Chapter 2

Likelihood-based approach

2.1 Introduction

In this chapter, we discuss the likelihood-based approach in the analysis of missing data. To do this, we first review the likelihood-based methods in the case of complete response. Let $\mathbf{y} = (y_1, y_2, \cdots y_n)$ be a realization of the random sample from an infinite population with density $f(y)$ so that $P(Y \in B) = \int_B f(y) d\mu(y)$ for any measurable set B where $\mu(y)$ is a σ-finite dominating measure. Assume that the true density $f(y)$ belongs to a parametric family of densities $\mathcal{P} = \{f(y;\theta) : \theta \in \Omega\}$ indexed by $\theta \in \Omega$. That is, there exists a $\theta_0 \in \Omega$ such that $f(y;\theta_0) = f(y)$ for all y. Once the parametric density is specified, the likelihood function and the maximum likelihood estimator can be defined formally as follows.

Definition 2.1. *The likelihood function of θ, denoted by $L(\theta)$, is defined as the probability density (mass) function of the observed data* $\mathbf{y}$ *considered as a function of θ. That is,*

$$L(\theta) = f(\mathbf{y};\theta)$$

where $f(\mathbf{y};\theta)$ is the joint density function of $\mathbf{y}$.

Definition 2.2. *Let $\hat{\theta}$ be the maximum likelihood estimator (MLE) of θ_0 if it satisfies*

$$L(\hat{\theta}) = \max_{\theta \in \Omega} L(\theta).$$

If $y_1, y_2, \ldots, y_n$ are independently and identically distributed (IID),

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

Also, if $\hat{\theta}$ is the MLE of θ_0, then $g(\hat{\theta})$ is the MLE of $g(\theta_0)$. The MLE is not necessarily unique. To guarantee uniqueness of the MLE, we require that the family of densities is identified. The definition of an identifiable distribution is given as follows:

Definition 2.3. *A parametric family of densities, given by $\mathcal{P} = \{f(y;\theta); \theta \in \Omega\}$, is called* identifiable (or identified) *if*

$$f(y;\theta_1) \neq f(y;\theta_2) \quad \textit{for every } \theta_1 \neq \theta_2$$

for all $\mathbf{y}$ *in the support of $\mathcal{P}$.*

Under the identifiability condition, the uniqueness of the MLE follows from the following lemma.

Lemma 2.1. *If $\mathcal{P} = \{f(y;\theta); \theta \in \Omega\}$ is identifiable and $E\{|\ln f(Y;\theta)|\} < \infty$ for all θ, then*

$$M(\theta) = -E_{\theta_0} \ln \left\{ \frac{f(y;\theta)}{f(y;\theta_0)} \right\} \geq 0 \tag{2.1}$$

with equality at $\theta = \theta_0$.

Proof. Let $Z = f(y;\theta)/f(y;\theta_0)$. Using the strict version of Jensen's inequality

$$-\ln\left\{E_{\theta_0}(Z)\right\} < E_{\theta_0}\left\{-\ln(Z)\right\}.$$

Because $E_{\theta_0}(Z) = \int f(y;\theta)\,d\mu(y) = 1$, we have $\ln\left\{E_{\theta_0}(Z)\right\} = 0$. □

In Lemma 2.1, $M(\theta)$ is called the *Kullback–Leibler divergence measure* of $f(y;\theta)$ from $f(y;\theta_0)$. It is often considered a measure of distance between two densities. If $\mathcal{P} = \{f(y;\theta);\theta \in \Omega\}$ is not identifiable, then $M(\theta)$ may not have a unique minimizer and $\hat{\theta}$ may not converge (in probability) to a single point.

The following two theorems present some asymptotic properties of the maximum likelihood estimator (MLE): (weak) consistency and asymptotic normality. To discuss the asymptotic properties, let $\hat{\theta}$ be any solution of

$$Q_n(\hat{\theta}) = \min_{\theta\in\Omega} Q_n(\theta)$$

where

$$Q_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \log f(y_i,\theta).$$

Also, define $Q(\theta) = -E_{\theta_0}\{\log f(Y;\theta)\}$ to be the probability limit of $Q_n(\theta)$ evaluated at θ_0. By Lemma 2.1, $Q(\theta)$ is minimized at $\theta = \theta_0$. Thus, under some conditions, we may expect that $\hat{\theta} = \arg\min Q_n(\theta)$ converges to $\theta_0 = \arg\min Q(\theta)$. The following theorem presents a formal result.

