Statistical Methods for Handling Incomplete Data
Chapter 2: Likelihood-based approach

Jae-Kwang Kim

Department of Statistics, Iowa State University

Outline

1. Introduction
2. Observed likelihood
3. Mean Score Approach
4. Observed Information

1. Introduction - Basic Setup (No missing data)

- $\mathbf{y} = (y_1, y_2, \cdots y_n)$ is a realization of the random sample from an infinite population with density $f(y)$.
- 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 exist $\theta_0 \in \Omega$ such that $f(y;\theta_0) = f(y)$ for all y.

Likelihood

Definitions for likelihood theory

- The likelihood function of θ is defined as

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

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

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

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

- A parametric family of densities, $\mathcal{P} = \{f(y;\theta);\theta\in\Theta\}$, is identifiable if for all y,

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

Lemma 2.1 Properties of identifiable distribution

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

$$M(\theta) = E_{\theta_0} \ln\left[\frac{f(y;\theta)}{f(y;\theta_0)}\right]$$

has a unique maximum at θ_0.

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

$$-\ln[E(Z)] < E[-\ln(Z)].$$

Because $E_{\theta_0}(Z) = \int f(y;\theta)\,dy = 1$, we have $\ln[E(Z)] = 0$.

Remark

1. 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 the two distributions.
2. Lemma 2.1 simply says that, under identifiability, $Q(\theta) = E_{\theta_0}\{\log f(Y;\theta)\}$ takes the (unique) maximum value at $\theta = \theta_0$.
3. Define

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

then the MLE $\hat{\theta}$ is the maximizer of $Q_n(\theta)$. Since $Q_n(\theta)$ converges in probability to $Q(\theta)$, can we say that the maximizer of $Q_n(\theta)$ converges to the maximizer of $Q(\theta)$?

Theorem 2.1: (Weak) consistency of MLE

Theorem 2.1

Let

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

and $\hat{\theta}$ be any solution of

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

Assume the following two conditions:

1. Uniform weak convergence:

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

   for some non-stochastic function $Q(\theta)$

2. Identification: $Q(\theta)$ is uniquely maximized at θ_0.

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

Remark

1. Convergence in probability:

$$\hat{\theta} \xrightarrow{p} \theta_0 \iff P\left\{|\hat{\theta} - \theta_0| > \epsilon\right\} \to 0 \text{ as } n \to \infty,$$

   for any $\epsilon > 0$.

2. If $\mathcal{P} = \{f(y;\theta); \theta \in \Omega\}$ is not identifiable, then $Q(\theta)$ may not have a unique maximum and $\hat{\theta}$ may not converge (in probability) to a single point.

3. If the true distribution $f(y)$ does not belong to the class $\mathcal{P} = \{f(y;\theta); \theta \in \Omega\}$, which point does $\hat{\theta}$ converge to?

Example: log-likelihood function for $N(\theta, 1)$

Log-likelihood

$$l_n(\theta) = \sum_{i=1}^{n} \log f(y_i; \theta)$$

Under assumption $N(\theta, 1)$:

Plot of $l_n(\theta)$ on θ:

Other properties of MLE

- Asymptotic normality (Theorem 2.2)
- Asymptotic optimality: MLE achieves Cramer-Rao lower bound
- Wilks' theorem:
$$2\{l_n(\hat{\theta}) - l_n(\theta_0)\} \xrightarrow{d} \chi_p^2$$

1 Introduction - Fisher information

Definition

1. Score function:
$$S(\theta) = \frac{\partial \log L(\theta)}{\partial \theta}$$
2. Fisher information = curvature of the log-likelihood:
$$\mathrm{I}(\theta) = -\frac{\partial^2}{\partial\theta\partial\theta^T} \log L(\theta) = -\frac{\partial}{\partial\theta^T} S(\theta)$$
3. Observed (Fisher) information: $\mathrm{I}(\hat{\theta}_n)$ where $\hat{\theta}_n$ is the MLE.
4. Expected (Fisher) information: $\mathcal{I}(\theta) = \mathrm{E}_\theta \{\mathrm{I}(\theta)\}$

- The observed information is always positive. The observed information applies to a single dataset.
- The expected information is meaningful as a function of θ across the admissible values of θ. The expected information is an average quantity over all possible datasets.
- $\mathcal{I}(\hat{\theta}) = \mathrm{I}(\hat{\theta})$ for exponential family.

Properties of score function

Theorem 2.3. Properties of score functions

Under the regularity conditions allowing the exchange of the order of integration and differentiation,

$$\mathrm{E}_{\theta}\left\{S(\theta)\right\} = 0 \quad \text{and} \quad \mathrm{V}_{\theta}\left\{S(\theta)\right\} = \mathcal{I}(\theta).$$

Proof

Remark

- Since $\hat{\theta} \xrightarrow{p} \theta_0$, we can apply a Taylor linearization on $S(\hat{\theta}) = 0$ to get

$$\hat{\theta} - \theta_0 \cong \{\mathcal{I}(\theta_0)\}^{-1} S(\theta_0).$$

  Here, we use the fact that $I(\theta) = -\partial S(\theta)/\partial \theta^T$ converges in probability to $\mathcal{I}(\theta)$.

- Thus, the (asymptotic) variance of MLE is

$$\begin{aligned} V(\hat{\theta}) &\doteq \{\mathcal{I}(\theta_0)\}^{-1} V\{S(\theta_0)\} \{\mathcal{I}(\theta_0)\}^{-1} \\ &= \{\mathcal{I}(\theta_0)\}^{-1}, \end{aligned}$$

  where the last equality follows from Theorem 2.3.

