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 de Jong
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.

Keywords: missing data, multilevel, clustering, **mice**, R.

1. Introduction

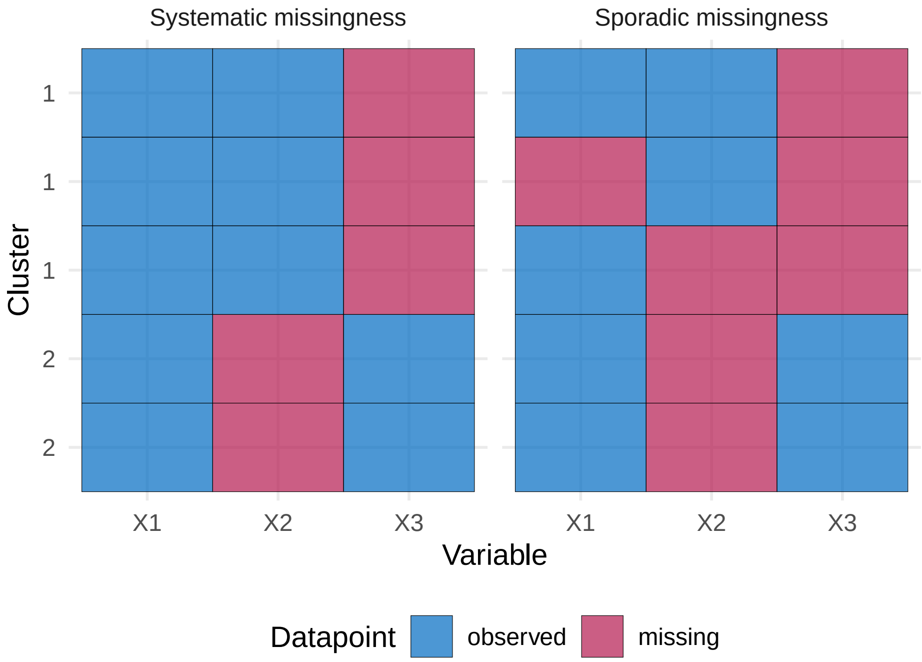
1.1. Multilevel data

- What is clustering/multilevel data? In this paper, we discuss grouped observations, not longitudinal data (within-patient clustering). -> ADD: timeseries also in Discussion section.
- 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 heterogeneity? I.e. variability within studies vs. variability between studies

- What does multilevel data look like? ADD: figure to show difference between patient-level datapoints vs cluster-level datapoints. Maybe also add different data frame formats (or just explain in text that there's long and wide formats).
- What methods are required to analyze multilevel data? Add references, e.g. **?**. At least explain difference random effects for intercept term, predictor effects, and/or variance residual error.

1.2. Missing data

- Why/where does missingness occur in multilevel data? 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. -> syst may vary or same for all patients (observations/participants).



- What kinds of missingness are there? ADD: missingness mechanisms here.
- Why are standard (ad hoc) missing data methods not well suited?
- What types of multilevel 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 difficulty that is addressed in this tutorial: MNAR data.

1.3. Aim of this paper

- Provide practical guidelines with code snippets for imputation of incomplete multilevel data.
- We focus on the workflow for conditional modeling (not JOMO) in `mice`. Refer to other packages: `mitml`, `miceadds`.
- Case study options: `metamisc::impact` (real data on traumatic brain injuries, IPD), `mice::popularity` (simulated data with MNAR/MAR mixture, schools). -> Check Gelman's data/NSRI data.
- Introduce case study and set scope of this tutorial: We're providing an overview of implementations. It's up-to the reader to decide which strategy suits their data. So we won't go into detail for the different methods (and equations). This paper is just a software tutorial. We'll keep it practical. -> ADD: some kind of help function that suggests a suitable pred matrix to the user, given a certain analysis model.

2. Workflows

2.1. Case study

- We'll use the IMPACT data (`metamisc::impact`) and a MAR/MNAR version of the `mice::popmis` data (i.e., a variation on the Hox (2010) popularity data, where the missingness in the variables is either missing at random (MAR) or missing not at random (MNAR)).
- `impact` is traumatic brain injury data with patients clustered in studies, $n_{\text{participants}} = 11022$ and $n_{\text{clusters}} = 15$, on the following 11 variables:
  - `name` Name of the study,
  - `type` Type of study (RCT: randomized controlled trial, OBS: observational cohort),
  - `age` Age of the patient,
  - `motor_score` Glasgow Coma Scale motor score,
  - `pupil` Pupillary reactivity,
  - `ct` Marshall Computerized Tomography classification,
  - `hypox` Hypoxia (0=no, 1=yes),
  - `hypots` Hypotension (0=no, 1=yes),
  - `tsah` Traumatic subarachnoid hemorrhage (0=no, 1=yes),
  - `edh` Epidural hematoma (0=no, 1=yes),
  - `mort` 6-month mortality (0=alive, 1=dead).

```
R> # load data
R> data("impact")
R> # # descriptive statistics
R> # by(impact, impact$name, summary)
R> # psych::describe(impact)[,c(2:5,8:9)]
R> # missingness
R> md_pat(impact)
```

