Introduction to handling missing data in multilevel modelling

James R. Carpenter

James.Carpenter@lshtm.ac.uk · J.Carpenter@ucl.ac.uk

London School of Hygiene and Tropical Medicine &

MRC Clinical Trials Unit at UCL

www.missingdata.lshtm.ac.uk

Support from ESRC, MRC, DFG, EU

13^{th} July 2020

Acknowledgements

This work is joint with:

Juan Carlos Bazo Alverez (UCL)

Tim Morris (UCL)

Harvey Goldstein (University of Bristol)

Irene Petersen (UCL)

Matteo Quartagno (LSHTM and MRC-CTU) — `jomo`

Overview

- What are multilevel models?
- Overview of missing data mechanisms
- When a complete records analysis is valid
- Handling missing outcome data
- Discussion

Multilevel data structures

Classical statistical modelling assumes that observations on the outcome (dependent) variable are independent of each other.

For example, for observations $i = 1, \ldots, n$ we might model lung function as:

$$\text{lung function}_i = \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{sex}_i + \epsilon_i$$

$$\epsilon_i \overset{iid}{\sim} N(0, \sigma_e^2)$$

Multilevel structures

Real life is typically more complicated:

- we may have repeated measures on individuals — *longitudinal* data
- we may have educational data on children, in classes, in schools, in educational authorities — *multilevel, hierarchical* data
- we may have longitudinal measures on patients within centres — both *longitduinal* and *hierarchical*
- and so on...

Typically, we will then have data relating to different levels of the hierarchy — e.g. children, teachers, schools.

Multilevel models

Multilevel data means observations are no longer independent.
Multilevel models account for this either:

- explicitly, by direct modelling of the correlation structure, or
- implicitly, through introducing random effects, or
- some combination of the two.

Example: asthma clinical trial

Consider data from an asthma clinical trial, where patients are randomised to placebo or active treatment, and lung function is recorded at clinic visits at baseline, 2, 4, 8 and 12 weeks.

The outcome is the treatment effect at 12 weeks.

If there were no missing data, we can simply regress the data from 12 weeks on baseline and treatment:

$$\text{lung function}_{i,12w} = \beta_0 + \beta_1 \text{lung function}_{i,baseline} + \beta_2 \text{treatment}_i + \epsilon_{i,12}$$

$$\epsilon_{i,12} \overset{iid}{\sim} N(0, \sigma^2_{12})$$

Longitudinal model

However, we may want to model all the follow-up data simultaneously. This turns out to be particularly useful if we have missing data — not all patients are seen at all follow-up times.

We can do this consistently with the previous model as follows:

$$\begin{pmatrix} \text{lung function}_{i,2w} \\ \text{lung function}_{i,4w} \\ \text{lung function}_{i,8w} \\ \text{lung function}_{i,12w} \end{pmatrix} = \begin{pmatrix} \beta_{0,2} + \beta_{1,2}\text{lung}_{i,base} + \beta_{2,2}\text{treat}_i + \epsilon_{i,2} \\ \beta_{0,4} + \beta_{1,4}\text{lung}_{i,base} + \beta_{2,4}\text{treat}_i + \epsilon_{i,4} \\ \beta_{0,8} + \beta_{1,8}\text{lung}_{i,base} + \beta_{2,8}\text{treat}_i + \epsilon_{i,8} \\ \beta_{0,12} + \beta_{1,12}\text{lung}_{i,base} + \beta_{2,12}\text{treat}_i + \epsilon_{i,12} \end{pmatrix}$$

$$(\epsilon_{i,2}, \epsilon_{i,4}, \epsilon_{i,8}, \epsilon_{i,12})^T \overset{iid}{\sim} N(0, \Sigma), \text{ where } \Sigma_{4,4} = \sigma_{12}^2 \text{ etc}$$

Note:
(i) with no missing data inference for the 12 week treatment effect is the same.
(ii) this uses an *unstructured* variance matrix; we can use simpler structures if appropriate.

