Insights into the use of Bayesian models for informative missing data

Alexina Mason[1], Nicky Best[1], Sylvia Richardson[1] and Ian Plewis[2]

[1] Department of Epidemiology and Public Health, Imperial College London, UK
[2] Centre for Census and Survey Research, University of Manchester, UK

Many studies are affected by missing data, which takes different forms and complicates subsequent analyses for researchers. Here, we are concerned with missing outcomes generated by a missingness mechanism that is informative. In this case, ad hoc approaches, such as complete-case analysis, are not suitable as they lead to bias and loss of precision [1]. If we wish to adequately model this type of missing data, we need to use 'statistically principled' methods which combine information in the observed data with assumptions about the missing value mechanism, and account for the uncertainty introduced by the missing data. These methods include Bayesian full probability modelling, in which a joint model consisting of a model of interest and a model for the missing data mechanism is specified, allowing realistic assumptions to be made about the missingness process, and sensitivity to these assumptions to be tested.
Using simulated data, we demonstrate the well known deficiencies of complete-case analysis when the response has missing values which are missing not at random [2], and explore the circumstances and the extent to which Bayesian methods can improve our parameter estimates. We find that the addition of a model of missingness to form a joint model generally improves the overall fit of the model of interest leading to better prediction, but the estimates of individual parameters can be adversely affected by skewness in the response variable. With real datasets, when the form of the missingness is unknown, we would like to have a diagnostic that indicates the amount of informativeness in the missing data given our assumptions about the model of interest and the form of the missing data mechanism. p_D is a measure of the dimensionality of a Bayesian model [3], and we explore the use of the scaled p_D of the model of missingness in this context. We find that it is useful for indicating how far our missing data departs from missing at random, but that it should not be used for choosing the 'best' model. These points are illustrated with simulated data and for real examples of longitudinal data taken from the British birth cohort studies and a clinical trial analysed by Diggle and Kenward [4].

References

[1] Roderick J. A. Little and Donald B. Rubin (2002) Statistical Analysis with Missing Data, 2nd edition. John Wiley and Sons

[2] Donald B. Rubin (1976) Inference and Missing Data. Biometrika 63:581-592

[3] David J. Spiegelhalter, Nicola G. Best, Bradley P. Carlin and Angelika van der Linde (2002) Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society, Series B (Statistical Methodology) 64:583-614

[4] P. Diggle and M. G. Kenward (1994) Informative Drop-out in Longitudinal Data Analysis. Journal of the Royal Statistical Society, Series C (Applied Statistics) 43:49-72