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 FC FACULDADE DE CIÊNCIAS UNIVERSIDADE DO PORTO

Formal description of the missing data mechanisms

Notation and terminology

- ↪ We have already informally introduced the concepts of MCAR, MAR, and MNAR.
- ↪ We will now look at more precise definitions of these mechanisms. To do so, we need to introduce some notation and terminology.
- ↪ The complete data consist of the values one would have obtained if there were no missing data and we denote it by $\mathbf{Y}$.
- ↪ The complete data is partially a hypothetical entity because some of its values might be missing.
- ↪ We write $\mathbf{Y} = (\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}})$, where $\mathbf{Y}_{\text{obs}}$ and $\mathbf{Y}_{\text{mis}}$ denote the observed components and the missing components of $\mathbf{Y}$, respectively.

Formal description of the missing data mechanisms

Notation and terminology

$\hookrightarrow$ Let **R** be the missingness indicator. Assuming $\mathbf{Y} \in \mathbb{R}^{n \times p}$ (assume n is the number of subjects and p the number of variables), **R** has also dimension $n \times p$ and it is defined as

$$R_{ij} = \begin{cases} 1 & \text{if } Y_{ij} \text{ has been observed,} \\ 0 & \text{if } Y_{ij} \text{ is missing.} \end{cases}$$

$\hookrightarrow$ The missing data model is a model for the conditional distribution of **R** given **Y**. Let $f(\mathbf{r} \mid \mathbf{y}, \psi)$ denote the probability that $\mathbf{R} = \mathbf{r}$ given that $\mathbf{Y} = \mathbf{y}$ according to this model, where ψ is an unknown parameter. Here, **r** and **y** are particular values that might be taken by **R** and **Y**.

Formal description of the missing data mechanisms

MCAR

- Data are said to be MCAR if

$$f(\mathbf{r} \mid \mathbf{y}, \psi) = f(\mathbf{r} \mid \psi), \quad \forall \mathbf{y}, \psi,$$

  that is, under MCAR the missing data model is completely unrelated to the data, observed or missing. It only depends on some parameter ψ, the overall probability of missingness.

- As it was already noted:
  - The essential feature of MCAR is that the observed data can be thought of as a random sample of the complete data.
  - The validity of MCAR can be checked from the data at hand against the alternative MAR, but we can never rule out MNAR.

Formal description of the missing data mechanisms

MAR

$\hookrightarrow$ Data are said to be MAR if

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

that is, under MAR the probability of the pattern of missing data only depends on the observed data.

$\hookrightarrow$ As it was already noted:

- $\hookrightarrow$ Within strata defined by $\mathbf{Y}_{\text{obs}}$, missingness is MCAR.
- $\hookrightarrow$ The validity of the MAR assumption cannot be checked from the data at hand against MNAR.

Formal description of the missing data mechanisms
MNAR

- Finally, data are said to be MNAR if

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

  that is, the probability of the missing data pattern depends on the unobserved data and may depend also on the observed data.

- A complicated form of MNAR is when missingness depends on a completely unobserved/unmeasured variable.

- For a concrete example, think again on the BMI/glucose level example. Suppose that the true missing mechanism for BMI is MAR, hence meaning that individuals with missing values of BMI may be more likely to have extreme blood glucose levels. However, the MAR missing values in BMI would become MNAR if we had no measurements of glucose at all.

Formal description of the missing data mechanisms

WARNING

- ↪ The previous definitions are widely used in the literature.
- ↪ However it is unclear whether the equations hold for the realised missing data pattern or for any realisation $(\mathbf{y}, \mathbf{r})$ of $(\mathbf{Y}, \mathbf{R})$ (i.e, for any missing patterns or observed data that could have been realised but were not), although it is widely understood as the latter.
- ↪ Seaman et al. (2013, Statistical Science) proposed two definitions of the MAR mechanism for which they differentiate if (i) the statements hold for any possible missing data pattern, which they denominate as *everywhere MAR*, or (ii) for the realised pattern, leading to what they denominate as *realised* MAR.

Formal description of the missing data mechanisms

Ignorability versus nonignorability

- The ψ parameter of the missing data model has little scientific interest (e.g., had the data been complete there would be no reason to worry about ψ) and is generally unknown.

- It would greatly simplify the analysis if we could just ignore this parameter. However, in some situations, this parameter may influence the estimate of the parameter of interest, the parameter, say $\boldsymbol{\theta}$, of the data model $f(\mathbf{y} \mid \boldsymbol{\theta})$.

- The practical importance of Rubin's distinction between MCAR, MAR, and MNAR is that it clarified the conditions that need to exist in order to accurately estimate $\boldsymbol{\theta}$ without the need to know ψ.

