INTRODUCTION TO MISSING DATA: PART 2

James Carpenter and Mike Kenward

Acknowledgements to

Geert Molenberghs (Universities of Hasselt and Leuven)

James Roger (LSHTM, formally GSK)

22nd November 2012

Recap

- When data that we intended to collect are missing, there is a selection, or missing value, mechanism operating.
- The nature of this mechanism has implications for
  1. bias in the analysis of completers;
  2. the methods we need to correct for such bias.
- The most important aspect is the dependence of the mechanism on on the outcome variable, Y.

- We may be able to lessen this dependence through *conditioning* or *adjustment* for other variables (included in the model or not).
- **But:** the data under analysis cannot tell us about the dependence of the missing value mechanism on Y.
- We can never know that bias has been removed: hence sensitivity analysis potentially plays an important role.
- Unless very few data are missing we should generally avoid *ad hoc* methods.

Rubin’s classification of missing value mechanisms

- Missing Completely at Random (MCAR)
- Missing at Random (MAR)

  (Conditionally random missingness)
- Missing Not at Random (MNAR)

Statistically Principled Methods

Goal:

- Improve efficiency under MCAR.
- Reduce bias when MCAR does not hold.
- Be principled (in a statistical sense) and make assumptions transparent.
- Keep the analysis *comparatively* simple and flexible.
- Use a framework that allows appropriate sensitivity analyses to be done in a relatively simple way.

- We need a statistical model for the data.
- What model?
- A natural starting place is the model that would apply to the *complete data*.
- This contains the quantities of interest ($\boldsymbol{\theta}$):

$$f(\mathbf{Y} \mid \mathbf{X}; \theta)$$

- We call this the **substantive** model.

SOME EXAMPLES:

- *Linear Regression*

$$Y_i \sim \mathrm{N}(\mathbf{x}_i\boldsymbol{\theta}, \sigma^2).$$

- *Logistic regression*

$$\mathrm{logit}\{\mathrm{P}(Y_i = 1)\} = \mathbf{x}_i\boldsymbol{\theta}.$$

- *Multivariate linear regression for longitudinal data*

$$\mathbf{Y}_i \sim \mathrm{N}(\mathbf{X}_i\theta, \boldsymbol{\Sigma}).$$

- *A structural equation model*

Likelihood and averaging.

The substantive model is defined in terms of $\mathbf{Y}$ and $\mathbf{X}$, but we only observe $\mathbf{Y}_O$, $\mathbf{X}_O$, and $\mathbf{R}$.

So how do we get a model (and likelihood) for $\mathbf{Y}_O$, $\mathbf{X}_O$, and $\mathbf{R}$ from the substantive model?

One way or another we need to *average* or *integrate* over the *conditional* distribution of the missing data *given* the observed.

This takes us from the likelihood for the full (but partly unseen) data to the likelihood for the observed data.

But this means we need a model for the *missing data* in terms of the observed.

If there are missing *covariates* among these then this model is a new *addition* to the original substantive model.

It follows that when starting from the substantive model, *under MAR*, the problem of missing data largely reduces to one of integration or averaging over the conditional distribution of the missing data (the conditional predictive distribution).

When using a likelihood (or Bayesian) based analysis under MAR, the missing data mechanism can be *ignored*.

The integration/averaging be done may ways:

- direct/indirect
- analytical/numerical
- Markov chain Monte Carlo
- Multiple imputation
- ...

Note an alternative route exists: **Inverse Probability**

TWO IMPORTANT SETTINGS

1. Longitudinal data with attrition/dropout
2. Missing covariates

(1) Longitudinal data with attrition/dropout

Setting: longitudinal studies with set measurement times.

(Sometimes called a *monotone* missing data pattern.)

	Times of measurement					
Unit	1	2	3	4	5	6
1	3.4	3.9	3.5	4.3	5.1	4.6
2	3.9	4.8	5.3	·	·	·
3	2.6	·	·	·	·	·
4	1.9	1.4	2.3	1.6	1.2	1.9
5	2.2	2.7	·	·	·	·
6	3.3	4.6	4.5	2.4	1.3	0.3
7	1.7	1.9	2.3	4.2	1.5	1.8
8	2.4	·	·	·	·	·

The substantive model may

- represent changes in the outcome over time in terms of baseline and time-dependent covariates, or more simply,
- model the outcome at some given time (often the last).

