Chapter 6

Maximum Likelihood Methods

6.1 Maximum Likelihood Estimation

In this chapter we develop statistical inference (estimation and testing) based on likelihood methods. We show that these procedures are asymptotically optimal under certain conditions (regularity conditions). Suppose that $X_1, \ldots, X_n$ are iid random variables with common pdf $f(x;\theta), \theta \in \Omega$. In general, we will use pdf rather than pmf, $p(x;\theta)$, but the results extend to the discrete case, also. For now, we assume that θ is a scalar but we will extend the results to vectors in Sections 6.4 and 6.5. The parameter θ is unknown. The basis of our inferential procedures is the likelihood function given by,

$$L(\theta;\mathbf{x}) = \prod_{i=1}^{n} f(x_i;\theta), \quad \theta \in \Omega, \tag{6.1.1}$$

where $\mathbf{x} = (x_1, \ldots, x_n)'$. Because we will treat L as a function of θ in this chapter, we have transposed the x_i and θ in the argument of the likelihood function. In fact we will often write it as $L(\theta)$. Actually, the log of this function is usually more convenient to work with mathematically. Denote the $\log L(\theta)$ by

$$l(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(x_i;\theta), \quad \theta \in \Omega. \tag{6.1.2}$$

Note that there is no loss of information in using $l(\theta)$ because the log is a one-to-one function. Most of our discussion in this chapter remains the same if X is a random vector. Although, we will generally consider X as a random variable, for several of our examples it will be a random vector.

To motivate the use of the likelihood function, we begin with a simple example and then provide a theoretical justification.

Example 6.1.1. Let $X_1, X_2, \ldots, X_n$ denote a random sample from the distribution with pmf

$$p(x) = \begin{cases} \theta^x(1-\theta)^{1-x} & x = 0, 1 \\ 0 & \text{elsewhere,} \end{cases}$$

where $0 \leq \theta \leq 1$. The probability that $X_1 = x_1,\ X_2 = x_2, \ldots, X_n = x_n$ is the joint pmf

$$\theta^{x_1}(1-\theta)^{1-x_1}\theta^{x_2}(1-\theta)^{1-x_2}\cdots\theta^{x_n}(1-\theta)^{1-x_n} = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i},$$

where x_i equals zero or 1, $i = 1, 2, \ldots, n$. This probability, which is the joint pmf of $X_1, X_2, \ldots, X_n$, as a function of θ is the likelihood function $L(\theta)$ defined above. That is,

$$L(\theta) = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}, \quad 0 \leq \theta \leq 1.$$

We might ask what value of θ would maximize the probability $L(\theta)$ of obtaining this particular observed sample $x_1, x_2, \ldots, x_n$. Certainly, this maximizing value of θ would seemingly be a good estimate of θ because it would provide the largest probability of this particular sample. Since the likelihood function $L(\theta)$ and its logarithm, $l(\theta) = \log L(\theta)$, are maximized for the same value of θ, either $L(\theta)$ or $l(\theta)$ can be used. Here

$$l(\theta) = \log L(\theta) = \left(\sum_1^n x_i\right)\log\theta + \left(n - \sum_1^n x_i\right)\log(1-\theta);$$

so we have

$$\frac{dl(\theta)}{d\theta} = \frac{\sum x_i}{\theta} - \frac{n - \sum x_i}{1-\theta} = 0,$$

provided that θ is not equal to zero or 1. This is equivalent to the equation

$$(1-\theta)\sum_1^n x_i = \theta\left(n - \sum_1^n x_i\right),$$

whose solution for θ is $\sum_1^n x_i/n$. That $\sum_1^n x_i/n$ actually maximizes $L(\theta)$ and $\log L(\theta)$ can be easily checked, even in the cases in which all of $x_1, x_2, \ldots, x_n$ equal zero together or one together. That is, $\sum_1^n x_i/n$ is the value of θ that maximizes $L(\theta)$. The corresponding statistic

