Lecture 7
Maximum Likelihood Estimation.

1 Attainability of Rao-Cramer bound

In the last lecture we derived the Rao-Cramer bound such that no unbiased estimator may have lower variance than this bound, at least under some regularity conditions. Now we can ask the question of whether we can always attain this bound in the sense that there exists an estimator whose variance equals the Rao-Cramer bound. To answer this question note that the crucial step in the derivation of the bound was the Cauchy-Schwarz inequality, i.e. $(\text{cov}(W(X), \partial l_n(\theta, X)/\partial\theta)^2 \leq V(W(X))I_n(\theta)$. Thus, an estimator $W(X)$ will attain the Rao-Cramer bound if and only if the inequality above holds as an equality which, in turns, happens if and only if $W(X)$ and $\partial l_n(\theta, X)/\partial\theta$ are linearly dependent. In this case there exist functions $a(\theta)$ and $b(\theta)$ such that $\partial l_n(\theta, X)/\partial l\theta = a(\theta)(W(X) - b(\theta))$. By the first information equality, $E[\partial l_n(\theta, X)/\partial l\theta] = 0$. So, since $W(X)$ is unbiased, it should be the case that $\partial l_n(\theta, X)/\partial\theta = a(\theta)(W(X) - \theta)$. Thus, there exists an unbiased estimator which attains the Rao-Cramer bound if and only if there exists some function $a(\theta)$ such that $(\partial l_n(\theta, X)/\partial l\theta)/a(\theta) + \theta$ is independent of θ.

As an example, let $X_1, ...X_n$ be a random sample from an $N(\mu, \sigma^2)$ distribution. Suppose that μ is known. Then the log-likelihood is

$$l_n = C - (n/2)\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

The first derivative is

$$\frac{\partial l_n}{\partial\sigma^2} = -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^4} = \frac{n}{2\sigma^4}\left(\sum_{i=1}^{n}(x_i - \mu)^2/n - \sigma^2\right).$$

Thus, $\hat{\sigma}^2 = \sum_{i=1}^{n}(X_i - \mu)^2/n$ attains the Rao-Cramer bound. If μ is unknown, then Rao-Cramer cannot be attained.

2 MLE

Let $f(\cdot|\theta)$ with $\theta \in \Theta$ be a parametric family. Let $X = (X_1, ..., X_n)$ be a random sample from distribution $f(\cdot|\theta_0)$ with $\theta_0 \in \Theta$. Then the joint pdf is $f_n(x|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$ where $x = (x_1, ..., x_n)$. The log-likelihood is

$l_n(\theta, x) = \sum_{i=1}^n \log f(x_i|\theta)$. The maximum likelihood estimator is, by definition, $\hat{\theta}_{ML} = \arg\max_{\theta\in\Theta} l_n(\theta, x)$. The FOC is $\sum_{i=1}^n \partial l(\hat{\theta}_{ML}, x_i)/\partial\theta/n = 0$. Note that the first information equality is $E[\partial l(\theta_0, X_i)] = 0$. Thus MLE is the method of moments estimator corresponding to the first information equality. So we can expect that the MLE is consistent. Indeed, the theorem below gives the consistency result for MLE:

Theorem 1 (MLE consistency)**.** *In the setting above, assume that (1) θ_0 is identifiable, i.e. for any $\theta \neq \theta_0$, there exists x such that $f(x|\theta) \neq f(x|\theta_0)$, (2) the support of $f(\cdot|\theta)$ does not depend on θ, and (3) θ_0 is an interior point of parameter space Θ. Then $\hat{\theta}_{ML} \to_p \theta_0$.*

The proof of MLE consistency will be given in 14.385.

Once we know that the estimator is consistent, we can think about the asymptotic distribution of the estimator. The next theorem gives the asymptotic distribution of MLE:

Theorem 2 (MLE asymptotic normality)**.** *In the setting above, assume that conditions (1)-(3) in the MLE consistency theorem hold. In addition, assume that (4) $f(x|\theta)$ is thrice differentiable with respect to θ and we can interchange integration with respect to x and differentiation with respect to θ, and (5) $|\partial^3 \log f(x|\theta)/\partial\theta^3| \leq M(x)$ and $E[M(X_i)] < \infty$. Then*

$$\sqrt{n}(\hat{\theta}_{ML} - \theta_0) \Rightarrow N(0, I^{-1}(\theta_0))$$

Proof. By definition, $\partial l_n(\hat{\theta}_{ML}, x)/\partial\theta = 0$. By the Taylor theorem with a remainder, there is some random variable $\tilde{\theta}$ with value between θ_0 and $\hat{\theta}_{ML}$ such that

$$\frac{\partial l_n(\hat{\theta}_{ML})}{\partial\theta} = \frac{\partial l_n(\theta_0)}{\partial\theta} + \frac{\partial^2 l_n(\tilde{\theta})}{\partial\theta^2}(\hat{\theta}_{ML} - \theta_0).$$

So,

$$\sqrt{n}(\hat{\theta}_{ML} - \theta_0) = \frac{-(1/\sqrt{n})\partial l_n(\theta_0)/\partial\theta}{(1/n)\partial^2 l_n(\tilde{\theta})/\partial\theta^2}.$$

Since $\hat{\theta}_{ML} \to_p \theta_0$ and $\tilde{\theta}$ is between θ_0 and $\hat{\theta}_{ML}$, $\tilde{\theta} \to_p \theta_0$ as well. From $\tilde{\theta} \to_p \theta_0$, one can prove that $(1/n)\partial^2 l_n(\tilde{\theta})/\partial\theta^2 - (1/n)\partial^2 l_n(\theta_0)/\partial\theta^2 \to_p 0$. We will not discuss this result here since it requires knowledge of the concept of asymptotic equicontinuity which we do not cover in this class. You will learn it in 14.385. Note, however, that this result does not follow from the Continuous mapping theorem since we have a sequence of functions l_n instead of just one function. Suppose we believe in this result. Then, by the Law of large numbers, $(1/n)\partial^2 l_n(\theta_0)/\partial\theta^2 \to_p E[\partial^2 l(\theta_0, X_i)/\partial\theta^2]$. By the Slutsky theorem, $(1/n)\partial^2 l_n(\tilde{\theta})/\partial\theta^2 \to_p E[\partial^2 l(\theta_0, X_i)/\partial\theta^2] = -I(\theta_0)$.