Why is dropout/attrition important in the missing value context?

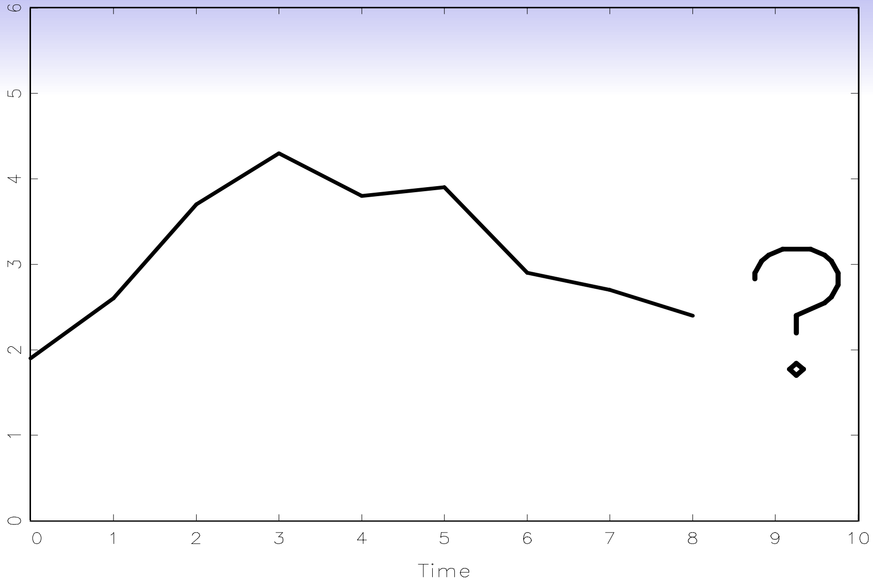
Longitudinal studies are very common in both trial and observational settings.

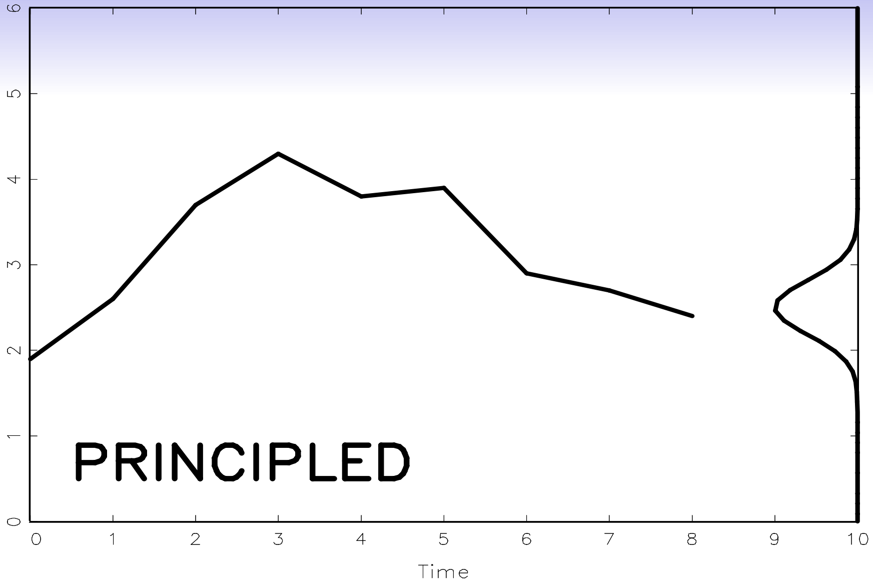
Dropout/attrition is almost ubiquitous in these and has potentially very serious implications for the conclusions drawn.

