Lecture 12 | 19.05.2025

Missing data
in longitudinal data

Missing data in general

Missing data occur in all kinds of statistical models but their particular importance is mainly relevant in longitudinal studies...

Formally, there are three main concepts/mechanisms distinguished...

- **Missing Completely at Random (MCAR)**
  - missingness structure is unrelated to observed or unobserved data
  - the least problematic (from the theoretical/practical point of view)

- **Missing Not at Random (MNAR)**
  - missingness depends on unobserved values themselves
  - theoretically/technically much more challenging concept to handle

Missing data in general

Missing data occur in all kinds of statistical models but their particular importance is mainly relevant in longitudinal studies...

Formally, there are three main concepts/mechanisms distinguished...

- **Missing Completely at Random (MCAR)**
  - missingness structure is unrelated to observed or unobserved data
  - the least problematic (from the theoretical/practical point of view)
- **Missing Not at Random (MNAR)**
  - missingness depends on unobserved values themselves
  - theoretically/technically much more challenging concept to handle
- **Missing at Random (MAR)**
  - missingness depends on the observed data but not on unobserved
  - in a sense, it is a bridge between the MCAR and MNAR concepts

Missing data in general

Missing data occur in all kinds of statistical models but their particular importance is mainly relevant in longitudinal studies...

Formally, there are three main concepts/mechanisms distinguished...

- **Missing Completely at Random (MCAR)**
  - missingness structure is unrelated to observed or unobserved data
  - the least problematic (from the theoretical/practical point of view)
- **Missing Not at Random (MNAR)**
  - missingness depends on unobserved values themselves
  - theoretically/technically much more challenging concept to handle
- **Missing at Random (MAR)**
  - missingness depends on the observed data but not on unobserved
  - in a sense, it is a bridge between the MCAR and MNAR concepts

$\Longrightarrow$ **but there is a difference between missing data and unbalanced data**

Missing data – formally

- let $\boldsymbol{Y} = (\boldsymbol{Y}_{(o)}^\top, \boldsymbol{Y}_{(m)}^\top)^\top$ be a vector off all possible values that can be observed with no missingness (i.e., a hypothetical quantity/vector)
- vector $\boldsymbol{Y}_{(o)}$ denotes the observations that are actually observed/available and $\boldsymbol{Y}_{(m)}$ denotes the vector of all unobserved/missing values
- let $\boldsymbol{R}$ denote the vector of indicator (0/1) random variables (with the same length as $\boldsymbol{Y}$) denoting which element of $\boldsymbol{Y}$ is missing (unobserved)

Missing data – formally

- let $\boldsymbol{Y} = (\boldsymbol{Y}_{(o)}^\top, \boldsymbol{Y}_{(m)}^\top)^\top$ be a vector off all possible values that can be observed with no missingness (i.e., a hypothetical quantity/vector)
- vector $\boldsymbol{Y}_{(o)}$ denotes the observations that are actually observed/available and $\boldsymbol{Y}_{(m)}$ denotes the vector of all unobserved/missing values
- let $\boldsymbol{R}$ denote the vector of indicator (0/1) random variables (with the same length as $\boldsymbol{Y}$) denoting which element of $\boldsymbol{Y}$ is missing (unobserved)

- All three concepts of missingness (MCAR, MNAR, and MAR) can be formalized using the joint distribution of the random vector $\boldsymbol{R}$
  - **MCAR** – the distribution of $\boldsymbol{R}$ is independent of $\boldsymbol{Y}$
  - **MAR** – the distribution of $\boldsymbol{R}$ is independent of $\boldsymbol{Y}_{(m)}$
  - **MNAR** – the distribution of $\boldsymbol{R}$ depends on $\boldsymbol{Y}_{(m)}$

Missing data – formally

- let $\boldsymbol{Y} = (\boldsymbol{Y}_{(o)}^\top, \boldsymbol{Y}_{(m)}^\top)^\top$ be a vector off all possible values that can be observed with no missingness (i.e., a hypothetical quantity/vector)
- vector $\boldsymbol{Y}_{(o)}$ denotes the observations that are actually observed/available and $\boldsymbol{Y}_{(m)}$ denotes the vector of all unobserved/missing values
- let $\boldsymbol{R}$ denote the vector of indicator (0/1) random variables (with the same length as $\boldsymbol{Y}$) denoting which element of $\boldsymbol{Y}$ is missing (unobserved)

