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

We talk of multilevel data when there is some kind of hierarchy or clustering in a dataset. In the typical case, individuals are nested within groups, but there are many different types of multilevel data. In the medical field, clustering occurs at e.g., the hospitals/center level in registry data, or at the study-level in meta-analyses (IPDMA). In the social sciences and official statistics we can find clustering e.g. at the country-level, or as imposed by the sampling design. In this paper, we will refer to the grouping variable as 'cluster', and the grouped variable as '(sample) unit'.[1] For reasons of brevity, we only discuss clustering between units, not within units (such as in timeseries or longitudinal data).

Analyzing multilevel data requires special care, compared to 'regular', single level data. The

[1] Add that we'll only discuss two levels?

cluster to which a unit belongs may influence the unit-level observations, and since clusters are made up of units, clusters depend on units as well (Hox, Moerbeek, and van de Schoot 2017). These relations can and should be taken into account when developing analysis models for multilevel data.[2] Multilevel models typically include separate intercepts for each cluster, which relieves one restriction imposed by single-level models: equal group means across clusters. Additionally, there may be random predictor effects and/or random error terms (residual error variances), see e.g. Hox *et al.* (2017) and de Jong, Moons, Eijkemans, Riley, and Debray (2021).[3] There are many names for models that take clustering into account. Some popular examples are 'multilevel models', 'hierarchical models', 'mixed effect models' and 'random effect models'.

1.2. Missing data

Multilevel data is not spared the ubiquitous problem of missing information. Just as in single level data, missingness may occur at the unit level. But with multiple levels of data comes the potential for missingness at multiple levels. Missingness in multilevel data can be categorized into two general patterns: systematical missingness and sporadic missingness, see Resche-Rigon, White, Bartlett, Peters, Thompson, and Group (2013). In figure 1, we show a dataset with units in the rows and variables in the columns, there are 5 units nested within 2 clusters, and 3 variables of interest. Variable `X1` is completely observed. Variable `X2` is systematically missing, `X3` is sporadically missing.[4] Systematic missingness can be further subdivided into unobserved constants (same value per cluster) and non-measured random variables (which may differ per unit within clusters). In Figure 1, the former implies that the unobserved values for units 4 and 5 on variable `X2` would be equal. With the latter, the values may differ. Depending on the missing data pattern, there are more or less optimal way of accommodating the missingness.

Ignoring missing data in research endeavors is almost never a good idea. Complete case analysis (i.e., excluding all units with one or more missing entries) can introduce bias in statistical inference and lowers statistical power. Instead, the missingness should be accommodated *before* or *within* the analysis of scientific interest. Especially the former is very generic and popular. Imputing (i.e., filling in) the missing values splits the missing data problem from the scientific problem. The R package **mice** has become the de-facto standard for imputation by chained equations, which solves the missingness one variable at a time, iteratively. **mice** is known to yield valid inferences under many different missing data circumstances (Van Buuren 2018). In this paper, we'll discuss how to use **mice** in the context of multilevel data, under varying missing data mechanisms.[5]

1.3. Aim of this paper

This papers serves as a tutorial for imputing incomplete multilevel data with **mice**. We provide

[2] Explain ICC here? The percentage of variance attributed to the cluster-level is expressed by the intra-class coefficient (ICC). The ICC can also be interpreted as the expected correlation between two randomly sampled units in same cluster. So if the ICC is high, a lot of variability in a variable is due to the clustering, which should be modeled accordingly.

[3] Add that heterogeneity refers to variability within clusters.

[4] Explain why.

[5] Discuss missingness mechanisms before this point, add references Yucel (2008) and Hox, van Buuren, and Jolani (2015).