Theorem 2.1. *Assume the following two conditions:*

1. *Identifiability: $Q(\theta)$ is uniquely minimized at θ_0. That is, for any $\varepsilon > 0$, there exists a $\delta > 0$ such that $\theta \notin B_\varepsilon(\theta_0)$ implies $Q(\theta) - Q(\theta_0) \geq \delta$, where $B_\varepsilon(\theta_0) = \{\theta \in \Omega; |\theta - \theta_0| < \varepsilon\}$.*
2. *Uniform weak convergence:*

$$\sup_{\theta\in\Omega} |Q_n(\theta) - Q(\theta)| \xrightarrow{p} 0$$

for some nonstochastic function $Q(\theta)$

Then, $\hat{\theta} \xrightarrow{p} \theta_0$.

Proof. For any $\varepsilon > 0$, we can find $\delta > 0$ such that

$$\begin{aligned} 0 \leq P\left[\hat{\theta} \notin B_\varepsilon(\theta_0)\right] &\leq P\left[Q(\hat{\theta}) - Q_n(\hat{\theta}) + Q_n(\hat{\theta}) - Q(\theta_0) \geq \delta\right] \\ &\leq P\left[Q(\hat{\theta}) - Q_n(\hat{\theta}) + Q_n(\theta_0) - Q(\theta_0) \geq \delta\right] \\ &\leq P\left[2\sup|Q_n(\theta) - Q(\theta)| \geq \delta\right] \to 0. \end{aligned}$$

□

In Theorem 2.1, $Q_n(\theta)$ is the negative log-likelihood of θ obtained from $f(y;\theta)$. The function $Q(\theta)$ is the uniform probability limit of $Q_n(\theta)$ and θ_0 is the unique minimizer of $Q(\theta)$. In Theorem 2.1, it is assumed that $\hat{\theta}$ is not necessarily uniquely determined, but θ_0 is. Uniform weak convergence of $Q_n(\theta)$ is stronger than pointwise convergence of $Q_n(\theta)$. One simple set of sufficient conditions for it is that $Q_n(\theta)$ converges pointwise to $Q(\theta)$ and that $Q_n(\theta)$ is continuous and has a unique minimizer at $\theta = \hat{\theta}$.

Theorem 2.2. *Assume the following regularity conditions:*

1. *θ_0 is in the interior of Ω.*
2. *$Q_n(\theta)$ is twice continuously differentiable on some neighborhood $\Omega_0(\subset \Omega)$ of θ_0 almost everywhere.*

3. The first-order partial derivative of Q_n satisfies

$$\sqrt{n}\frac{\partial}{\partial\theta}Q_n(\theta_0) \xrightarrow{d} N(0,A_0)$$

for some positive definite A_0.

4. The second-order partial derivative of Q_n satisfies

$$\sup_{\theta\in\Omega_o} \| \frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\theta) - B(\theta) \| \xrightarrow{p} 0$$

for some $B(\theta)$ continuous at θ_0 and $B_0 = B(\theta_0)$ is nonsingular.

Furthermore, assume that $\hat{\theta}$ satisfies

5. $\hat{\theta} \xrightarrow{p} \theta_0$.

6. $\sqrt{n}\,\partial Q_n(\hat{\theta})/\partial\theta = o_p(1)$.

Then

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right) \xrightarrow{d} N\left(0, B_0^{-1}A_0B_0^{-1'}\right).$$

Proof. By assumption 6 and the mean value theorem,

$$\begin{aligned} \partial Q_n(\hat{\theta})/\partial\theta &= \partial Q_n(\theta_0)/\partial\theta + \left[\partial^2 Q_n(\theta^*)/\partial\theta\partial\theta'\right]\left(\hat{\theta}-\theta_0\right) \\ &= o_p\left(n^{-1/2}\right) \end{aligned}$$

for some θ^* between $\hat{\theta}$ and θ_0. By assumption 5, $\theta^* \xrightarrow{p} \theta_0$. Hence, by assumptions 2 and 4,

$$\frac{\partial^2 Q_n}{\partial\theta\partial\theta'}(\theta^*) = B(\theta_0) + o_p(1).$$

Thus, by the invertibility of B_0,

$$o_p(1) = B_0^{-1}\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial\theta} + [1+o_p(1)]\sqrt{n}\left(\hat{\theta}-\theta_0\right)$$