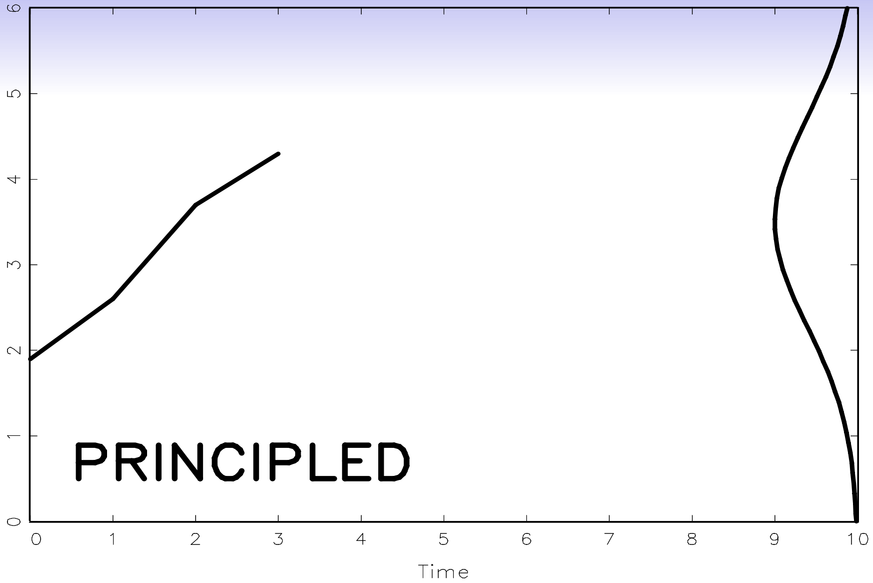
There are important aspects of the problem that are amenable to methods not applicable to non-monotone missingness.

Principled Analyses

- What does a statistical model do in these settings?
- It ensures that each subject/unit makes the appropriate contribution based on the observations seen.


0
1
2
3
4
5
6
0
1
2
3
4
5
6
7
8
9
10
Time
?


PRINCIPLED
0
1
2
3
4
5
6
0
1
2
3
4
5
6
7
8
9
10
Time


0
1
2
3
4
5
6
0
1
2
3
4
5
6
7
8
9
10
Time
PRINCIPLED

Modelling Approaches

- Often in these settings a main concern is the missing outcomes for those who are lost to follow-up.
- In such settings, for simpler longitudinal models, we can often do MAR based analyses using standard software, *e.g.* Stata xtmixed, MLwiN, SAS PROC GLIMMIX, and so on.
- For more complex settings, structural equation models (likelihood or Bayes) might be considered.
- These do the necessary integration for us, we don’t need to “add” to the existing substantive model.
- It is important to remember the underlying assumptions: the past of a subject who drops out contributes information about values in the future.
- What about deaths, or other situations where future values are impossible?

One way of expressing the MAR assumption with dropout:

the future statistical behaviour of a subject, conditional on the history, is the same whether the subject drops out or not in the future.

When using such methods we need to think carefully about whether such an assumption is appropriate.

[Note: the MAR assumption is far less readily interpreted with intermediate/non-monotone missingness.]

(2) Missing covariates

With observational data, the setting considered at the beginning is very common, missing values are scattered across many covariates:

	Variables							
Unit	1	2	3	4	5	6	7	…
1	1	4	1	3.4	5.67	A	8.251	…
2	1	3	·	·	5.67	B	9.253	…
3	1	2	1	2.7	5.72	B	12.812	…
4	1	1	1	3.6	5.13	·	13.614	…
5	2	·	1	·	·	A	11.442	…
6	2	2	1	3.4	5.61	A	9.241	…
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱

It is assumed that we have a substantive model in which the covariates are treated as “fixed”, *e.g.* logistic regression:

$$\mathrm{logit}\{\mathrm{P}(Y_i = 1)\} = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \ldots + \beta_p X_{i,p}.$$

The distribution of the covariates is strictly irrelevant at this point – we *condition* on the observed values of these.

But, if some covariates are missing for a subject, and we want to include this subject in the analysis, then we must introduce a model for the missing covariates.

e.g. if X_1 and X_2 are incomplete we need to *introduce* a model for these:

$$f(X_1, X_2 \mid Y, X_3, X_4, \ldots, X_p, \mathbf{R}).$$

Depending on the context we call this the **conditional predictive distribution** or **imputation model**.