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^n X_i = \overline{X},$$

is called the *maximum likelihood estimator* of θ. As formally defined below, we will call $\sum_1^n x_i/n$ the *maximum likelihood estimate* of θ. For a simple example, suppose that $n = 3$, and $x_1 = 1,\ x_2 = 0,\ x_3 = 1$, then $L(\theta) = \theta^2(1-\theta)$ and the observed $\hat{\theta} = \frac{2}{3}$ is the maximum likelihood estimate of θ. ■

Let θ_0 denote the *true value* of θ. Theorem 6.1.1 gives a theoretical reason for maximizing the likelihood function. It says that the maximum of $L(\theta)$ asymptotically separates the true model at θ_0 from models at $\theta \neq \theta_0$. To prove this theorem, we will assume certain assumptions, usually called *regularity conditions.*

Assumptions 6.1.1. (Regularity Conditions).

(R0): *The pdfs are distinct; i.e.,* $\theta \neq \theta' \Rightarrow f(x_i;\theta) \neq f(x_i;\theta')$.

(R1): *The pdfs have common support for all* θ.

(R2): *The point* θ_0 *is an interior point in* Ω.

The first assumption states that the parameters identify the pdfs. The second assumption implies that the support of X_i does not depend on θ. This is restrictive and some examples and exercises will cover models where (R1) is not true.

Theorem 6.1.1. *Let* θ_0 *be the true parameter. Under assumptions (R0) and (R1),*

$$\lim_{n\to\infty} P_{\theta_0}[L(\theta_0, \mathbf{X}) > L(\theta, \mathbf{X})] = 1, \quad \textit{for all } \theta \neq \theta_0. \tag{6.1.3}$$

Proof: By taking logs, the inequality $L(\theta_0, \mathbf{X}) > L(\theta, \mathbf{X})$ is equivalent to

$$\frac{1}{n}\sum_{i=1}^{n} \log\left[\frac{f(X_i;\theta)}{f(X_i;\theta_0)}\right] < 0.$$

Because the summands are iid with finite expectation and the function $\phi(x) = -\log(x)$ is strictly convex, it follows from the Law of Large Numbers (Theorem 4.2.1) and Jensen's inequality, (Theorem 1.10.5), that, when θ_0 is the true parameter,

$$\frac{1}{n}\sum_{i=1}^{n} \log\left[\frac{f(X_i;\theta)}{f(X_i;\theta_0)}\right] \xrightarrow{P} E_{\theta_0}\left[\log\frac{f(X_1;\theta)}{f(X_1;\theta_0)}\right] < \log E_{\theta_0}\left[\frac{f(X_1;\theta)}{f(X_1;\theta_0)}\right].$$

But

$$E_{\theta_0}\left[\frac{f(X_1;\theta)}{f(X_1;\theta_0)}\right] = \int \frac{f(x;\theta)}{f(x;\theta_0)} f(x;\theta_0)\, dx = 1.$$

Because $\log 1 = 0$, the theorem follows. Note that common support is needed to obtain the last equalities. ■

Theorem 6.1.1 says that asymptotically the likelihood function is maximized at the true value θ_0. So in considering estimates of θ_0, it seems natural to consider the value of θ which maximizes the likelihood.

Definition 6.1.1 (Maximum Likelihood Estimator). *We say that* $\widehat{\theta} = \widehat{\theta}(\mathbf{X})$ *is a* **maximum likelihood estimator** *(mle) of* θ *if*

$$\widehat{\theta} = \text{A}rgmax\, L(\theta; \mathbf{X}); \tag{6.1.4}$$

The notation Argmax means that $L(\theta; \mathbf{X})$ *achieves its maximum value at* $\widehat{\theta}$.

As in the example above, to determine the mle, we often take the log of the likelihood and determine its critical value; that is, letting $l(\theta) = \log L(\theta)$ the mle solves the equation

$$\frac{\partial l(\theta)}{\partial \theta} = 0. \tag{6.1.5}$$

This is an example of an **estimating equation** which we will often label as an EE. This is the first of several EEs in the text.

There is no guarantee that the mle exists or if it does whether it is unique. This is often clear from the application as in the next three examples. Other examples are given in the exercises.

