Lecture 9: Likelihood approach for incomplete data

Likelihood function when there are missing data

\( y \): a variable or a vector of variables of interest.
\( x \): a vector of covariates.
\( Y = (y_1, \ldots, y_n) \): the complete data when there is no missing.
\( X = (x_1, \ldots, x_n) \): the observed covariates.

If there is no missing data, then the likelihood function is

\[ \ell(\theta | Y, X) = f_\theta(Y | X), \]

the joint probability density of \( Y \), given \( X \).

When there are missing data, this likelihood function cannot be used, because some values in \( Y \) are not observed.

\( Y_o \): the observed data
\( Y_m \): missing data
\( A \): the set of indicators of observing \( Y \)

\( f_\psi(A | Y, X) \): the probability density of \( A \) given \( Y \) and \( X \), where \( \psi \) is an unknown parameter vector.

The joint probability density of \( Y \) and \( A \) given \( X \) is

\[ f_{\theta, \psi}(Y, A | X) = f_\psi(A | Y, X)f_\theta(Y | X) \]
Lecture 9: Likelihood approach for incomplete data

Likelihood function when there are missing data

**y**: a variable or a vector of variables of interest.

**x**: a vector of covariates.

**Y** = \((y_1, ..., y_n)\): the complete data when there is no missing.

**X** = \((x_1, ..., x_n)\): the observed covariates.

If there is no missing data, then the likelihood function is

\[ \ell(\theta | Y, X) = f_\theta (Y | X), \]

the joint probability density of **Y**, given **X**.

When there are missing data, this likelihood function cannot be used, because some values in **Y** are not observed.

**Y_o**: the observed data

**Y_m**: missing data

**A**: the set of indicators of observing **Y**

\[ f_\psi (A | Y, X) \]: the probability density of **A** given **Y** and **X**, where \( \psi \) is an unknown parameter vector.

The joint probability density of **Y** and **A** given **X** is

\[ f_{\theta, \psi} (Y, A | X) = f_\psi (A | Y, X) f_\theta (Y | X) \]
Missing at random and likelihood analysis

Missing at random (MAR) or Ignorable Missingness:

\[ f_\psi(A|Y_o, Y_m, X) = f_\psi(A|Y_o, X). \]

The observed likelihood under MAR is

\[
\ell(\theta, \psi|Y_o, A, X) = \int f_{\theta, \psi}(Y_o, Y_m, A|X) dY_m
\]

\[
= \int f_\psi(A|Y_o, Y_m, X)f_\theta(Y_o, Y_m|X) dY_m
\]

\[
= f_\psi(A|Y_o, X) \int f_\theta(Y_o, Y_m|X) dY_m
\]

When \( \theta \) and \( \psi \) are unrelated parameter vectors, maximizing \( \ell(\theta, \psi|Y_o, A, X) \) over \((\theta, \psi)\) can be done by separately maximizing \( \int f_\theta(Y_o, Y_m|X) dY_m \) over \( \theta \) (ignoring \( A \) and \( f_\psi(A|Y_o, X) \)) and maximizing \( f_\psi(A|Y_o, X) \) over \( \psi \).

If \( y \) is discrete then the integral should be replaced by summation. Other parameters can be estimated based on \( \hat{\theta} \) and \( \hat{\psi} \).
Estimation for univariate \( y \) under MAR

Assume that \((y_1, x_1), \ldots, (y_n, x_n)\) are iid.

Assuming without loss of generality that \(y_1, \ldots, y_{n_1}\) are observed and \(y_{n_1+1}, \ldots, y_n\) are missing, we obtain that

\[
\ell(\theta \mid Y_0, X) = \int f_\theta (Y \mid X) dY_m
\]

\[
= \int f_\theta (y_1 \mid x_1) \cdots f_\theta (y_n \mid x_n) dY_m
\]

\[
= f_\theta (y_1 \mid x_1) \cdots f_\theta (y_{n_1} \mid x_{n_1})
\]

\[
\times \int f_\theta (y_{n_1+1} \mid x_{n_1+1}) dy_{n_1+1} \cdots \int f_\theta (y_n \mid x_n) dy_n
\]

\[
= f_\theta (y_1 \mid x_1) \cdots f_\theta (y_{n_1} \mid x_{n_1}).
\]

This means that the maximum likelihood estimator of \( \theta \) can be obtained by simply maximizing

\[
f_\theta (y_1 \mid x_1) \cdots f_\theta (y_{n_1} \mid x_{n_1}).
\]

The incomplete “data” \((y_{n_1+1}, x_{n_1+1}), \ldots, (y_n, x_n)\) are ignored.
Example 1: Normal distributions