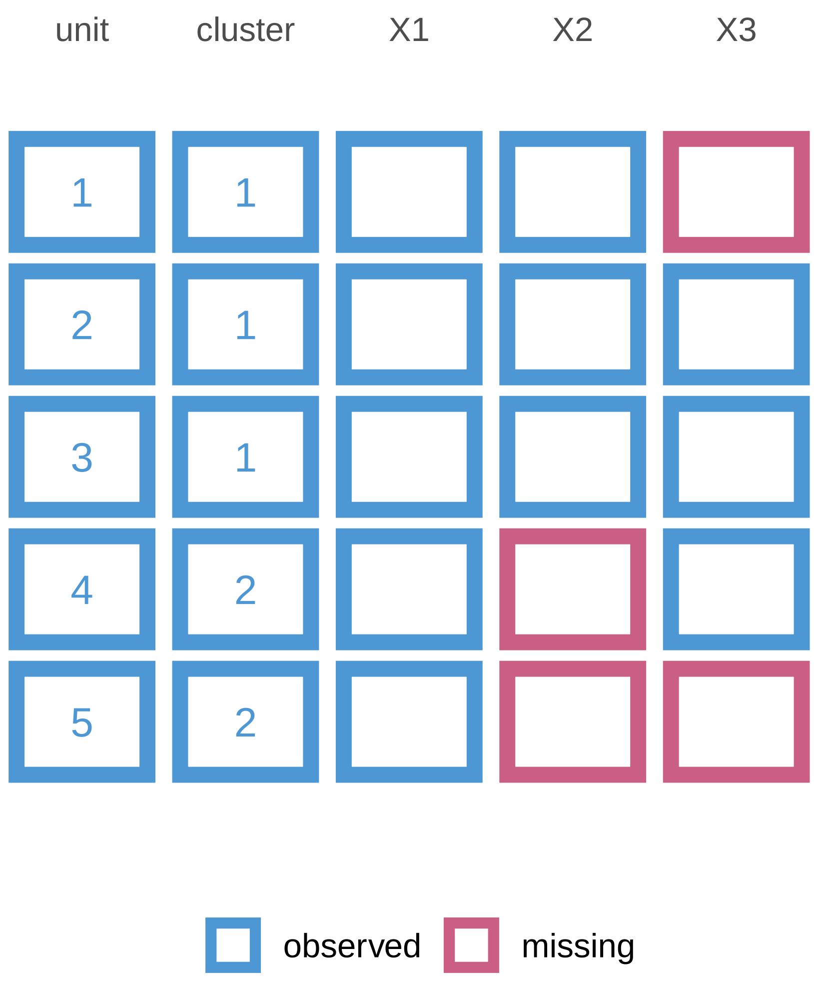


Figure 1: Missingness in multilevel data

practical guidelines and code snippets for different missing data situations. For reasons of brevity, we focus on imputation by chained equations, although JOMO is available in **mice** as well. Other useful packages for incomplete multilevel data include: **mitml**, **miceadds**, **mdmb**.[6]

We structure this tutorial around three case studies:

- `mice::popmis` (simulated data on school kids, with MNAR/MAR mixture);
- `metamisc::impact` (real IPD on traumatic brain injuries, without `NA`s);
- `GJRM::hiv` (simulated patient data on HIV, without `NA`s)

For each case study we focus on a different aspect to illustrate how to impute incomplete multilevel data.

2. Workflows

We introduce three case studies to illustrate the workflow. In the `mice::popmis` data, we show the advantages of including the multilevel structure of the data into the imputation model. In the `metamisc::impact` data we'll show how to induce missingness and solve it in real-world data. In the `GJRM::hiv` we provide novel methodology[7] for imputing MNAR missingness according to the Heckman model.

For each case study we'll look at least at: 1) the incomplete data; 2) the imputation model; 3) the imputed data; and 4) how to obtain pooled estimates for the analysis of scientific interest.

2.1. Case Study I: Popularity

`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.

Incomplete data

The popularity data is created such that there are strong relations between the incomplete variables and the clustering variable `class`. We can express this using the intra-class correlation (ICC). For `popular` the ICC is 0.33. For `popteach` it is 0.31. It would thus be wise to use multilevel modeling.

[6]Rephrase: Some level of knowledge on multilevel models is assumed. We're providing an overview of implementations. It's up-to the reader to decide which multilevel strategy suits their data. So we won't go into detail for the different methods (and equations). Refer to Meng (1994), an Audigier paper, and a paper by Grund on congeniality and random slopes. This paper is just a software tutorial. We'll keep it practical.

[7]not really, the methods exist already, but how to show that this is something new and exciting?