Note that on the right hand side we have all complete covariates, the missing value indicator, *and the outcome variable Y*.

Usually this will be some type of joint regression model.

This model is an *addition* to the original substantive model.

Under MAR we can drop **R**, but what does this imply when different subjects have very different patterns of missing data?

Structure (*e.g.* longitudinal, clustered) in the original substantive model must carry-over to the imputation model.

Such models can become very complex and difficult to handle: they are potentially multivariate, multilevel, with mixtures of variable types (continuous, binary, ordinal, nominal).

One approach is to build a single overarching model for the substantive and imputation components: essentially the incomplete covariates become additional outcome variables.

Again, for some settings SEM's and Bayesian analyses in WinBUGS can be used for this.

An important alternative solution is to use a clearly separate imputation model, retaining the original substantive model.

Multiple Imputation (MI) is the standard, and widely used, framework for this.

MI allows us to separate out (to some degree) the *substantive* and *imputation* models.

The imputation model may contain *auxiliary* variables: variables *not* in the substantive model that are predictive of the missing data and missingness itself.

Auxiliary variables may be on the causal pathway, *e.g.* post-randomization measurements in a trial.

Multiple imputation is a convenient way of conducting sensitivity analyses.

There is plenty of software: Stata, SAS, MLwiN, REALCOM,...

SENSITIVITY ANALYSIS

Most of the principled methods so far described assume MAR (explicitly or implicitly).

These form a sensible starting point, but this cannot be justified from the data under analysis.

It may be worth exploring the effect of departures from MAR for particular incomplete variables.

In principle there are many ways in which this can be done.

Some possibilities (there are overlaps):

- Sensitivity parameters
- Pattern-mixture/selection models
- Likelihood based/full enumeration regions
- Multiple imputation based
- Incorporating prior belief (Bayes)

Very broadly we can distinguish two important ways of approaching sensitivity analysis for missing data

- **Route 1:** modify directly the behaviour of the missing data. *i.e.* modify
$$f(Z_M \mid \mathbf{Z}_O, R = 0)$$
- **Route 2:** consider an explicit MNAR missing value mechanism.

  i.e. modify
$$\mathrm{P}(R \mid \mathbf{Z}_O, Z_M)$$

- Multiple imputation is well-suited to Route 1: we can intervene directly in the imputation process, or use weighted imputations.
- For Route 2, fully Bayesian analyses are particularly convenient.
- There are many exceptions and alternatives to these connections however.

EXAMPLE: LONGITUDINAL CLINICAL TRIAL WITH DROPOUT/WITHDRAWAL

Recall from James's talk: **Aims of the study/analysis**

de jure

Estimate the treatment effect under the best case scenario.

Does the treatment work under the best case scenario?

[Per-Protocol, PP, efficacy]

de facto

Estimate the effect seen in practice if this treatment were applied to the population defined by the trial inclusion criteria.

Is an effect seen in practice if this treatment is applied to the population defined by the trial inclusion criteria?

[Intention To Treat, ITT, effectiveness]

In both cases, the analysis needs to reflect the treatment taken by a patient, that is consistent with the trial (analysis) aims [estimands].

Only in the special settings of

- *no dropout* and a *de facto* hypothesis; and
- *no dropout*, *no withdrawal* and a *de jure* hypothesis;

are there plausible definitive analyses that do not rest on additional untestable assumptions about the statistical behaviour of unobserved measurements.

Otherwise, to construct an analysis that is consistent with the trial aims some assumptions must be made about

treatment use following withdrawal;

and/or

future statistical behaviour of outcomes from dropouts and/or withdrawals.

How does MAR fit in with the *de jure* and *de facto* questions?

Recall, MAR dropout implies

the future statistical behaviour of a subject, conditional on the history, is the same whether the subject drops out or not in the future.

That is, both outcome and *future treatment use*, follow the same conditional distributions for all subjects.

- This would lead to the testing of a *de jure* hypothesis if treatment compliance before dropout matches the protocol.
- Or, extending this, if there were withdrawal, all post-withdrawal outcomes were set to missing.
- Or, if future treatment compliance after dropout matched that expected in use in the population, this would lead to the testing of a *de facto* hypothesis.