- All three concepts of missingness (MCAR, MNAR, and MAR) can be formalized using the joint distribution of the random vector $\boldsymbol{R}$
  - **MCAR** – the distribution of $\boldsymbol{R}$ is independent of $\boldsymbol{Y}$
  - **MAR** – the distribution of $\boldsymbol{R}$ is independent of $\boldsymbol{Y}_{(m)}$
  - **MNAR** – the distribution of $\boldsymbol{R}$ depends on $\boldsymbol{Y}_{(m)}$

$\Longrightarrow$ Little & Rubin (1987). **Statistical Analysis with Missing Data**
`https://onlinelibrary.wiley.com/doi/book/10.1002/9781119013563`

Likelihood methods for missing data

- **AIM:** Statistical inference (accounting for possibly missing/incomplete observations) based on the classical likelihood principles...
- **Starting point:** the joint density $(\boldsymbol{Y}_{(o)}^\top, \boldsymbol{Y}_{(m)}^\top, \boldsymbol{R}^\top)^\top \sim f(\boldsymbol{y}_o, \boldsymbol{y}_m, \boldsymbol{y}_r)$

Likelihood methods for missing data

- **AIM:** Statistical inference (accounting for possibly missing/incomplete observations) based on the classical likelihood principles...
- **Starting point:** the joint density $(\boldsymbol{Y}_{(o)}^\top, \boldsymbol{Y}_{(m)}^\top, \boldsymbol{R}^\top)^\top \sim f(\boldsymbol{y}_o, \boldsymbol{y}_m, \boldsymbol{y}_r)$
- Standard factorization in terms of conditioning gives

$$f(\boldsymbol{y}_o, \boldsymbol{y}_m, \boldsymbol{y}_r) = f(\boldsymbol{y}_o, \boldsymbol{y}_m) \cdot f(\boldsymbol{y}_r | \boldsymbol{y}_o, \boldsymbol{y}_m)$$

where for the likelihood-based inference we need the joint distribution/density of the observed part of the data (i.e., $(\boldsymbol{Y}_{(o)}^\top, \boldsymbol{R}^\top)^\top$)

$$f(\boldsymbol{y}_o, \boldsymbol{y}_r) = \int f(\boldsymbol{y}_o, \boldsymbol{y}_m) \cdot f(\boldsymbol{y}_r | \boldsymbol{y}_o, \boldsymbol{y}_m) \mathrm{d}\boldsymbol{y}_m$$

Likelihood methods for missing data

- **AIM:** Statistical inference (accounting for possibly missing/incomplete observations) based on the classical likelihood principles...
- **Starting point:** the joint density $(\boldsymbol{Y}_{(o)}^\top, \boldsymbol{Y}_{(m)}^\top, \boldsymbol{R}^\top)^\top \sim f(\boldsymbol{y}_o, \boldsymbol{y}_m, \boldsymbol{y}_r)$
- Standard factorization in terms of conditioning gives

$$f(\boldsymbol{y}_o, \boldsymbol{y}_m, \boldsymbol{y}_r) = f(\boldsymbol{y}_o, \boldsymbol{y}_m) \cdot f(\boldsymbol{y}_r | \boldsymbol{y}_o, \boldsymbol{y}_m)$$

  where for the likelihood-based inference we need the joint distribution/density of the observed part of the data (i.e., $(\boldsymbol{Y}_{(o)}^\top, \boldsymbol{R}^\top)^\top$)

$$f(\boldsymbol{y}_o, \boldsymbol{y}_r) = \int f(\boldsymbol{y}_o, \boldsymbol{y}_m) \cdot f(\boldsymbol{y}_r | \boldsymbol{y}_o, \boldsymbol{y}_m) \mathrm{d}\boldsymbol{y}_m$$

- If the **missing concept is random**, the density function $f(\boldsymbol{y}_r | \boldsymbol{y}_o, \boldsymbol{y}_m)$ does not depend on the argument $\boldsymbol{y}_m$, therefore

$$f(\boldsymbol{y}_o, \boldsymbol{y}_r) = f(\boldsymbol{y}_r | \boldsymbol{y}_o) \cdot f(\boldsymbol{y}_o)$$

$\Longrightarrow$ thus, in some literature, the MCAR and MAR concepts are not distinguished that much carefully (random vs. informative missing)

