Maximum-Likelihood Methods

In statistics a methodology of estimating the parameters of a statistical model is extremely valuable. When a statistical model is specified for a data set, maximum-likelihood estimation (MLE) provides a systematic way to find a statistic or statistics which estimates a parameter or parameters of the model.

Maximum likelihood estimate (MLE). Having observed a random sample $(X_1, \ldots, X_n) = (x_1, \ldots, x_n)$ from an underlying pdf $f(x;\theta)$ with parameter θ, we can construct the *likelihood function*

$$L(\theta, \boldsymbol{x}) = \prod_{i=1}^{n} f(x_i;\theta),$$

and consider it as a function of θ, where a vector $\boldsymbol{x} = (x_1, \ldots, x_n)$ is viewed as unknown constant. Then the *maximum likelihood estimate (MLE)* $\hat{\theta}$ is the value $\theta = \theta^*$ which "maximizes" the likelihood function $L(\theta, \boldsymbol{x})$. It is usually easier to maximize the *log likelihood*

$$\ln L(\theta, \boldsymbol{x}) = \sum_{i=1}^{n} \ln f(x_i;\theta)$$

with respect to the parameter θ.

MLE for Bernoulli trials. Let $f(x;\theta) = \theta^x(1-\theta)^{1-x}$, $x = 0, 1$, be the Bernoulli frequency function with parameter θ of success probability. By solving the equation

$$\frac{d \ln L(\theta, \boldsymbol{x})}{d\theta} = (\textstyle\sum_{i=1}^{n} x_i)\dfrac{1}{\theta} - (n - \sum_{i=1}^{n} x_i)\dfrac{1}{1-\theta} = 0,$$

we obtain $\theta^* = \sum_{i=1}^{n} x_i/n$ which minimizes $\ln L(\theta, \boldsymbol{x})$. Observe that $\theta^* = \sum_{i=1}^{n} x_i/n$ is a function of $\boldsymbol{x}$. By replacing $\boldsymbol{x}$ with a random sample $(X_1, \ldots, X_n)$, it becomes a statistic $\hat{\theta} = \bar{X}$; thus, we are able to construct the MLE $\hat{\theta}$ for θ.

MLE for normal distribution. Let $X_1, \ldots, X_n$ be a random sample from a normal distribution with parameter (μ, σ^2). Then the log likelihood function of parameter (μ, σ^2) is given by

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

Then we can obtain the MLE's $\hat{\mu}$ and $\hat{\sigma}^2$ as a solution to satisfy

$$\frac{\partial \ln L(\mu, \sigma^2)}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = 0;$$

$$\frac{\partial \ln L(\mu, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} (x_i - \mu)^2 = 0,$$

MLE for normal distribution, continued. The solution can be derived as follows:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X};$$
$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2.$$

We will see later that the MLE $\hat{\sigma}^2$ of σ^2 is consistent; however, we can immediately find that this point estimate $\hat{\sigma}^2$ is not equal to the sample variance S^2, and that it is *not* unbiased.

MLE for Poisson distribution. Let $X_1, \ldots, X_n$ be a random sample from a Poisson distribution with parameter λ. Then the log likelihood function of parameter λ is given by

$$\ln L(\lambda) = \ln \lambda \sum_{i=1}^{n} x_i - n\lambda - \sum_{i=1}^{n} \ln x_i!.$$

By solving

$$\frac{d \ln L(\lambda)}{d\lambda} = \frac{1}{\lambda}\sum_{i=1}^{n} x_i - n = 0,$$

we can obtain the MLE

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}.$$

Example 1. Let $X_1, \ldots, X_n$ be a random sample from an exponential density function $f(x;\theta) = \theta e^{-\theta x}$, $x \geq 0$, with parameter $\theta > 0$. Find the MLE $\hat{\theta}$ of θ.