Example 6.1.2 (Exponential Distribution). Suppose the common pdf is the exponential(θ) density given by (3.3.2). The log of the likelihood function is given by,

$$l(\theta) = -n \log \theta - \theta^{-1} \sum_{i=1}^{n} x_i.$$

For this example, differentiable calculus leads directly to the mle. The first partial of the log-likelihood with respect to θ is:

$$\frac{\partial l}{\partial \theta} = -n\theta^{-1} + \theta^{-2} \sum_{i=1}^{n} x_i.$$

Setting this partial to 0 and solving for θ we obtain the solution $\overline{x}$. There is only one critical value and, furthermore, the second partial of the log likelihood evaluated at $\overline{x}$ is strictly negative, verifying that it is indeed a maximum. Hence, for this example the statistic $\widehat{\theta} = \overline{X}$ is the mle of θ. ■

Example 6.1.3 (Laplace Distribution). Let $X_1, \ldots, X_n$ be iid with density

$$f(x;\theta) = \frac{1}{2} e^{-|x-\theta|}, \quad -\infty < x < \infty, -\infty < \theta < \infty. \tag{6.1.6}$$

This pdf is referred to as either the *Laplace* or the *double exponential distribution.* The log of the likelihood simplifies to

$$l(\theta) = -n \log 2 - \sum_{i=1}^{n} |x_i - \theta|.$$

The first partial derivative is

$$l'(\theta) = \sum_{i=1}^{n} \text{sgn}(x_i - \theta), \tag{6.1.7}$$

where $\text{sgn}(t) = 1, 0,$ or -1 depending on whether $t > 0, t = 0,$ or $t < 0$. Note that we have used $\frac{d}{dt}|t| = \text{sgn}(t)$ which is true unless $t = 0$. Setting the equation (6.1.7) to 0, the solution for θ is $\text{med}\{x_1, x_2, \ldots, x_n\}$, because the median will make half the terms of the sum in expression (6.1.7) nonpositive and half nonnegative. Recall that we denote the median of a sample by Q_2 (the second quartile of the sample). Hence $\widehat{\theta} = Q_2$ is the mle of θ for the Laplace pdf (6.1.6). ■

Example 6.1.4 (Logistic Distribution). Let $X_1, \ldots, X_n$ be iid with density

$$f(x;\theta) = \frac{\exp\{-(x-\theta)\}}{(1+\exp\{-(x-\theta)\})^2}, \quad -\infty < x < \infty, \ -\infty < \theta < \infty. \tag{6.1.8}$$

The log of the likelihood simplifies to

$$l(\theta) = \sum_{i=1}^{n} \log f(x_i;\theta) = n\theta - n\overline{x} - 2\sum_{i=1}^{n} \log(1+\exp\{-(x_i-\theta)\}).$$

Using this, the first partial derivative is

$$l'(\theta) = n - 2\sum_{i=1}^{n} \frac{\exp\{-(x_i-\theta)\}}{1+\exp\{-(x_i-\theta)\}}. \tag{6.1.9}$$

Setting this equation to 0 and rearranging terms results in the equation,

$$\sum_{i=1}^{n} \frac{\exp\{-(x_i-\theta)\}}{1+\exp\{-(x_i-\theta)\}} = \frac{n}{2}. \tag{6.1.10}$$

Although this does not simplify, we can show that equation (6.1.10) has a unique solution. The derivative of the left side of equation (6.1.10) simplifies to,

$$(\partial/\partial\theta)\sum_{i=1}^{n} \frac{\exp\{-(x_i-\theta)\}}{1+\exp\{-(x_i-\theta)\}} = \sum_{i=1}^{n} \frac{\exp\{-(x_i-\theta)\}}{(1+\exp\{-(x_i-\theta)\})^2} > 0.$$

Thus the left side of equation (6.1.10) is a strictly increasing function of θ. Finally, the left side of (6.1.10) approachs 0 as $\theta \to -\infty$ and approachs n as $\theta \to \infty$. Thus, the equation (6.1.10) has a unique solution. Also the second derivative of $l(\theta)$ is strictly negative for all θ; so the solution is a maximum.

