Incomplete Data Analysis

Vanda Inácio

University of Edinburgh

Semester 1, 2018/2019

Likelihood based inference with incomplete data

- ↪ As Little and Rubin (2002) stated: “In one formal sense there is no difference between ML or Bayes inference for incomplete data and ML or Bayes inference for complete data. The likelihood for the parameters based on the incomplete data is derived, ML estimates are found by solving the likelihood equation, and the posterior distribution is obtained by incorporating a prior distribution and performing the necessary integrations.”

- ↪ As before let $\mathbf{y}$ denote the data that would occur in the absence of missing values. We assume $\mathbf{y} = (y_{ij})$ is a rectangular dataset, $i = 1, \ldots, n$ individuals and $j = 1, \ldots, p$ variables.

- ↪ We write $\mathbf{y} = (\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}})$, where $\mathbf{y}_{\text{obs}}$ denotes the observed values and $\mathbf{y}_{\text{mis}}$ denotes the missing values.

- ↪ Let $\mathbf{r} = (r_{ij})$ be the missing data indicator matrix, defined as

$$r_{ij} = \begin{cases} 1, & y_{ij} \text{ is observed} \\ 0, & y_{ij} \text{ is missing.} \end{cases}$$

Likelihood based inference with incomplete data

- Let θ be the parameters of the model for $\mathbf{y}$ and ψ the parameters for $\mathbf{r}$.
- Then, the joint model of the full data is

$$f(\mathbf{y}, \mathbf{r} \mid \theta, \psi) = f(\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}}, \mathbf{r} \mid \theta, \psi).$$

- The joint model $f(\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}}, \mathbf{r} \mid \theta, \psi)$ cannot be evaluated in the usual way because it depends on missing data.
- However, the marginal distribution of the observed data can be obtained by integrating out the missing data

$$f(\mathbf{y}_{\text{obs}}, \mathbf{r} \mid \theta, \psi) = \int f(\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}}, \mathbf{r} \mid \theta, \psi) d\mathbf{y}_{\text{mis}}.$$

Likelihood based inference with incomplete data

↪ Two factorisations of the joint model are commonly used:

1. Selection model factorisation

$$f(\mathbf{y}, \mathbf{r} \mid \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\mathbf{r} \mid \mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\psi}) f(\mathbf{y} \mid \boldsymbol{\theta}, \boldsymbol{\psi})$$

This factorisation involves directly the full data density $f(\mathbf{y} \mid \boldsymbol{\theta}, \boldsymbol{\psi})$ (our model of interest) and the missingness mechanism $f(\mathbf{r} \mid \mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\psi})$. Selection models were first used by Rubin (1976), and according to Molenberghs and Kenward (2007, Chapter 3) the terminology was coined in the econometric literature.

2. Pattern mixture factorisation

$$f(\mathbf{y}, \mathbf{r} \mid \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\mathbf{y} \mid \mathbf{r}, \boldsymbol{\theta}, \boldsymbol{\psi}) f(\mathbf{r} \mid \boldsymbol{\theta}, \boldsymbol{\psi}).$$

The pattern mixture factorisation can be viewed as incorporating the density of the full data for given patterns of missingness weighted by the probability of each pattern. Pattern mixture models were first proposed by Little (1993).

Likelihood based inference with incomplete data

- An advantage of the selection model factorisation is that it includes the model of interest term directly.
- On the other hand, the pattern mixture model corresponds more directly to what is actually observed, i.e., the distribution of the data within subgroups having different missing data patterns.
- We will focus on the selection model factorisation and demonstrate how it forms the basis for deriving likelihood-based inference using the observed data.
- Assuming the data and missingness model have distinct parameters, the selection model simplifies to

$$\begin{aligned} f(\mathbf{y}, \mathbf{r} \mid \theta, \psi) &= f(\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}}, \mathbf{r} \mid \theta, \psi) \\ &= f(\mathbf{r} \mid \mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}}, \theta, \psi) f(\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}} \mid \theta, \psi) \\ &= f(\mathbf{r} \mid \mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}}, \psi) f(\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}} \mid \theta) \end{aligned}$$