Solution.

$$\ln L(\theta) = -\left(\sum_{i=1}^{n} x_i\right)\theta + n \ln \theta.$$

By solving

$$\frac{d}{d\theta} \ln L(\theta) = -\left(\sum_{i=1}^{n} x_i\right) + \frac{n}{\theta} = 0,$$

we obtain $\hat{\theta} = \dfrac{n}{\sum_{i=1}^{n} x_i}$.

Example 2. Let $X_1, \ldots, X_n$ be a random sample from the density function $f(x;\theta) = \theta L^{\theta} x^{-\theta-1}$, $x \geq L$, where $L > 0$ is given and $\theta > 1$ (Pareto distribution). Find the MLE $\hat{\theta}$ of θ.

Solution.

$$\ln L(\theta) = -\left(\sum_{i=1}^{n} \ln x_i\right)(\theta + 1) + n \ln \theta + n\theta \ln L.$$

By solving

$$\frac{d}{d\theta}\ln L(\theta) = -\left(\sum_{i=1}^{n}\ln x_i\right) + \frac{n}{\theta} + n\ln L = 0,$$

we obtain $\hat{\theta} = \dfrac{n}{\sum_{i=1}^{n}\ln x_i - n\ln L}$.

Consistency property of MLE. One of the important attributes of point estimate is unbiasedness. Since a statistic $\hat{\theta}$ is a random variable, we can consider the expectation $E(\hat{\theta})$ of $\hat{\theta}$. Then the point estimate $\hat{\theta}$ of θ is unbiased if it satisfies $E(\hat{\theta}) = \theta$. In the case of normal distribution, the MLE $\bar{X}$ for μ is unbiased since $E(\bar{X}) = \mu$. However, the MLE $\hat{\sigma}^2$ for σ^2 is not unbiased, since

$$E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = \frac{n-1}{n}\sigma^2$$

Note that the point estimate $\hat{\theta}$ of θ is also dependent on the sample size n. We say that $\hat{\theta}$ is *consistent* if $\hat{\theta}$ converges in probability to the true parameter θ as $n \to \infty$. For example, the above MLE's $\bar{X}$ and $\hat{\sigma}^2$ are both consistent by the weak law of large number. In general, the MLE is consistent under appropriate conditions.

Invariance property of MLE. Suppose that $h(\theta)$ is a one-to-one function of θ. Then it is clearly seen that $\hat{\theta}$ is the MLE for θ if and only if $h(\hat{\theta})$ is the MLE for $h(\theta)$. Even if $h(\theta)$ is not one-to-one, $h(\hat{\theta})$ will be viewed as the MLE which corresponds to the maximum likelihood.

Example 3. We already know the MLE $\hat{\theta}$ for a random sample $X_1, \ldots, X_n$ from the exponential density function $f(x;\theta) = \theta e^{-\theta x}$, $x \geq 0$. Then find the MLE $\hat{\beta}$ for a density function $f(x;\beta) = \frac{1}{\beta}e^{-x/\beta}$, $x \geq 0$.

Solution. We know the MLE $\hat{\theta} = 1/\bar{X}$. Since $\beta = h(\theta) = \frac{1}{\theta}$, we obtain the MLE $\hat{\beta} = \bar{X}$.

Fisher information. Let $f(\boldsymbol{x};\theta)$ be a joint density function for a random sample $\boldsymbol{X} = (X_1, \ldots, X_n)$. Furthermore, we assume that (i) $A = \{\boldsymbol{x} : f(\boldsymbol{x};\theta) > 0\}$ does not depend on θ, and (ii) $\frac{\partial}{\partial\theta}\ln f(\boldsymbol{x};\theta)$ exists; we will call the conditions in (i)–(ii) the *regularity assumptions*. For example, a joint density $f(\boldsymbol{x};\theta)$ of exponential distribution $f(x;\theta) = \theta x^{-\theta x}$, $x \geq 0$, satisfies the regularity assumptions. By observing that $E\left[\frac{\partial}{\partial\theta}\ln f(\boldsymbol{X};\theta)\right] = 0$, we can define the *Fisher information* $I(\theta)$ by

$$I(\theta) = E\left[\left(\frac{\partial}{\partial\theta}\ln f(\boldsymbol{X};\theta)\right)^2\right] = \mathrm{Var}\left(\frac{\partial}{\partial\theta}\ln f(\boldsymbol{X};\theta)\right).$$

In particular we can show that $I(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\ln f(\boldsymbol{X};\theta)\right]$.

