Missing Data

Outline

1. Notation and Definitions

2. Modeling Strategies
   (a) Selection models
   (b) Pattern mixture models

3. Assumptions: MCAR, MAR, Ignorable

4. Missing covariates
   (a) Selection model approach
   (b) Computational strategies

5. Example: Human Fertility
Notation:

\[ Y_i \overset{iid}{\sim} f_Y(\theta) \text{ for } i = 1, \ldots, n \]

\[ M_i = \begin{cases} 
1 & \text{if } Y_i \text{ is observed} \\
0 & \text{if } Y_i \text{ is missing}
\end{cases} \]

Interest: Inference on \( \theta \)
How to proceed?

Complete-case analysis: Focus on observed $Y$’s

Assumption: Complete cases are random subsample

Alternative: Model joint distribution of $Y$ & $M$
Suppose we send out a questionnaire to 100 Duke professors

We request information on their current body weight (bw)

Our interest is in estimating the bw distribution

Only 55 of the profs responded to the questionnaire

There are 45 missing observations
Estimated body weight distribution

![Graph showing body weight distribution with density on the y-axis and body weight (lbs) on the x-axis. The graph compares 'All profs' and 'Profs providing data' with a peak at around 200 lbs.]
Modeling the Missing Data Mechanism

Selection model:

\[ f_{Y,M}(y_i, m_i \mid \theta, \psi) = f_Y(y_i \mid \theta) f_{M|Y}(m_i \mid y_i, \psi) \]

\[ f_Y(y_i \mid \theta) = \text{Complete-data model for } Y \]

\[ f_{M|Y}(m_i \mid y_i, \psi) = \text{Model for missing data mechanism} \]
Pattern-mixture model:

\[ f_{Y,M}(y_i, m_i \mid \varphi, \pi) = f_{Y\mid M}(y_i \mid m_i, \varphi) f_M(m_i \mid \pi) \]

\[ f_{Y\mid M}(y_i \mid m_i, \varphi) = \text{conditional distribution of } Y \]
Application to Body Weight Data

Let $y_1, \ldots, y_{100}$ denote the body weights for the 100 profs surveyed.

Let $m_1, \ldots, m_{100}$ denote 0/1 indicators that the survey was returned.

**Selection Model Likelihood:**

$$
\left(2\pi\sigma^2\right)^{-n/2} \exp\left\{ - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right\} \left[ \prod_{i=1}^{n} h(y_i; \psi)^{m_i} \{1 - h(y_i; \psi)\}^{1-m_i} \right],
$$

where $h(y_i; \psi) = \Pr(M = 1 \mid Y = y_i)$.

Can we choose a non-informative prior for $\theta = (\sigma, \mu)$?
Point 1: For $m_i = 1$, $y_i$ is unknown (i.e., a latent variable)

Point 2: $\psi$ is typically not identified from the data

Point 3: Analyses with missing data must rely on mechanistic assumptions or informative priors
We simulated the prof weight data under the selection model with

\[
\begin{align*}
\mu &= 185 \\
\sigma &= 20
\end{align*}
\]

\[
Pr(M = 1 \mid Y = y_i) = h(y_i; \psi) = \frac{\exp\left\{ -1 + 2(y_i - \mu)/\sigma \right\}}{1 + \exp\left\{ -1 + 2(y_i - \mu)/\sigma \right\}}
\]

Under this model, the probability of not returning the questionnaire increases with body weight.

For example, the heavier profs may be embarassed to respond (or may respond inaccurately, which is a different issue).
Such **informative missingness** occurs commonly in biomedical studies & can lead to bias

If the missingness model is known, it is straightforward to adjust for bias

Potentially, sensitivity analyses can be conducted using a range of models
Pattern-Mixture Likelihood:

\[
\left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu - m_i\alpha)^2 \right\} \left[ \prod_{i=1}^{n} \pi^{m_i}(1 - \pi)^{1-m_i} \right].
\]

The normal mean depends on whether or not the observation is missing, which is not the case for the selection model.

For \( \alpha = 0 \), random subsample assumption holds & complete-case analysis is appropriate.

Under the random subsample assumption, the pattern-mixture & selection model likelihoods are equivalent.