Likelihood based inference with incomplete data

- $f(\mathbf{y}_{obs}, \mathbf{y}_{mis} \mid \theta)$ is the usual likelihood we would specify if all the data had been observed.
- $f(\mathbf{r} \mid \mathbf{y}_{obs}, \mathbf{y}_{mis}, \psi)$ represents the missing data mechanism and describes the way in which the probability of an observation being missing depends on other variables (measured or not) and on its own values.
- Remember that for some types of missing data, the form of the conditional distribution of $\mathbf{r}$ can be simplified.
- Recall we wish to integrate out the missingness

$$\begin{aligned} f(\mathbf{y}_{obs}, \mathbf{r} \mid \theta, \psi) &= \int f(\mathbf{y}_{obs}, \mathbf{y}_{mis}, \mathbf{r} \mid \theta, \psi) d\mathbf{y}_{mis} \\ &= \int f(\mathbf{r} \mid \mathbf{y}_{obs}, \mathbf{y}_{mis}, \psi) f(\mathbf{y}_{obs}, \mathbf{y}_{mis} \mid \theta) d\mathbf{y}_{mis} \end{aligned}$$

Likelihood based inference with incomplete data

↪ MAR missingness depends only on observed data, i.e.,

$$f(\mathbf{r} \mid \mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}}, \psi) = f(\mathbf{r} \mid \mathbf{y}_{\text{obs}}, \psi).$$

↪ So,

$$f(\mathbf{y}_{\text{obs}}, \mathbf{r} \mid \theta, \psi) = \int f(\mathbf{r} \mid \mathbf{y}_{\text{obs}}, \psi) f(\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}} \mid \theta) \mathrm{d}\mathbf{y}_{mis}.$$

↪ Since $f(\mathbf{r} \mid \mathbf{y}_{\text{obs}}, \psi)$ does not depend on $\mathbf{y}_{mis}$ it can be regarded as a constant when integrating with respect to $\mathbf{y}_{mis}$. Thus,

$$\begin{aligned} f(\mathbf{y}_{\text{obs}}, \mathbf{r} \mid \theta, \psi) &= f(\mathbf{r} \mid \mathbf{y}_{\text{obs}}, \psi) \int f(\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}} \mid \theta) \mathrm{d}\mathbf{y}_{mis} \\ &= f(\mathbf{r} \mid \mathbf{y}_{\text{obs}}, \psi) f(\mathbf{y}_{\text{obs}} \mid \theta) \end{aligned}$$

Likelihood based inference with incomplete data

- MCAR missingness is a special case of MAR that does not even depend on the observed data

$$f(\mathbf{r} \mid \mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}}, \psi) = f(\mathbf{r} \mid \psi).$$

- Hence,

$$f(\mathbf{y}_{\text{obs}}, \mathbf{r} \mid \theta, \psi) = f(\mathbf{r} \mid \psi) f(\mathbf{y}_{\text{obs}} \mid \theta).$$

- Rewriting in terms of likelihoods

$$L(\theta, \psi \mid \mathbf{y}_{\text{obs}}, \mathbf{r}) = f(\mathbf{r} \mid \psi) L(\theta \mid \mathbf{y}_{\text{obs}}).$$

- If the missingness mechanism is MAR (or MCAR) and θ and ψ are distinct then the likelihood based inferences for θ from $L(\theta, \psi \mid \mathbf{y}_{\text{obs}}, \mathbf{r})$ will be the same as likelihood based inferences for θ from $L(\theta \mid \mathbf{y}_{\text{obs}})$, i.e.,

$$\begin{aligned} \widehat{\theta}_{\text{MLE}} &= \arg\max_{\theta \in \Theta} L(\theta, \psi \mid \mathbf{y}_{\text{obs}}, \mathbf{r}) \\ &= \arg\max_{\theta \in \Theta} L(\theta \mid \mathbf{y}_{\text{obs}}). \end{aligned}$$

Likelihood based inference with incomplete data

↪ The likelihood function

$$L(\boldsymbol{\theta} \mid \mathbf{y}_{\text{obs}}) = f(\mathbf{y}_{\text{obs}} \mid \boldsymbol{\theta}),$$

is called the likelihood ignoring the missing data mechanism or observed data likelihood.

↪ A missing data mechanism is ignorable for likelihood inference if

1. the missing data are MAR (or MCAR) and
2. the parameter ψ (missingness mechanism) and $\boldsymbol{\theta}$ (data model) are distinct, in the sense that the joint parameter space of $(\psi, \boldsymbol{\theta})$ is the product of the parameter spaces Ψ and Θ.