2. Observed likelihood

Basic Setup

1. Let $\mathbf{y} = (y_1, y_2, \cdots y_n)$ be a realization of the random sample from an infinite population with density $f(y; \theta_0)$.

2. Let δ_i be an indicator function defined by

$$\delta_i = \begin{cases} 1 & \text{if } y_i \text{ is observed} \\ 0 & \text{otherwise} \end{cases}$$

   and we assume that

$$Pr(\delta_i = 1 \mid y_i) = \pi(y_i, \phi)$$

   for some (unknown) parameter ϕ and $\pi(\cdot)$ is a known function.

3. Thus, instead of observing (δ_i, y_i) directly, we observe $(\delta_i, y_{i,obs})$

$$y_{i,obs} = \begin{cases} y_i & \text{if } \delta_i = 1 \\ * & \text{if } \delta_i = 0. \end{cases}$$

4. What is the (marginal) density of $(y_{i,obs}, \delta_i)$?

Motivation (Change of variable technique)

1. Suppose that z is a random variable with density $f(z)$.
2. Instead of observing z directly, we observe only $y = y(z)$, where the mapping $z \to y(z)$ is known.
3. The density of y is

$$g(y) = \int_{\mathcal{R}(y)} f(z)\, dz$$

where $\mathcal{R}(y) = \{z; y(z) = y\}$.

Derivation

1. The original distribution for $z = (y, \delta)$ is

$$f(y, \delta) = f_1(y)\, f_2(\delta \mid y)$$

where $f_1(y)$ is the density of y and $f_2(\delta \mid y)$ is the conditional density of δ conditional on y and is given by $f_2(\delta \mid y) = \{\pi(y)\}^{\delta} \{1 - \pi(y)\}^{1-\delta}$, where $\pi(y) = Pr(\delta = 1 \mid y)$.

2. Instead of observing $z = (y, \delta)$ directly, we observe only (y_{obs}, δ) where $y_{\text{obs}} = y_{\text{obs}}(y, \delta)$ and the mapping $(y, \delta) \to y_{\text{obs}}$ is known.

3. The (marginal) density of (y_{obs}, δ) is

$$g(y_{\text{obs}}, \delta) = \int_{\mathcal{R}(y_{\text{obs}}, \delta)} f(y, \delta)\, dy$$

where $\mathcal{R}(y_{\text{obs}}, \delta) = \{y; y_{\text{obs}}(y, \delta) = y_{\text{obs}}\}$.

Likelihood under missing data (=Observed likelihood)

The observed likelihood is the likelihood obtained from the marginal density of $(y_{i,\mathrm{obs}}, \delta_i)$, $i = 1, 2, \cdots, n$, and can be written as, under the IID setup,

$$\begin{aligned} L_{\mathrm{obs}}(\theta, \phi) &= \prod_{\delta_i=1} \left[f_1(y_i;\theta) f_2(\delta_i \mid y_i;\phi) \right] \times \prod_{\delta_i=0} \left[\int f_1(y_i;\theta) f_2(\delta_i \mid y_i;\phi)\, dy_i \right] \\ &= \prod_{\delta_i=1} \left[f_1(y_i;\theta) \pi(y_i;\phi) \right] \times \prod_{\delta_i=0} \left[\int f_1(y;\theta) \{1 - \pi(y;\phi)\}\, dy \right], \end{aligned}$$

where $\pi(y;\phi) = P(\delta = 1 \mid y;\phi)$.

Example 2.2 (Censored regression model, or Tobit model)

$$z_i = x_i'\beta + \epsilon_i \qquad \epsilon_i \sim N\left(0, \sigma^2\right)$$

$$y_i = \begin{cases} z_i & \text{if } z_i \geq 0 \\ 0 & \text{if } z_i < 0. \end{cases}$$

The observed log-likelihood is

$$l\left(\beta, \sigma^2\right) = -\frac{1}{2}\sum_{y_i>0}\left[\ln 2\pi + \ln \sigma^2 + \frac{\left(y_i - x_i'\beta\right)^2}{\sigma^2}\right] + \sum_{y_i=0}\ln\left[1 - \Phi\left(\frac{x_i'\beta}{\sigma}\right)\right]$$

where $\Phi\left(x\right)$ is the cdf of the standard normal distribution.

Example 2.3

Let $t_1, t_2, \cdots, t_n$ be an IID sample from a distribution with density $f_\theta(t) = \theta e^{-\theta t} I(t > 0)$. Instead of observing t_i, we observe (y_i, δ_i) where

$$y_i = \begin{cases} t_i & \text{if } \delta_i = 1 \\ c & \text{if } \delta_i = 0 \end{cases}$$

and

$$\delta_i = \begin{cases} 1 & \text{if } t_i \le c \\ 0 & \text{if } t_i > c, \end{cases}$$

where c is a known censoring time. The observed likelihood for θ can be derived as

$$\begin{aligned} L_{obs}(\theta) &= \prod_{i=1}^{n} \left[\{f_\theta(t_i)\}^{\delta_i} \{P(t_i > c)\}^{1-\delta_i} \right] \\ &= \theta^{\sum_{i=1}^{n} \delta_i} \exp(-\theta \sum_{i=1}^{n} y_i). \end{aligned}$$

