Guidelines for handling missing data in Social Science Research

James Carpenter and Mike Kenward

www.missingdata.org.uk

Acknowledgement

James Carpenter is supported by ESRC research methods project grant 'Missing data in multilevel models' H333 25 0047.

Overview

Missing data are ubiquitous in social science research. This document is a guideline for researchers faced with analysing partially observed datasets describes the issues that need to be considered. Technical details, however, will vary considerably between analyses, so these are not discussed here. Further information and references can be obtained from www.missingdata.org.uk, or by emailing James.Carpenter@lshtm.ac.uk

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. Considerable care over this aspect of design will pay a substantial dividend when the study is analysed.

Inevitably, however, missing data will 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 (see the 'Getting Started' section of www.missingdata.org.uk). In other words, investigators should consider which variables are likely to prove difficult to collect. Then they should 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

Make sure you are familiar with the issues raised by missing data; see for instance the documents in the 'Getting Started' section of www.missingdata.org.uk

The next stage is to 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. See, for example, the document 'Exploring the reasons for missingness' in the 'Getting started' section of www.missingdata.org.uk 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. Simply speaking, this approach finds a subset of the data is found with similar observed values to the unit with missing data, and the 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. The 'Links' page under www.missingdata.org.uk lists alternative software.

Note that, 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 (see the document on *ad-hoc* methods in the 'Getting started' section of www.missingdata.org.uk):

- 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

Sensitivity analysis

When analysing missing data, additional assumptions about the reasons for the missing data have to be made. Unfortunately, these cannot be validated definitively from the data at hand. Therefore, some form of sensitivity analysis is advisable. Ideally, this should be closely linked to the substantive problem under analysis.

Note that if (i) the analysis valid under missing at random (MAR) gives similar results to analysing the completely observed data, and (ii) there are substantive reasons to believe the MAR mechanism is plausible, then it is usually reasonable to conclude that the missing observations are unlikely to alter the conclusions.

However, if this is not the case, some form of sensitivity analysis should be undertaken. Broadly speaking, there are two approaches:

1. *Explicitly model based*
   Here, a model for the non-response is formulated and fitted using appropriate software, to see how the conclusions vary. The study steering group, or other experts, may help to identify plausible models. Such models can be fitted by maximum likelihood, which usually requires some form of numerical integration, but often it requires less specialist programming to use *WinBUGS*, as illustrated in the 'Example analyses' section of www.missingdata.org.uk The plausibility of such analyses can be enhanced if the analysis formally incorporates prior information from experts (White and Carpenter, 2004). The goal is to see how the assumptions built in to the postulated missing value mechanism affect the conclusions drawn.
2. *Imputation driven*
   Here, we do not have an explicit model for the dropout mechanism. Rather, we see how the conclusions change as we work through a range of different behaviours for the missing observations. For example, a well-known method if the missing data are binary is to explore the range of conclusions obtained by replacing missing observations by 0's, and then 1's. More generally, we can replace missing values with 'extreme' values, and see how the conclusions vary. Clearly, this is most convenient with categorical variables, but can be done with continuous variables by placing explicit bounds on possible values. Note, however, that this approach does not necessarily return the most extreme parameter values, or inferences that are consistent with the

constraints; however it will often reveal the sensitivity of the parameter estimates to assumptions about the behaviour of the missing data. A flexible generalization of this idea can be incorporated conveniently into procedures that use multiple imputations. Instead of accepting all the random imputations, we choose according to some rule that reflects in a direct way some form of postulated selection bias. The sensitivity of the conclusions to this selection process can then be assessed.

We are currently developing a general approach for use with *MLwiN*, that can be applied after a conventional multiple imputation analysis has been performed. This will be posted on the web site as soon as it is available.

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. For an example of a trial report with missing data, see Schroter *et al.* (2004).

Additional advice

We are happy to provide additional advice. Please visit our website, www.missingdata.org.uk or email James.Carpenter@lshtm.ac.uk

References

Schroter, S. Black, N., Evans, S, Carpenter, J., Godlee, F. and Smith, R. (2004). Effects of training on quality of peer review: randomised controlled trial. *British Medical Journal*, **328**, 673-637.

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; download from* www.missingdata.org.uk