Formal description of the missing data mechanisms

Ignorability versus nonignorability

↪ Rubin showed that likelihood based analyses (e.g., maximum likelihood) and multiple imputation do not require information about $\boldsymbol{\psi}$ if:

1. the data are MAR or MCAR, and
2. the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are distinct, in the sense that the joint parameter space of $(\boldsymbol{\psi}, \boldsymbol{\theta})$ is the product of the parameter space of $\boldsymbol{\psi}$ and the parameter space of $\boldsymbol{\theta}$.

↪ Schafer (1997, p.11) says that in many situations the second condition is, at least, reasonable from an intuitive point of view, given that knowing $\boldsymbol{\theta}$ will provide little information about $\boldsymbol{\psi}$ and vice-versa.

↪ For this reason, missing data literature often describes MAR (and MCAR!) data as ignorable. Although strictly speaking, we still need (2), not only (1). We will study this more carefully later in the course.

Further considerations on missingness meachnisms

A little more on checking MCAR versus MAR

- ↪ MCAR is the only missingness mechanism that, to a certain extent, yields testable propositions.
- ↪ Several tests have been proposed in the literature to check whether missingness is consistent with the MCAR assumption.
- ↪ These tests are not routinely used and their practical value is still not clear. Remember that we can never rule out the possibility of MNAR.
- ↪ Enders (2010, pp. 17–21) contains a good discussion of this topic.

Further considerations on missingness meachnisms

A little more on checking MCAR versus MAR

- ↪ One popular and simple option is to perform a series of *t*-tests.
- ↪ This approach separates the missing and observed values on a particular variable and uses a *t*-test to examine group mean differences in the two groups induced by such splitting in the other variables in the dataset.
- ↪ The MCAR mechanism implies that such two groups should be similar on average.
- ↪ As a consequence, a significant *t*-test (i.e., rejecting the null hypothesis that the means of the two groups are equal) provides evidence *against* MCAR.
- ↪ The main advantage for implementing the *t*-test approach is to identify (auxiliar) variables that we can later adjust for in the missing data handling procedure.
- ↪ Alternatively, one may use density plots and boxplots to visualise the distributions of the two groups.

Further considerations on missingness meachnisms

A little more on checking MCAR versus MAR

- ↪ As the number of variables grows, computing the t-test statistics can be cumbersome.
- ↪ Little (1988) proposed a multivariate version of the *t*-test that simultaneously evaluates mean differences on every variable in the data set. It is a global test of MCAR that applies to the entire dataset.
- ↪ For details see Little (1988) or Enders (2010, pp. 19–20). This can be carried out by the `LittleMCAR` function in the `BaylorEdPsych` `R` package.

Further considerations on missingness meachnisms

How to prevent MNAR missingness?

- ↪ As we had already quoted, the ideal solution to the missing data problem would be to have none.
- ↪ Missing data prevention requires a careful experiment's design and a very careful execution as well.
- ↪ Most of the methods we will cover assume MAR data. However, we cannot be sure whether the data are really missing at random, or whether the missingness depends on unobserved variables or the missing data themselves.
- ↪ The idea is to start the study with a data collection strategy that will turn MNAR missingness into MAR missingness.
- ↪ This, so called inclusive analysis strategy, incorporates variables that are known to be correlated with the missing prone variables. Then, missing values will be more likely to be MAR than MNAR.
- ↪ These correlated variables are called auxiliary variables in the missing data literature.

Further considerations on missingness meachnisms

How to prevent MNAR missingness?

↪ Note that auxiliary variables might not be of substantial interest in the sense that they would not have been included in the analysis had the data been complete.

↪ Theory and past research, as well as the MCAR tests/visualisation checks, can help to identify auxiliary variables.

↪ Note that the inclusion of auxiliary variables *per se* does not guarantee that the MAR assumption is satisfied, but it certainly improves the chances of it.

↪ For instance, it may be a strong assumption that nonresponse to an income question in a survey depends only on gender, race and education, but this is certainly a lot more plausible than assuming the probability of nonresponse is constant, or that it depends only on one of these variables.

Further considerations on missingness meachnisms

Planned missing data designs

- ↪ In the first week, we have seen an example (nutrition study) where missing, instead of out of the researcher control, was ‘induced’ on purpose.
- ↪ The idea of planned missing data design is to intentionally generate MCAR or MAR data.
- ↪ For example, in a randomised study with two treatments (e.g., active treatment vs. placebo), each individual has a hypothetical score on both treatments, but participants only provide a response to their assigned treatment. The unobserved response to the other treatment is MCAR.
- ↪ A classic example of intentional MAR data occurs in selection designs where values on one variable determine whether respondents provide data on a second variable.
- ↪ For instance, universities often use exam(s)’ marks as a selection tool for admissions, so first year marks are subsequently missing for students who scored below the admissions threshold.