Multivariate extension

Basic Setup

- Let $\mathbf{y} = (y_1, \ldots, y_p)$ be a p-dimensional random vector with probability density function $f(\mathbf{y}; \theta)$.
- Let δ_{ij} be the response indicator function of y_{ij} with $\delta_{ij} = \begin{cases} 1 & \text{if } y_{ij} \text{ is observed} \\ 0 & \text{otherwise.} \end{cases}$
- $\boldsymbol{\delta}_i = (\delta_{i1}, \cdots, \delta_{ip})$: p-dimensional random vector with density $\mathrm{P}(\boldsymbol{\delta} \mid \mathbf{y})$ assuming $\mathrm{P}(\boldsymbol{\delta}|\mathbf{y}) = \mathrm{P}(\boldsymbol{\delta}|\mathbf{y}; \phi)$ for some ϕ.
- Let $(\mathbf{y}_{i,obs}, \mathbf{y}_{i,mis})$ be the observed part and missing part of $\mathbf{y}_i$, respectively.
- Let $\mathcal{R}(\mathbf{y}_{obs}, \boldsymbol{\delta}) = \{\mathbf{y}; \mathbf{y}_{obs}(\mathbf{y}_i, \boldsymbol{\delta}_i) = \mathbf{y}_{i,obs}, i = 1, \ldots, n\}$ be the set of all possible values of $\mathbf{y}$ with the same realized value of $\mathbf{y}_{obs}$, for given $\boldsymbol{\delta}$, where $\mathbf{y}_{obs}(\mathbf{y}_i, \boldsymbol{\delta}_i)$ is a function that gives the value of y_{ij} for $\delta_{ij} = 1$.

Definition: Observed likelihood

Under the above setup, the observed likelihood of (θ, ϕ) is

$$L_{obs}(\theta, \phi) = \int_{\mathcal{R}(\mathbf{y}_{obs}, \boldsymbol{\delta})} f(\mathbf{y}; \theta) \mathrm{P}(\boldsymbol{\delta} | \mathbf{y}; \phi) d\mathbf{y}.$$

Under IID setup: The observed likelihood is

$$L_{obs}(\theta, \phi) = \prod_{i=1}^{n} \left[\int f(\mathbf{y}_i; \theta) \mathrm{P}(\boldsymbol{\delta}_i | \mathbf{y}_i; \phi) d\mathbf{y}_{i,mis} \right],$$

where it is understood that, if $\mathbf{y}_i = \mathbf{y}_{i,obs}$ and $\mathbf{y}_{i,mis}$ is empty then there is nothing to integrate out.

- In the special case of scalar y, the observed likelihood is

$$L_{obs}(\theta, \phi) = \prod_{\delta_i=1} [f(y_i; \theta)\pi(y_i; \phi)] \times \prod_{\delta_i=0} \left[\int f(y; \theta) \{1 - \pi(y; \phi)\} dy \right],$$

where $\pi(y; \phi) = \mathrm{P}(\delta = 1 | y; \phi)$.

Missing At Random

Definition: Missing At Random (MAR)

$\mathrm{P}(\boldsymbol{\delta}|\mathbf{y})$ is the density of the conditional distribution of $\boldsymbol{\delta}$ given $\mathbf{y}$. Let $\mathbf{y}_{obs} = \mathbf{y}_{obs}(\mathbf{y}, \boldsymbol{\delta})$ where

$$y_{i,obs} = \begin{cases} y_i & \text{if } \delta_i = 1 \\ * & \text{if } \delta_i = 0. \end{cases}$$

The response mechanism is MAR if $\mathrm{P}\left(\boldsymbol{\delta}|\mathbf{y}_1\right) = \mathrm{P}\left(\boldsymbol{\delta}|\mathbf{y}_2\right)$ $\{$ or $\mathrm{P}(\boldsymbol{\delta}|\mathbf{y}) = \mathrm{P}(\boldsymbol{\delta}|\mathbf{y}_{obs})\}$ for all $\mathbf{y}_1$ and $\mathbf{y}_2$ satisfying $\mathbf{y}_{obs}(\mathbf{y}_1, \boldsymbol{\delta}) = \mathbf{y}_{obs}(\mathbf{y}_2, \boldsymbol{\delta})$.

- MAR: the response mechanism $\mathrm{P}(\boldsymbol{\delta}|\mathbf{y})$ depends on $\mathbf{y}$ only through $\mathbf{y}_{obs}$.
- Let $\mathbf{y} = (\mathbf{y}_{obs}, \mathbf{y}_{mis})$. By Bayes theorem,

$$\mathrm{P}\left(\mathbf{y}_{mis}|\mathbf{y}_{obs}, \boldsymbol{\delta}\right) = \frac{\mathrm{P}(\boldsymbol{\delta}|\mathbf{y}_{mis}, \mathbf{y}_{obs})}{\mathrm{P}(\boldsymbol{\delta}|\mathbf{y}_{obs})} \mathrm{P}\left(\mathbf{y}_{mis}|\mathbf{y}_{obs}\right).$$

- MAR: $\mathrm{P}(\mathbf{y}_{mis}|\mathbf{y}_{obs}, \boldsymbol{\delta}) = \mathrm{P}(\mathbf{y}_{mis}|\mathbf{y}_{obs})$. That is, $\mathbf{y}_{mis} \perp \boldsymbol{\delta} \mid \mathbf{y}_{obs}$.
- MAR: the conditional independence of $\boldsymbol{\delta}$ and $\mathbf{y}_{mis}$ given $\mathbf{y}_{obs}$.

Remark

- MCAR (Missing Completely at random): $P(\boldsymbol{\delta} \mid \mathbf{y})$ does not depend on $\mathbf{y}$.
- MAR (Missing at random): $P(\boldsymbol{\delta} \mid \mathbf{y}) = P(\boldsymbol{\delta} \mid \mathbf{y}_{obs})$
- NMAR (Not Missing at random): $P(\boldsymbol{\delta} \mid \mathbf{y}) \neq P(\boldsymbol{\delta} \mid \mathbf{y}_{obs})$
- Thus, MCAR is a special case of MAR.

