lambda} = 1/\bar{x}$. What can we say about the large sample properties of this estimator? Using the law of large numbers we have

$$\bar{x} \xrightarrow{p} E[X] = 1/\lambda^*,$$

so

$$\hat{\lambda} = 1/\bar{x} \xrightarrow{p} 1/E[X] = \lambda^*.$$

Using the central limit theorem we also have

$$\sqrt{N} \cdot (\bar{x} - 1/\lambda^*) \xrightarrow{d} \mathcal{N}(0, 1/\lambda^{*2}).$$

Then we can use the delta method to establish that

$$\sqrt{N} \cdot (g(\bar{x}) - g(1/\lambda^*)) \xrightarrow{d} \mathcal{N}(0, g'(1/\lambda^*)^2/\lambda^{*2}).$$

Applying this with $g(a) = 1/a$, and thus $g'(a) = -1/a^2$, we get

$$\sqrt{N} \cdot (1/\bar{x}) - \lambda^* \xrightarrow{d} \mathcal{N}(0, \lambda^{*4}/\lambda^{*2}) = \mathcal{N}(0, \lambda^{*2}).$$

Hence, approximately,

$$\hat{\lambda} \sim \mathcal{N}(\lambda^*, \lambda^{*2}/N).$$

So, approximately, in large samples, this maximum likelihood estimator is unbiased, and has variance approximately equal to the Cramér-Rao bound. This is true in general for maximum likelihood estimators. \square

Result 1

Let X_1, \dots, X_n be a random sample from $f_X(x; \theta^*)$. Assume that the regularity conditions in CB 10.6.2 hold, and let $\hat{\theta}$ be the maximum likelihood estimator:

$$\hat{\theta} = \operatorname{argmax}_{\theta} \sum_{i=1}^N \ln f_X(x_i; \theta).$$

Then $\hat{\theta}$ is consistent for θ^* :

$$\hat{\theta} \xrightarrow{p} \theta^*,$$

and $\hat{\theta}$ has asymptotically a normal distribution:

$$\sqrt{N}(\hat{\theta} - \theta^*) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}(\theta^*)^{-1}),$$

where $\mathcal{I}(\theta^*)$ is the single observation information matrix:

$$\mathcal{I}(\theta^*) = E \left[\left(\frac{\partial \ln f_X}{\partial \theta}(X; \theta^*) \right)^2 \right] = -E \left[\frac{\partial^2 \ln f_X}{\partial \theta^2}(X; \theta^*) \right].$$

□

First let us interpret this result using the Cramer–Rao bound. The CR bound implies that no unbiased estimator has a variance smaller than

$$\mathcal{I}(\theta^*)^{-1}/N.$$

The maximum likelihood estimator has a limiting normal distribution

$$\sqrt{N}(\hat{\theta} - \theta^*) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}(\theta^*)^{-1}),$$

implying that for fixed, large N ,

$$\sqrt{N}(\hat{\theta} - \theta^*) \approx \mathcal{N}(0, \mathcal{I}(\theta^*)^{-1}).$$

This in turn implies that

$$\hat{\theta}_{mle} \approx \mathcal{N}(\theta^*, \mathcal{I}(\theta^*)^{-1}/N).$$

Now, if this was the exact distribution of the MLE, it would be the minimum variance unbiased estimator. Although this is only the approximate distribution in large samples, it seems reasonable to think of the MLE as “approximately optimal.”¹

Example

To illustrate what this means consider an example we have looked at before, where the maximum likelihood estimator differs from the minimum variance unbiased estimator. Suppose X_1, \dots, X_N are a random sample from a normal distribution with unknown mean μ and unknown variance σ^2 . We are interested in the variance σ^2 . The minimum variance unbiased estimator is

$$W_1 = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2.$$

The maximum likelihood estimator is

$$W_2 = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2 = \frac{N-1}{N} \cdot W_1.$$

As the sample gets large, the two estimators get close to each other. They are both consistent and have the same large sample distributions.

$$\sqrt{N} \cdot (W_1 - \sigma^2) \xrightarrow{d} \mathcal{N}(0, 2 \cdot \sigma^4),$$

¹This reasoning can be made more formal. One such result is a statement that any other estimator that is asymptotically unbiased has higher asymptotic variance, at almost all points in the parameter space.

and

$$\sqrt{N} \cdot (W_2 - \sigma^2) \xrightarrow{d} \mathcal{N}(0, 2 \cdot \sigma^4),$$

□

Sketch of Proof of Result 1:

For each value of θ , we can apply a law of large numbers so that

$$\frac{1}{N}L(\theta) = \frac{1}{N} \sum_{i=1}^N \ln f_X(X_i; \theta) \xrightarrow{p} E[\ln f_X(X; \theta)].$$

In addition we know from Jensen's inequality that

$$\theta^* = \operatorname{argmax} E[\ln f_X(X; \theta)].$$

To get the result that

$$\operatorname{argmax} \frac{1}{N}L(\theta) = \operatorname{argmax} E \left[\frac{1}{N}L(\theta) \right] = \theta^*,$$

we need that the convergence is not just pointwise, but uniform in θ , that is,

$$\sup_{\theta} \left| \frac{1}{N}L(\theta) - E \left[\frac{1}{N}L(\theta) \right] \right| \xrightarrow{p} 0.$$