By Slutsky's theorem,

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right) = -B_0^{-1}\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial\theta} + o_p(1)$$

and the asymptotic normality follows from assumption 3. □

In assumption 3, the partial derivative of the log-likelihood is called the *score function*, denoted by

$$S(\theta) = \frac{\partial}{\partial\theta}\ln L(\theta).$$

Under the model $y_1,\cdots,y_n \overset{i.i.d.}{\sim} f(y;\theta_0)$, condition 3 in Theorem 2.2 can be expressed as

$$n^{-1/2}S(\theta_0) \xrightarrow{d} N(0,A_0)$$

where $A_0 = n^{-1}E_{\theta_0}\{S(\theta)S(\theta)'\}$. In assumption 4, $B(\theta)$ is essentially the probability limit of the second-order partial derivatives of the log-likelihood. This is the expected Fisher information matrix based on a single observation, defined by

$$\mathcal{I}(\theta_0) = -E_{\theta_0}\left\{\frac{\partial^2}{\partial\theta\partial\theta'}\ln f(y;\theta) \mid \theta=\theta_0\right\}.$$

That is, $B_0 = \mathcal{I}(\theta_0)$. We now summarize the definitions associated with the score function.

Definition 2.4. *1. Score function:*

$$S(\theta) = \frac{\partial}{\partial \theta} \ln L(\theta)$$

2. Fisher information (representing curvature of the log-likelihood function)

$$I(\theta) = -\frac{\partial^2}{\partial \theta \partial \theta'} \ln L(\theta) = -\frac{\partial}{\partial \theta'} S(\theta)$$

3. Observed (Fisher) information: $I(\hat{\theta})$, where $\hat{\theta}$ is the MLE.

4. Expected (Fisher) information: $\mathcal{I}(\theta) = E_\theta \{I(\theta)\}$.

Because of the definition of the MLE, the observed Fisher information is always positive. The expected information is meaningful as a function of θ across the admissible values of θ, but $I(\theta)$ is only meaningful in the neighborhood of $\hat{\theta}$. The observed information applies to a single dataset. In contrast, the expected information is an average quantity over all possible datasets generated at the true value of the parameter. For exponential families of distributions, we have $\mathcal{I}(\hat{\theta}) = I(\hat{\theta})$. In general, $I(\hat{\theta})$ is preferred for variance estimation of $\hat{\theta}$. The use of the observed information in assessing the accuracy of the MLE is advocated by Efron and Hinkley (1978).

Example 2.1. *1. Let $x_1, \cdots, x_n$ be an IID sample from $N(\theta, \sigma^2)$ with σ^2 known. We have*

$$\begin{aligned} V_\theta \{S(\theta)\} &= V_\theta \left\{ \sum_{i=1}^n (x_i - \theta)/\sigma^2 \right\} = n/\sigma^2 \\ I(\theta) &= -\partial S(\theta)/\partial \theta = n/\sigma^2. \end{aligned}$$

In this case, $\mathcal{I}(\theta) = I(\theta)$, a happy coincidence in any exponential family model with canonical parameter θ.

2. Let $x_1, \cdots, x_n$ be an IID sample from Poisson(θ). In this case,

$$\begin{aligned} V_\theta \{S(\theta)\} &= V_\theta \left\{ \sum_{i=1}^n (x_i - \theta)/\theta \right\} = n/\theta \\ I(\theta) &= -\partial S(\theta)/\partial \theta = n\bar{x}/(\theta^2). \end{aligned}$$

Thus, $\mathcal{I}(\theta) \neq I(\theta)$, but $\mathcal{I}(\hat{\theta}) = I(\hat{\theta})$ at $\hat{\theta} = \bar{x}$. This is true for the exponential family. It means that we can estimate the variance of the score function by either $\mathcal{I}(\hat{\theta})$ or $I(\hat{\theta})$.

3. If $x_1, \cdots, x_n$ are an IID sample from Cauchy(θ), then

$$\begin{aligned} \mathcal{I}(\theta) &= n/2 \\ I(\theta) &= -\sum_{i=1}^n \frac{2\{(x_i - \theta)^2 - 1\}}{\{(x_i - \theta)^2 + 1\}^2} \end{aligned}$$

and so $\mathcal{I}(\hat{\theta}) \neq I(\hat{\theta})$.

We now expand what we have found out through the above examples and present two important equalities of the score function. The equality in (2.3) is often called the (second-order) *Bartlett identity*.