Likelihood vs. GEE

- Likelihood based methods are derived from the (overall) joint distribution of the observed data and, therefore, the random missingness concept (MCAR or MAR) is enough to ensure a valid statistical inference

- Less restricted inference techniques which do not utilize the whole distribution (GEE for instance) require a slightly stronger assumption to guarantee a valid statistical inference (MCAR only)

Likelihood vs. GEE

- Likelihood based methods are derived from the (overall) joint distribution of the observed data and, therefore, the random missingness concept (MCAR or MAR) is enough to ensure a valid statistical inference

- Less restricted inference techniques which do not utilize the whole distribution (GEE for instance) require a slightly stronger assumption to guarantee a valid statistical inference (MCAR only)

- **MCAR**
  – it is generally easier (i.e., less problematic) to delete the rows with missing observations—the analysis of the remaining cases should remain unbiased (imputation methods can be used as an option)
- **MAR**
  – deleting the rows with missing observations may introduce an additional bias—imputation methods become more important (moreover, unlike MCAR, it can not be statistically tested for)

Dropouts vs. intermittents

- **dropouts** – if some Y_j is missing, so are all observations Y_t, where $t \geq j$
  - mostly due to the loss of the whole follow-up process (for whatever reason but very often it is directly related to the main question of interest)
  - any relationship between the measurement process (the process of the main interest) and the dropout process can be problematic
  - complex methods incorporating missing data are often used

Dropouts vs. intermittents

- **dropouts** – if some Y_j is missing, so are all observations Y_t, where $t \geq j$
  - mostly due to the loss of the whole follow-up process (for whatever reason but very often it is directly related to the main question of interest)
  - any relationship between the measurement process (the process of the main interest) and the dropout process can be problematic
  - complex methods incorporating missing data are often used

- **intermittents** – occasional missingness that is not classified as dropout
  - huge variety of different reasons but very typically not related to the main question of interest
  - very often the the reasons for missingness are well-known (the patient is still followed and other information can be collected)
  - methods for unbalanced data can be often used directly

Dealing with missing observations

- **Simple/naive methods**
  - last observations carried forward
    (not only in terms of the values of Y)
  - complete case analysis
    (loss of some substantial information)

- **Imputation methods**
  - likelihood-based techniques
    (Bayesian techniques, EM algorithm)
  - model-based methods
    (mean/median/regression imputation, K-NN)

Statistical inference on missingness

MCAR

- exploratory and confirmatory analysis is possible
- various statistical approaches can be constructed (typically by comparing the means of observed vs. missing groups for multiple variables)
- Little's MCAR test

MAR

- no formal statistical test for MAR because MAR is, by its definition, about the unobserved data only
- logistic regression is typically used for predicting the missing observations

MNAR

- can't be tested nor explored directly from the data alone
- assumption imposed from the theory, design experiment, data collection mechanism, etc.

GEE extension for MAR

- GEE models are considered to be very general modeling techniques as they only require the correct mean specification for consistency
- The final model is given as a solution to the estimating equations

$$\sum_{i=1}^{N} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \beta} \right)^{\top} \left[Var\, \boldsymbol{Y}_i \right]^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0}$$

GEE extension for MAR

- GEE models are considered to be very general modeling techniques as they only require the correct mean specification for consistency
- The final model is given as a solution to the estimating equations

$$\sum_{i=1}^{N} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \beta} \right)^\top \left[Var\, \boldsymbol{Y}_i \right]^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0}$$

- Let $p_{ij} \in (0,1)$ denotes the probability that subject $i \in \{1, \dots, N\}$ drops out at the time point $j \in \{1, \dots, n_i\}$ (given the subject's history)
- Modified GEE consistent under MAR is given by the estimating equations

$$\sum_{i=1}^{N} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \beta} \right)^\top \left[Var\, \boldsymbol{Y}_i \right]^{-1} \mathbb{P}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0}$$

with the diagonal matrix $\mathbb{P}_i = diag\{p_{i1}, \dots, p_{in_i}\}$

Statistical models for MNAR

- There are also some ideas to deal with the informative missing data...
- Common approaches involve selection models and pattern mixture models

Statistical models for MNAR

- There are also some ideas to deal with the informative missing data...
- Common approaches involve selection models and pattern mixture models