Consider the case where \( f_\theta(y|x) \) is the normal distribution with mean \( \alpha + \beta x \) and variance \( \sigma^2 \) and \( \theta = (\alpha, \beta, \sigma^2) \). Then,

\[
\ell(\theta|Y_o, X) = f_\theta(y_1|x_1) \cdots f_\theta(y_{n_1}|x_{n_1}) = \frac{1}{(2\pi\sigma^2)^{n_1/2}} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n_1} (y_i - \alpha - \beta x_i)^2\right\}
\]

\[
\log \ell(\theta|Y_o, X) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n_1} (y_i - \alpha - \beta x_i)^2 - \frac{n_1}{2} \log \sigma^2.
\]

Maximizing \( \ell(\theta|Y_o, X) \) over \( \theta \) is the same as maximizing \( \log \ell(\theta|Y_o, X) \) over \( \theta \), which yields the solution

\[
\hat{\beta} = \frac{\sum_{i=1}^{n_1} (x_i - \bar{x}_1)y_i}{\sum_{i=1}^{n_1} (x_i - \bar{x}_1)^2}, \quad \hat{\alpha} = \bar{y}_1 - \hat{\beta} \bar{x}_1,
\]

\[
\hat{\sigma}^2 = \frac{1}{n_1} \sum_{i=1}^{n_1} (y_i - \hat{\alpha} - \hat{\beta} x_i)^2, \quad \bar{y}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} y_i, \quad \bar{x}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} x_i.
\]
Two dimensional Case

\[ Y_o = (y_{11}, \ldots, y_{n1}, y_{12}, \ldots, y_{n2}) \quad \text{and} \quad Y_m = (y_{(n_2+1)1}, \ldots, y_{n2}) \]

After removing \( X \) in the likelihood \( \ell(\theta|Y_o, X) \), we obtain that

\[
\ell(\theta|Y_o) = \int f_{\theta}(Y_o, Y_m) dY_m \\
= f_{\theta}(y_{11}, y_{12}) \cdots f_{\theta}(y_{n21}, y_{n22}) \int f_{\theta}(y_{(n_2+1)1}, y_{(n_2+1)2}) \cdots f_{\theta}(y_{n1}, y_{n2}) dY_m \\
= f_{\theta}(y_{11}, y_{12}) \cdots f_{\theta}(y_{n21}, y_{n22}) f_{\vartheta}(y_{(n_2+1)1}) \cdots f_{\vartheta}(y_{n1}),
\]

where \( \vartheta \) is a function of \( \theta \).

Using \( f_{\theta}(y_{i1}, y_{i2}) = f_{\varphi}(y_{i2}|y_{i1}) f_{\vartheta}(y_{i1}) \), where \((\vartheta, \varphi)\) is a one-to-one function of \( \theta \), we get

\[
\ell(\theta|Y_o) = f_{\vartheta}(y_{11}) \cdots f_{\vartheta}(y_{n1}) f_{\varphi}(y_{12}|y_{11}) \cdots f_{\varphi}(y_{n22}|y_{n21})
\]

Thus, \( \theta = (\vartheta, \varphi) \) can be estimated by seperately maximizing

\[
f_{\vartheta}(y_{11}) \cdots f_{\vartheta}(y_{n1}) \quad \text{and} \quad f_{\varphi}(y_{12}|y_{11}) \cdots f_{\varphi}(y_{n22}|y_{n21})
\]
Example 2: Bivariate normal data

Suppose that $f_\theta(y)$ is the density of the bivariate normal distribution with mean vector $(\mu_1, \mu_2)$ and covariance matrix

$$
\begin{pmatrix}
\sigma_1^2 & \sigma_1 \sigma_2 \rho \\
\sigma_1 \sigma_2 \rho & \sigma_2^2
\end{pmatrix}
$$

Then

$$
\theta = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho).
$$

From the properties of the normal distribution, $f_\vartheta(y_{i1})$ is the normal distribution with mean $\mu_1$ and variance $\sigma_1^2$, where $\vartheta = (\mu_1, \sigma_1^2)$, and $f_\varphi(y_{i2}|y_{i1})$ is the normal distribution with mean $\alpha + \beta y_{i1}$ and variance $\tau^2$, where

$$
\alpha = \mu_2 - \frac{\sigma_2 \rho \mu_1}{\sigma_1}, \quad \beta = \frac{\sigma_2 \rho}{\sigma_1}, \quad \tau^2 = \sigma_2^2 (1 - \rho^2), \quad \varphi = (\alpha, \beta, \tau^2).
$$