- But this excludes many common settings.
- So, often it is likely that we need to consider NMAR models.
- How should this be done?
- A secondary analyses? As part of sensitivity analyses?

The generic sensitivity problem in the dropout setting

1. Define the specific questions, and consequent quantities to be estimates, and hypotheses, that we wish to test;
2. define the nomenclature for departures from protocol;
3. frame the relevant accessible assumptions, and
4. formulate the procedures for estimation and inference.

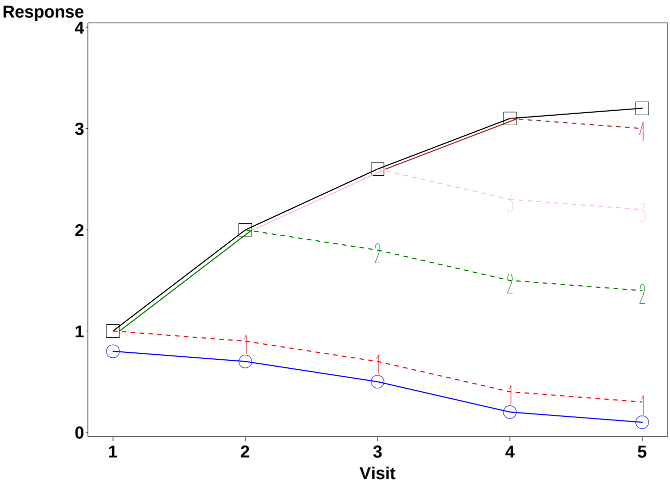
An example of a MNAR defacto scenario: Copy Difference from Control

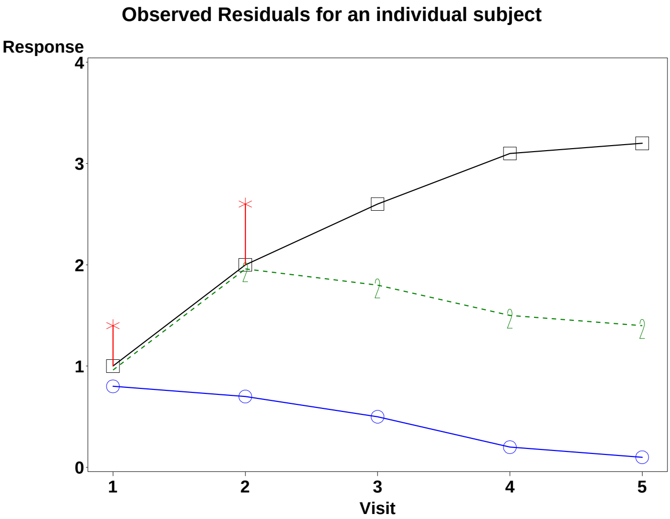
The subject takes no further treatment.

Instead their post-dropout mean increments copy those from the reference group (typically, but not necessarily, control patients).

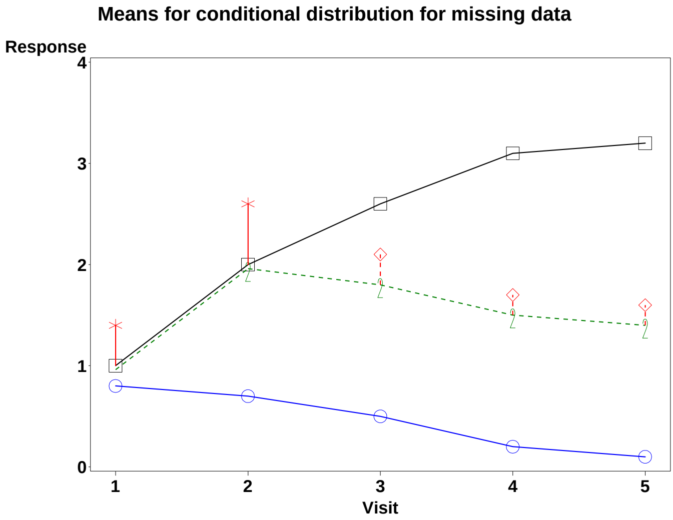
If the reference is chosen to be the control arm, the subject profile following withdrawal tracks that of those in the control arm, but starting from the benefit already obtained.

