Incomplete Data in Clinical and Epidemiological Studies

Geert Molenberghs
I-BioStat, Universiteit Hasselt & Katholieke Universiteit Leuven, Belgium.

In many longitudinal and multivariate settings, not all measurements planned are taken in actual practice. It is important to reflect on the nature and implications of such incompleteness, or missingness, and properly accommodate it in the modeling process.

When referring to the missing-value process we will use terminology of Little and Rubin (2002, Chapter 6). A non-response process is said to be *missing completely at random* (MCAR) if the missingness is independent of both unobserved and observed data and *missing at random* (MAR) if, conditional on the observed data, the missingness is independent of the unobserved measurements. A process that is neither MCAR nor MAR is termed *non-random* (MNAR).

Given MAR, a valid analysis that ignores the missing value mechanism can be obtained, within a likelihood or Bayesian framework, provided the parameters describing the measurement process are functionally independent of the missingness model parameters, the so-called parameter distinctness condition. This situation is termed ignorable by Rubin (1976) and Little and Rubin (2002) and leads to considerable simplification in the analysis (Verbeke and Molenberghs 2000). There is a strong trend, nowadays, to prefer this kind of analyses, in the likelihood context also termed *direct-likelihood* analysis, over *ad hoc* methods such as *last observation carried forward* (LOCF), *complete case analysis* (CC), or simple forms of imputation (Molenberghs and Kenward 2007). Practically, it means conventional tools for longitudinal and multivariate data, such as the linear and generalized linear mixed-effects models (Verbeke and Molenberghs 2000, Molenberghs and Verbeke 2005) can be used in exactly the same way as with complete data. Software tools like the SAS procedures MIXED, NLMIXED, and GLIMMIX facilitate this paradigm shift.

In spite of direct likelihood's elegance, fundamental model assessment and model selection issues remain. Such issues, occurring under MAR and even more under MNAR, are the central theme of this paper.

Indeed, one can never fully rule out MNAR, in which case the missingness mechanism needs to be modeled alongside the mechanism generating the responses. In the light of this, one approach could be to estimate from the available data the parameters of a model representing a MNAR mechanism. It is typically difficult to justify the particular choice of missingness model, and the data do not necessarily contain information on the parameters of the particular model chosen (Molenberghs and Kenward 2007). For example, different MNAR models may fit the observed data equally well, but have quite different implications for the unobserved measurements, and hence for the conclusions to be drawn from the respective analyses. Without additional information one can only distinguish between such models using their fit to the observed data, and so goodness-of-fit tools typically do not provide a relevant means of choosing between such models. It follows that there is an important role for sensitivity analysis in assessing inferences from incomplete data (Verbeke and Molenberghs 2000, Molenberghs and Verbeke 2005, and Molenberghs and Kenward 2007).

References

Little, R.J.A. and Rubin, D.B. (2002) *Statistical Analysis with Missing Data* (2nd ed.). New York: John Wiley & Sons.

Molenberghs, G. and Kenward, M.G. (2007) *Missing Data in Clinical Studies.* Chichester: John Wiley

& Sons.

Molenberghs, G. and Verbeke, G. (2005) *Models for Discrete Longitudinal Data.* New York: Springer.

Rubin, D.B. (1976) Inference and missing data. *Biometrika*, **63**, 581–592.

Verbeke, G. and Molenberghs, G. (2000) *Linear Mixed Models for Longitudinal Data.* New York: Springer.

Acknowledgment

Based on an article from Lovric, Miodrag (2011), International Encyclopedia of Statistical Science. Heidelberg: Springer Science +Business Media, LLC