Handling missing data

James R. Carpenter & Harvey Goldstein
London School of Hygiene & Tropical Medicine / University of Bristol
james.carpenter@lshtm.ac.uk / h.goldstein@bristol.ac.uk

www.missingdata.org.uk / www.cmm.bristol.ac.uk

Support for JRC from ESRC

March 17, 2009

Acknowledgements

John Carlin, Lyle Gurrin, Helena Romaniuk, Kate Lee (Melbourne)
Mike Kenward, Harvey Goldstein (LSHTM)
Geert Molenberghs (Limburgs University, Belgium)
James Roger (GlaxoSmithKline Research)
Sara Schroter (BMJ, London)
Jonathan Sterne, Michael Spratt, Rachael Hughes (Bristol)
Stijn Vansteelandt (Ghent University, Belgium)
Ian White (MRC Biostatistics Unit, Cambridge)

Plan

Introduction — James Carpenter

Multilevel multiple imputation — Harvey Goldstein

Introduction

1. Principles
2. Missing data mechanisms
3. Brief outline of multiple imputation
4. What may be gained using multiple imputation

A starting point: the E9 guideline on conducting RCTs, 1999

The International Conference on Harmonisation (ICH) issued the E9 guideline on statistical aspects of carrying out and reporting trials in 1999 [5]; see also www.ich.org.

With regard to missing data, in summary it says:

- Missing data are a potential source of bias
- Avoid if possible (!)
- With missing data, a trial[study] may still be regarded as valid if the methods are *sensible*, and preferably *predefined*
- There can be no universally applicable method of handling missing data
- The sensitivity of conclusions to methods should thus be investigated, particularly if there are a large number of missing observations

The same principles apply to observational research.

The question is, how do we apply them in practice?

Study validity and sensible analysis

Data are sometimes missing by design, but our focus is on observations we intended to make but did not.

The sampling process involves both the selection of the units, and the process by which observations on those units [i.e. the *items*] become missing — the *missingness mechanism*.

Thus for sensible inference, we need to take account of the missingness mechanism

From a frequentist standpoint, by *sensible* we mean that nominal properties hold. Eg:

estimators consistent; confidence intervals attain nominal coverage.

Why there can be no universal method:

In contrast with the sampling process, which is usually known, the missingness mechanism is usually unknown.

The data alone cannot usually definitively tell us the sampling process.

Likewise, the missingness pattern, and its relationship to the observations, cannot identify the missingness mechanism.

With missing data, extra assumptions are thus required for analysis to proceed.

The validity of these assumptions cannot be determined from the data at hand.

Assessing the sensitivity of the conclusions to the assumptions should therefore play a central role.

Key points for analysis

- the question (i.e. the hypothesis under investigation)
- the information in the observed data
- the reason for missing data

With missing data, information is lost: the value of what remains depends on:

1. whether we can identify plausible reasons for the data being missing (called *missingness mechanisms*), and
2. the sensitivity of the conclusions to different missingness mechanisms.

A possible systematic approach is as follows:

A systematic approach

Investigators discuss possible missingness mechanisms, say A–E, possibly informed by a (blind) review of the data, and consider their plausibility. Then

1. Under most plausible mechanism A, perform valid analysis, draw conclusions
2. Under similar mechanisms, B–C, perform valid analysis, draw conclusions
3. Under least plausible mechanisms, D–E, perform valid analysis, draw conclusions

Investigators discuss the implications, and arrive at a valid interpretation of the study in the light of the possible mechanisms causing the missing data.

For trialists, this approach broadly agrees with the E9 guideline.

Missing data mechanisms (see [2], ch. 1)

It follows from this that the missing data mechanism plays a central role in informing the analysis.

Fortunately, it turns out that there are three broad classes of mechanism, each with distinct implications for the analysis.

In practice, to obtain sensible answers, we therefore have to:

1. postulate a missingness mechanism;
2. identify its class, and
3. perform a valid analysis for that class of missingness mechanism.

We now consider these three classes.

I: Missing completely at random

If the missingness mechanism is unrelated to any inference we wish to draw, missing observations (items) are *Missing Completely at Random* (MCAR).

Eg: missing observations because a page of the questionnaire was missing; missing data because of a data processing error; missing data because of a change in data collection procedure.

In this case analysing only those with observed data gives sensible results.

Of course, results are less precise than when full data are observed.

Data are randomly missing

II: Missing at random

If, given the observed data, the missingness mechanism does not depend on the unseen data, then we say the missing observations are *Missing at Random* (MAR).

For example, the probability of a missing observation may depend on an earlier observation. After accounting for the earlier observation, the chance of seeing the missing observation is independent of its value.

