

METHODS BRIEFING 5

Missing Data Methodology for Multilevel Models

Michael G Kenward and James Carpenter,

Medical Statistics Unit, London School of Hygiene and Tropical Medicine

November 2002 - November 2004

A series of methods briefings from projects funded by ESRC as part of the Research Methods Programme.

The Programme aims to develop qualitative and quantitative methods within the context of substantive research. It also aims to encourage effective dissemination of good practice.

Further copies are available from:
CCSR
2nd Floor
Crawford House
Manchester
M13 9PL

0161 275 4891

www.ccsr.ac.uk/methods/

Missing data is ubiquitous in social science research. For inferences to be valid, assumptions about the process by which the data became missing have to be made. As the appropriateness of these assumptions cannot be determined from the data, the sensitivity of the conclusions to the assumptions should be examined.

Website

We have established a website www.missingdata.org.uk which contains guidelines for researchers; material to enable non-technical researchers to 'get started' with handling missing data; example analyses, including code; our macro for multilevel multiple imputation in *MLwiN*; a frequently asked questions section; a discussion board; an extensive bibliography; preprints of papers submitted to journals, and links to other missing data sites worldwide.

Literature review

Our review of existing approaches for the analysis of partially observed datasets in a social science context concluded that the most appropriate and practicable method was a form of multiple imputation. This approach imputes missing data from the estimated distribution of the missing data given the observed data.

The process creates several 'complete' datasets. Each of these datasets is then analysed in turn, as though it were complete, and the results combined according to certain rules that take proper statistical account of the imputation process.

Guidelines

Plain English guidelines for handling missing data are available at www.missingdata.org.uk. These include the following points:

Design issues

It is important to consider the issues raised by missing data at the research design stage. As unplanned missing data inevitably introduce ambiguity into the inferences that can be drawn from a study, the design should be carefully scrutinised to minimise the scope for missing data to arise.

Ambiguity in the analysis can be reduced if the chance of the data being missing depends only on observed data; the so-called 'missing at random' scenario. Investigators should consider which variables are likely to prove difficult to collect and then see whether there are variables they could reliably collect which are likely to predict the chance of observing the difficult to collect variables.

To illustrate, people may be reluctant to divulge their income, but it may be easy to obtain their property band. If property band is a good predictor of the chance of people divulging their income (technically, if within each property band we observe a random sample of incomes) then collecting property band, and making appropriate adjustments in the analysis, will allow valid inferences to be drawn.

Longitudinal studies should consider which subgroups of individuals are likely to be lost to follow-up, and consider strategies for keeping in touch with representative samples of these groups. Ensuring there is sufficient funding, and a careful strategy, for following up initial non-responders greatly increases the credibility of the conclusions.

Finally, if you suspect missing data is likely to be a substantial issue in the analysis, budget for statistical advice on handling it.

Strategy for analysis of partially observed data set

First make sure you are familiar with the issues raised by missing data. Then familiarise yourself with the data. A natural starting point is an analysis of the fully observed data; note that with missing data this is only the starting point! At this stage you should clearly identify (if you have not done so already) (i) the hypotheses of interest (ii) the models that you are going to use to explore them and (iii) the variables that you are going to use, including any that are partially observed. Note that variables that are apparently unrelated in the subset of observed data may become important later on!

Next, explore as much as you can the reasons for the missing data. This should be done a variable at a time, or a wave at a time, in a longitudinal study.

Discussions about the reason for missing data should also include the study steering group, who may have useful insights. If no variables are predictive of missingness, then it may be plausible that the observed data are a random sample of the data you intended to collect (note, however, that you can never be sure of this). Nevertheless, unless only response data are missing, it is usually more efficient to carry out a missing at random analysis. If you are working with a regression model, and the responses are missing, then, provided you include the variables predictive of a missing response as covariates, the analysis will be valid. Note, however, that the model's interpretation is conditional on these covariates.

However, usually a combination of responses and covariates are missing. In this case, the most practical approach is some form of imputation. In a large data set, this could take the form of 'hot-deck' imputation. This approach finds a subset of the data with similar observed values to the unit with missing data, and then samples from this subset to impute the missing observations.

In practice, multiple imputation is currently the only practical, generally applicable, approach for substantial datasets. Methods for doing this are discussed on www.missingdata.org.uk; in particular, imputations that respect the multilevel nature of the data can be carried out using our macros with *MLwiN*. No specialist experience with imputation is necessary to use these. Note that ignoring the multilevel aspect of the data in imputation can lead to biases.

With partially observed data, conclusions are often far more sensitive to model choice. This is because, even under missing at random, different models make quite different predictions about the missing data. It is wise to examine carefully the predictions for the missing data before choosing a final model.

Methods to avoid

We strongly recommend avoiding the following *ad-hoc* approaches, which can give unpredictable results, and are not underpinned by statistical theory.

