Maximum Likelihood Estimation, Logistic Regression

Course: Data Mining

Professor: Dr. Tahaei

Herbod Pourali, Ilya Farhangfar

Subject: Maximum Likelihood Estimation

December 2024

In statistics and estimation theory, we might allude to estimating characteristics of the distribution from the corresponding ones of the sample, hoping that the latter would be reasonably close to the former. For example, the sample mean $\bar{x}$ can be thought of as an estimate of the distribution mean μ, and the sample variance s^2 can be used as an estimate of the distribution variance σ^2. Even the relative frequency histogram associated with a sample can be taken as an estimate of the pdf of the underlying distribution.

But how good are these estimates? What makes an estimate good? Can we say anything about the closeness of an estimate to an unknown parameter?

In this section, we consider random variables for which the functional form of the pmf or pdf is known, but the distribution depends on an unknown parameter (say, θ) that may have any value in a set (say, Θ) called the parameter space. For example, perhaps it is known that

$$f(x;\theta) = \frac{1}{\theta}e^{-x/\theta}, \quad 0 < x < \infty, \quad \theta \in \Theta = \{\theta : 0 < \theta < \infty\}.$$

In certain instances, it might be necessary for the experimenter to select precisely one member of the family $\{f(x;\theta), \theta \in \Theta\}$ as the most likely pdf of the random variable. That is, the experimenter needs a point estimate of the parameter θ, namely, the value of the parameter that corresponds to the selected pdf.

In one common estimation scenario, we take a random sample from the distribution to elicit some information about the unknown parameter θ. That is, we repeat the experiment n independent times, observe the sample, $X_1, X_2, \ldots, X_n$, and try to estimate the value of θ by using the observations $x_1, x_2, \ldots, x_n$. The function of $X_1, X_2, \ldots, X_n$ used to estimate θ—say, the statistic $u(X_1, X_2, \ldots, X_n)$—is called an estimator of θ. We want it to be such that the computed estimate $u(x_1, x_2, \ldots, x_n)$ is usually close to θ.

Since we are estimating one member of Θ, such an estimator is often called a *point estimator*. The following example should help motivate one principle that is often used in finding point estimates:

Suppose that X is $b(1,p)$, so that the pmf of X is

$$f(x;p) = p^x(1-p)^{1-x}, \quad x = 0, 1, \quad 0 \le p \le 1.$$

We note that $p \in \Theta = \{p : 0 \le p \le 1\}$, where Θ represents the parameter space—that is, the space of all possible values of the parameter p. Given a random sample $X_1, X_2, \ldots, X_n$, the problem is to find an estimator $u(X_1, X_2, \ldots, X_n)$ such that $u(x_1, x_2, \ldots, x_n)$ is a good point estimate of p, where $x_1, x_2, \ldots, x_n$ are the observed values of the random sample.

Now, the probability that $X_1, X_2, \ldots, X_n$ takes these particular values is (with $\sum x_i$ denoting $\sum_{i=1}^n x_i$):

$$P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum x_i}(1-p)^{n-\sum x_i},$$

which is the joint pmf of $X_1, X_2, \ldots, X_n$ evaluated at the observed values.

One reasonable way to proceed toward finding a good estimate of p is to regard this probability (or joint pmf) as a function of p and find the value of p that maximizes it. That is, we find the p value most likely to have produced these sample values. The joint pmf, when regarded as a function of p, is frequently called the *likelihood function*. Thus, here the likelihood function is

$$L(p) = L(p; x_1, x_2, \ldots, x_n) = f(x_1;p)f(x_2;p)\cdots f(x_n;p) = p^{\sum x_i}(1-p)^{n-\sum x_i}, \quad 0 \le p \le 1.$$

If $\sum_{i=1}^n x_i = 0$, then

$$L(p) = (1-p)^n,$$

which is maximized over $p \in [0,1]$ by taking $\hat{p} = 0$. If, on the other hand, $\sum_{i=1}^n x_i = n$, then

$$L(p) = p^n,$$

and this is maximized over $p \in [0,1]$ by taking $\hat{p} = 1$. If $\sum_{i=1}^n x_i$ equals neither 0 nor n, then $L(0) = L(1) = 0$ while $L(p) > 0$ for all $p \in (0,1)$; thus, in this case it suffices to maximize $L(p)$ for $0 < p < 1$, which we do by standard methods of calculus.

