

Multiple imputation tailored for a specific database

N.T. Longford*, M.G. Pittau**

*SNTL and Departament d'Economia i Empresa, Universitat Pompeu Fabra, Ramon Trias Fargas 25–27, 08005 Barcelona, Spain. NTL@sntl.co.uk

**Dip. di Statistica, Probabilità e Statistiche Applicate, Univ. “La Sapienza”, Roma. grazia.pittau@uniroma1.it

Since the pioneering work on dealing with missing data by Rubin in the 1970's, encompassing the EM algorithm and multiple imputation, the general ideas have been implemented in many substantive applications, and numerous extensions have been devised. Despite the proliferating modules for multiple imputation in various packages, a principled application to large-scale datasets remains a formidable challenge. In our perspective, such modules are close to reaching the limits how general and universal they can be, and a tailor-made customised software may be much more effective for the task of multiply completing the database of a large scale survey, thus enabling the community of secondary users to conduct efficient analyses without requiring any expertise in the methods for incomplete data.

The presentation will describe a suite of functions written in R for dealing with the missing values in the General Household Survey for the 2006 cross-section.

The suite includes a function for rearranging the variables into an order best suited for the sequence of univariate imputations (chained equations) used to generate the sets of plausible values (replacements for missing values). The model-fitting core of the suite comprises functions for ordinary regression, sequence of logistic regressions for categorical variables, and a combination of ordinary and logistic regressions for semi-continuous data. Dealing with ill-conditioning (near-confounding) can be dealt with automatically, but good understanding of how the data are recorded may spare the analyst a lot of frustration. Exploration of the response patterns is very useful for this.

The chained equations involve cycles over the incompletely recorded variables. There are several options for assessing whether a sufficient number of cycles has been run. We describe some approaches based on subsets of the data. The completed datasets can be stored in the extended format, as an R-list of datasets, but storing only the sets of imputed values is an economic alternative.

Extensions for longitudinal (panel) data are based on the concept of group-monotone response patterns, assuming that dropping out is the dominant response pattern, in addition to varied response patterns within a wave (time point). The suite is available from the authors, with the understanding that it is unlikely to work flawlessly and effectively for a

different database without a suitable adaptation.

References

- Rubin, D.B. (2002). Multiple Imputation for Nonresponse in *Surveys*, 2nd ed. Wiley and Sons, New York
- Little, R.J.A., and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. Wiley and Sons, New York
- Longford, N.T. (2005). *Missing Data and Small-Area Estimation. Modern Analytical Equipment for the Survey Statistician*. Springer-Verlag, New York.
- van Buuren, S., Boshuizen, H.C., and Knook, D.L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine* 18, 681-694.