Example: educational data

Let i index pupil, and j school, and let $Y_{i,j}$ be the 16+ exam points score. A typical multilevel model is:

$$Y_{i,j} = \beta_0 + u_j + \beta_1 \text{age 12 reading score}_{i,j} + \text{sex}_{i,j} + 1[\text{single sex school}]_j + \epsilon_{i,j}$$

$$u_j \overset{iid}{\sim} N(0, \sigma_u^2)$$

$$e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2).$$

Note:
(i) we have a 'residual' or 'random intercept' for school, as well as the residuals for pupil within school
(ii) some variables are defined at the school level: such variables arise naturally in the multilevel setting.

The random intercept gives a nice interpretation to the model (each school has its own regression line) and also accounts for the correlation of children's scores within school.

Missing data mechanisms

Before discussing how we handle missing data, it is important to think about the likely reasons for missing data (the *missingness mechanism*) as this will have implications for how we do the analysis.

While there are many reasons why missing data may arise, [1] introduced three broad classifications, with distinct implications for the analysis. We now review this informally:

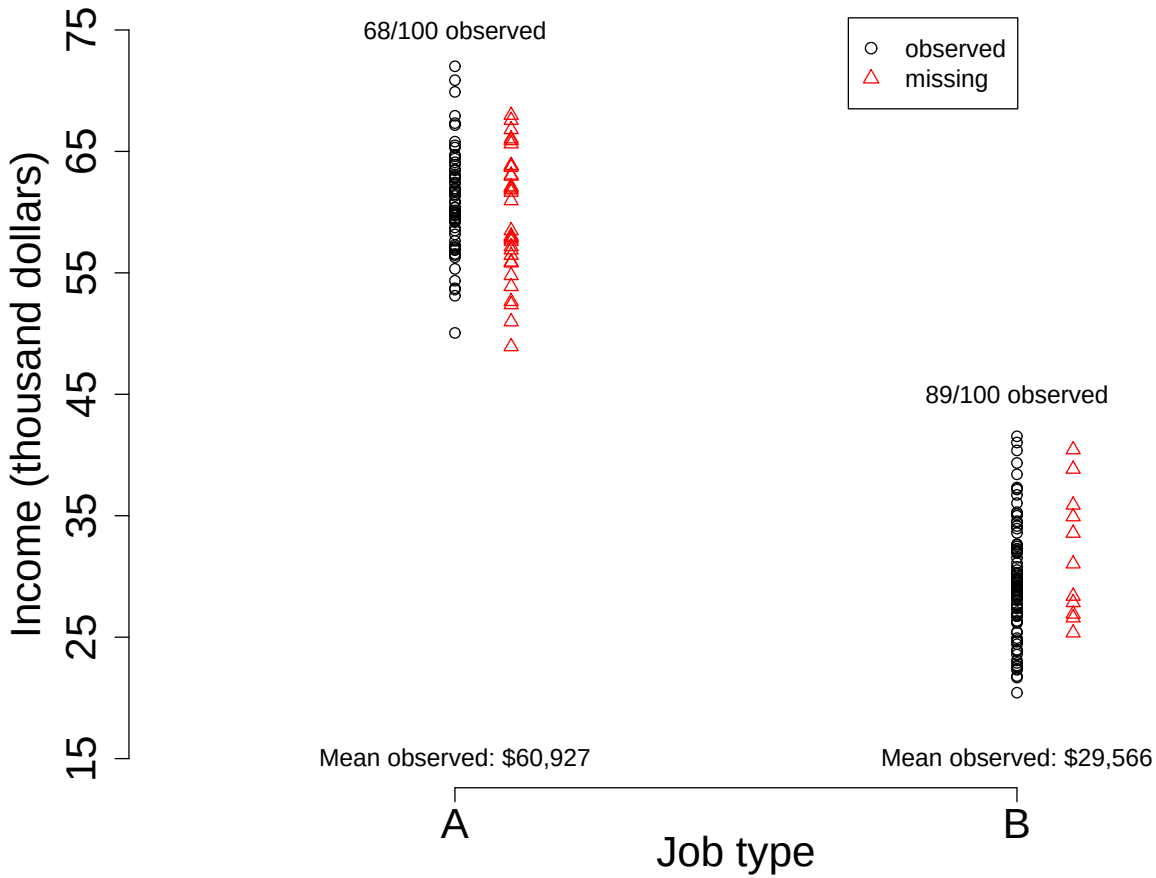
Rubin's missing data taxonomy

- Missing Completely At Random (MCAR) — the chance for the missing data is not related to anything to do with our analysis: restriction to the complete records will therefore give valid results.
- Missing At Random (MAR) — the chance of missing data depends on the underlying value **BUT** this association can be broken by observed variables.
- Missing Not At Random (MNAR) — the chance of missing data depends on the underlying value and this association **cannot** be broken by observed variables.