In this case simply analysing the observed data is invalid: we have two threats:

- bias — the fully observed subset of data is not representative, and
- loss of information — we have thrown away information on cases with even 1 missing observation.

Thus simple summary statistics are invalid as estimates of population parameters.

How to proceed

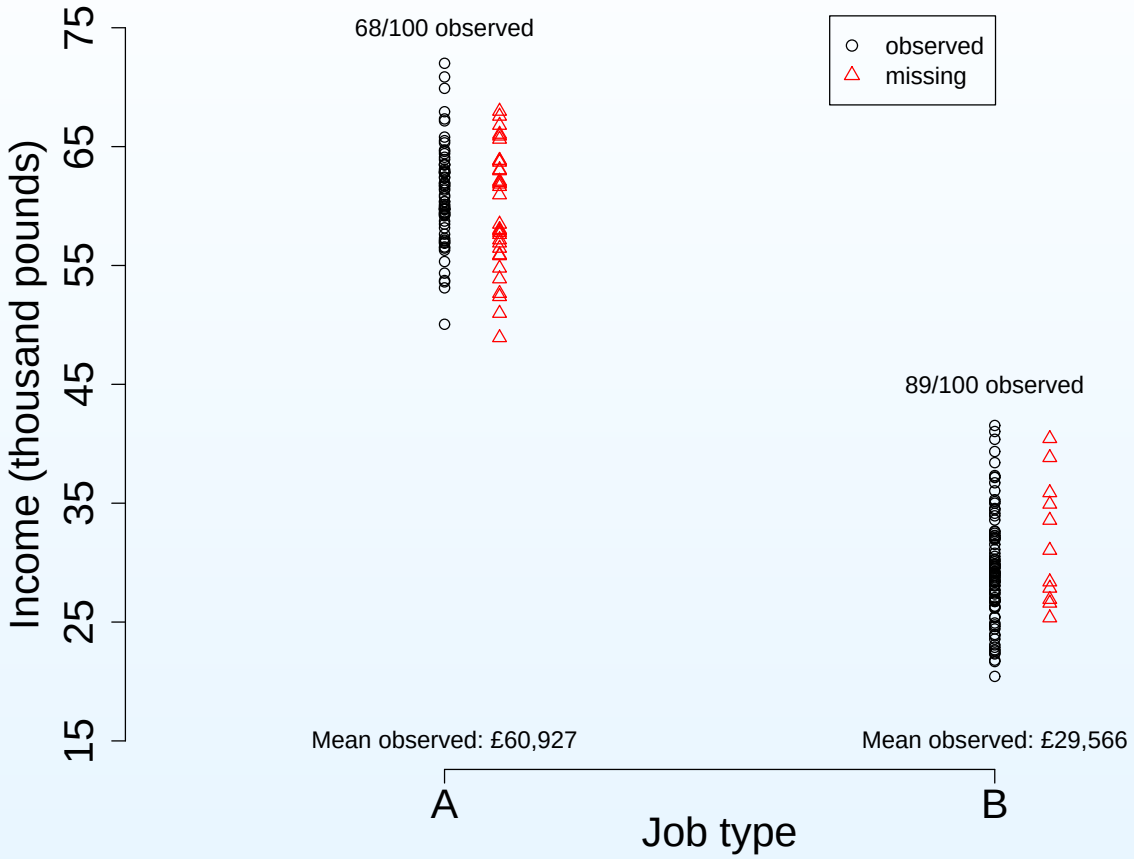
To obtain valid estimates, we have to include in the analysis the variables predictive of non-response.

For example, we may condition on them, eg. as covariates in a regression.

Of course, with several partially observed variables the issues are more complex.

‘Missing At Random’ means Data are Conditionally Randomly Missing

Example: true mean income £45,000



Observed income: $£43,149.$

MAR estimate: $$\frac{100 \times 60,927 + 100 \times 29,566}{200} = £45,246$$

III: Missing Not At Random

If data are neither MCAR nor MAR, we say they are Missing Not at Random (MNAR).

The missingness mechanism depends on the unobserved data, *even after taking into account all the information in the observed data*.

Under MNAR, we have to model both:

1. the response of interest, and
2. the missingness mechanism.

This is considerably harder! Often there is little to choose between various models for (2), but they may give quite different conclusions. The ‘pattern mixture’ approach is sometimes a convenient way to proceed — see Session 4.

Summary

Why MI?

There are a number of methods for the analysis of partially observed data under MAR:

1. Direct likelihood (not always possible)
2. EM algorithm
3. Mean score algorithm
4. Bayesian analysis

Multiple imputation can be viewed as a 2-step approximation to a Bayesian analysis.