It follows from the same derivation as that in Example 1 that the maximum likelihood estimator of $\vartheta$ can be obtained by maximizing
Example 2: Bivariate normal data

\[ f_\theta(y_{11}) \cdots f_\theta(y_{n1}) = \frac{1}{(2\pi \sigma_1^2)^{n/2}} \exp \left\{ -\frac{1}{2\sigma_1^2} \sum_{i=1}^{n} (y_{i1} - \mu_1)^2 \right\} \]

The maximum likelihood estimator of \( \theta \) is

\[ \hat{\theta} = (\hat{\mu}_1, \hat{\sigma}_1^2), \quad \hat{\mu}_1 = \frac{1}{n} \sum_{i=1}^{n} y_{i1}, \quad \hat{\sigma}_1^2 = \frac{1}{n} \sum_{i=1}^{n} (y_{i1} - \hat{\mu}_1)^2 \]

The maximum likelihood estimator of \( \phi \) can be obtained by maximizing

\[ f_\phi(y_{12}|y_{11}) \cdots f_\phi(y_{n2|y_{n1}}) = \frac{1}{(2\pi \tau^2)^{n_2/2}} \exp \left\{ -\frac{1}{2\tau^2} \sum_{i=1}^{n_2} (y_{i2} - \alpha - \beta y_{i1})^2 \right\} \]

The maximum likelihood estimator of \( \phi \) is

\[ \hat{\phi} = (\hat{\alpha}, \hat{\beta}, \hat{\tau}^2), \quad \hat{\beta} = \frac{\sum_{i=1}^{n_2} (y_{i1} - \bar{y}_1)(y_{i2} - \bar{y}_2)}{\sum_{i=1}^{n_2} (y_{i1} - \bar{y}_1)^2}, \quad \bar{y}_k = \frac{1}{n_2} \sum_{i=1}^{n_2} y_{ik} \]
Example 2: Bivariate normal data

\[ \hat{\alpha} = \bar{y}_2 - \hat{\beta} \bar{y}_1, \quad \hat{\tau}^2 = \frac{1}{n_2} \sum_{i=1}^{n_2} (y_{i2} - \hat{\alpha} - \hat{\beta} y_{i1})^2. \]

Since

\[ \mu_2 = \alpha + \beta \mu_1, \quad \sigma_2^2 = \tau^2 + \sigma_1^2 \beta^2, \quad \rho = \sigma_1 \beta / \sigma_2, \]

estimators of \( \mu_2, \sigma_2^2, \) and \( \rho \) can be obtained by replacing \( \alpha, \beta, \tau^2, \mu_1 \) and \( \sigma_1^2 \) by their maximum likelihood estimators.

General case of monotone missing data

Let \( x \) be a covariate vector and \( y \) be an \( m \)-dimensional vector of variables having missing values. Suppose that missing is monotone in the sense that, for the \( t \)th component, \( y_{1t}, \ldots, y_{nt} \) are observed and \( y_{(n_t+1)t}, \ldots, y_{nt} \) are missing, and \( n \geq n_1 \geq n_2 \geq \cdots \geq n_m \geq 2 \). Then
Example 2: Bivariate normal data

\[ \hat{\alpha} = \bar{y}_2 - \hat{\beta} \bar{y}_1, \quad \hat{\tau}^2 = \frac{1}{n_2} \sum_{i=1}^{n_2} (y_{i2} - \hat{\alpha} - \hat{\beta} y_{i1})^2. \]

Since

\[ \mu_2 = \alpha + \beta \mu_1, \quad \sigma_2^2 = \tau^2 + \sigma_1^2 \beta^2, \quad \rho = \sigma_1 \beta / \sigma_2, \]

estimators of \( \mu_2, \sigma_2^2, \) and \( \rho \) can be obtained by replacing \( \alpha, \beta, \tau^2, \mu_1 \) and \( \sigma_1^2 \) by their maximum likelihood estimators.

