Incomplete Data Analysis

V. Inácio de Carvalho & M. de Carvalho

University of Edinburgh

PRR Plano de Recuperação e Resiliência

GOVERNO DE PORTUGAL

Financiado pela União Europeia
NextGenerationEU

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Likelihood based inference with incomplete data

- ↪ Let $\mathbf{y}$ denote the complete data that we would have observed in the absence of missing values. We assume that $\mathbf{y} = (y_{ij})$ is a rectangular data set, with $i = 1, \ldots, n$ individuals and $j = 1, \ldots, p$ variables.

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

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

- ↪ We write $\mathbf{y} = (\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}})$, where $\mathbf{y}_{\text{obs}}$ is the vector containing those y_{ij} for which $r_{ij} = 1$ and, analogously, $\mathbf{y}_{\text{mis}}$ is the vector containing those y_{ij} for which $r_{ij} = 0$.

Likelihood based inference with incomplete data

- ↪ The full data $(\mathbf{y}, \mathbf{r})$ consist of the complete data together with the missing values indicators.

- ↪ Unless all components of $\mathbf{r}$ are equal to one, the full data components are never jointly observed but rather one observes $\mathbf{y}_{\text{obs}}$ together with the missing data indicators $\mathbf{r}$.

- ↪ Let $\boldsymbol{\theta}$ be the parameters of the model for $\mathbf{y}$ and $\boldsymbol{\psi}$ the parameters for $\mathbf{r}$.

- ↪ Then, the joint model of the full data is

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

- ↪ The joint model $f(\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}}, \mathbf{r} \mid \boldsymbol{\theta}, \boldsymbol{\psi})$ cannot be evaluated in the usual way because it depends on missing data.

- ↪ However, the marginal distribution of $(\mathbf{y}_{\text{obs}}, \mathbf{r})$ can be obtained by integrating out the missing data

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

Likelihood based inference with incomplete data

↪ Two factorisations of the joint model are commonly used:

1. Selection model factorisation

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

   This factorisation involves directly the model for the complete data $f(\mathbf{y} \mid \boldsymbol{\theta})$ (our model of interest) and the missingness mechanism $f(\mathbf{r} \mid \mathbf{y}, \boldsymbol{\psi})$. 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

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

   The pattern mixture factorisation can be viewed as incorporating the density of the complete 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, $f(\mathbf{y} \mid \boldsymbol{\theta})$, 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.

- ↪ Note that $\boldsymbol{\theta}$ and ψ have different interpretations in the selection and in the pattern mixture model, but we use the same symbols for convenience.

- ↪ The parameter $\boldsymbol{\theta}$ in the pattern mixture model is not the parameter ordinarily of interest, that governs the assumed model for the complete data. In fact, such quantity is not directly represented in the pattern mixture model.

- ↪ We will focus on the selection model factorisation and demonstrate how it forms the basis for deriving likelihood-based inference using the observed data.

Likelihood based inference with incomplete data

- ↪ Recall that, under the selection model factorisation we have

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

- ↪ $f(\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}} \mid \boldsymbol{\theta})$ is the usual likelihood we would specify if all the data had been observed.

- ↪ $f(\mathbf{r} \mid \mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}}, \boldsymbol{\psi})$ 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 **r** can be simplified.

- ↪ Recall we wish to integrate out the missingness

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

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,

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

- ↪ 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,

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

Likelihood based inference with incomplete data

- MCAR missingness is a special case of MAR that does not even depend on the observed data
$$f(\mathbf{r} \mid \mathbf{y}_{\mathrm{obs}}, \mathbf{y}_{\mathrm{mis}}, \boldsymbol{\psi}) = f(\mathbf{r} \mid \boldsymbol{\psi}).$$

- Hence,
$$f(\mathbf{y}_{\mathrm{obs}}, \mathbf{r} \mid \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\mathbf{r} \mid \boldsymbol{\psi}) f(\mathbf{y}_{\mathrm{obs}} \mid \boldsymbol{\theta}).$$

- Rewriting in terms of likelihoods, for the general MAR case
$$L(\boldsymbol{\theta}, \boldsymbol{\psi} \mid \mathbf{y}_{\mathrm{obs}}, \mathbf{r}) = f(\mathbf{r} \mid \mathbf{y}_{\mathrm{obs}}, \boldsymbol{\psi}) L(\boldsymbol{\theta} \mid \mathbf{y}_{\mathrm{obs}}).$$

- If the missingness mechanism is MAR (or MCAR) and, additionally, $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are distinct, then the likelihood based inferences for $\boldsymbol{\theta}$ from $L(\boldsymbol{\theta}, \boldsymbol{\psi} \mid \mathbf{y}_{\mathrm{obs}}, \mathbf{r})$ will be the same as likelihood based inferences for $\boldsymbol{\theta}$ from $L(\boldsymbol{\theta} \mid \mathbf{y}_{\mathrm{obs}})$, i.e.,
$$\begin{aligned}\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}} &= \arg\max_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}, \boldsymbol{\psi} \mid \mathbf{y}_{\mathrm{obs}}, \mathbf{r}) \\ &= \arg\max_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta} \mid \mathbf{y}_{\mathrm{obs}}).\end{aligned}$$

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/disjoint, in the sense that the joint parameter space of $(\psi, \boldsymbol{\theta})$ is the product of the parameter spaces $\boldsymbol{\Psi}$ and $\boldsymbol{\Theta}$ (separability condition).