- Last observation carried forward
- Creating an extra category for the missing variable
- Replacing missing observations by the mean of the variable
- Mean imputation using regression

Reporting the analysis

The proportion of missing data in key variables should be stated clearly, and possible reasons discussed. This information should motivate an analysis valid under the ‘missing at random’ assumption, whose conclusions should be preferred to a ‘complete case’ analysis (which may also need to be presented). The sensitivity of the conclusions to the possibility of ‘not missing at random’ should also be reported, how plausible dropout mechanisms influence the conclusions.

Software

Our user-friendly macro for multiple-imputation in multilevel models using *MLwiN* is available from www.missingdata.org.uk. We show an example below. The left panel of

Figure 1 shows a model relating literacy at the end of reception year (*nlitpost*) to literacy at the beginning of reception year (*nlitpre*) adjusting for free school meals (*fsmn*) gender (*gend*) and term of entry (*tentry*). However, 1741 pre-reception year literacy scores are missing, representing a considerable loss of information. Note in particular, the coefficient for gender (-0.022) is equal to its standard error.

To use our macro you simply type ‘obey mi’. The program then (i) records the initial model; (ii) sets up an appropriate multilevel imputation model (iii) imputes several completed data sets (iv) refits the initial model to each of these data sets, and (v) combines the results using the appropriate statistical rules, presenting them to the user in the format shown in the right panel of Figure 1. This takes about 30 seconds. Imputation is carried out under the ‘missing at random’ assumption (see the sensitivity analysis section below).

The coefficient for gender is now -0.048, over twice its standard error. As gender is 1 for boys and 0 for girls, the multiple imputation has recovered evidence that boys are doing worse than girls, which was hidden when only the subset of 3132 fully observed cases was analysed.

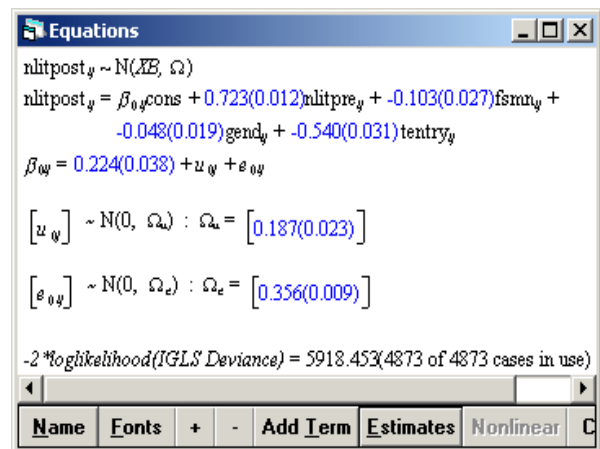
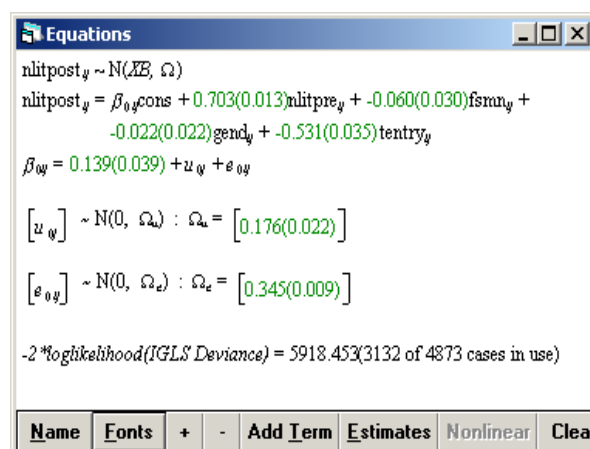


Figure 1: Analysis of class size data. Left panel: analysis of observed data only. Right panel, analysis of multiply imputed data sets.

Sensitivity analysis

Missing data introduce an inherent uncertainty into statistical analysis. Therefore, there can be no definitive analysis and assumptions must be made about the nature of the missing data. It is therefore important to examine the sensitivity of the conclusions to these assumptions.

We have developed two new methods for doing this, with the aim of enabling sensitivity analysis to be carried out more routinely by researchers.

Our first method allows the results of multiple imputation to be reused. Our algorithm re-weights them to explore how conclusions vary as the ‘missing at random’ assumption is relaxed. This assumption states that all the

information on the reasons that led to the data being unobserved is available in the observed data. Unlike many other approaches, our method avoids specialist statistical software. Instead, we are preparing an *MLwiN* macro to implement it.

Our second approach is to obtain quantitative information from experts about possible differences between missing and observed outcomes and provide a quantitative synthesis of the data with this expert opinion. Our Bayesian method uses simple formulae, making this approach accessible and practical.

For more information on all these topics go to: www.missingdata.org.uk



Key publications

Carpenter, J. and Kenward, M. (2004) A comparison of multiple imputation and inverse probability weighting for analyses with missing data. *Submitted to Journal of the Royal Statistical Society, Series A.*

White, I., Carpenter, J., Evans, S. and Schroter, S. (2004) Eliciting and using expert opinions about dropout bias in randomised controlled clinical trials. *Submitted to Clinical Trials.*

Further details are available from:
James.Carpenter@lshtm.ac.uk
Phone: +44 (0)20 7927 2033
www.missingdata.org.uk