Selection models

- potential outcomes $\boldsymbol{Y}^\star = (Y_1^\star, \dots, Y_n^\star)^\top$, observed $\boldsymbol{Y} = (Y_1, \dots, Y_n)^\top$, dropout time $D \in \{1, \dots, n\}$, such that $Y_j = Y_j^\star$, for $j < D$
- the model for dropouts is selected from the observed history
- factorization $\quad P(\boldsymbol{Y}^\star, D) = P(\boldsymbol{Y}^\star) \cdot P(D|\boldsymbol{Y}^\star)$

Pattern mixture models

- potential outcomes $\boldsymbol{Y}^\star = (Y_1^\star, \dots, Y_n^\star)^\top$, observed $\boldsymbol{Y} = (Y_1, \dots, Y_n)^\top$, dropout time $D \in \{1, \dots, n\}$, such that $Y_j = Y_j^\star$, for $j < D$
- dropout process is predetermined and $\boldsymbol{Y}^\star$ is modeled given the dropouts
- factorization $\quad P(\boldsymbol{Y}^\star, D) = P(D) \cdot P(\boldsymbol{Y}^\star|D)$

Statistical models for MNAR

- There are also some ideas to deal with the informative missing data...
- Common approaches involve selection models and pattern mixture models

Selection models

- potential outcomes $\boldsymbol{Y}^\star = (Y_1^\star, \dots, Y_n^\star)^\top$, observed $\boldsymbol{Y} = (Y_1, \dots, Y_n)^\top$, dropout time $D \in \{1, \dots, n\}$, such that $Y_j = Y_j^\star$, for $j < D$
- the model for dropouts is selected from the observed history
- factorization $\quad P(\boldsymbol{Y}^\star, D) = P(\boldsymbol{Y}^\star) \cdot P(D|\boldsymbol{Y}^\star)$

Pattern mixture models

- potential outcomes $\boldsymbol{Y}^\star = (Y_1^\star, \dots, Y_n^\star)^\top$, observed $\boldsymbol{Y} = (Y_1, \dots, Y_n)^\top$, dropout time $D \in \{1, \dots, n\}$, such that $Y_j = Y_j^\star$, for $j < D$
- dropout process is predetermined and $\boldsymbol{Y}^\star$ is modeled given the dropouts
- factorization $\quad P(\boldsymbol{Y}^\star, D) = P(D) \cdot P(\boldsymbol{Y}^\star|D)$

$\Longrightarrow$ various simplifications assumptions and different model structures used for the quantities $P(\boldsymbol{Y}^\star), P(D), P(D|\boldsymbol{Y}^\star)$, and $P(\boldsymbol{Y}^\star|D)$

To conclude...

MCAR

- Missingness is independent of both observed and unobserved data

$$P(\boldsymbol{Y}|\boldsymbol{Y}_{(o)}, \boldsymbol{Y}_{(m)}) = P(\boldsymbol{R})$$

- Missing data may be (generally) ignored
- Different options for handling the missing cases (including deletion)
- Generally do not induce estimation bias and can be statistically tested for

MAR

- Missingness depends only on observed data

$$P(\boldsymbol{Y}|\boldsymbol{Y}_{(o)}, \boldsymbol{Y}_{(m)}) = P(\boldsymbol{Y}|\boldsymbol{Y}_{(o)})$$

- Missing data can be in some cases ignored
- All kinds of imputation methods are proposed to handle missing cases
- The model estimates can be unbiased if the model is correctly proposed

MNAR

- Missingness depends on unobserved data (including the missing values)

$$P(\boldsymbol{Y}|\boldsymbol{Y}_{(o)}, \boldsymbol{Y}_{(m)}) \neq P(\boldsymbol{Y}|\boldsymbol{Y}_{(o)})$$

- Missing data can not be ignored
- Selection models, Pattern mixture models, or some sensitivity analysis
- The estimates of the model are very often substantially biased

NMST422 – exam terms

1. **22.05.2025 (Tuesday)** (starting at 10:40 at K4)
2. **29.05.2025 (Thursday)** (starting at 09:00 at K11)
3. **02.06.2025 (Monday)** (starting at 09:00 at K11)
4. **09.06.2025 (Monday)** (starting at 09:00 at K11)
5. **26.06.2025 (Thursday)** (starting at 09:00 at K11)

6. **One more exam term in September** (TBD)