Fisher information, continued. Suppose that $f(\boldsymbol{x};\theta)$ is a joint density of random sample by

$$f(\boldsymbol{x};\theta) = f(x_1;\theta) \times \cdots \times f(x_n;\theta).$$

Then we can introduce

$$I_1(\theta) = \mathrm{Var}\left(\frac{\partial}{\partial\theta}\ln f(X_1;\theta)\right)$$

for the single random variable X_1 with common density $f(x_1;\theta)$, and express $I(\theta)$ by

$$I(\theta) = n \times \mathrm{Var}\left(\frac{\partial}{\partial\theta}\ln f(X_1;\theta)\right) = nI_1(\theta).$$

Cauchy-Schwarz inequality. Let X and Y be random variables, and let a be a real number. Then we have $\mathrm{Var}(aX-Y) = a^2\mathrm{Var}(X) - 2a\mathrm{Cov}(X,Y) + \mathrm{Var}(Y) \geq 0$. By substituting $a = \mathrm{Cov}(X,Y)/\mathrm{Var}(X)$, we obtain

$$[\mathrm{Cov}(X,Y)]^2 \leq \mathrm{Var}(X)\cdot\mathrm{Var}(Y).$$

This inequality is a special case of *Cauchy-Schwarz inequality*, and will be referred as such.

Cramér-Rao lower bound. Let $u(\boldsymbol{X})$ be a statistic. Then $E[u(\boldsymbol{X})]$ is a function of θ, say $\psi(\theta) = E[u(\boldsymbol{X})]$. Here we can show that

$$\psi'(\theta) = \mathrm{Cov}\left(u(\boldsymbol{X}), \frac{\partial}{\partial\theta}\ln f(\boldsymbol{X};\theta)\right).$$

By applying the Cauchy-Schwarz inequality we obtain the *Cramér-Rao inequality*

$$\mathrm{Var}(u(\boldsymbol{X})) \geq \frac{[\psi'(\theta)]^2}{I(\theta)}. \tag{9.1}$$

When $u(\boldsymbol{X})$ is an unbiased statistic, the right-hand side of (9.1) becomes $1/I(\theta)$, and is called the *Cramér-Rao lower bound.* If a statistic $u(\boldsymbol{X})$ achieves the Cramér-Rao lower bound, we call $u(\boldsymbol{X})$ an *efficient* estimator. An efficient and unbiased statistic is a minimum variance unbiased estimator.

Approximation of MLE. Suppose that $\hat\theta$ is the MLE and consistent, and that it satisfies $\dfrac{\partial \ln L(\hat\theta, \boldsymbol{X})}{\partial\theta} = 0$. By the Taylor expansion we have

$$\frac{\partial \ln L(\theta,\boldsymbol{X})}{\partial\theta} + (\hat\theta-\theta)\frac{\partial^2 \ln L(\theta,\boldsymbol{X})}{\partial\theta^2} \approx \frac{\partial \ln L(\hat\theta,\boldsymbol{X})}{\partial\theta} = 0.$$

Since $\hat\theta$ is close to θ by consistency, the approximation is valid.

