Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *doi: 10.18637/jss.v000.i00*

Imputation of Incomplete Multilevel Data with mice

Hanne Oberman
Utrecht University

Johanna Munoz Avila
University Medical Center Utrecht

Valentijn
University Medical Center Utrecht

Gerko Vink
Utrecht University

Thomas Debray
University Medical Center Utrecht

Abstract

Tutorial paper on imputing incomplete multilevel data with **mice**. Including methods for ignorable and non-ignorable missingness. AKA missing on so many levels.

Keywords: missing data, multilevel, clustering, R.

1. Introduction

1.1. Multilevel data

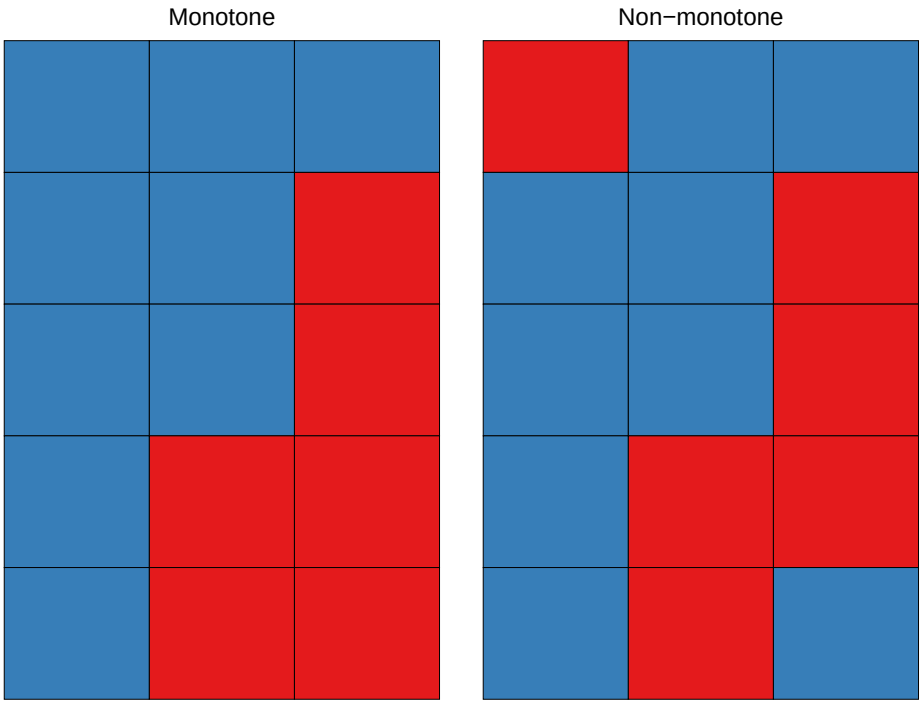
- What is clustering/multilevel data? In this paper, we discuss grouped observations, not longitudinal data (within-patient clustering). What do we mean by clustering? In the medical field: Clustering by studies (IPDMA), hospitals in registries, multi-center studies etc. In other fields: e.g. official stats clustering at country-level, or social sciences clustering at school-level (related to the sampling design).

- What is heterogeniety (i.e. variability within studies vs. variability between studies)? ADD: Figure to show difference between patient-level datapoints vs cluster-level datapoints. And different data frame formats.

- What methods are required to analyze multilevel data? Random effects for intercept term, predictor effects, and/or variance residual error. Add references, e.g. **?**.

1.2. Missing data

- Why/where does missingness occur? I.e., not only patient-level but also cluster-level.
- How can we categorize this? Systematic vs sporadic missingness, see Resche-Rigon, White, Bartlett, Peters, Thompson, and Group. ADD: visualization of systematic vs sporadic missingness. Within systematic we have always missing (same value per cluster) and non-measured variables (may differ per patient). TODO: adjust md pattern to match text.



- Why are standard (ad hoc) missing data methods not well suited?
- What other methods are available? General overview of approaches, see Audigier, White, Jolani, Debray, Quartagno, Carpenter, van Buuren, and Resche-Rigon Grund, Lüdtke, and Robitzsch. E.g., imputation of study level versus patient-level covariates, and one-stage imputation versus two-stage imputation methods.
- Additional problem addressed in this paper: handling MNAR.

1.3. Aim of this paper

- Provide practical guidelines with code snippets for imputation of incomplete multilevel data.
- Case study options: metamisc::impact (IPD), GREAT (IPD).

2. Workflows

2.1. Modeling choices

- ADD: Section with more in-depth info on different data types and suitable methods. Add some equations, but not too much -> grow in complexity.
- Ideally we impute with random everything and heteroscedastic errors: most generic method (no worry about congeneiality) –> explain predictor matrix. Just mention that we should think about random intercept/slope, but just refer to other papers. This paper is just a software tutorial. –> Typically, at least random intercepts, but often random slopes as well. Refer to other papers for background, we'll focus just on the software implementation of the situations mentioned there. Sometimes there's little reason to assume some variable is affected by heterogeniety. We won't focus on congeneiality, that's too complex for 2-3 papers.

2.2. Conditional models

- How to define the imputation model(s) in mice?
- What do the different implementations look like?

2.3. Pooling

- Pooling 'regular' parameters vs more 'exotic' parameters (SE of residual errors, or autocorreletion)
- ADD: export mids objects to other packages like lme4 or coxme(?)

3. Think about adding

- JOMO in mice –> on the side for now
- Additional levels of clustering

References

Audigier V, White IR, Jolani S, Debray TPA, Quartagno M, Carpenter J, van Buuren S, Resche-Rigon M (????). "Multiple Imputation for Multilevel Data with Continuous and Binary Variables." **33**(2), 160–183. ISSN

0883-4237, 2168-8745. doi:10.1214/18-STS646. 1702.00971, URL https://projecteuclid.org/journals/statistical-science/volume-33/issue-2/Multiple-Imputation-for-Multilevel-Data-with-Continuous-and-Binary-Variables/10.1214/18-STS646.full.

Grund S, Lüdtke O, Robitzsch A (????). "Multiple Imputation of Missing Data for Multilevel Models: Simulations and Recommendations." **21**(1), 111–149. ISSN 1094-4281. doi:10.1177/1094428117703686. URL https://doi.org/10.1177/1094428117703686.

Resche-Rigon M, White IR, Bartlett JW, Peters SA, Thompson SG, Group obotPIS (????). "Multiple Imputation for Handling Systematically Missing Confounders in Meta-Analysis of Individual Participant Data." **32**(28), 4890–4905. ISSN 1097-0258. doi:10.1002/sim.5894. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.5894.

Affiliation:

Hanne Oberman
Utrecht University
Padualaan 14
3584 CH Utrecht
E-mail: h.i.oberman@uu.nl
URL: https://hanneoberman.github.io/