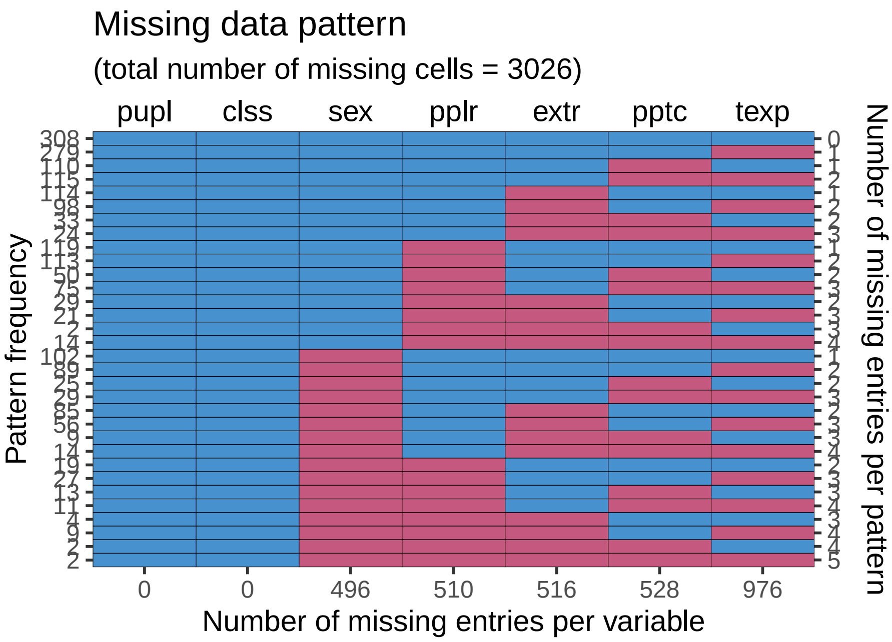
```
 /\     /\
{  `---'  }
{  O   O  }
==>  V <==  No need for mice. This data set is completely observed.
 \  \|/  /
  `-----'
```

```
      name type age motor_score pupil ct hypox hypots tsah edh mort
11022    1    1   1           1     1  1     1      1    1   1    1 0
         0    0   0           0     0  0     0      0    0   0    0 0
```

-> Why are there no missings? According to the vignette, the data is already imputed (Steyerberg et al, 2008).

- popNCR is a simulated dataset with pupils clustered in classes, $n_{\text{participants}} = 2000$, $n_{\text{clusters}} = 100$, on 7 variables:
  - `pupil` Pupil number within class,
  - `class` Class number,
  - `extrav` Pupil extraversion,
  - `sex` Pupil gender,
  - `texp` Teacher experience (years),
  - `popular` Pupil popularity,
  - `popteach` Teacher popularity.

```
R> # load data
R> pop <- readRDS("../Data/popNCR.RDS")
R> # missingness
R> md_pat(pop)
```



2.2. Modeling choices

- Which models will we discuss? We'll build the model to grow in complexity. The final model is the most complex but also the most versatile.
- Note on model complexity: Typically, we should at least use random intercepts, but often random slopes as well. Ideally we impute with random everything and heteroscedastic errors: most generic method (no worry about congeniality, but don't mention the term) -> 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 heterogeneity. -> Refer to Meng, Vincent, and a paper by Grund on congeniality and random slopes.
- Step 0: study as a predictor, AKA multilevel imputation for dummies. Doesn't work for syst missing.

2.3. Conditional models

- How to define the imputation model(s) in `mice`?
- What do the different implementations look like?
- Step 1: Intercept
- Step 2: Slope

- Step 3: Residuals
- Heckman model for MNAR

2.4. Pooling

- Analysis of scientific interest.
- Pooling using `mitml`.
- 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. Discussion

- JOMO in `mice` –> on the side for now
- Additional levels of clustering
- Timeseries: and polynomial relationship in the 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/

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
MMMMMM YYYY, Volume VV, Issue II *Submitted:* yyyy-mm-dd
doi:10.18637/jss.v000.i00 *Accepted:* yyyy-mm-dd