Maximum Likelihood Estimation

February 14, 2008

As before, we begin with a sample $X = (X_1, \dots, X_n)$ of random variables chosen according to one of a family of probabilities P_θ where θ is element from the parameter space Θ.

In addition, $\mathbf{f}(\mathbf{x}|\theta)$, $\mathbf{x} = (x_1, \dots, x_n)$ will be used to denote the density function for the data when θ is the state of nature.

Definition 1. *The* **likelihood function** *is the density function regarded as a function of θ.*

$$\mathbf{L}(\theta|\mathbf{x}) = \mathbf{f}(\mathbf{x}|\theta), \ \theta \in \Theta. \tag{1}$$

The **maximum likelihood estimator (MLE)**,

$$\hat{\theta}(\mathbf{x}) = \arg\max_\theta \mathbf{L}(\theta|\mathbf{x}). \tag{2}$$

Note that if $\hat{\theta}(\mathbf{x})$ is a maximum likelihood estimator for θ, then $g(\hat{\theta}(\mathbf{x}))$ is a maximum likelihood estimator for $g(\theta)$.

Typically, maximizing $\ln \mathbf{L}(\theta|\mathbf{x})$ will be easier.

Example 2 (Bernoulli trials)**.** *If the experiment consists of n Bernoulli trial with success probability θ, then*

$$\mathbf{L}(\theta|\mathbf{x}) = \theta^{x_1}(1-\theta)^{(1-x_1)} \cdots \theta^{x_n}(1-\theta)^{(1-x_n)} = \theta^{(x_1+\cdots+x_n)}(1-\theta)^{n-(x_1+\cdots+x_n)}.$$

$$\ln \mathbf{L}(\theta|\mathbf{x}) = \ln\theta(\sum_{i=1}^n x_i) + \ln(1-\theta)(n - \sum_{i=1}^n x_i) = n\bar{x}\ln\theta + n(1-\bar{x})\ln(1-\theta).$$

$$\frac{\partial}{\partial\theta}\ln \mathbf{L}(\theta|\mathbf{x}) = n\left(\frac{\bar{x}}{\theta} - \frac{1-\bar{x}}{1-\theta}\right).$$

This equals zero when $\theta = \bar{x}$. Check that this is a maximum. Thus,

$$\hat{\theta}(\mathbf{x}) = \bar{x}.$$

1 Examples

Example 3 (Normal data)**.** *Maximum likelihood estimation can be applied to a vector valued parameter. For a simple random sample of n normal random variables,*

$$\mathbf{L}(\mu, \sigma^2|\mathbf{x}) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\frac{-(x_1-\mu)^2}{2\sigma^2}\right)\cdots\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\frac{-(x_n-\mu)^2}{2\sigma^2}\right) = \frac{1}{\sqrt{(2\pi\sigma^2)^n}}\exp -\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2.$$

$$\ln \mathbf{L}(\mu, \sigma^2|\mathbf{x}) = -\frac{n}{2} \ln 2\pi\sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2.$$

$$\frac{\partial}{\partial \mu} \ln \mathbf{L}(\mu, \sigma^2|\mathbf{x}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = . \frac{1}{\sigma^2} n(\bar{x} - \mu)$$

Because the second partial derivative with respect to μ is negative,

$$\hat{\mu}(\mathbf{x}) = \bar{x}$$

is the maximum likelihood estimator.

$$\frac{\partial}{\partial \sigma^2} \ln \mathbf{L}(\mu, \sigma^2|\mathbf{x}) = -\frac{n}{\sigma^2} + \frac{1}{(\sigma^2)^2} \sum_{i=1}^{n} (x_i - \mu)^2 = \frac{n}{(\sigma^2)^2} \left(\sigma^2 - \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2 \right).$$

Recalling that $\hat{\mu}(\mathbf{x}) = \bar{x}$, we obtain

$$\hat{\sigma}^2(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{x})^2.$$

Note that the maximum likelihood estimator is a biased estimator.

Example 4 (Uniform random variables)**.** *If our data $X = (X_1, \ldots, X_n)$ are a simple random sample drawn from uniformly distributed random variable whose maximum value θ is unknown, then each random variable has density*

$$f(x|\theta) = \begin{cases} 1/\theta & \text{if } 0 \le x \le \theta, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, the likelihood

$$\mathbf{L}(\theta|\mathbf{x}) = \begin{cases} 1/\theta^n & \text{if, for all } i,\ 0 \le x_i \le \theta, \\ 0 & \text{otherwise.} \end{cases}$$

Consequently, to maximize $\mathbf{L}(\theta|\mathbf{x})$, we should minimize the value of θ^n in the first alternative for the likelihood. This is achieved by taking

$$\hat{\theta}(\mathbf{x}) = \max_{1 \le i \le n} x_i.$$

However,

$$\hat{\theta}(X) = \max_{1 \le i \le n} X_i < \theta$$

and the maximum likelihood estimator is biased.

For $0 \le x \le \theta$, the distribution of $X_{(n)} = \max_{1 \le i \le n} X_i$ is

$$F_{(n)}(x) = P\{\max_{1 \le i \le n} X_i \le x\} = P\{X_1 \le x\}^n = (x/\theta)^n.$$

Thus, the density

$$f_{(n)}(x) = \frac{n x^{n-1}}{\theta^n}.$$

The mean

$$E_\theta X_{(n)} = \frac{n}{n+1} \theta.$$

and thus

$$d(X) = \frac{n+1}{n} X_{(n)}$$

is an unbiased estimator of θ.

As can be seen in this and in other examples, the maximum likelihood estimator has some problems. The attraction of this estimation techniques is in its application to large simple random samples.

2 Asymptotic Properties

If θ_0 is the state of nature, then

$$\mathbf{L}(\theta_0|X) > \mathbf{L}(\theta|X)$$

if and only if

$$\frac{1}{n}\sum_{i=1}^{n} \ln \frac{f(X_i|\theta_0)}{f(X_i|\theta)} > 0.$$

By the strong law of large numbers, this sum converges to

$$E_{\theta_0}\left[\ln \frac{f(X_1|\theta_0)}{f(X_1|\theta)}\right].$$

which is greater than 0. From this, we obtain

$$\hat{\theta}(X) \to \theta_0 \quad \text{as } n \to \infty.$$

We call this property of the estimator **consistency**.

Under some assumptions that is meant to insure some regularity, a central limit theorem holds in this context. Here we have

$$\sqrt{n}(\hat{\theta}(X) - \theta_0)$$

converges in distribution as $n \to \infty$ to a normal random variable with mean 0 and variance $1/I(\theta_0)$, the Fisher information for one observation. Thus

$$\mathrm{Var}_{\theta_0}(\hat{\theta}(X)) \approx \frac{1}{nI(\theta_0)},$$

the lowest possible under the Crámer-Rao lower bound. This property is called **asymptotic efficiency**.