General case of monotone missing data

Let \( x \) be a covariate vector and \( y \) be an \( m \)-dimensional vector of variables having missing values. Suppose that missing is monotone in the sense that, for the \( t \)th component, \( y_{1t}, \ldots, y_{nt} \) are observed and \( y_{(nt+1)t}, \ldots, y_{nt} \) are missing, and \( n \geq n_1 \geq n_2 \geq \cdots \geq n_m \geq 2 \). Then
General case of monotone missing data

\[ \ell(\theta | Y_o, X) \]

\[ = \int \prod_{i=1}^{n} f_\theta(\mathbf{y}_{i1}, \ldots, \mathbf{y}_{im} | \mathbf{x}_i) d\mathbf{y}_m \]

\[ = \int \prod_{i=1}^{n} f_\theta(\mathbf{y}_{i1} | \mathbf{x}_i) f_\theta(\mathbf{y}_{i2} | \mathbf{y}_{i1}, \mathbf{x}_i) \cdots f_\theta(\mathbf{y}_{im} | \mathbf{y}_{i1}, \ldots, \mathbf{y}_{i(m-1)}, \mathbf{x}_i) d\mathbf{y}_m \]

\[ = \int \prod_{i=1}^{n} \prod_{t=1}^{m} f_\theta(\mathbf{y}_{it} | \mathbf{y}_{i1}, \ldots, \mathbf{y}_{i(t-1)}, \mathbf{x}_i) d\mathbf{y}_m \]

\[ = \prod_{t=1}^{m} \int \prod_{i=1}^{n} f_\theta(\mathbf{y}_{it} | \mathbf{y}_{i1}, \ldots, \mathbf{y}_{i(t-1)}, \mathbf{x}_i) dy_{(nt+1)t} \cdots dy_{nt} \]

\[ = \prod_{t=1}^{m} \prod_{i=1}^{nt} f_\theta(\mathbf{y}_{it} | \mathbf{y}_{i1}, \ldots, \mathbf{y}_{i(t-1)}, \mathbf{x}_i) \prod_{i=nt+1}^{n} \int f_\theta(\mathbf{y}_{it} | \mathbf{y}_{i1}, \ldots, \mathbf{y}_{i(t-1)}, \mathbf{x}_i) dy_{it} \]

\[ = \prod_{t=1}^{m} \prod_{i=1}^{nt} f_\theta(\mathbf{y}_{it} | \mathbf{y}_{i1}, \ldots, \mathbf{y}_{i(t-1)}, \mathbf{x}_i) \]
General case of monotone missing data

The MLE of $\theta$ can be obtained by the following iterative method:

1. Calculate the MLE of $\theta_1$ in the linear regression between $y_{i1}$ and $x_i$, $i = 1, \ldots, n_1$.

2. Calculate the MLE of $\theta_2$ in the linear regression between $y_{i2}$ and $(y_{i1}, x_i)$, $i = 1, \ldots, n_2$.

......

m. Calculate the MLE of $\theta_m$ in the linear regression between $y_{im}$ and $(y_{i1}, \ldots, y_{i(m-1)}, x_i)$, $i = 1, \ldots, n_m$.

final. Obtain the MLE of $\theta$ using the function relationship between $\theta$ and $(\theta_1, \ldots, \theta_m)$.

The EM algorithm under MAR

If missing is not monotone, maximizing $\ell(\theta | Y_o, X)$ can be very difficult or impossible. The well-known EM algorithm can be applied to partially solve this problem (see Little and Rubin, 2002). The EM algorithm consists of an E step (expectation step) and an M step (maximization step) and we carry out iterations between the E and M steps.
General case of monotone missing data

The MLE of $\theta$ can be obtained by the following iterative method:

1. Calculate the MLE of $\theta_1$ in the linear regression between $y_{i1}$ and $x_i$, $i = 1, \ldots, n_1$.
2. Calculate the MLE of $\theta_2$ in the linear regression between $y_{i2}$ and $(y_{i1}, x_i)$, $i = 1, \ldots, n_2$.
   .......

$m$. Calculate the MLE of $\theta_m$ in the linear regression between $y_{im}$ and $(y_{i1}, \ldots, y_{i(m-1)}, x_i)$, $i = 1, \ldots, n_m$.

final. Obtain the MLE of $\theta$ using the function relationship between $\theta$ and $(\theta_1, \ldots, \theta_m)$.

The EM algorithm under MAR

If missing is not monotone, maximizing $\ell(\theta | Y_o, X)$ can be very difficult or impossible. The well-known EM algorithm can be applied to partially solve this problem (see Little and Rubin, 2002). The EM algorithm consists of an E step (expectation step) and an M step (maximization step) and we carry out iterations between the E and M steps.
E step at the $t$th iteration

Let $\theta^{(t-1)}$ be the estimate of $\theta$ at the $(t-1)$th iteration of the EM algorithm. The E step at the $t$th iteration calculates the expectation

$$Q(\theta|\theta^{(t-1)}) = E_{\theta^{(t-1)}}[\log f_\theta(Y_o, Y_m|X)|Y_o, X]$$

$$= \int [\log f_\theta(Y_o, Y_m|X)] f_{\theta^{(t-1)}}(Y_m|Y_o, X) dY_m$$

M step at the $t$th iteration

We maximize $Q(\theta|\theta^{(t-1)})$ over $\theta$, i.e., we find a $\theta^{(t)}$ that satisfies

$$Q(\theta^{(t)}|\theta^{(t-1)}) = \max_\theta Q(\theta|\theta^{(t-1)})$$

Why does EM algorithm work?