Having shown that the mle exists and is unique, we can use a numerical method to obtain the solution. In this case, Newton's procedure is useful. We discuss this in general in the next section at which time we will reconsider this example. ■

Even though we did not get the mle in closed form in the last example, in all three of these examples standard differential calculus methods led us to the solution. For the next example, the support of the random variable involves θ and, hence, does not satisfy the regularity conditions. For such cases, differential calculus may not be useful.

Example 6.1.5 (Uniform Distribution). Let $X_1, \ldots, X_n$ be iid with the uniform $(0,\theta)$ density, i.e., $f(x) = 1/\theta$ for $0 < x \leq \theta$, 0 elsewhere. Because θ is in the support, differentiation is not helpful here. The likelihood function can be written as

$$L(\theta) = \theta^{-n} I(\max\{x_i\}, \theta); \quad \text{for all } \theta > 0,$$

where $I(a,b)$ is 1 or 0 if $a \leq b$ or $a > b$, respectively. This function is a decreasing function of θ for all $\theta \geq \max\{x_i\}$ and is 0, otherwise, (please sketch it). Hence, the maximum occurs at the smallest value of θ; i.e., the mle is $\widehat{\theta} = \max\{X_i\}$. ■

Example 6.1.6. In Example 6.1.1, we discussed the mle of the probability of success θ for a random sample $X_1, X_2, \ldots, X_n$ from the Bernoulli distribution with pmf

$$p(x) = \begin{cases} \theta^x (1-\theta)^{1-x} & x = 0, 1 \\ 0 & \text{elsewhere,} \end{cases}$$

where $0 \leq \theta \leq 1$. Recall that the mle is $\overline{X}$, the proportion of sample successes. Now suppose that we know in advance that, instead of $0 \leq \theta \leq 1$, θ is restricted by the inequalities $0 \leq \theta \leq 1/3$. If the observations were such that $\overline{x} > 1/3$, then $\overline{x}$ would not be a satisfactory estimate. Since $\frac{\partial l(\theta)}{\partial \theta} > 0$, provided $\theta < \overline{x}$, under the restriction $0 \leq \theta \leq 1/3$, we can maximize $l(\theta)$ by taking $\widehat{\theta} = \min\{\overline{x}, \frac{1}{3}\}$. ■

The following is an appealing property of maximum likelihood estimates.

Theorem 6.1.2. *Let $X_1, \ldots, X_n$ be iid with the pdf $f(x;\theta), \theta \in \Omega$. For a specified function g, let $\eta = g(\theta)$ be a parameter of interest. Suppose $\widehat{\theta}$ is the mle of θ. Then $g(\widehat{\theta})$ is the mle of $\eta = g(\theta)$.*

Proof: First suppose g is a one-to-one function. The likelihood of interest is $L(g(\theta))$, but because g is one-to-one,

$$\max L(g(\theta)) = \max_{\eta = g(\theta)} L(\eta) = \max_{\eta} L(g^{-1}(\eta)).$$

But the maximum occurs when $g^{-1}(\eta) = \widehat{\theta}$; i.e., take $\widehat{\eta} = g(\widehat{\theta})$.

Suppose g is not one-to-one. For each η in the range of g, define the set (preimage),

$$g^{-1}(\eta) = \{\theta : g(\theta) = \eta\}.$$

The maximum occurs at $\widehat{\theta}$ and the domain of g is Ω which covers $\widehat{\theta}$. Hence, $\widehat{\theta}$ is in one of these preimages and, in fact, it can only be in one preimage. Hence to maximize $L(\eta)$, choose $\widehat{\eta}$ so that $g^{-1}(\widehat{\eta})$ is that unique preimage containing $\widehat{\theta}$. Then $\widehat{\eta} = g(\widehat{\theta})$. ■

In Example 6.1.5, it might be of interest to estimate $\text{Var}(X) = \theta^2/12$. Hence by Theorem 6.1.2, the mle is $\max\{X_i\}^2/12$. Next, consider Example 6.1.1, where $X_1, \ldots, X_n$ are iid Bernoulli random variables with probability of success p. As shown in the example, $\widehat{p} = \overline{X}$ is the mle of p. Recall that in the large sample confidence interval for p, (5.4.8), an estimate of $\sqrt{p(1-p)}$ is required. By Theorem 6.1.2, the mle of this quantity is $\sqrt{\widehat{p}(1-\widehat{p})}$.

