Multiple Imputation for Multilevel Data

M. Quartagno[1 2]

[1]Department of Medical Statistics
London School of Hygiene and Tropical Medicine

[2]Clinical Trial Unit
Medical Research Council, London

Centre for Statistical Methodology
Celebrating the first International Statistics Prize Winner,
Prof Sir David Cox
March 2017

Introduction

Outline of the presentation:

- Missing data: why worry?
- Multiple Imputation: how does it work? What are main advantages? And the main issues?
- Multilevel Multiple Imputation: additional difficulties compared to single level case;
- Substantive model compatible imputation: what do we do with non-linearities/interactions/survival data?
- Summary and future work.

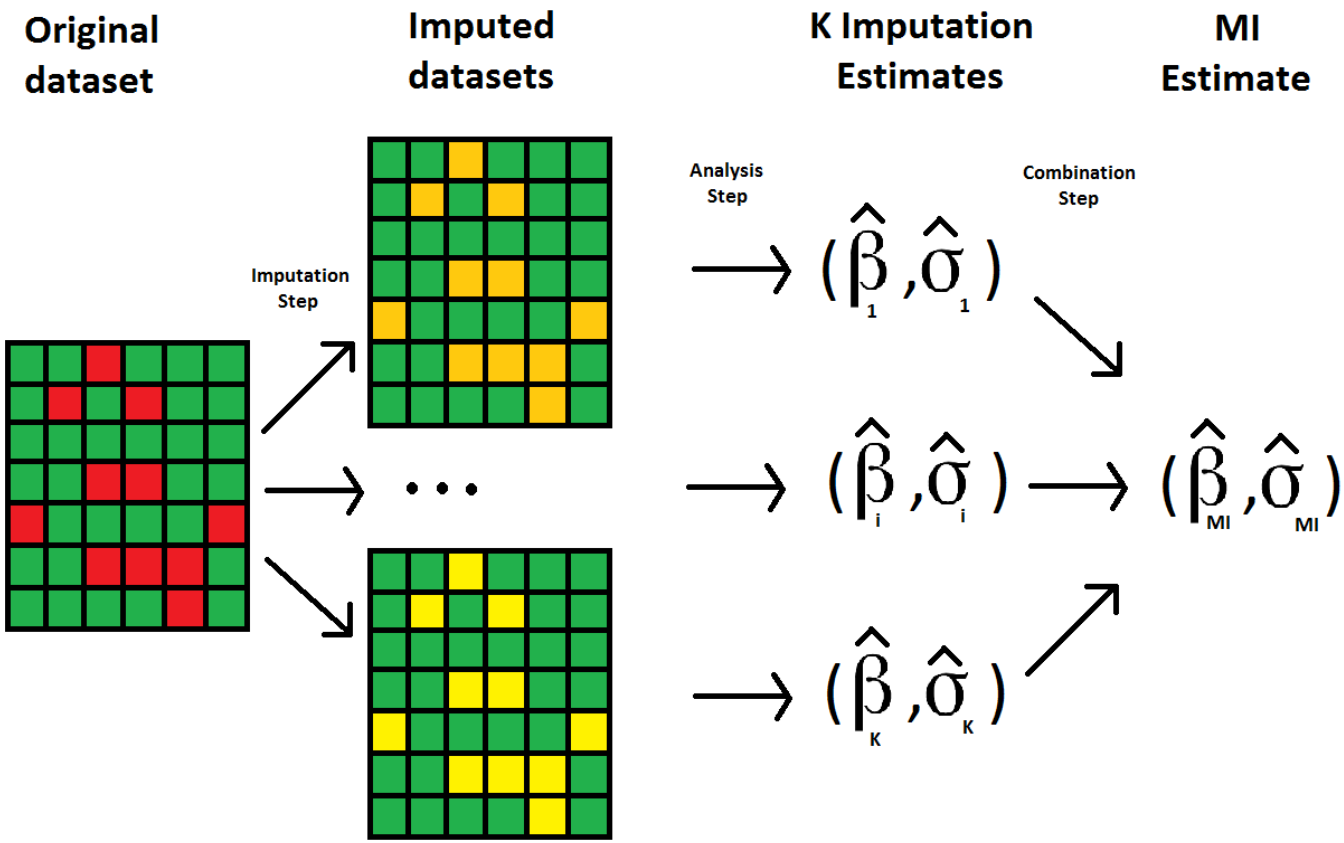
Missing data

- Missing data are extremely common; but why worry?
  1. They cause loss of power;
  2. When inappropriately handled, may even introduce bias in inferences;
- Why are data missing?
  - Missing data Mechanisms: MCAR, MAR, MNAR
- Three classes of valid methods:
  1. Likelihood-based methods;
  2. Inverse probability weighting methods;
  3. Multiple Imputation.