Likelihood factorization theorem

Theorem 2.4 (Rubin, 1976)

$\mathrm{P}_{\phi}(\boldsymbol{\delta}|\mathbf{y})$ is the joint density of $\boldsymbol{\delta}$ given $\mathbf{y}$ and $f_{\theta}(\mathbf{y})$ is the joint density of $\mathbf{y}$. Under conditions

1. the parameters θ and ϕ are distinct and
2. MAR condition holds,

the observed likelihood can be written as

$$L_{obs}(\theta, \phi) = L_1(\theta) L_2(\phi),$$

and the MLE of θ can be obtained by maximizing $L_1(\theta)$.

Thus, we do not have to specify the model for response mechanism. The response mechanism is called ignorable if the above likelihood factorization holds.

Example 2.4

- Bivariate data (x_i, y_i) with pdf $f(x, y) = f_1(y \mid x) f_2(x)$.
- x_i is always observed and y_i is subject to missingness.
- Assume that the response status variable δ_i of y_i satisfies

$$P(\delta_i = 1 \mid x_i, y_i) = \Lambda_1(\phi_0 + \phi_1 x_i + \phi_2 y_i)$$

  for some function $\Lambda_1(\cdot)$ of known form.
- Let θ be the parameter of interest in the regression model $f_1(y \mid x; \theta)$. Let α be the parameter in the marginal distribution of x, denoted by $f_2(x_i; \alpha)$. Define $\Lambda_0(x) = 1 - \Lambda_1(x)$.
- Three parameters
  - θ: parameter of interest
  - α and ϕ: nuisance parameter

Example 2.4 (Cont'd)

- Observed likelihood

$$
\begin{aligned}
L_{obs}(\theta, \alpha, \phi) &= \left[\prod_{\delta_i = 1} f_1(y_i \mid x_i; \theta) f_2(x_i; \alpha) \Lambda_1(\phi_0 + \phi_1 x_i + \phi_2 y_i) \right] \\
&\times \left[\prod_{\delta_i = 0} \int f_1(y \mid x_i; \theta) f_2(x_i; \alpha) \Lambda_0(\phi_0 + \phi_1 x_i + \phi_2 y)\, dy \right] \\
&= L_1(\theta, \phi) \times L_2(\alpha)
\end{aligned}
$$

  where $L_2(\alpha) = \prod_{i=1}^{n} f_2(x_i; \alpha)$.

- Thus, we can safely ignore the marginal distribution of x if x is completely observed.

Example 2.4 (Cont'd)

- If $\phi_2 = 0$, then MAR holds and

  $$L_1(\theta, \phi) = L_{1a}(\theta) \times L_{1b}(\phi)$$

  where

  $$L_{1a}(\theta) = \prod_{\delta_i=1} f_1(y_i \mid x_i; \theta)$$

  and

  $$L_{1b}(\phi) = \prod_{\delta_i=1} \Lambda_1(\phi_0 + \phi_1 x_i) \times \prod_{\delta_i=0} \Lambda_0(\phi_0 + \phi_1 x_i).$$

- Thus, under MAR, the MLE of θ can be obtained by maximizing $L_{1a}(\theta)$, which is obtained by ignoring the missing part of the data.

Example 2.4 (Cont'd)

- Instead of y_i subject to missingness, if x_i is subject to missingness, then the observed likelihood becomes

$$
\begin{aligned}
L_{obs}(\theta, \phi, \alpha) &= \left[\prod_{\delta_i = 1} f_1(y_i \mid x_i; \theta) f_2(x_i; \alpha) \Lambda_1(\phi_0 + \phi_1 x_i + \phi_2 y_i) \right] \\
&\times \left[\prod_{\delta_i = 0} \int f_1(y_i \mid x; \theta) f_2(x; \alpha) \Lambda_0(\phi_0 + \phi_1 x + \phi_2 y_i) \, dx \right] \\
&\neq L_1(\theta, \phi) \times L_2(\alpha) .
\end{aligned}
$$

- If $\phi_1 = 0$ then

$$L_{obs}(\theta, \alpha, \phi) = L_1(\theta, \alpha) \times L_2(\phi)$$

  and MAR holds. Although we are not interested in the marginal distribution of x, we still need to specify the model for the marginal distribution of x.

3. Mean Score Approach

3 Mean Score Approach

- The observed likelihood is the marginal density of $(\mathbf{y}_{obs}, \boldsymbol{\delta})$.
- The observed likelihood is

$$L_{obs}(\eta) = \int_{\mathcal{R}(\mathbf{y}_{obs}, \boldsymbol{\delta})} f(\mathbf{y}; \theta) \mathrm{P}(\boldsymbol{\delta} | \mathbf{y}; \phi) d\mu(\mathbf{y}) = \int f(\mathbf{y}; \theta) \mathrm{P}(\boldsymbol{\delta} | \mathbf{y}; \phi) d\mu(\mathbf{y}_{mis})$$

  where $\mathbf{y}_{mis}$ is the missing part of $\mathbf{y}$ and $\eta = (\theta, \phi)$.

- Observed score equation:

$$S_{obs}(\eta) \equiv \frac{\partial}{\partial \eta} \log L_{obs}(\eta) = 0$$

- Computing the observed score function can be computationally challenging because the observed likelihood is an integral form.

Theorem 2.5: Mean Score Theorem (Fisher, 1922)

Under some regularity conditions, the observed score function equals to the mean score function. That is,

$$S_{obs}(\eta) = \bar{S}(\eta)$$

where

$$\begin{aligned} \bar{S}(\eta) &= \mathrm{E}\{S_{com}(\eta) | \mathbf{y}_{obs}, \boldsymbol{\delta}\} \\ S_{com}(\eta) &= \frac{\partial}{\partial \eta} \log f(\mathbf{y}, \boldsymbol{\delta}; \eta), \\ f(\mathbf{y}, \boldsymbol{\delta}; \eta) &= f(\mathbf{y}; \theta) \mathrm{P}(\boldsymbol{\delta} | \mathbf{y}; \phi). \end{aligned}$$

- The mean score function is computed by taking the conditional expectation of the complete-sample score function given the observation.
- The mean score function is easier to compute than the observed score function.