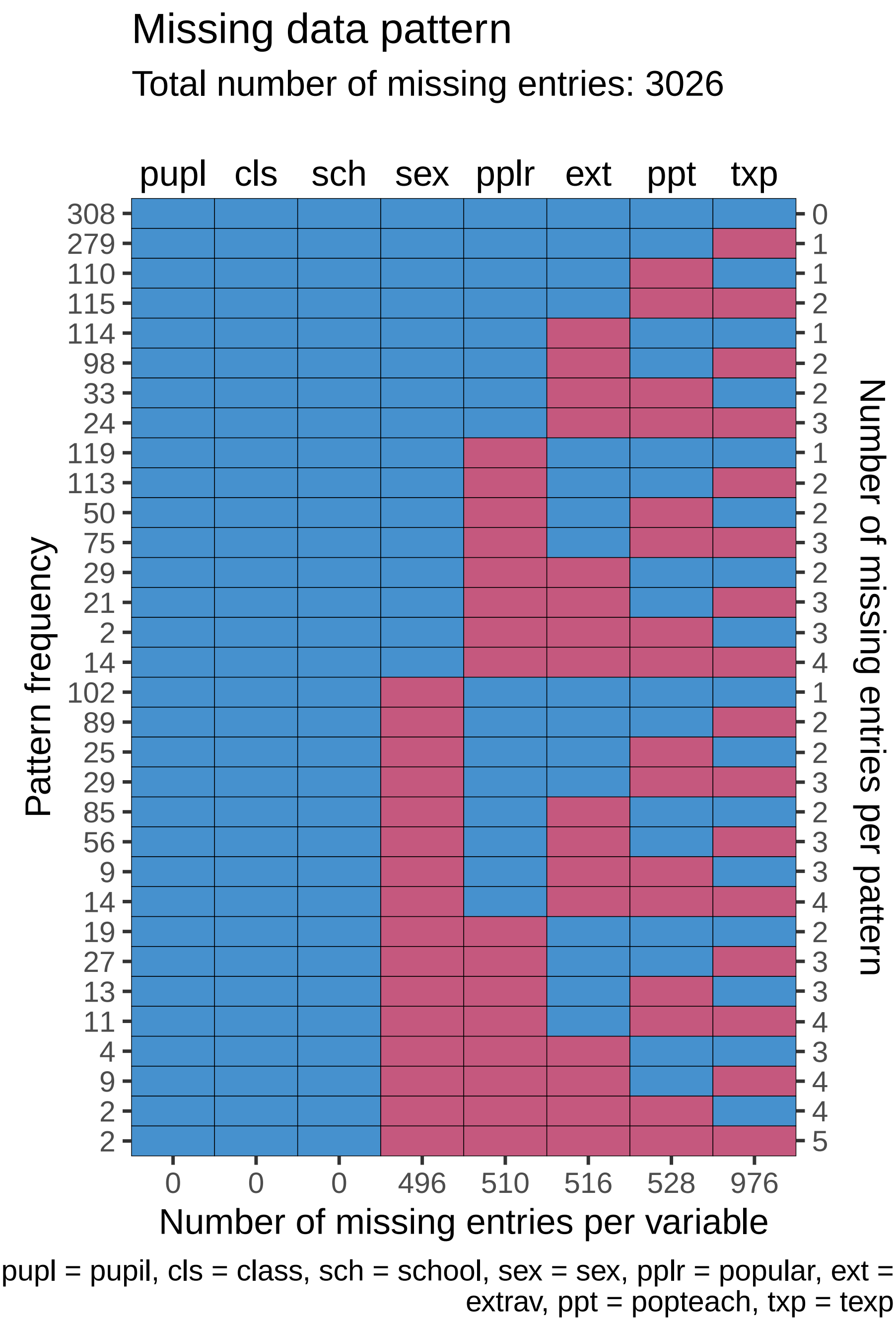


Figure 2: Missing data pattern in the popularity data

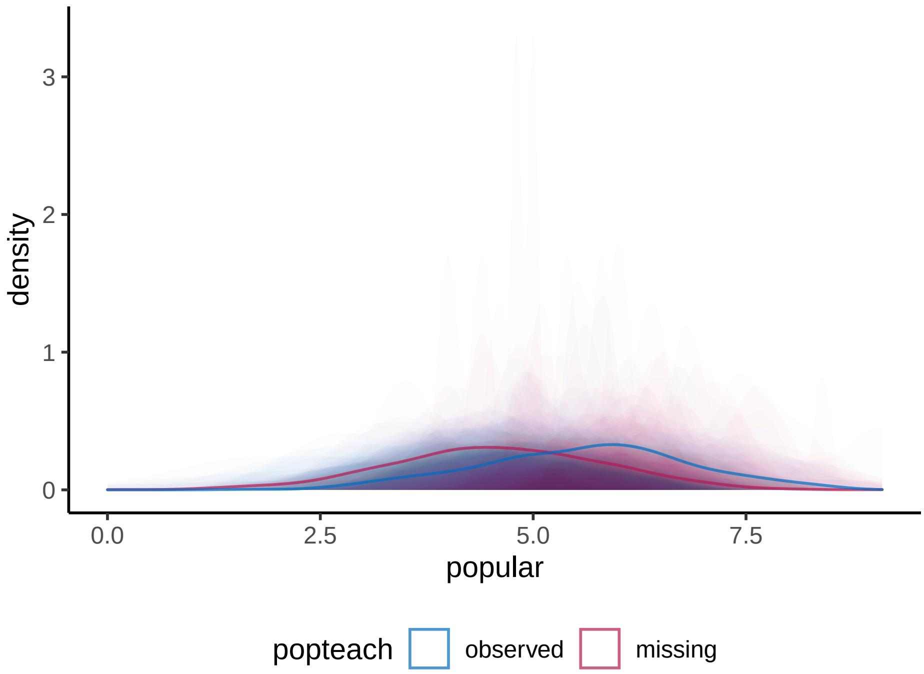


Figure 3: Conditional distributions in the popularity data

The missingness in this dataset is induced conform MAR and MNAR mechanisms. The missing data pattern, Figure 2, shows the systematic nature of the missingness.

To develop the best imputation model, we need to know whether the missingness in one variable depends on the observed values of other variables. Visual inspection usually suffices. We'll highlight only two variables to illustrate, but ideally one would inspect all relations. The questions we'll ask are: 'Does the missing data of `popular` depend on `popteach`?' and 'Does the missingness in teacher popularity depend on pupil popularity?' We'll evaluate this by making a histogram of `popteach` separately for the pupils with known popularity and missing popularity, and the other way around.