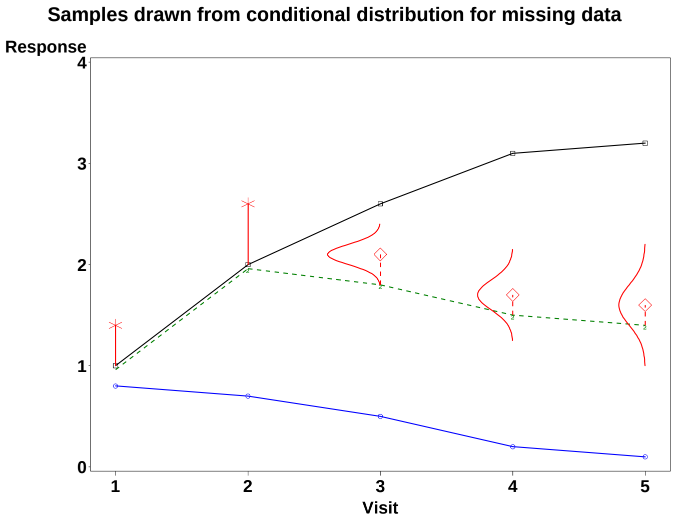
Copy Difference from Control (CDC)





Observed Residuals for an individual subject


Means for conditional distribution for missing data
Response
0
1
2
3
4
1
2
3
4
5
Visit


Samples drawn from conditional distribution for missing data
Response
4
3
2
1
0
1
2
3
4
5
Visit

SUMMING UP

- The implications for missing data on the subsequence analysis, and conclusions, depend crucially on the relationship between the missing data mechanism and the outcome, whether observed or not.
- This *cannot* be directly assessed from the data under analysis.
- It is rarely (if ever) sensible to treat “impossible” outcomes (*e.g.* quality of life after death) as missing data.
- Rubins’s classification underpins the statistical handling of missing data (MCAR, MAR, MNAR).

- Analysis of completers only may well be inefficient, and will be biased in general under MAR and MNAR.
- Likelihood based analyses (including Bayesian), and some special weighted analyses, are valid under MAR.

  Other methods, like unweighted GEE are not.
- MAR becomes more plausible if we can incorporate observed variables known to be predictive of missingness.
- We may not want to condition on these however.
- Assuming MAR is likely to be “better” than ignoring the missing value problem completely, even when MNAR is holds.

  There *are* exceptions to this however.

- When covariates are missing, and we want to incorporate information from “incompleters”, we need to introduce a (joint) model for the missing variables.

- We can do this using a more general all-embracing model, *i.e.* treat the missing covariates as additional outcome variables (*e.g.* use SEM’s, full Bayes in WinBUGS).

- Or, a distinct imputation model can be introduced using Multiple Imputation (*e.g.* stata ice, SAS proc MI).

Because MNAR cannot be ruled out from the observed data, carefully directed sensitivity analysis is sensible: a good focus for this is departures from MAR.

e.g. examine the impact of allowing the missing value process for key incomplete variables to depend on the values those variables take (whether observed or not).

SOME REFERENCES

Allison PD (1991) *Missing Data.* London: Sage Publications.

Carpenter JR and Kenward MG (2012) *Multiple Imputation and its Application.* Chichester: Wiley. (in press)

Little RJA and Rubin DB (2002) *Statistical Analysis with Missing data. Second Edition* Chichester: Wiley.

Molenberghs G and Kenward MG (2007) *Missing Data in Clinical Studies.* Chichester: Wiley.

Rubin DB (1987) *Multiple Imputation for Nonresponse in Surveys.* Chichester: Wiley.

Schafer JL (1997) *Analysis of Incomplete Multivariate Data.* London: Chapman & Hall/CRC.

Van Buuren S (2012) *Flexible Imputation of Missing Data.* London: Chapman & Hall/CRC.