This implies that the convergence to the limit is not much weaker for some values of θ than for others. It requires stronger regularity conditions than pointwise convergence. (Sufficient but not necessary is that $\ln f_X(x; \theta) \leq k(x)$, with $E[k(X)] < \infty$.) In large samples at the maximum likelihood estimator the derivative of the log likelihood function must be equal to zero:

$$\frac{\partial L}{\partial \theta}(\hat{\theta}) = 0.$$

Now expand the derivative of the log likelihood function around the true value of theta:

$$0 = \frac{\partial L}{\partial \theta}(\hat{\theta}) = \frac{\partial L}{\partial \theta}(\theta^*) + \frac{\partial^2 L}{\partial \theta^2}(\tilde{\theta}) \cdot (\hat{\theta} - \theta^*),$$

for some $\tilde{\theta}$ between θ^* and $\hat{\theta}$. In large samples $\hat{\theta} \rightarrow \theta^*$, and therefore $\tilde{\theta} \rightarrow \theta^*$. Rearranging this gives

$$\hat{\theta} - \theta = \left[\frac{\partial^2 L}{\partial \theta^2}(\tilde{\theta}) \right]^{-1} \cdot \frac{\partial L}{\partial \theta}(\theta^*),$$

or

$$\sqrt{N} \cdot (\hat{\theta} - \theta) = \left[\frac{\partial^2 L}{\partial \theta^2}(\tilde{\theta}) / N \right]^{-1} \cdot \left[\frac{\partial L}{\partial \theta}(\theta^*) / \sqrt{N} \right].$$

In large samples

$$-\frac{\partial^2 L}{\partial \theta^2}(\tilde{\theta}) / N \approx -\frac{1}{N} \sum_{i=1}^N \frac{\partial^2 \ln f_X}{\partial \theta^2}(x_i; \tilde{\theta}) \xrightarrow{p} \mathcal{I}(\theta^*),$$

converges in probability to the information matrix $\mathcal{I}(\theta^*)$. The second part,

$$\frac{\partial L}{\partial \theta}(\theta^*) / \sqrt{N} = \frac{1}{\sqrt{N}} \sum_{i=1}^N \frac{\partial \ln f_X}{\partial \theta}(x_i; \theta^*) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}(\theta^*)),$$

because it satisfies a central limit theorem with variance equal to the information matrix. This completes the argument.

Random Vectors and Multiple Parameters

In many models, the parameter θ may be a vector. For example, in the normal model with mean μ and variance σ^2 , we can think of the parameter as a 2-vector $\theta = (\mu, \sigma)'$. It turns out that everything extends very easily to the case with a vector parameter, but we need to introduce a bit of additional notation and a few extensions of our previous results.

Suppose that X is a K -dimensional random vector. The CDF of the random vector X can be defined as before. For $x \in \mathbb{R}^k$,

$$F_X(x) := P(X \leq x),$$

where now $X \leq x$ means that the inequality holds for every element: $X_i \leq x_i$ for each i .

Now consider a *sequence* of random vectors X_1, X_2, \dots . (Be careful of notation: now each X_n is a k -dimensional random vector.) Then X_n converges in distribution to a random vector X if our previous definition holds:

$$F_{X_n}(x) \rightarrow F_X(x),$$

at each continuity point of F_X .

For convergence in probability, we need to modify our previous definition only slightly. For a vector $x \in \mathbb{R}^k$, its length is defined as

$$\|x\| := \left(\sum_{i=1}^k x_i^2 \right)^{1/2}.$$

This is just the usual “Euclidean length” of a vector. Now, a sequence of random vectors X_n converges in probability to a constant vector $c \in \mathbb{R}^k$ if, for every $\epsilon > 0$,

$$P(\|X_n - c\| > \epsilon) \rightarrow 0.$$

The standard Law of Large Numbers and the Central Limit Theorem extend to the vector case. For example, if X_1, X_2, \dots are IID random vectors with mean μ and variance matrix Σ (note: this allows elements within a vector to be nonindependent and have different distributions), then

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

where $N(0, \Sigma)$ is the multivariate normal distribution with mean $(0, 0, \dots, 0)'$ and variance matrix Σ .

Finally, having developed this extra notation, we can consider the large sample properties of MLE. Suppose that our model $f(x; \theta)$ now depends on a vector parameter $\theta = (\theta_1, \dots, \theta_k)'$. Result 1 extends to this case as follows:

Result 2

Let X_1, \dots, X_n be a random sample from $f_X(x; \theta^*)$, where θ^* is $k \times 1$. Let $\hat{\theta}$ be the maximum

likelihood estimator:

$$\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_k)' = \arg \max_{\theta} \sum_{i=1}^N \ln f_X(x_i; \theta).$$

Then $\hat{\theta}$ is consistent for θ^* :

$$\hat{\theta} \xrightarrow{p} \theta^*,$$

and $\hat{\theta}$ is asymptotically normally distributed:

$$\sqrt{N}(\hat{\theta} - \theta^*) \xrightarrow{d} N(0, \mathcal{I}(\theta^*)^{-1}),$$

where $\mathcal{I}(\theta^*)$ is the single observation information matrix:

$$\mathcal{I}(\theta^*) = E \left[\frac{\partial \ln f(X; \theta^*)}{\partial \theta} \cdot \frac{\partial \ln f(X; \theta^*)}{\partial \theta'} \right] = -E \left[\frac{\partial^2 \ln f(X; \theta^*)}{\partial \theta \partial \theta'} \right].$$

(Note: the term $\frac{\partial \ln f(X; \theta^*)}{\partial \theta}$ is $k \times 1$ and the information matrix $\mathcal{I}(\theta)$ is $k \times k$.)