Next, by the first information equality, $E[\partial l(\theta_0, X_i)/\partial\theta] = 0$. Thus, by the Central limit theorem,

$$(1/\sqrt{n})\partial l_n(\theta_0)/\partial\theta = (1/\sqrt{n})\sum_{i=1}^n \partial l(\theta_0, X_i)/\partial\theta \Rightarrow N(0, E[(\partial l(\theta_0, X_i)/\partial\theta)^2]) = N(0, I(\theta_0)).$$

Finally, by the Slutsky theorem again,

$$\sqrt{n}(\hat{\theta}_{ML} - \theta_0) \Rightarrow N(0, I^{-1}(\theta_0)).$$

□

Example Let $X_1, ..., X_n$ be a random sample from a distribution with pdf $f(x|\lambda) = \lambda \exp(-\lambda x)$. This distribution is called exponential. Its loglikelihood is $l(\lambda, x_i) = \log \lambda - \lambda x_i$. So $\partial l(\lambda, x_i)/\partial \lambda = 1/\lambda - x_i$ and $\partial^2 \lambda(\lambda, x_i) = -1/\lambda^2$. So Fisher information is $I(\lambda) = 1/\lambda^2$. Let us find the MLE for λ. The joint loglikelihood is $l_n(\theta, x) = n \log \lambda - \lambda \sum_{i=1}^n x_i$. The FOC is $n/\hat{\lambda}_{ML} - \sum_{i=1}^n x_i = 0$. So $\hat{\lambda}_{ML} = \frac{1}{\overline{X}_n}$. Its asymptotic distribution is given by $\sqrt{n}(\hat{\lambda}_{ML} - \lambda) \Rightarrow N(0, \lambda^2)$.

Example A word of caution. For asymptotic normality of MLE, we should have common support. Let us see what might happen otherwise. Let $X_1, ..., X_n$ be a random sample from $U[0, \theta]$. Then $\hat{\theta}_{ML} = X_{(n)}$. So $\sqrt{n}(\hat{\theta}_{ML} - \theta)$ is always nonpositive. So it does not converge to mean zero normal distribution. In fact, $E[X_{(n)}] = (n/(n+1))\theta$ and $V(X_{(n)}) = \theta^2 n/((n+1)^2(n+2)) \approx \theta^2/n^2$. On the other hand, if the theorem worked, we would have $V(X_{(n)}) \approx 1/(nI(\theta))$.

Example Now, let us consider what might happen if the true parameter value θ_0 were on the boundary of Θ. Let $X_1, ..., X_n$ be a random sample from distribution $N(\mu, 1)$ with $\mu \geq 0$. As an exercise, check that $\hat{\mu}_{ML} = \overline{X}_n$ if $\overline{X}_n \geq 0$ and 0 otherwise. Suppose that $\mu_0 = 0$. Then $\sqrt{n}(\hat{\mu}_{ML} - \mu_0)$ is always nonnegative. So it does not converge to mean zero normal distribution.

Example Finally, note that it is implicitly assumed both in the consistency and asymptotic normality theorems that parameter space Θ is fixed, i.e. independent of n. In particular, the number of parameters should not depend on n. Indeed, let

$$X_i = \begin{pmatrix} X_{1i} \\ X_{2i} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_i \\ \mu_i \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix} \right)$$

for $i = 1, ..., n$, and $X_1, ..., X_n$ be mutually independent. Then

$$f(x_i|\mu_i, \sigma^2) = \frac{1}{2\pi\sigma^2} \exp\left\{ -\frac{1}{2\sigma^2}[(x_{1i} - \mu_i)^2 + (x_{2i} - \mu_i)^2] \right\}$$

and

$$l_n = C - n \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n [(x_{1i} - \mu_i)^2 + (x_{2i} - \mu_i)^2].$$

So the FOC with respect to μ_i is $[(x_{1i} - \hat{\mu}_i) + (x_{2i} - \hat{\mu}_i)] = 0$. So $\hat{\mu}_i = (X_{1i} + X_{2i})/2$. The FOC with respect to σ^2 is

$$-\frac{n}{\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4} \sum_{i=1}^n [(x_{1i} - \hat{\mu}_i)^2 + (x_{2i} - \hat{\mu}_i)^2] = 0,$$

or, equivalently,

$$-\frac{n}{\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4} \sum_{i=1}^n [(x_{1i} - x_{2i})^2/4 + (x_{2i} - x_{1i})^2/4] = 0,$$

since $x_{1i} - \hat{\mu}_i = (x_{1i} - x_{2i})/2$. Thus, $\hat{\sigma}^2 = \sum_{i=1}^{n}(X_{1i} - X_{2i})^2/(4n)$. This estimator is not consistent for σ^2 since $E[(X_{1i} - X_{2i})^2] = 2\sigma^2$ and $E[\hat{\sigma}^2] = \sigma^2/2$.

14.381 Statistical Method in Economics
Fall 2013