In Figure **??** we see that the distribution for the missing `popular` is further to the right than the distribution for observed `popular`. This would indicate a right-tailed MAR missingness. In fact, this is exactly what happens, because the missingness in these data was created manually. Now, we've made it observable by examining the relations between the missingness in popular and the observed data in `popteach`. There is also a dependency between the missingness in teacher popularity and pupil popularity. The relation seems to be right-tailed as well.

Imputation model

The first imputation model that we'll use is likely to be invalid. In this model, we ignore the multilevel structure of the data, despite the high ICCs. This is purely to illustrate the effects of ignoring the clustering in our imputation effort.

We'll use predictive mean matching to impute the continuous variables and logistic regression to impute the binary variable `sex`. We do not use the observation identifier `pupil` or cluster

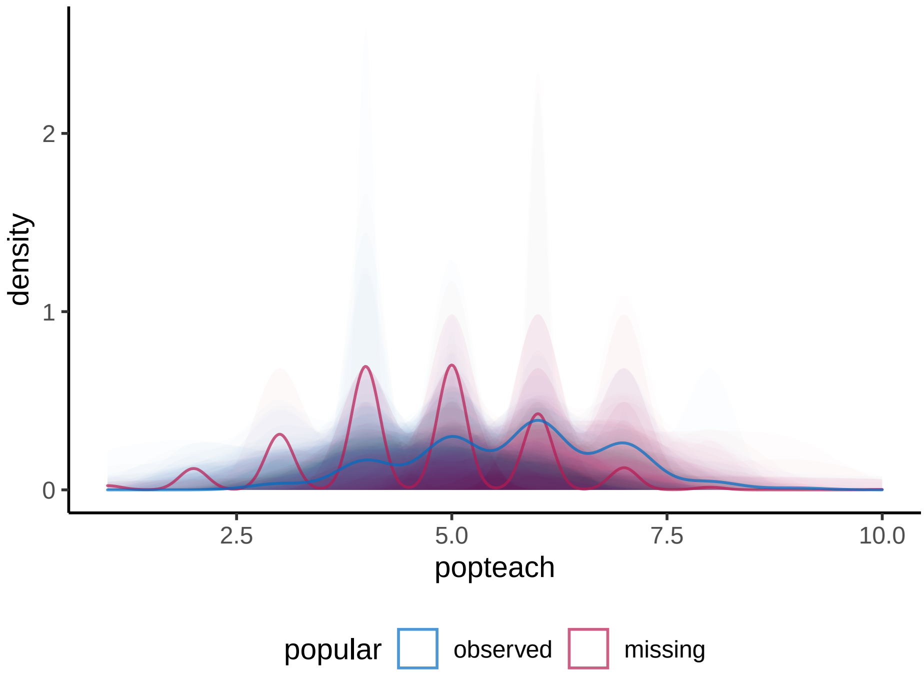
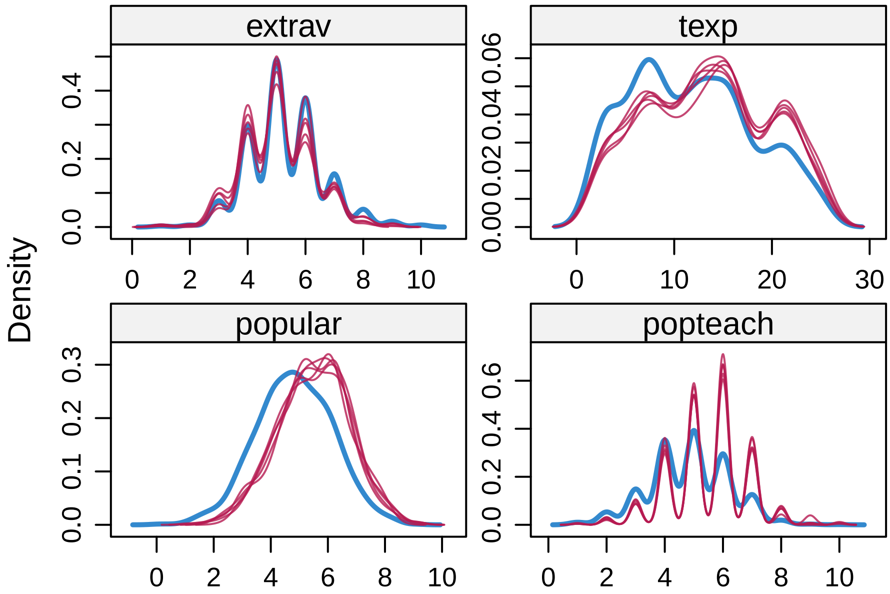