Likelihood based inference with incomplete data

- ↪ The first condition (MAR) is typically regarded as the more important condition.
- ↪ If MAR does not hold, then the maximum likelihood estimator based on the observed data likelihood can be seriously biased.
- ↪ If the data are MAR but distinctness does not hold, inference based on the observed data likelihood $L(\boldsymbol{\theta} \mid \mathbf{y}_{\text{obs}})$ is still valid but not fully efficient.
- ↪ Further few remarks apply. First, with a likelihood analysis, the observed information should be used rather than the expected one (Kenward and Molenberghs, 1998).
- ↪ Second, ignoring the missing data mechanism assumes there is no scientific interest attached to it. When this is untrue, the analyst can fit appropriate models to the missing data indicators, although in the vast majority of the times, this is not a trivial task.
- ↪ Third, regardless of the appeal of an ignorable analysis, remember that as we already discussed, MNAR can almost never be ruled out as a mechanism, and therefore one should also consider the possible impact of such mechanism.

Likelihood based inference with incomplete data

Example: incomplete univariate (normal) data

- Let us assume that the data $\mathbf{y} = (y_1, \ldots, y_n)$ comes from a normal random sample with mean μ and variance σ^2.

- Further suppose that, possibly after reordering, only the first m observations $(y_1, \ldots, y_m)$ are observed, with the remainder $n - m$ observations $(y_{m+1}, \ldots, y_n)$ being missing. In what follows we assume MAR (or MCAR) and distinctness of the parameters.

- The observed data likelihood is

$$
\begin{aligned}
L(\theta \mid \mathbf{y}_{\text{obs}}) &= \int f(\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}} \mid \boldsymbol{\theta}) \mathrm{d}\mathbf{y}_{mis} \\
&= \int \ldots \int \prod_{i=1}^{m} f(y_i \mid \theta) \prod_{i=m+1}^{n} f(y_i \mid \theta) \mathrm{d}y_{m+1} \ldots \mathrm{d}y_n \\
&= \prod_{i=1}^{m} f(y_i \mid \theta) \int \ldots \int \prod_{i=m+1}^{n} f(y_i \mid \theta) \mathrm{d}y_{m+1} \ldots \mathrm{d}y_n \\
&= \prod_{i=1}^{m} f(y_i \mid \theta)
\end{aligned}
$$

Likelihood based inference with incomplete data

Example: incomplete univariate (normal) data

- Thus the observed data likelihood is a complete data likelihood based on the reduced sample $(y_1, \ldots, y_m)$.

- Maximisation of $L(\theta \mid \mathbf{y}_{\text{obs}})$ for $\theta = (\mu, \sigma^2)$ leads to the following maximum likelihood estimates

$$\widehat{\mu} = \frac{1}{m}\sum_{i=1}^{m} y_i = \bar{y}_{(m)}, \qquad \widehat{\sigma}^2 = \frac{1}{m}\sum_{i=1}^{m}(y_i - \bar{y}_{(m)})^2.$$

Likelihood based inference with incomplete data

Example: non-distinct parameters

- Let us suppose that $Y_i \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$ and $R_i \mid Y_i \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$, $i = 1, \ldots, n$. Missing data generated this way is clearly MCAR. Suppose further that we only observe, possibly after reordering, the first m observations.
- In this case, the parameters for the data and missingness model are the same (and thus, obviously, not distinct!!).
- The joint likelihood of $\mathbf{y}_{\text{obs}}$ and $\mathbf{r}$ is

$$\begin{aligned} L(\theta \mid \mathbf{y}_{\text{obs}}, \mathbf{r}) &= f(\mathbf{r} \mid \theta) L(\theta \mid \mathbf{y}_{obs} = (y_1, \ldots, y_m)) \\ &= \prod_{i=1}^{n} \theta^{r_i} (1-\theta)^{1-r_i} \prod_{i=1}^{m} \theta^{y_i} (1-\theta)^{1-y_i} \\ &= \theta^{\sum_{i=1}^{n} r_i} (1-\theta)^{n-\sum_{i=1}^{n} r_i} \theta^{\sum_{i=1}^{m} y_i} (1-\theta)^{m-\sum_{i=1}^{m} y_i} \\ &= \theta^{m+\sum_{i=1}^{m} y_i} (1-\theta)^{n-\sum_{i=1}^{m} y_i}, \end{aligned}$$

  note that since there are m observed units/individuals, we thus have $\sum_{i=1}^{n} r_i = m$.
- This likelihood leads to the following mle estimator

$$\widehat{\theta}_{\text{MLE}} = \frac{m + \sum_{i=1}^{m} Y_i}{m+n}.$$

Likelihood based inference with incomplete data

Example: non-distinct parameters

↪ If we ignore the missing data mechanism, the observed data likelihood is

$$L(\theta \mid \mathbf{y}_{\text{obs}}) = \prod_{i=1}^{m} \theta^{y_i}(1-\theta)^{1-y_i},$$

