Incomplete Data Analysis

Vanda Inácio

University of Edinburgh

Semester 1, 2020/2021

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 $\mathbf{Y} = (\mathbf{Y}_{obs}, \mathbf{Y}_{mis})$, where $\mathbf{Y}_{obs}$ and $\mathbf{Y}_{mis}$ denote the observed components and the missing components of $\mathbf{Y}$, respectively.

Formal description of the missing data mechanisms

Notation and terminology

↪ 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

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

↪ 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

↪ 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.

↪ As it was already noted:

- ↪ Within strata defined by $\mathbf{Y}_{\text{obs}}$, missingness is MCAR.
- ↪ 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.

Ignorability versus nonignorability

- ↪ The ψ parameter of the missing data model have no 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 θ, of the data model $f(\mathbf{y} \mid \theta)$.
- ↪ 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 θ without the need to know ψ.

Ignorability versus nonignorability

↪ Rubin showed that likelihood based analyses (e.g., maximum likelihood) and multiple imputation do not require information about ψ if:

1. the data are MAR or MCAR, and
2. the parameters θ and ψ are distinct, in the sense that the joint parameter space of (ψ, θ) is the product of the parameter space of ψ and the parameter space of θ.

↪ 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 θ will provide little information about ψ 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.