Identifiability of \( \alpha \) necessarily relies on an informative prior.
**Identifying assumption**: Missing Completely At Random (MCAR):

$M$ is independent of $Y$:

$$f_Y(y_i | \theta) = f_{Y|M}(y_i | m_i, \varphi) \quad \theta = \varphi$$

$$f_{M|Y}(m_i | y_i, \psi) = f_M(m_i | \pi) \quad \psi = \pi$$

Under MCAR, Selection & Pattern-mixture models equivalent

**Alternative**: Informative prior

**Note**: When data are multivariate, MCAR assumption can be relaxed without sacrificing identifiability
Now suppose that we measure covariates \( \mathbf{X} = (X_1, \ldots, X_p)' \) for each subject

Consider the extended selection model incorporating \( \mathbf{X} \),

\[
f_{Y,M|X}(y_i, m_i | \mathbf{x}_i, \theta, \psi) = f_{Y|X}(y_i | \mathbf{x}_i, \theta) f_{M|Y,X}(m_i | y_i, \mathbf{x}_i, \psi)
\]

If \( M \) is conditionally independent of \( Y \) given \( X \), then

\[
f_{M|Y,X}(m_i | y_i, \mathbf{x}_i, \psi) = f_{M|X}(m_i | \mathbf{x}_i, \psi)
\]

and the data are said to be **Missing At Random** (MAR)
Under MAR, whether or not an observation is missing depends only on the observed data.

Under MAR, it is appropriate to **ignore** the missingness process when making inference on $\theta$ if $\theta$ is distinct from $\psi$. 
Ignorability

The observed data likelihood $L(\theta, \psi \mid y_{obs}, m, x)$ is

$$
\prod_{i=1}^{n} \int \left[ f_{Y \mid X}(y_i \mid x_i, \theta) f_{M \mid Y, X}(0 \mid y_i, x_i, \psi) \right]^{1-m_i} \\
\times \left[ f_{Y \mid X}(y_{mis} \mid x_i, \theta) f_{M \mid Y, X}(1 \mid y_{mis}, x_i, \psi) \right]^{m_i} dy_{mis}
$$

Under MAR, this reduces to

$$
\prod_{i=1}^{n} f_{M \mid Y, X}(m_i \mid x_i, \psi) f_{Y \mid X}(y_i \mid x_i, \theta)^{1-m_i} \int f_{Y \mid X}(y_{mis} \mid x_i, \theta) dy_{mis}.
$$

Hence, as long as $\theta$ and $\psi$ are distinct (i.e., have disjoint sample spaces), inference on $\theta$ can be based on the likelihood proportional to

$$
\prod_{i: m_i = 0} f_{Y \mid X}(y_i \mid x_i, \theta).
$$
Example

Suppose we know the age ($X_1$) & gender ($X_2$) of all 100 profs

Complete-data model:

$$y_i = \beta_1 + \beta_2 x_{i1} + \beta_3 x_{i2} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

Missing-data model:

$$\text{logit } \Pr(M = 1 \mid Y = y_i, X_1 = x_{i1}, X_2 = x_{i2}) = \psi_1 + \psi_2 x_{i1} + \psi_3 x_{i2}$$

Under this model, missingness is conditionally independent of weight given age & gender
• Suppose older male profs are heavier & less likely to respond.

• Then integrating $X_1, X_2$ out of the above expressions leads to a positive association between non-response & weight (as in the earlier model)

• If we are willing to assume that violations of MCAR are explained by age & gender, we can focus on the complete-case regression model.
Missing Covariates

Suppose interest focuses on relating a response $Y$ to covariates $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p)'$

- It is **extremely** common to be missing data on a subset of $\mathbf{X}$ for some of the subjects under study

- For example, suppose we included questions about medical history, demographic characteristics, and diet in the questionnaire sent to the profs. Some questions would likely be left blank.

- Standard software discards all information collected on a subject having any missing data.

- This complete-case approach can be very inefficient.
Selection Model Approach

Let \( \mathbf{m}_i = (m_{i1}, \ldots, m_{ip})' \) denote the missing data indicators for covariates \( X_1, \ldots, X_p \), respectively.

**Model components:**

1. **Conditional likelihood of response:**
   \[
   f_{Y|X}(y_i | x_i, \theta)
   \]

2. **Sampling density of \( X \):**
   \[
   f_X(x_i | \tau)
   \]

3. **Missing-data model:**
   \[
   f_{M|Y,X}(m_i | y_i, x_i, \psi)
   \]

If there were no missing data, we would focus on 1.

Typically, unless some sort of probability weighting scheme is used, \( f_X \) needs to be modeled even if interest is in \( \theta \).
Imputing the Missing Values

A Simple Bootstrap Approach:

1. Fill in the missing X’s by drawing a random sample with replacement from the empirical distribution of the observed X’s.
2. Repeat the process multiple times to produce multiple “complete” data sets.
3. Fit each complete data set using standard software.
4. Average the parameter estimates & obtain a variance estimate that accounts for within- and between-imputation uncertainty.