which leads to the mle estimator

$$\widehat{\theta}_{\text{MLE}} = \frac{\sum_{i=1}^{m} Y_i}{m}.$$

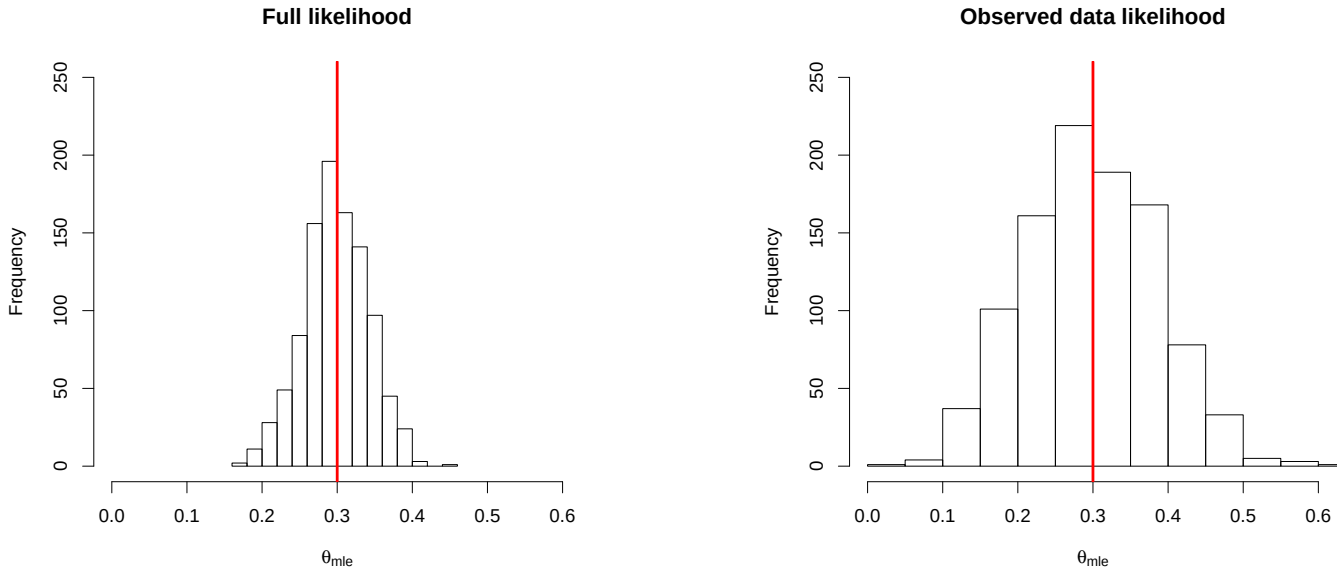
↪ Let us conduct a simulation experiment to check to which extent is the estimate of θ impacted by ignoring the missing data mechanism.

↪ We consider the following setting: $n = 100$, $\theta = 0.3$, and $nsim = 1000$, where $nsim$ denotes the number of generated datasets.

Likelihood based inference with incomplete data

Example: non-distinct parameters

↪ Below we show the histogram of $\widehat{\theta}_{\text{MLE}}$ across the 1000 simulated datasets in both scenarios (taking into account and ignoring the missing data mechanism). The solid vertical red line denotes the true value $\theta = 0.3$.



↪ As can be observed, in both cases there is a concentration of the maximum likelihood estimates around 0.3, but ignoring the missingness mechanism leads, as expected, to more variability around the true value.

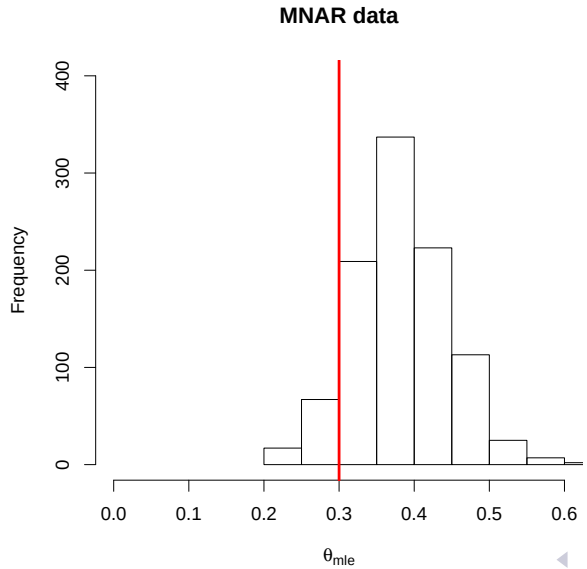
Likelihood based inference with incomplete data

Example: MNAR data

- We now investigate what happens when the MAR condition is violated.
- We assume

$$Y_i \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta), \quad \text{and} \quad \text{logit}\{\Pr(R_i = 1 \mid Y_i)\} = Y_i.$$

- We are thus violating the MAR assumption since data generating this way are MNAR (but parameters are distinct).
- We conduct a similar simulation study to check the performance of $\widehat{\theta}_{\text{MLE}} = \sum_{i=1}^{m} Y_i/m$ in this setting.