The derivative of $L(p)$ is

$$L'(p) = (\sum x_i) p^{\sum x_i - 1}(1-p)^{n-\sum x_i} - (n - \sum x_i) p^{\sum x_i}(1-p)^{n-\sum x_i - 1}.$$

Setting this first derivative equal to zero gives us, with the restriction that $0 < p < 1$,

$$p^{\sum x_i}(1-p)^{n-\sum x_i}\left(\frac{\sum x_i}{p} - \frac{n-\sum x_i}{1-p}\right) = 0.$$

Since $0 < p < 1$, the preceding equation equals zero when

$$\frac{\sum x_i}{p} - \frac{n-\sum x_i}{1-p} = 0.$$

Multiplying each member of the equation by $p(1-p)$ and simplifying, we obtain

$$\sum_{i=1}^n x_i - np = 0$$

or, equivalently,

$$p = \frac{\sum_{i=1}^n x_i}{n} = \bar{x}.$$

It can be shown that $L''(\bar{x}) < 0$, so that $L(\bar{x})$ is a maximum. The corresponding statistic, namely, $\frac{\sum_{i=1}^n X_i}{n} = \bar{X}$, is called the *maximum likelihood estimator* and is denoted by $\hat{p}$; that is,

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

When finding a maximum likelihood estimator, it is often easier to find the value of the parameter that maximizes the natural logarithm of the likelihood function rather than the value of the parameter that maximizes the likelihood function itself. Because the natural logarithm function is a strictly increasing function, the solutions will be the same.

To see this, note that for $0 < p < 1$, the example we have been considering gives us

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

To find the maximum, we set the first derivative equal to zero to obtain

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

which is the same as the earlier equation. Thus, the solution is $p = \bar{x}$ and the maximum likelihood estimator for p is $\hat{p} = \bar{X}$.

Motivated by the preceding example, we present the formal definition of maximum likelihood estimators (this definition is used in both the discrete and continuous cases).

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution that depends on one or more unknown parameters $\theta_1, \theta_2, \ldots, \theta_m$ with pmf or pdf denoted by $f(x; \theta_1, \theta_2, \ldots, \theta_m)$. Suppose that $(\theta_1, \theta_2, \ldots, \theta_m)$ is restricted to a given parameter space Θ. Then, the joint pmf or pdf of $X_1, X_2, \ldots, X_n$, namely,

$$L(\theta_1, \theta_2, \ldots, \theta_m) = f(x_1; \theta_1, \ldots, \theta_m) f(x_2; \theta_1, \ldots, \theta_m) \cdots f(x_n; \theta_1, \ldots, \theta_m),$$

where $(\theta_1, \theta_2, \ldots, \theta_m) \in \Theta$, is called the *likelihood function* when regarded as a function of $\theta_1, \theta_2, \ldots, \theta_m$.

Suppose $[u_1(x_1, \ldots, x_n), u_2(x_1, \ldots, x_n), \ldots, u_m(x_1, \ldots, x_n)]$ is the m-tuple in Θ that maximizes $L(\theta_1, \theta_2, \ldots, \theta_m)$. Then

$$\hat{\theta}_1 = u_1(X_1, \ldots, X_n), \quad \hat{\theta}_2 = u_2(X_1, \ldots, X_n), \quad \ldots, \quad \hat{\theta}_m = u_m(X_1, \ldots, X_n)$$

are called the *maximum likelihood estimators* of $\theta_1, \theta_2, \ldots, \theta_m$, respectively. The corresponding observed values of these statistics, namely,

$$u_1(x_1, \ldots, x_n), \quad u_2(x_1, \ldots, x_n), \quad \ldots, \quad u_m(x_1, \ldots, x_n),$$

are referred to as *maximum likelihood estimates.*

In many practical cases, these estimators (and estimates) are unique. For many applications, there is just one unknown parameter. In such cases, the likelihood function is given by

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