Very easy to implement & does not require specification of $f_X$ -
Unfortunately, does require MCAR

Numerous imputation schemes have been proposed in the literature.
Data Augmentation within MCMC

Choose a prior for $\theta, \tau, \psi$.

Consider the unknown $X$’s as latent data

Apply the following MCMC algorithm:

1. Impute the missing $X$’s by sampling from their full conditional distribution.

2. Conditional on the completed data, follow standard Gibbs sampling (or other) steps for updating the parameters $\theta, \tau, \psi$.

3. Repeat steps 1-2 for a large number of iterations.

Often, the conditional distributions required for implementation of this algorithm follow a simple form.
Human Fertility Modeling Example

Data collected for women attempting pregnancy

\[ y_{ij} = \begin{cases} 
1 & \text{if conception occurs in cycle } j \\
0 & \text{otherwise} 
\end{cases} \]

\[ x_{ijk} = \begin{cases} 
1 & \text{if intercourse on day } k \text{ of cycle } j \\
0 & \text{otherwise} 
\end{cases} \]

\[ m_{ijk} = \begin{cases} 
1 & \text{if missing intercourse data for day } k \\
0 & \text{otherwise} 
\end{cases} \]

**Conception Probability Model:**

\[ \Pr(y_{ij} = 1 \mid x_{ij}) = \omega \{1 - \prod_{k=1}^{K} (1 - p_k)^{x_{ijk}}\} \]

**Intercourse Occurrence Model:**

\[ \Pr(x_{ijk} = 1) = \pi_k, \quad \text{for } k = 1, \ldots, K \]

**Missing Data Model:**

\[ \Pr(m_{ijk} = 1 \mid x_{ijk} = 0) = \lambda_0, \quad \Pr(m_{ijk} = 1 \mid x_{ijk} = 1) = \lambda_1 \]
Once the missing intercourse indicators are filled in, the Gibbs sampling algorithm of Dunson & Zhou (2000, JASA) can be used for posterior computation.

To extend the Dunson & Zhou algorithm, we simply add a step to impute the missing intercourse indicators by sampling from their full conditional distribution, which is

\[
[m_{ijk} \mid y_{ij}, x_{ijk}, \cdot] \sim \text{Bernoulli} \left( \frac{\Pr(Y = y_{ij} \mid x_{ijk} = 1, \cdot) \lambda_k^*}{\Pr(Y = y_{ij} \mid x_{ijk} = 1, \cdot) \lambda_k^* + \Pr(Y = y_{ij} \mid x_{ijk} = 0, \cdot)(1 - \lambda_k^*)} \right),
\]

where

\[
\lambda_k^* = \frac{\lambda_1 \pi_k}{\lambda_1 \pi_k + \lambda_0(1 - \pi_k)}.
\]

A similar strategy can be used to accommodate informative missingness in a broad variety of applications.
Estimated Pregnancy Probabilities

Pr(Pregnancy | Cycle Viable)

Day Relative to Ovulation

Uncorrected
Corrected
Summary

- Selection and Pattern mixture models can be used to accommodate missing data.

- Under MCAR or MAR assumptions, the missing data mechanism is typically ignorable.

- In missing covariate applications, there can be a substantial loss of efficiency associated with discarding of subjects with missing observations - even under MCAR or MAR.

- To avoid restrictive assumptions, one often needs to model the missing data mechanism.

- Given lack of identifiability, informative priors should be used & sensitivity analyses conducted.
Assignment:

1. Complete the following exercise: Suppose $y_i \sim N(\mu, \sigma^2)$ and let $m_i$ be an indicator that $y_i$ is missing, for $i = 1, \ldots, n$. Suppose that it is reasonable to assume that

$$\text{logit} \Pr(m_i = 1 \mid y_i, x_i) = \alpha + x_i'\beta,$$

where $x_i = (x_{i1}, \ldots, x_{ip})'$ is a vector of known covariates.

(a) Generate data under the above model with $y_i \sim N(x_i'\theta, 1)$, where $x_i \sim N_5(\mathbf{0}, I_{5 \times 5})$, $\theta = \beta = (1, 1, 1, 1, 1)'$, $\alpha = -1$, and $n = 250$.

(b) Estimate $\mu$ by maximum likelihood using a complete case analysis.

(c) Does MCAR hold? Why or why not?

(d) Does MAR hold? Why or why not?

(e) Is the complete case estimate unbiased?

(f) Propose a strategy for posterior computation of $\mu, \sigma$:
   i. With the sampling density of $x_i$ assumed MVN.
   ii. With the sampling density of $x_i$ unspecified.