The EM algorithm maximizes

$$Q(\theta|\theta) = \int [\log f_\theta(Y_o, Y_m|X)] f_{\theta}(Y_m|Y_o, X) dY_m$$

How does this relate to maximizing $\ell(\theta|Y_o, X)$?
E step at the $t$th iteration

Let $\theta^{(t-1)}$ be the estimate of $\theta$ at the $(t-1)$th iteration of the EM algorithm. The E step at the $t$th iteration calculates the expectation

$$Q(\theta|\theta^{(t-1)}) = E_{\theta^{(t-1)}}[\log f_\theta(Y_o, Y_m|X)|Y_o, X]$$

$$= \int [\log f_\theta(Y_o, Y_m|X)]f_{\theta^{(t-1)}}(Y_m|Y_o, X)dY_m$$

M step at the $t$th iteration

We maximize $Q(\theta|\theta^{(t-1)})$ over $\theta$, i.e., we find a $\theta^{(t)}$ that satisfies

$$Q(\theta^{(t)}|\theta^{(t-1)}) = \max_\theta Q(\theta|\theta^{(t-1)})$$

Why does EM algorithm work?

The EM algorithm maximizes

$$Q(\theta|\theta) = \int [\log f_\theta(Y_o, Y_m|X)]f_\theta(Y_m|Y_o, X)dY_m$$

How does this relate to maximizing $\ell(\theta|Y_o, X)$?
E step at the $t$th iteration

Let $\theta^{(t-1)}$ be the estimate of $\theta$ at the $(t-1)$th iteration of the EM algorithm. The E step at the $t$th iteration calculates the expectation

$$Q(\theta|\theta^{(t-1)}) = E_{\theta^{(t-1)}}[\log f_\theta(Y_o, Y_m|X)|Y_o, X]$$

$$= \int \left[ \log f_\theta(Y_o, Y_m|X) \right] f_\theta^{(t-1)}(Y_m|Y_o, X) dY_m$$

M step at the $t$th iteration

We maximize $Q(\theta|\theta^{(t-1)})$ over $\theta$, i.e., we find a $\theta^{(t)}$ that satisfies

$$Q(\theta^{(t)}|\theta^{(t-1)}) = \max_\theta Q(\theta|\theta^{(t-1)})$$

Why does EM algorithm work?

The EM algorithm maximizes

$$Q(\theta|\theta) = \int \left[ \log f_\theta(Y_o, Y_m|X) \right] f_\theta(Y_m|Y_o, X) dY_m$$

How does this relate to maximizing $\ell(\theta|Y_o, X)$?
Convergence of EM-algorithm

Since \( \ell(\theta | Y_o, X) = \int f_\theta(Y_o, Y_m | X) dY_m = f_\theta(Y_o, Y_m | X) / f_\theta(Y_m | Y_o, X) \),

\[ \log \ell(\theta | Y_o, X) = Q(\theta | \theta) - H(\theta | \theta) \]

where \( H(\theta | \theta) = \int [\log f_\theta(Y_m | Y_o, X)] f_\theta(Y_m | Y_o, X) dY_m. \)

By Jensen’s inequality, for any \( t \), \( H(\theta | \theta(t)) \leq H(\theta(t) | \theta(t)). \)

Hence, at the \( t \)th iteration,

\[
\begin{align*}
\log \ell(\theta(t) | Y_o, X) - \log \ell(\theta(t-1) | Y_o, X) &= Q(\theta(t) | \theta(t-1)) - Q(\theta(t-1) | \theta(t-1)) - H(\theta(t) | \theta(t-1)) + H(\theta(t-1) | \theta(t-1)) \\
&\geq Q(\theta(t) | \theta(t-1)) - Q(\theta(t-1) | \theta(t-1)) \geq 0
\end{align*}
\]

with equality holds if and only if \( Q(\theta(t) | \theta(t-1)) = Q(\theta(t-1) | \theta(t-1)) \).

This means the change from \( \theta(t-1) \) to \( \theta(t) \) increases the likelihood.

For given observed \( (Y_o, X) \), the EM algorithm produces a sequence \( \theta(t), t = 1, 2, \ldots \). Under certain conditions, this sequence converges and the limit is considered as the EM estimator of \( \theta \).