Theorem 2.3. *Under regularity conditions allowing the exchange of the order of integration and differentiation,*

$$E_\theta \{S(\theta)\} = 0 \tag{2.2}$$

and

$$V_\theta \{S(\theta)\} = \mathcal{I}(\theta). \tag{2.3}$$

Proof.

$$\begin{aligned} E_\theta\{S(\theta)\} &= E_\theta\left\{\frac{\partial}{\partial\theta}\ln L(\theta)\right\} \\ &= E_\theta\left\{\frac{\partial f(\mathbf{y};\theta)/\partial\theta}{f(\mathbf{y};\theta)}\right\} \\ &= \int\frac{\partial}{\partial\theta}f(\mathbf{y};\theta)\,d\mu(\mathbf{y}). \end{aligned}$$

By assumption,

$$\int\frac{\partial}{\partial\theta}f(\mathbf{y};\theta)\,d\mu(\mathbf{y}) = \frac{\partial}{\partial\theta}\int f(\mathbf{y};\theta)\,d\mu(\mathbf{y}).$$

Since $\int f(\mathbf{y};\theta)\,d\mu(\mathbf{y}) = 1$,

$$\frac{\partial}{\partial\theta}\int f(\mathbf{y};\theta)\,d\mu(\mathbf{y}) = 0$$

and (2.2) is proven.

To prove (2.3), note that since $E_\theta\{S(\theta)\} = 0$, equality (2.3) is equivalent to

$$E_\theta\left\{S(\theta)S(\theta)'\right\} = -E_\theta\left\{\frac{\partial}{\partial\theta'}S(\theta)\right\}. \tag{2.4}$$

To show (2.4), taking the partial derivative of (2.2) with respect to θ, we get

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta'}\int S(\theta;\mathbf{y})f(\mathbf{y};\theta)\,d\mu(\mathbf{y}) \\ &= \int\left\{\frac{\partial}{\partial\theta'}S(\theta;\mathbf{y})\right\}f(\mathbf{y};\theta)\,d\mu(\mathbf{y}) + \int S(\theta;\mathbf{y})\left\{\frac{\partial}{\partial\theta'}f(\mathbf{y};\theta)\right\}d\mu(\mathbf{y}) \\ &= E_\theta\left\{\frac{\partial}{\partial\theta'}S(\theta)\right\} + E_\theta\{S(\theta)S(\theta)'\}, \end{aligned}$$

and we have shown (2.3). □

Equality (2.3) is a special case of the general equality

$$Cov\{g(\mathbf{y};\theta),S(\theta)\} = -E\{\partial g(\mathbf{y};\theta)/\partial\theta'\} \tag{2.5}$$

for any $g(\mathbf{y};\theta)$ such that $E\{g(\mathbf{y};\theta)\} = 0$. Under the model $y_1,\cdots,y_n \overset{i.i.d.}{\sim} f(y;\theta_0)$, Theorem 2.2 states that the limiting distribution of the MLE is

$$\sqrt{n}\left(\hat\theta-\theta_0\right) \xrightarrow{d} N\left(0, n\mathcal{I}(\theta_0)^{-1}A_0\mathcal{I}(\theta_0)^{-1}\right)$$

where $A_0 = E_{\theta_0}\{S(\theta_0)S(\theta_0)'\}$. By Theorem 2.3, $A_0 = \mathcal{I}(\theta_0)$, then the limiting distribution of the MLE is

$$\sqrt{n}\left(\hat\theta-\theta_0\right) \xrightarrow{d} N\left(0, n\mathcal{I}^{-1}(\theta_0)\right).$$

The MLE also satisfies

$$-2\ln\left[\frac{L(\theta_0)}{L(\hat\theta)}\right] \xrightarrow{d} \chi^2_p$$

which can be used to develop likelihood-ratio (LR) confidence intervals for θ_0. The level α LR confidence intervals (CI) are constructed by

$$\left\{\theta; L(\theta) > k_\alpha \times L(\hat\theta)\right\}$$

for some k_α which is the upper α quantile of the chi-square distribution with p degrees of freedom. The LR confidence interval is more attractive than the Wald confidence interval in two aspects: (i) A Wald CI can often produce interval estimates beyond the parameter space. (ii) A LR interval is invariant with respect to parameter transformation. For example, if (θ_L,θ_U) is the 95% CI for θ, then $(g(\theta_L),g(\theta_U))$ is the 95% CI for a monotone increasing function $g(\theta)$.