It follows that:
If data are MCAR, simple summary statistics using the observed values (e.g. mean) are valid.
If data are MAR, simple summary statistics using the observed values will be biased, *BUT* we can correct this using the observed data.

Schematic example

True mean income $45,000$



Observed income: $\$43,149$.

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

Some comments:

1. Missing At Random is an assumption
   If you think about the previous slide, you see that MAR is an assumption we use to estimate the mean income.
   We cannot test it, because to do so we would need to have seen the missing data (red triangles).
2. Missing At Random means — given observed data — we can avoid specifying how the distribution of the observed and missing values differ: because they are assumed to be the same!
3. Missing Not At Random means we *must* specify how — given the observed data — the distribution of missing and observed values differ.

This last point makes MNAR much more challenging, because the diferences can occur in many different ways!

In applications, I therefore often first assume MAR, then explore the robustness of my conclusions to departures from MAR.

What is the appropriate strategy[3]?

Given our scientific question, and corresponding analysis model, a broad strategy is:

1. Explore patterns and likely reasons for missing data (both using the data and in discussion with colleagues)
2. Decide if a *complete records* (or more typically for longitudinal data, *all observed data* analysis is valid for the primary analysis (see below)
3. If we need to go beyond this, assume MAR and use multiple imputation (consistent with the multilevel structure — see next presentation) to obtain parameter estimates, standard errors and confidence intervals.
4. Perform secondary, sensitivity analyses, to explore the robustness of conclusions to departures from the MAR assumption. Again, multiple imputation can be helpful [2].

When is complete records valid?

Consider a (multilevel) regression of outcome **Y** on covariates **X**.

If it is reasonable to assume (we cannot know for sure) that the probability of a complete record depends on **X**, and given **X** not on **Y**, then a complete records analysis is valid (though it may be very inefficient).

The consequence of this is that:

1. If missing values are in **Y**, then a complete records (observed data) analysis is valid if **Y** is MAR given **X**.
2. If missing values are in **X** (or both) a complete records analysis will be valid even if some data are MNAR.

In case (1), we can get valid inferences without multiple imputation, by fitting our multilevel model to the observed data. In case (2), my preferred approach is to assume MAR and use multilevel multiple imputation.

Terminology: complete cases and observed data

With cross-sectional data, a complete records analysis omits any individual with one or more missing values in the variables in the analysis model.

With longitudinal (multilevel) data, a strict definition of a complete records analysis would exclude an individual if their longitudinal outcome measures were not complete (e.g. in the asthma study if they do not have data at 2, 4, 8 and 12 weeks).

However, an *observed data* analysis includes all the observed outcomes on each individual, provided their covariates are complete.

So if we observed patient i at 2 and 8 weeks only, and their covariates (baseline data) are complete, then their 2 and 8 week data would be included in the analysis.

Loss of information with missing outcomes/covariates

- In standard regression: if an individual's outcome (dependent) variable is missing — and we assume it is MAR given covariates in the model — then that individual's record contributes no information about the parameters in the model.

- In longitudinal (multilevel) models: if an individual's outcome (dependent) variable at a particular follow-up time is missing — and we assume it is MAR given covariates and other outcomes in the model — then that observation contributes no information about the parameters in the model[1].

- For both: If an outcome is observed, but some covariates are missing, then the record can contribute information — use multiple imputation to bring this back in

[1] Unless we have auxiliary variables, not in our model, that are good predictors of the missing outcome. In which case use MI and include them

Application: asthma study

Now consider the asthma study. Recall we have two treatment arms, and lung function (forced expiratory volume, FEV) measured at baseline, and then planned to be measured at 2, 4 8 and 12 weeks.

The treatment effect at 12 weeks is the primary outcome.

We have:

- all baseline covariates observed
- for individual patients some outcomes are missing

Therefore — consistent with our arguments above — if we assume missing outcomes (lung function at clinic visits) are missing at random given baseline and observed lung function data, then we can get valid inference by fitting a longitudinal (multilevel) model to the observed data. Multiple imputation is not needed here.