Missing data

- Missing data are extremely common; but why worry?
  1. They cause loss of power;
  2. When inappropriately handled, may even introduce bias in inferences;
- Why are data missing?
  - Missing data Mechanisms: MCAR, MAR, MNAR
- Three classes of valid methods:
  1. Likelihood-based methods;
  2. Inverse probability weighting methods;
  3. **Multiple Imputation**.

Multiple Imputation



Multiple Imputation

- Advantages of Multiple Imputation:
  - Extremely flexible, compared to other valid methods;
  - We still use same substantive model we would have used were we able to observe all intended data;
  - Easy to include auxiliary variables to recover information on missing data;
  - Straightforward to perform sensitivity analysis to different missing data assumptions.
- Difficulties with Multiple Imputation:
  1. Additional distributional assumptions: imputation model needs to be (at least approximately) correctly specified;
  2. Issues of compatibility between imputation and analysis model.

Multiple Imputation: challenges.

- There are simple situations in which specification of imputation model and compatibility do not cause particular problems;
  - Example: Missing data in a single variable, analysis model is a simple linear regression model with no interactions/non-linearities;
- But what if...
  1. ... we had missing data in multiple variables?
  2. ... **we had data with a multilevel structure?**
  3. ... **we had interactions/non-linearities in the analysis model?**

Multiple Imputation: Joint Modelling.

1) What if we had missing data in multiple variables?

- There are both parametric and non-parametric methods;
- Among parametric methods, two main strategies: Joint Modelling and Full Conditional Specification;
  1. Joint Modelling Imputation:
     - it consists in defining a multivariate joint model for partially observed variables given fully observed;
     - Gibbs sampling with data augmentation is used to fit imputation model and generate imputed datasets;
     - Specification: difficult to define sensible joint models for mixed data types;
     - Compatibility: it is often the case that substantive analysis model is simply derivable from joint imputation model by conditioning over covariates;

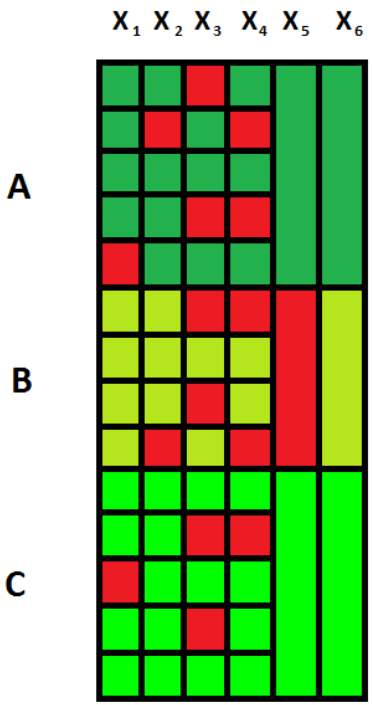
Multilevel data structure.

2) What if partially observed data had a multilevel structure?

- Standard example: pupils (level 1) clustered in schools (level 2);
- We need to reflect this structure in imputation model, similarly to what we do for analysis model;
- For my Ph.D, we decided to use a Joint Modelling Imputation approach, defining a joint multilevel imputation model (Schafer and Yucel, Carpenter et. al).

Multilevel data structure

Multilevel Joint Model:



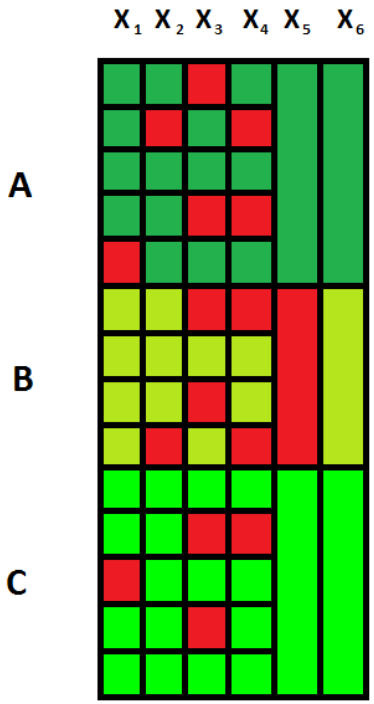
- 3 Clusters: A,B and C

- 4 Level-1 variables: X1, X2, X3 and X4

-2 Level-2 variables: X5, X6

Multilevel data structure

Multilevel Joint Model:



$$\begin{cases} X_{i,j,1} = \alpha_1 + u_{j,1} + e_{i,j,1} \\ X_{i,j,2} = \alpha_2 + u_{j,2} + e_{i,j,2} \\ X_{i,j,3} = \alpha_3 + u_{j,3} + e_{i,j,3} \\ X_{i,j,4} = \alpha_4 + u_{j,4} + e_{i,j,4} \\ X_{i,j,5} = \alpha_5 + u_{j,5} \\ X_{i,j,6} = \alpha_6 + u_{j,6} \end{cases}$$

$$\mathbf{e}_{i,j} = \begin{pmatrix} e_{i,j,1} \\ e_{i,j,2} \\ e_{i,j,3} \\ e_{i,j,4} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \Omega_e \right)$$

$$\mathbf{u}_j = \begin{pmatrix} u_{j,1} \\ u_{j,2} \\ u_{j,3} \\ u_{j,4} \\ u_{j,5} \\ u_{j,6} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \Omega_u \right).$$

Multilevel data structure.

2) What if partially observed data had a multilevel structure?

- Specification: we found that latent normals are good way to incorporate binary/categorical variables in joint model;
- Compatibility: we found (and Resche-Rigon and White proved algebraically) that heteroscedastic models, i.e. cluster-specific covariance matrices, performed better than homoscedastic ones; however, still not perfectly compatible with random slopes;
- I created an R package, called **jomo**, to perform Multilevel Multiple Imputation.

Multilevel data structure.

2) What if partially observed data had a multilevel structure?

- For my Ph.D I applied these methods to individual patient data **meta-analysis**; level 1 are individual observations and level 2 study;
- Two other methods have been developed at the same time, based on FCS (Resche-Rigon and White, Jolani et al.);
- These methods have been compared recently, our method seems to be preferable with binary data;
- Need to be careful with small datasets, priors play a major role with few clusters or few observations per cluster.

Substantive Model Compatible Imputation

3) What if partially observed variables were included in interactions/non-linearities in the imputation model?

- Up to 3/4 years ago, ad-hoc methods with several limitations: passive imputation, JAV...
- Bartlett et al. (FCS) and Goldstein et al. (JM) developed the so-called substantive model compatible imputation method.
- For single level data, R package SMC-FCS;

Substantive Model Compatible Imputation

3) What if partially observed variables were included in interactions/non-linearities in the imputation model?

- Example: Y, X continuous variables;
- Substantive Analysis model is a linear regression with quadratic effect:

$$Y \sim X + X^2;$$

- With missing data in both Y and X, difficult to set up appropriate joint model for X and Y;
- Alternative: factor model in two terms:

$$X \sim N(\beta, \sigma) \qquad Y \sim N(\alpha_0 X + \alpha_1 X^2, \omega)$$

Substantive Model Compatible Imputation

3) What if partially observed variables were included in interactions/non-linearities in the imputation model?

- When imputing missing values in X, we cannot simply use marginal model, still need to impute compatibly with substantive model for Y;
- We can't use Gibbs sampler, we therefore rely either on rejection sampling or Metropolis-Hastings;
- Specification: we can easily accommodate any sort of substantive model, linear regression, logistic regression, or even Cox model.
- Compatibility: no issues any more with compatibility, at the only cost of having to know the form of the substantive model prior to imputation;

Conclusions

- MI very flexible method to handle missing data, quickly became gold standard;
- Valid under MAR assumption, with possibility to perform sensitivity analyses to different assumptions;
- Imputation model needs to be correctly specified and reasonably compatible with substantive model;
- We proposed a way to handle multilevel structures in the imputation model;
- We are also working on software allowing for substantive model compatible imputation;

Future Work

- Extend software to allow for imputation of covariates of survival models;
- How should we include weights in the imputation model?
- With longitudinal data, methods for imputing taking into account different correlation structures.