- Ignorability basically means that when doing inference, i.e., to get the maximum likelihood estimates of the parameters from an incomplete dataset, one can simply maximise the observed likelihood. Note that we have not *ignored* the missing data mechanism at all, instead, we have made explicit constraints on it.

Likelihood based inference with incomplete data

- ↪ The first condition (MAR) is typically regarded as the most important condition.
- ↪ If MAR does not hold (i.e., if data are MNAR), 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, *Statistical Science*, 1998).
- ↪ Second, ignoring the missing data mechanism assumes there is no scientific interest attached to it. When this is not true, 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.

- Let $\mathbf{r} = (r_1, \ldots, r_n)$, where $r_i = 1$, for $i = 1, \ldots, m$, and $r_i = 0$, for $i = m + 1, \ldots, n$.

- Suppose that each unit is observed with probability ψ (distinct of $\theta = (\mu, \sigma^2)$), so that

$$f(\mathbf{r} \mid \mathbf{y}_{\text{obs}}, \mathbf{y}_{mis}, \psi) = f(\mathbf{r} \mid \psi) = \prod_{i=1}^{n} \psi^{r_i}(1 - \psi)^{r_i} = \psi^{m}(1 - \psi)^{n-m}.$$

  The mechanism is clearly MCAR.

- Therefore,

$$\begin{aligned} L(\theta, \psi \mid \mathbf{y}_{\text{obs}}, \mathbf{r}) &= f(\mathbf{r} \mid \psi) L(\theta \mid \mathbf{y}_{\text{obs}}) \\ &= \{\psi^{m}(1 - \psi)^{n-m}\} \left\{ (2\pi\sigma^2)^{-m/2} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{m} (y_i - \mu)^2 \right] \right\}. \end{aligned}$$

Likelihood based inference with incomplete data

Example: incomplete univariate (normal) data

$\hookrightarrow$ Because the missing data are MCAR and ψ and θ are distinct, then inference about θ can be based simply on the observed likelihood $L(\theta \mid \mathbf{y}_{\text{obs}})$.

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

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

$\hookrightarrow$ Note that for this specific and simple example, results coincides with those from a maximum likelihood analysis applied to the complete cases. However, this is more the exception than the rule. We will see more challenging examples (e.g., involving missing data in a bivariate normal distribution).

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 $\mathbf{y}_{\text{obs}} = (y_1, \ldots, y_m)$ and $\mathbf{r}$ is

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

  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

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

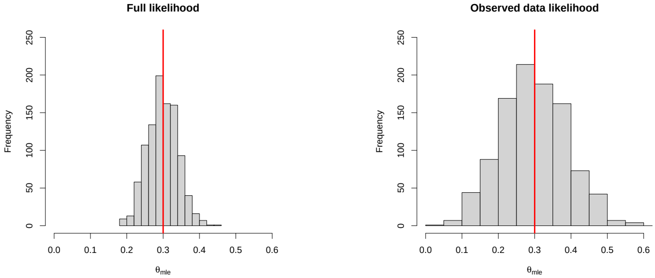
- ↪ Let us conduct a simulation study to check to which extent the estimate of θ is impacted by ignoring the missing data mechanism.

- ↪ 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 $\theta = 0.3$.

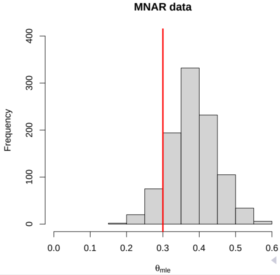


- As can be observed, in both cases there is a concentration of the maximum likelihood estimates around the true value 0.3, but ignoring the missingness mechanism leads, as expected, to more variability around the true value.

Likelihood based inference with incomplete data

Example: MNAR data

- We now investigate what happens when the MAR condition is violated.
- We assume
$$Y_i \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta), \quad \text{and} \quad \Pr(R_i = 1 \mid Y_i) = \frac{e^{Y_i}}{1 + e^{Y_i}}.$$
- We are thus violating the MAR assumption since data generated this way are MNAR (but parameters are distinct).
- We conduct a similar simulation study to check the performance of the mle estimator arising from the observed likelihood, $\widehat{\theta}_{\text{MLE}} = \sum_{i=1}^{m} Y_i/m$, in this setting.


MNAR data
Frequency
θ_{mle}