Assuming the model of interest is known, once the imputation model has been decided upon the process is almost automatic.

This includes the estimation of the standard errors, which rely on a relatively simple yet general formula: an attraction compared to.

Together, these points make MI an attractive practical method in many settings.

MI: The basic idea

Consider two variables X, Y with some Y values MAR given X.

Under the assumption that data are MAR, using only units with both observed we can get valid estimates of the regression of Y on X.

However, inference based on observed values of Y alone (eg sample mean, variance) is typically biased.

This suggests the following idea

1. Fit the regression of Y on X
2. Use this to impute the missing Y
3. With this completed data set, calculate our statistic of interest (eg sample mean, variance, regression of X on Y).

As we can only ever know the *distribution* of missing data (given observed), steps 2,3 have to be repeated, and the results averaged in some way.

MI: what we do

All methods for MI fit (explicitly or implicitly) a joint model to the observed data, and impute the missing data from this, taking full account of the uncertainty in the estimated parameters of the joint model.

Often this joint model can take the form of a (multivariate) regression, with partially observed variables on the left. Under MAR this joint model can be fitted simply by including the observed data (full and partial observations).

We then impute the missing data from this model multiple times, as follows:

1. Draw parameters from the sampling distribution of the joint model
2. Given the values drawn in (1) and the observed data, draw from the distribution of the missing given the observed to create a ‘complete’ data set

Step 1 is important: it makes the calculation of the variance relatively simple

Using the imputed data

Fit the model of interest to each of K imputed data set, giving estimates $\hat{\theta}_1, \ldots, \hat{\theta}_K$ and their standard errors $\hat{\sigma}_1, \ldots, \hat{\sigma}_K$.
Let the multiple imputation estimator of θ be $\hat{\Theta}_{MI}$. Then

$$\hat{\theta}_{MI} = \frac{1}{K} \sum_{k=1}^{K} \hat{\theta}_k.$$

Further define the within imputation and between imputation components of variance by

$$\hat{\sigma}_w^2 = \frac{1}{K} \sum_{i=1}^{K} \hat{\sigma}_k^2, \quad \text{and} \quad \hat{\sigma}_b^2 = \frac{1}{K-1} \sum_{k=1}^{K} (\hat{\theta}_k - \hat{\theta}_{MI})^2,$$

Then

$$\hat{\sigma}_{MI}^2 = \left(1 + \frac{1}{K}\right) \hat{\sigma}_b^2 + \hat{\sigma}_w^2.$$

Tests use $t-$distribution to compensate for finite number of imputations.

Comments

Once we have chosen the imputation model, the process is automatic.

Users thus need to think hard about the imputation model.

This will usually include extra variables, not in the model of interest to (i) increase the plausibility of the MAR assumption and (ii) recover information on partially observed variables.

Correcting bias: missing response values

Consider a regression of Y on two covariates X, Z

Suppose only Y has missing data

CC (Complete Cases) will be unbiased when:

- Y MCAR
- Y MAR given X, Z.
- Y MAR given some W, but W independent of $[Y, X, Z]$.

CC biased when

- Y MAR given W, and W dependent on $[Y, X, Z]$.
- Y MNAR

Implication: Variables predictive of Y being missing, and associated with variables in the analysis, should be included in the imputation model.

Correcting bias - missing covariate values

Consider a regression of Y on two covariates X, Z

Suppose only X has missing data

CC will be unbiased when:

- X is MCAR
- X is MAR given Z (but not Y)
- X is MAR given some W, but W independent of $[Y, X, Z]$.
- X is MNAR (dependent on X, possibly Z, but not Y)

Missing covariate values (ctd)

CC biased when

- X MAR, and mechanism depends on Y
- X is MAR, and mechanism depends on some W, and W not independent of $[Y, X, Z]$.

Implication: Variables predictive of X being missing, and associated with variables in the model, should be included in the imputation model.

Warning: If covariates MNAR (mechanism unrelated to response), then MI may be biased (since it requires MAR to be unbiased) while CC would not be.

More discussion in White & Carlin (2009) (under review with Statistics in Medicine)

When is bias correction most likely with MI?

We assume that we have variables such that data are MAR.

In general the simpler the model of interest, the more likely that we have omitted a variable predictive of missingness, and correlated with response and covariates. Thus the more likely the CC analysis is biased.

The simplest ‘model’ is the sample mean, sample variance etc.

Example

In clinical trials with partially observed longitudinal follow-up, marginal means are often very biased.

Suppose now the response is MAR given treatment, baseline response and baseline age.