We close this section by showing that maximum likelihood estimators, under regularity conditions, are consistent estimators. Recall that $\mathbf{X}' = (X_1, \ldots, X_n)$.

Theorem 6.1.3. *Assume that $X_1, \ldots, X_n$ satisfy the regularity conditions (R0) - (R2), where θ_0 is the true parameter, and further that $f(x;\theta)$ is differentiable with respect to θ in Ω. Then the likelihood equation,*

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

or equivalently

$$\frac{\partial}{\partial\theta}l(\theta) = 0$$

has a solution $\widehat{\theta}_n$ *such that* $\widehat{\theta}_n \xrightarrow{P} \theta_0$.

Proof: Because θ_0 is an interior point in Ω, $(\theta_0 - a, \theta_0 + a) \subset \Omega$, for some $a > 0$. Define S_n to be the event

$$S_n = \{\mathbf{X} : l(\theta_0; \mathbf{X}) > l(\theta_0 - a; \mathbf{X})\} \cap \{\mathbf{X} : l(\theta_0; \mathbf{X}) > l(\theta_0 + a; \mathbf{X})\}.$$

By Theorem 6.1.1, $P(S_n) \to 1$. So we can restrict attention to the event S_n. But on S_n, $l(\theta)$ has a local maximum say, $\widehat{\theta}_n$ such that $\theta_0 - a < \widehat{\theta}_n < \theta_0 + a$ and $l'(\widehat{\theta}_n) = 0$. That is,

$$S_n \subset \left\{\mathbf{X} : |\widehat{\theta}_n(\mathbf{X}) - \theta_0| < a\right\} \cap \left\{\mathbf{X} : l'(\widehat{\theta}_n(\mathbf{X})) = 0\right\}.$$

Therefore,

$$1 = \lim_{n\to\infty} P(S_n) \leq \overline{\lim_{n\to\infty}}\, P\left[\left\{\mathbf{X} : |\widehat{\theta}_n(\mathbf{X}) - \theta_0| < a\right\} \cap \left\{\mathbf{X} : l'(\widehat{\theta}_n(\mathbf{X})) = 0\right\}\right] \leq 1;$$

see Remark 4.3.3 for discussion on $\overline{\lim}$. It follows that for the sequence of solutions $\widehat{\theta}_n$, $P[|\widehat{\theta}_n - \theta_0| < a] \to 1$.

The only contentious point in the proof is that the sequence of solutions might depend on a. But we can always choose a solution "closest" to θ_0 in the following way. For each n, the set of all solutions in the interval is bounded, hence the infimum over solutions closest to θ_0 exists. ■

Note that this theorem is vague in that it discusses solutions of the equation. If, however, we know that the mle is the unique solution of the equation $l'(\theta) = 0$, then it is consistent. We state this as a corollary:

Corollary 6.1.1. *Assume that* $X_1, \ldots, X_n$ *satisfy the regularity conditions (R0) - (R2), where* θ_0 *is the true parameter, and that* $f(x;\theta)$ *is differentiable with respect to* θ *in* Ω. *Suppose the likelihood equation has the unique solution* $\widehat{\theta}_n$. *Then* $\widehat{\theta}_n$ *is a consistent estimator of* θ_0.

EXERCISES

6.1.1. Let $X_1, X_2, \ldots, X_n$ be a random sample from a $N(\theta, \sigma^2)$ distribution, $-\infty < \theta < \infty$ with σ^2 known. Determine the mle of θ.

6.1.2. Let $X_1, X_2, \ldots, X_n$ be a random sample from a $\Gamma(\alpha = 3, \beta = \theta)$ distribution, $0 < \theta < \infty$. Determine the mle of θ.

6.1.3. Let $X_1, X_2, \ldots, X_n$ represent a random sample from each of the distributions having the following pdfs or pmfs:

(a) $f(x;\theta) = \theta^x e^{-\theta}/x!$, $x = 0, 1, 2, \ldots$, $0 \leq \theta < \infty$, zero elsewhere, where $f(0;0) = 1$.