Proof of Theorem 2.5

Since $L_{\rm obs}(\eta) = f(\mathbf{y}, \boldsymbol{\delta}; \eta) / f(\mathbf{y}, \boldsymbol{\delta} \mid \mathbf{y}_{\rm obs}, \boldsymbol{\delta}; \eta)$, we have

$$\frac{\partial}{\partial \eta} \ln L_{\rm obs}(\eta) = \frac{\partial}{\partial \eta} \ln f(\mathbf{y}, \boldsymbol{\delta}; \eta) - \frac{\partial}{\partial \eta} \ln f(\mathbf{y}, \boldsymbol{\delta} \mid \mathbf{y}_{\rm obs}, \boldsymbol{\delta}; \eta),$$

taking a conditional expectation of the above equation over the conditional distribution of $(\mathbf{y}, \boldsymbol{\delta})$ given $(\mathbf{y}_{\rm obs}, \boldsymbol{\delta})$, we have

$$\begin{aligned} \frac{\partial}{\partial \eta} \ln L_{\rm obs}(\eta) &= E\left\{ \frac{\partial}{\partial \eta} \ln L_{\rm obs}(\eta) \mid \mathbf{y}_{\rm obs}, \boldsymbol{\delta} \right\} \\ &= E\left\{ S_{com}(\eta) \mid \mathbf{y}_{\rm obs}, \boldsymbol{\delta} \right\} - E\left\{ \frac{\partial}{\partial \eta} \ln f(\mathbf{y}, \boldsymbol{\delta} \mid \mathbf{y}_{\rm obs}, \boldsymbol{\delta}; \eta) \mid \mathbf{y}_{\rm obs}, \boldsymbol{\delta} \right\}. \end{aligned}$$

Here, the first equality holds because $L_{\rm obs}(\eta)$ is a function of $(\mathbf{y}_{\rm obs}, \boldsymbol{\delta})$ only. The last term is equal to zero by Theorem 2.3, which states that the expected value of the score function is zero and the reference distribution in this case is the conditional distribution of $(\mathbf{y}, \boldsymbol{\delta})$ given $(\mathbf{y}_{\rm obs}, \boldsymbol{\delta})$.

Example 2.6

1. Suppose that the study variable y follows from a normal distribution with mean $\mathbf{x}'\beta$ and variance σ^2. The score equations for $\boldsymbol{\beta}$ and σ^2 under complete response are

$$S_1(\boldsymbol{\beta}, \sigma^2) = \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right) \mathbf{x}_i / \sigma^2 = \mathbf{0}$$

and

$$S_2(\boldsymbol{\beta}, \sigma^2) = -n/(2\sigma^2) + \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right)^2 / (2\sigma^4) = 0.$$

2. Assume that y_i are observed only for the first r elements and the MAR assumption holds. In this case, the mean score function reduces to

$$\bar{S}_1(\boldsymbol{\beta}, \sigma^2) = \sum_{i=1}^{r} \left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right) \mathbf{x}_i / \sigma^2$$

and

$$\bar{S}_2(\boldsymbol{\beta}, \sigma^2) = -n/(2\sigma^2) + \sum_{i=1}^{r} \left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right)^2 / (2\sigma^4) + (n - r)/(2\sigma^2).$$

Example 2.6 (Cont'd)

3. The maximum likelihood estimator obtained by solving the mean score equations is

$$\hat{\beta} = \left(\sum_{i=1}^{r} \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \sum_{i=1}^{r} \mathbf{x}_i y_i$$

and

$$\hat{\sigma}^2 = \frac{1}{r} \sum_{i=1}^{r} \left(y_i - \mathbf{x}_i' \hat{\beta} \right)^2 .$$

Thus, the resulting estimators can be also obtained by simply ignoring the missing part of the sample, which is consistent with the result in Example 2.4 (for $\phi_2 = 0$).

Discussion of Example 2.6

- We are interested in estimating θ for the conditional density $f(y \mid x; \theta)$.
- Under MAR, the observed likelihood for θ is

$$L_{obs}(\theta) = \prod_{i=1}^{r} f(y_i \mid x_i; \theta) \times \prod_{i=r+1}^{n} \int f(y \mid x_i; \theta) dy = \prod_{i=1}^{r} f(y_i \mid x_i; \theta).$$

- The same conclusion can follow from the mean score theorem. Under MAR, the mean score function is

$$\begin{aligned} \bar{S}(\theta) &= \sum_{i=1}^{r} S(\theta; x_i, y_i) + \sum_{i=r+1}^{n} E\{S(\theta; x_i, Y) \mid x_i\} \\ &= \sum_{i=1}^{r} S(\theta; x_i, y_i) \end{aligned}$$

  where $S(\theta; x, y)$ is the score function for θ and the second equality follows from Theorem 2.3.

Example 2.5

1. Suppose that the study variable y is randomly distributed with Bernoulli distribution with probability of success p_i, where

$$p_i = p_i(\beta) = \frac{\exp(\mathbf{x}_i'\beta)}{1 + \exp(\mathbf{x}_i'\beta)}$$

   for some unknown parameter β and $\mathbf{x}_i$ is a vector of the covariates in the logistic regression model for y_i. We assume that 1 is in the column space of $\mathbf{x}_i$.