Asthma data

Dropout pattern	Mean FEV_1 (litres) measured at week 0	2	4	8	12	Number	Percent
Placebo arm							
1	2.11	2.14	2.07	2.01	2.06	37	40
2	2.31	2.18	1.95	2.13	—	15	16
3	1.96	1.73	1.84	—	—	22	24
4	1.84	1.72	—	—	—	16	17
All patients (Mean)	2.11	1.97	1.98	2.04	2.06	90	100
All patients (Std.)	0.57	0.67	0.56	0.58	0.55		
Lowest Active arm							
1	2.03	2.22	2.23	2.24	2.23	71	78
2	1.93	1.91	2.01	2.14	—	8	9
3	2.28	2.10	2.29	—	—	8	9
4	2.24	1.84	—	—	—	3	3
All patients (Mean)	2.03	2.17	2.22	2.23	2.23	90	100
All patients (Std.)	0.65	0.75	0.80	0.85	0.81		

Model

$$\begin{pmatrix} \text{lung function}_{i,2w} \\ \text{lung function}_{i,4w} \\ \text{lung function}_{i,8w} \\ \text{lung function}_{i,12w} \end{pmatrix} = \begin{pmatrix} \beta_{0,2} + \beta_{1,2}\text{lung}_{i,base} + \beta_{2,2}\text{treat}_i + \epsilon_{i,2} \\ \beta_{0,4} + \beta_{1,4}\text{lung}_{i,base} + \beta_{2,4}\text{treat}_i + \epsilon_{i,4} \\ \beta_{0,8} + \beta_{1,8}\text{lung}_{i,base} + \beta_{2,8}\text{treat}_i + \epsilon_{i,8} \\ \beta_{0,12} + \beta_{1,12}\text{lung}_{i,base} + \beta_{2,12}\text{treat}_i + \epsilon_{i,12} \end{pmatrix}$$

$$(\epsilon_{i,2}, \epsilon_{i,4}, \epsilon_{i,8}, \epsilon_{i,12})^T \overset{iid}{\sim} N(0, \Sigma), \text{ where } \Sigma_{4,4} = \sigma_{12}^2 \text{ etc}$$

We fit the model to *all the observed follow-up data*. So, if a patient only has data at 2 weeks, we just fit this. If the patient only has data at 2, 4 weeks, we just fit this.

Fortunately, all software for longitudinal (multilevel) models does this by default.

We use an unstructured variance matrix here (10 parameters); we could use a structured variance model, eg. compound symmetry, AR(1), random intercepts and slopes — but none fit so well.

Results

Method	Point estimate	Standard error
Regression of 12 week data on baseline and treatment (n=108)	0.247	0.1005
Longitudinal (multilevel) model n=180, all observed data	0.345	0.1013
MI–MAR (n=180, not needed here)	0.323	0.1031

We see:
(i) more information and stronger effect estimate under MAR
(ii) very similar (theoretically equivalent) results using multiple imputation

Summary

- Most data are multilevel - and they need multilevel models.
- If outcomes are MAR given covariates, times with missing outcome data contribute no information to the analysis. So:
  - If missing values are (almost all) in the outcome, and we can reasonably assume they are MAR given covariates, then fitting the multilevel model to the observed data will give valid results — MI is not necessary.[2]
  - Note that the choice of covariance/correlation structure is very important with missing data — since it controls how information is shared across time.

[2] However MI can recover information in this setting if we have auxiliary variables, predictive of outcome, not in the substantive model

- If missing values are in the covariates, then (assuming MAR) multilevel multiple imputation can correct bias and recover information.

- The multilevel aspect of imputation is important, so that we handle missing data in covariates at each level of the model (e.g. pupil, school) correctly.

- We should explore the robustness of our conclusions to departures from MAR. This can be simply done with MI, see [2].

References I

[1] D B Rubin.
Inference and missing data.
Biometrika, 63:581–592, 1976.

[2] James Carpenter.
Multiple Imputation-Based Sensitivity Analysis, pages 1–18.
02 2019.

[3] J R Carpenter.
Missing data: a framework for practice.
Biometrical journal (revision under review), xx:yyyy–zzzz, 2020.