Approximation of MLE, continued. The random variables $\left(\frac{\partial \ln f(X_i;\theta)}{\partial\theta}\right)$'s are iid with mean 0 and variance $I_1(\theta) = \mathrm{Var}\left(\frac{\partial}{\partial\theta}\ln f(X_1;\theta)\right)$. By the central limit theorem,

$$Z_n = \frac{\dfrac{\partial \ln L(\theta,\boldsymbol{X})}{\partial\theta}}{\sqrt{nI_1(\theta)}} = \frac{\displaystyle\sum_{i=1}^n \frac{\partial \ln f(X_i;\theta)}{\partial\theta}}{\sqrt{nI_1(\theta)}}$$

converges to $N(0,1)$ in distribution as $n\to\infty$.

Approximation of MLE, continued. The random variables $\left(\frac{\partial^2 \ln f(X_i;\theta)}{\partial\theta^2}\right)$'s are iid with mean $(-I_1(\theta))$ and finite variance. By the weak law of large number,

$$W_n = \frac{1}{n}\frac{\partial^2 \ln L(\theta,\boldsymbol{X})}{\partial\theta^2} = \frac{1}{n}\sum_{i=1}^n \frac{\partial^2 \ln f(X_i;\theta)}{\partial\theta^2}$$

converges to $-I_1(\theta)$ in probability as $n \to \infty$.

Asymptotic normality of MLE. Together with Slutsky's theorem we can find that

$$\sqrt{nI_1(\theta)}(\hat{\theta} - \theta) \approx \frac{Z_n}{W_n / (-I_1(\theta))}$$

converges to $N(0,1)$ in distribution as $n \to \infty$. Hence, the MLE $\hat{\theta}$ has "approximately" a normal distribution with mean θ and variance $\frac{1}{nI_1(\theta)} = \frac{1}{I(\theta)}$ if n is large, where $I(\theta)$ is the Fisher information for the random sample $\boldsymbol{X}$. This suggests that

$$n \times \mathrm{Var}(\hat{\theta}) \to 1/I_1(\theta) \quad \text{as } n \to \infty.$$

Then we call $\hat{\theta}$ *asymptotically efficient* (cf. Bickel and Doksum, "Mathematical Statistics," Chapter 4).

Confidence intervals for MLE. Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random sample from $f(\boldsymbol{x}; \theta)$, and let $\hat{\theta}$ be the MLE. Then $\sqrt{nI_1(\theta)}(\hat{\theta} - \theta)$ has approximately a standard normal distribution when n is large. By using the critical point $z_{\alpha/2}$, we obtain

$$\begin{aligned} 1 - \alpha &\approx P\left(\left|\sqrt{nI_1(\theta)}(\hat{\theta} - \theta)\right| \le z_{\alpha/2}\right) \\ &= P\left(\hat{\theta} - \frac{z_{\alpha/2}}{\sqrt{nI_1(\theta)}} \le \theta \le \hat{\theta} + \frac{z_{\alpha/2}}{\sqrt{nI_1(\theta)}}\right) \end{aligned}$$

Confidence intervals for MLE, continued. When n is large, we can also approximate $I_1(\theta)$ by $I_1(\hat{\theta})$. Thus, $(1-\alpha)$ level confidence interval for the MLE $\hat{\theta}$ can be approximated by

$$\left(\hat{\theta} - \frac{z_{\alpha/2}}{\sqrt{nI_1(\hat{\theta})}},\ \hat{\theta} + \frac{z_{\alpha/2}}{\sqrt{nI_1(\hat{\theta})}}\right)$$

Exercises

Problem 1. Let $X_1, \ldots, X_n$ be a random sample having the common pdf $f(x;\theta) = e^{-(x-\theta)}$, $\theta \le x < \infty$, with parameter $-\infty < \theta < \infty$.

(a) Show that $L(\theta) = \exp\left(n\theta - \sum_{i=1}^n X_i\right) I_{(-\infty, X_{(1)}]}(\theta)$, where $X_{(1)}$ is the first order statistic and

$$I_{(-\infty, X_{(1)}]}(\theta) = \begin{cases} 1 & \text{if } \theta \le X_{(1)}; \\ 0 & \text{otherwise.} \end{cases}$$

(b) Find the MLE $\hat{\theta}$ of θ using the definition of MLE.