2. Under complete response, the score function for β is

$$S_1(\beta) = \sum_{i=1}^{n} (y_i - p_i(\beta))\,\mathbf{x}_i.$$

Example 2.5 (Cont'd)

3. Let δ_i be the response indicator function for y_i with distribution *Bernoulli*(π_i) where
$$\pi_i = \frac{\exp\left(\mathbf{x}_i'\phi_0 + y_i\phi_1\right)}{1 + \exp\left(\mathbf{x}_i'\phi_0 + y_i\phi_1\right)}.$$
We assume that x_i is always observed, but y_i is missing if $\delta_i = 0$.

4. Under missing data, the mean score function for β is
$$\bar{S}_1(\beta, \phi) = \sum_{\delta_i=1} \{y_i - p_i(\beta)\} \mathbf{x}_i + \sum_{\delta_i=0} \sum_{y=0}^{1} w_i(y; \beta, \phi) \{y - p_i(\beta)\} \mathbf{x}_i,$$
where $w_i(y; \beta, \phi)$ is the conditional probability of $y_i = y$ given x_i and $\delta_i = 0$:
$$w_i(y; \beta, \phi) = \frac{P_\beta(y_i = y \mid x_i) P_\phi(\delta_i = 0 \mid y_i = y, x_i)}{\sum_{z=0}^{1} P_\beta(y_i = z \mid x_i) P_\phi(\delta_i = 0 \mid y_i = z, x_i)}$$
Thus, $\bar{S}_1(\beta, \phi)$ is also a function of ϕ.

Example 2.5 (Cont'd)

5. If the response mechanism is MAR so that $\phi_1 = 0$, then

$$w_i\left(y;\beta,\phi\right) \quad = \quad \frac{P_\beta\left(y_i = y \mid x_i\right)}{\sum_{z=0}^{1} P_\beta\left(y_i = z \mid x_i\right)} = P_\beta\left(y_i = y \mid x_i\right)$$

   and so

$$\bar{S}_1\left(\beta,\phi\right) = \sum_{\delta_i=1}\left\{y_i - p_i\left(\beta\right)\right\}x_i = \bar{S}_1\left(\beta\right).$$

6. If MAR does not hold, then $(\hat{\beta},\hat{\phi})$ can be obtained by solving $\bar{S}_1\left(\beta,\phi\right) = 0$ and $\bar{S}_2(\beta,\phi) = 0$ jointly, where

$$\begin{aligned}\bar{S}_2\left(\beta,\phi\right) \quad &= \quad \sum_{\delta_i=1}\left\{\delta_i - \pi\left(\phi;x_i,y_i\right)\right\}\left(\mathbf{x}_i,y_i\right)\\ &\quad + \sum_{\delta_i=0}\sum_{y=0}^{1} w_i\left(y;\beta,\phi\right)\left\{\delta_i - \pi_i\left(\phi;\mathbf{x}_i,y\right)\right\}\left(\mathbf{x}_i,y\right).\end{aligned}$$

Discussion of Example 2.5

- We may not have a unique solution to $\bar{S}(\eta)=0$, where $\bar{S}(\eta)=\left[\bar{S}_1(\beta,\phi),\bar{S}_2(\beta,\phi)\right]$ when MAR does not hold, because of the non-identifiability problem associated with non-ignorable missing.
- To avoid this problem, often a reduced model is used for the response model.

  $$Pr(\delta=1 \mid \mathbf{x},y)=Pr(\delta=1 \mid \mathbf{u},y)$$

  where $\mathbf{x}=(\mathbf{u},\mathbf{z})$. The reduced response model introduces a smaller set of parameters and the over-identified situation can be resolved. (More discussion will be made in Chapter 6.)
- Computing the solution to $\bar{S}(\eta)$ is also difficult. EM algorithm, which will be discussed in Chapter 3, is a useful computational tool.

2.4 Observed information

Definition

1. Observed score function: $S_{obs}(\eta) = \frac{\partial}{\partial \eta} \log L_{obs}(\eta)$
2. Fisher information from observed likelihood: $\mathrm{I}_{obs}(\eta) = -\frac{\partial^2}{\partial \eta \partial \eta^T} \log L_{obs}(\eta)$
3. Expected (Fisher) information from observed likelihood: $\mathcal{I}_{obs}(\eta) = \mathrm{E}_\eta \left\{ \mathrm{I}_{obs}(\eta) \right\}$.

Theorem 2.6

Under regularity conditions,

$$\mathrm{E}\{S_{obs}(\eta)\} = 0, \quad \text{and} \quad \mathrm{V}\{S_{obs}(\eta)\} = \mathcal{I}_{obs}(\eta),$$

where $\mathcal{I}_{obs}(\eta) = \mathrm{E}_\eta\{\mathrm{I}_{obs}(\eta)\}$ is the expected information from the observed likelihood.

4 Observed information

- Under missing data, the MLE $\hat{\eta}$ is the solution to $\bar{S}(\eta) = 0$.
- Under some regularity conditions, $\hat{\eta}$ converges in probability to η_0 and has the asymptotic variance $\{\mathcal{I}_{obs}(\eta_0)\}^{-1}$ with

$$\mathcal{I}_{obs}(\eta) = \mathrm{E}\left\{-\frac{\partial}{\partial \eta^T} S_{obs}(\eta)\right\} = \mathrm{E}\left\{S_{obs}^{\otimes 2}(\eta)\right\} = \mathrm{E}\left\{\bar{S}^{\otimes 2}(\eta)\right\} \quad \& \quad B^{\otimes 2} = BB^T.$$

- For variance estimation of $\hat{\eta}$, may use $\{\mathcal{I}_{obs}(\hat{\eta})\}^{-1}$.
- In general, $\mathrm{I}_{obs}(\hat{\eta})$ is preferred to $\mathcal{I}_{obs}(\hat{\eta})$ for variance estimation of $\hat{\eta}$.