Likelihood based inference with incomplete data

Example: bivariate normal data with one variable subject to missingness

- We wil now consider an example with bivariate normal data with one variable subject to missingness.
- Before we proceed, I introduce some facts about the bivariate normal distribution.
- Proofs of the results stated here can be found in almost every multivariate analysis book (or somewhere in the web!). I personally like the following book Rencher, A.C. (2002) *Methods of Multivariate Analysis*, Wiley (see Chapter 4, pp.82–111, about the multivariate normal distribution).
- Let

$$\mathbf{Y} = (Y_1, Y_2)^T \sim \mathrm{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

with

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix},$$

where

$$\rho = \frac{\sigma_{12}}{\sigma_1\sigma_2},$$

is the correlation between Y_1 and Y_2.

Likelihood based inference with incomplete data

Example: bivariate normal data with one variable subject to missingness

↪ The density of $\mathbf{Y}$ is given by

$$f(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{2\pi}|\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right\}.$$

↪ The marginal distribution of Y_1 is

$$Y_1 \sim \mathrm{N}(\mu_1, \sigma_1^2).$$

↪ The conditional distribution of Y_2 given $Y_1 = y_1$ is

$$Y_2 \mid Y_1 = y_1 \sim \mathrm{N}(b_0 + b_1 y_1, \sigma^2_{2|1}),$$

with parameters

$$b_0 = \mu_2 - b_1\mu_1,$$
$$b_1 = \frac{\sigma_{12}}{\sigma_1^2},$$
$$\sigma^2_{2|1} = \sigma_2^2 - \frac{\sigma_{12}^2}{\sigma_1^2}.$$

Likelihood based inference with incomplete data

Example: bivariate normal data with one variable subject to missingness

- Our context is the following, while variable Y_1 is completely observed, only m out of n values (without loose of generality, the first m values) of Y_2 are observed.

- We have $Y_i = (Y_{1i}, Y_{2i}) \overset{\text{iid}}{\sim} \text{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, for $i = 1, \ldots, n$.

- The observed data likelihood is given by

$$
\begin{aligned}
L(\boldsymbol{\theta} \mid \mathbf{y}_{\text{obs}}) &= \prod_{i=1}^{m} f(y_{1i}, y_{2i} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \prod_{i=m+1}^{n} f(y_{1i} \mid \mu_1, \sigma_1^2) \\
&= \left\{ \prod_{i=1}^{m} f\left(y_{1i} \mid \mu_1, \sigma_1^2\right) f\left(y_{2i} \mid y_{1i}, b_0, b_1, \sigma_{2|1}^2\right) \right\} \prod_{i=m+1}^{n} f\left(y_{1i} \mid \mu_1, \sigma_1^2\right) \\
&= \prod_{i=1}^{m} f\left(y_{2i} \mid y_{1i}, b_0, b_1, \sigma_{2|1}^2\right) \prod_{i=1}^{n} f\left(y_{1i} \mid \mu_1, \sigma_1^2\right),
\end{aligned}
$$

where $\boldsymbol{\theta} = (\mu_1, \sigma_1^2, b_0, b_1, \sigma_{1|1}^2)$.

Likelihood based inference with incomplete data

Example: bivariate normal data with one variable subject to missingness

↪ After some calculations, one obtains

$$\widehat{\mu}_1 = \frac{1}{n}\sum_{i=1} y_{1i},$$

$$\widehat{\sigma}_1^2 = \frac{1}{n}\sum_{i=1}(y_{1i} - \widehat{\mu}_1)^2,$$

$$\widehat{b}_1 = \frac{s_{12}}{s_1^2},$$

$$\widehat{b}_0 = \bar{y}_2 - \widehat{b}_1\bar{y}_1,$$

$$\widehat{\sigma}_{2|1}^2 = \frac{1}{m}\sum_{i=1}^{m}(y_{2i} - (\widehat{b}_0 - \widehat{b}_1 y_{1i}))^2$$