Problem 2. Let $X_1, \ldots, X_n$ be a random sample having the common pdf $f(x;\theta) = 2x/\theta^2$, $0 < x \le \theta$, with parameter $0 < \theta$.

(a) Show that

$$L(\theta) = \left(\prod_{i=1}^n X_i\right)\left(\frac{2}{\theta^2}\right)^n I_{[X_{(n)}, \infty)}(\theta)$$

where $X_{(n)}$ is the n-th order statistic and

$$I_{[X_{(n)}, \infty)}(\theta) = \begin{cases} 1 & \text{if } X_{(n)} \le \theta; \\ 0 & \text{otherwise.} \end{cases}$$

(b) Find the MLE $\hat{\theta}$ of θ using the definition of MLE.

(c) Find the density function for $\hat{\theta}$.

(d) Is $\hat{\theta}$ unbiased?

Problem 3. Let $X_1, \ldots, X_n$ be a random sample having the common pdf $f(x;\theta) = \frac{1}{\Gamma(4)\theta^4} x^3 e^{-x/\theta}$.

(a) Find the Fisher information $I_1(\theta)$ for X_1.

(b) Show that the MLE $\hat{\theta}$ of θ is an efficient estimator of θ.

(c) What is the aymptotic distribution of $\sqrt{n}(\hat{\theta} - \theta)$?

Problem 4. Let $X_1, \ldots, X_n$ be a random sample having $N(0, \theta)$ with parameter $0 < \theta < \infty$.

(a) Find the Fisher information $I_1(\theta)$ for X_1.

(b) Show that the MLE $\hat{\theta}$ of θ is an efficient estimator of θ.

(c) What is the aymptotic distribution of $\sqrt{n}(\hat{\theta} - \theta)$?

Answers to exercises

Problem 1. (b) $L(\theta)$ is an increasing function, and achieves the maximum at $\theta = X_{(1)}$. Thus, MLE becomes $\hat{\theta} = X_{(1)}$.

Problem 2. (b) The likelihood function $L(\theta)$ is an decreasing function, and achieves the maximum at $\theta = X_{(n)}$. Thus, $\hat{\theta} = X_{(n)}$ is the MLE.

(c) $X_{(n)}$ has the density function $f(x) = \frac{2n}{\theta^{2n}} x^{2n-1}$, $0 < x \leq \theta$.

(d) We can calculate $E[X_{(n)}] = \frac{2n}{2n+1}\theta$. Thus, $X_{(n)}$ is not an unbiased estimator of θ.

Problem 3. (a) $I_1(\theta) = \frac{4}{\theta^2}$.

(b) $\hat{\theta} = \frac{1}{4n}\sum_{i=1}^{n} X_i$ is the MLE. Observe that $\sum_{i=1}^{n} X_i$ has a gamma distribution with $\alpha = 4n$ and $\lambda = \frac{1}{\theta}$. Then we can find that $\hat{\theta}$ is unbiased, and that $\mathrm{Var}(\hat{\theta}) = \frac{\theta^2}{4n}$. Since $I(\theta) = nI_1(\theta) = \frac{4n}{\theta^2}$, $\hat{\theta}$ is efficient.

(c) $N(0, \frac{\theta^2}{4})$.

Problem 4. (a) $I_1(\theta) = \frac{1}{2\theta^2}$.

(b) $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i^2$ is the MLE. Observe that $\sum_{i=1}^{n} \frac{X_i^2}{\theta}$ has a χ^2-distribution with n degrees of freedom. Then we can find that $\hat{\theta}$ is unbiased, and that $\mathrm{Var}(\hat{\theta}) = \frac{2\theta^2}{n}$. Since $I(\theta) = nI_1(\theta) = \frac{n}{2\theta^2}$, $\hat{\theta}$ is efficient.

(c) $N(0, 2\theta^2)$.