Return to Example 2.3

- Observed log-likelihood

$$\ln L_{obs}(\theta) = \sum_{i=1}^{n} \delta_i \log(\theta) - \theta \sum_{i=1}^{n} y_i$$

- MLE for θ:

$$\hat{\theta} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} \delta_i}$$

- Fisher information: $I_{obs}(\theta) = \sum_{i=1}^{n} \delta_i / \theta^2$
- Expected information: $\mathcal{I}_{obs}(\theta) = \sum_{i=1}^{n} (1 - e^{-\theta c}) / \theta^2 = n(1 - e^{-\theta c}) / \theta^2$.

Which one do you prefer ?

Motivation

- $L_{com}(\eta) = f(\mathbf{y}, \boldsymbol{\delta}; \eta)$: complete-sample likelihood with no missing data
- Fisher information associated with $L_{com}(\eta)$:

$$\mathrm{I}_{com}(\eta) = -\frac{\partial}{\partial \eta^T} S_{com}(\eta) = -\frac{\partial^2}{\partial \eta \partial \eta^T} \log L_{com}(\eta)$$

- $L_{obs}(\eta)$: the observed likelihood
- Fisher information associated with $L_{obs}(\eta)$:

$$\mathrm{I}_{obs}(\eta) = -\frac{\partial}{\partial \eta^T} S_{obs}(\eta) = -\frac{\partial^2}{\partial \eta \partial \eta^T} \log L_{obs}(\eta)$$

- How to express $\mathrm{I}_{obs}(\eta)$ in terms of $\mathrm{I}_{com}(\eta)$ and $S_{com}(\eta)$?

Louis' formula

Theorem 2.7 (Louis, 1982; Oakes, 1999)

Under regularity conditions allowing the exchange of the order of integration and differentiation,

$$\begin{aligned} \mathrm{I}_{obs}(\eta) &= \mathrm{E}\{I_{com}(\eta)|\mathbf{y}_{obs}, \boldsymbol{\delta}\} - \left[\mathrm{E}\{S_{com}^{\otimes 2}(\eta)|\mathbf{y}_{obs}, \boldsymbol{\delta}\} - \bar{S}(\eta)^{\otimes 2}\right] \\ &= \mathrm{E}\{I_{com}(\eta)|\mathbf{y}_{obs}, \boldsymbol{\delta}\} - V\{S_{com}(\eta)|\mathbf{y}_{obs}, \boldsymbol{\delta}\}, \end{aligned}$$

where $\bar{S}(\eta) = \mathrm{E}\{S_{com}(\eta)|\mathbf{y}_{obs}, \boldsymbol{\delta}\}$.

Proof of Theorem 2.7

By Theorem 2.5, the observed information associated with $L_{\text{obs}}(\eta)$ can be expressed as

$$I_{\text{obs}}(\eta) = -\frac{\partial}{\partial \eta'} \bar{S}(\eta)$$

where $\bar{S}(\eta) = E\left\{S_{\text{com}}(\eta) \mid \mathbf{y}_{\text{obs}}, \boldsymbol{\delta}; \eta\right\}$. Thus, we have

$$\begin{aligned}
\frac{\partial}{\partial \eta'} \bar{S}(\eta) &= \frac{\partial}{\partial \eta'} \int S_{\text{com}}(\eta; \mathbf{y}) f(\mathbf{y}, \boldsymbol{\delta} \mid \mathbf{y}_{\text{obs}}, \boldsymbol{\delta}; \eta)\, d\mu(\mathbf{y}) \\
&= \int \left\{ \frac{\partial}{\partial \eta'} S_{\text{com}}(\eta; \mathbf{y}) \right\} f(\mathbf{y}, \boldsymbol{\delta} \mid \mathbf{y}_{\text{obs}}, \boldsymbol{\delta}; \eta)\, d\mu(\mathbf{y}) \\
&\quad + \int S_{\text{com}}(\eta; \mathbf{y}) \left\{ \frac{\partial}{\partial \eta'} f(\mathbf{y}, \boldsymbol{\delta} \mid \mathbf{y}_{\text{obs}}, \boldsymbol{\delta}; \eta) \right\} d\mu(\mathbf{y}) \\
&= E\left\{ \partial S_{\text{com}}(\eta) / \partial \eta' \mid \mathbf{y}_{\text{obs}}, \boldsymbol{\delta} \right\} \\
&\quad + \int S_{\text{com}}(\eta; \mathbf{y}) \left\{ \frac{\partial}{\partial \eta'} \log f(\mathbf{y}, \boldsymbol{\delta} \mid \mathbf{y}_{\text{obs}}, \boldsymbol{\delta}; \eta) \right\} f(\mathbf{y}, \boldsymbol{\delta} \mid \mathbf{y}_{\text{obs}}, \boldsymbol{\delta}; \eta)\, d\mu(\mathbf{y}).
\end{aligned}$$

The first term is equal to $-E\left\{I_{\text{com}}(\eta) \mid \mathbf{y}_{\text{obs}}, \boldsymbol{\delta}\right\}$ and the second term is equal to

$$\begin{aligned}
E\left\{S_{\text{com}}(\eta) S_{\text{mis}}(\eta)' \mid \mathbf{y}_{\text{obs}}, \boldsymbol{\delta}\right\} &= E\left[\left\{\bar{S}(\eta) + S_{\text{mis}}(\eta)\right\} S_{\text{mis}}(\eta)' \mid \mathbf{y}_{\text{obs}}, \boldsymbol{\delta}\right] \\
&= E\left\{S_{\text{mis}}(\eta) S_{\text{mis}}(\eta)' \mid \mathbf{y}_{\text{obs}}, \boldsymbol{\delta}\right\}
\end{aligned}$$

because $E\left\{\bar{S}(\eta) S_{\text{mis}}(\eta)' \mid \mathbf{y}_{\text{obs}}, \boldsymbol{\delta}\right\} = \mathbf{0}$.