As we bring these terms into the model we reduce the bias.

Directional Acyclic Graphs (DAGs) can be useful for highlighting likely biases.

Recovering information

Even if the CC analysis is approximately unbiased, MI can recover information.

Given the cost of collecting the data, versus the cost of MI, this alone is sufficient to justify its use.

With MI, broadly speaking, information is recovered through two routes:

1. bring cases with response and almost all variables observed into analysis, and
2. bring in information on missing values through additional variables correlated with them.

Implication: Include variables predictive of partially observed variables in the imputation model (even if they are not predictive of missingness).

Warning: If the principal missing data patterns have a missing response, information only comes in by route (2) above.

Structuring the imputation model

In order to do multiple imputation, it suffices to fit a model where partially observed variables are responses, and fully observed covariates.

This is tricky in general!

Thus, people have started with the assumption of multivariate normality, and tried to build out from that. Implicit in that the regression of any one variable on the others is linear.

Skew variables can be transformed to (approximate) normality before imputation and then back transformed afterwards.

With an unstructured multivariate normal distribution, it doesn't matter whether we condition on fully observed variables or have them as additional responses: so most software treat them as responses.

Software taxonomy: methods derived from multivariate normal

	Complexity		
Response type	Normal		Mixed response
Data structure	Independent	Multilevel	Multilevel
Package			
Standalone	NORM	PAN	REALCOM
SAS	NORM-port	—	—
Stata	NORM-port	—	—
R/S+	NORM-port	—	—
MLwiN	MCMC algorithm emulates PAN		+ 1–2 binary

All methods: General missingness pattern; fitting by Markov Chain Monte Carlo (MCMC) or data augmentation algorithm (see references on later slides).

Relationships essentially normal/linear (except MLwiN).

Interactions must be handled by imputing separately in each group.

Schafer has a general location model package, relatively little used.

Some references

Schafer (1997)[10] — Key book giving details of data augmentation and MI methods in many models.

Rubin (1987)[9] — Book bringing together the theory in a fairly accessible way.

Rubin(1996)[8] — review of the use of MI after $\sim$ 18 years.

Horton and Lipsitz (2001)[4] — Comparison of software packages.

Allison (2000)[1] — a cautionary tale!

Kenward & Carpenter (2007) [6]

Carpenter & Kenward (2008) [2] — freely available monograph, focusing on clinical trial issues.

Key points:

- Missing data introduce ambiguity into the analysis, beyond the familiar sampling imprecision.
- Extra assumptions about the missingness mechanism are needed; these assumptions can rarely be verified from the data at hand.
- Under the MAR assumption, multiple imputation is an attractive method for analysing the data.
- However, as MI requires joint modelling of the data, setting up appropriate imputation models requires careful thought:
  - about the variables to include
  - about the structure of the data

References

[1] P D Allison. Multiple imputation for missing data: a cautionary tale. *Sociological methods and Research*, 28:301–309, 2000.

[2] James R Carpenter and Michael G Kenward. *Missing data in clinical trials — a practical guide*. Birmingham: National Health Service Co-ordinating Centre for Research Methodology. Free from http://www.pcpoh.bham.ac.uk/publichealth/methodology/projects/RM03_JH17_MK.shtml, 2008.

[3] A-W Chan and Douglas G Altman. Epidemiology and reporting of randomised trials published in PubMed journals. *The Lancet*, 365:1159–1162, 2005.

[4] N J Horton and S R Lipsitz. Multiple imputation in practice: comparison of software packages for regression models with missing variables. *The American Statistician*, pages 244–254, 2001.

[5] ICH E9 Expert Working Group. Statistical Principles for Clinical Trials: ICH Harmonised Tripartite Guideline. *Statistics in Medicine*, 18:1905–1942, 1999.

[6] Michael G Kenward and James R Carpenter. Multiple imputation: current perspectives. *Statistical Methods in Medical Research*, pages 199–218, 2007.

[7] M A Klebanoff and S R Cole. Use of multiple imputation in the epidemiologic literature. *American Journal of Epidemiology*, 168:355–357, 2008.

[8] D Rubin. Multiple imputation after 18 years. *Journal of the American Statistical Association*, 91:473–490, 1996.

[9] D B Rubin. *Multiple imputation for nonresponse in surveys*. New York: Wiley, 1987.

[10] J L Schafer. *Analysis of incomplete multivariate data*. London: Chapman and Hall, 1997.

[11] Angela M Wood, Ian R White, and Simon G Thompson. Are missing outcome data adequately handled? a review of published randomized controlled trials in major medical journals. *Clinical Trials*, 1:368–376, 2004.