Missing Information Principle

- $S_{mis}(\eta)$: the score function with the conditional density $f(\mathbf{y}, \boldsymbol{\delta} | \mathbf{y}_{obs}, \boldsymbol{\delta})$.
- Expected missing information: $\mathcal{I}_{mis}(\eta) = \mathrm{E}\left\{-\frac{\partial}{\partial \eta^T} S_{mis}(\eta)\right\}$ satisfying $\mathcal{I}_{mis}(\eta) = \mathrm{E}\left\{S_{mis}(\eta)^{\otimes 2}\right\}$.
- Missing information principle (Orchard and Woodbury, 1972):

  $$\mathcal{I}_{mis}(\eta) = \mathcal{I}_{com}(\eta) - \mathcal{I}_{obs}(\eta),$$

  where $\mathcal{I}_{com}(\eta) = \mathrm{E}\left\{-\partial S_{com}(\eta)/\partial \eta^T\right\}$ is the expected information with complete-sample likelihood .
- An alternative expression of the missing information principle is

  $$\mathrm{V}\{S_{mis}(\eta)\} = \mathrm{V}\{S_{com}(\eta)\} - \mathrm{V}\{\bar{S}(\eta)\}.$$

  Note that $\mathrm{V}\{S_{com}(\eta)\} = \mathcal{I}_{com}(\eta)$ and $\mathrm{V}\{S_{obs}(\eta)\} = \mathcal{I}_{obs}(\eta)$.

Example 2.7

1. Consider the following bivariate normal distribution:

$$\begin{pmatrix} y_{1i} \\ y_{2i} \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}\right],$$

for $i = 1, 2, \cdots, n$. Assume for simplicity that σ_{11}, σ_{12} and σ_{22} are known constants and $\boldsymbol{\mu} = (\mu_1, \mu_2)'$ be the parameter of interest.

2. The complete sample score function for $\boldsymbol{\mu}$ is

$$S_{\text{com}}(\boldsymbol{\mu}) = \sum_{i=1}^{n} S_{\text{com}}^{(i)}(\boldsymbol{\mu}) = \sum_{i=1}^{n} \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}^{-1} \begin{pmatrix} y_{1i} - \mu_1 \\ y_{2i} - \mu_2 \end{pmatrix}.$$

The information matrix of $\boldsymbol{\mu}$ based on the complete sample is

$$\mathcal{I}_{\text{com}}(\boldsymbol{\mu}) = n \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}^{-1}.$$

Example 2.7 (Cont'd)

3. Suppose that there are some missing values in y_{1i} and y_{2i} and the original sample is partitioned into four sets:

$$\begin{aligned} H &= \text{both } y_1 \text{ and } y_2 \text{ respond} \\ K &= \text{only } y_1 \text{ is observed} \\ L &= \text{only } y_2 \text{ is observed} \\ M &= \text{both } y_1 \text{ and } y_2 \text{ are missing.} \end{aligned}$$

   Let n_H, n_K, n_L, n_M represent the size of H, K, L, M, respectively.

4. Assume that the response mechanism does not depend on the value of (y_1, y_2) and so it is MAR. In this case, the observed score function of $\boldsymbol{\mu}$ based on a single observation in set K is

$$\begin{aligned} E\left\{ S_{\mathrm{com}}^{(i)}(\boldsymbol{\mu}) \mid y_{1i}, i \in K \right\} &= \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}^{-1} \begin{pmatrix} y_{1i} - \mu_1 \\ E(y_{2i} \mid y_{1i}) - \mu_2 \end{pmatrix} \\ &= \begin{pmatrix} \sigma_{11}^{-1}(y_{1i} - \mu_1) \\ 0 \end{pmatrix}. \end{aligned}$$

Example 2.7 (Cont'd)

5. Similarly, we have

$$E\left\{S_{\text{com}}^{(i)}(\boldsymbol{\mu}) \mid y_{2i}, i \in L\right\} = \begin{pmatrix} 0 \\ \sigma_{22}^{-1}(y_{2i} - \mu_2) \end{pmatrix}.$$

6. Therefore, the observed information matrix of $\boldsymbol{\mu}$ is

$$\mathcal{I}_{\text{obs}}(\boldsymbol{\mu}) = n_H \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}^{-1} + n_K \begin{pmatrix} \sigma_{11}^{-1} & 0 \\ 0 & 0 \end{pmatrix} + n_L \begin{pmatrix} 0 & 0 \\ 0 & \sigma_{22}^{-1} \end{pmatrix}$$

and the asymptotic variance of the MLE of $\boldsymbol{\mu}$ can be obtained by the inverse of $\mathcal{I}_{\text{obs}}(\boldsymbol{\mu})$.

REFERENCES

Fisher, R. A. (1922), 'On the mathematical foundations of theoretical statistics', *Philosophical Transactions of the Royal Society of London A* **222**, 309–368.

Louis, T. A. (1982), 'Finding the observed information matrix when using the EM algorithm', *Journal of the Royal Statistical Society: Series B* **44**, 226–233.

Oakes, D. (1999), 'Direct calculation of the information matrix via the em algorithm', *Journal of the Royal Statistical Society: Series B* **61**, 479–482.

Orchard, T. and M.A. Woodbury (1972), A missing information principle: theory and applications, *in* 'Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability', Vol. 1, University of California Press, Berkeley, California, pp. 697–715.

Rubin, D. B. (1976), 'Inference and missing data', *Biometrika* **63**, 581–592.