Figure 4: Conditional distributions in the popularity data

identifier `class` as predictors to impute other variables.

```
R> # dry run to get imputation parameters
R> ini <- mice(pop, maxit = 0)
R>
R> # extract predictor matrix and adjust
R> pred <- ini$pred
R> pred[, c("class", "pupil")] <- 0
R>
R> # impute the data, ignoring the cluster structure
R> imp_ignored <- mice(pop, maxit = 10, pred = pred, print = FALSE)
```

Imputed data




```
      vars incomplete   ignored
1  popular  0.3280070 0.2958571
2 popteach  0.3138658 0.2561198
3     texp  1.0000000 0.4696024
```

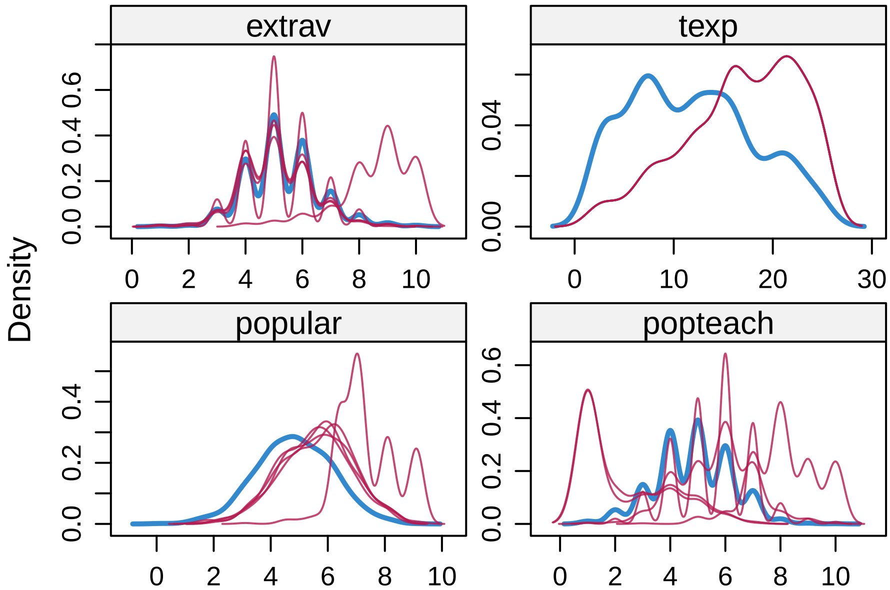
As the original ICCs show, 100% of the variance in `texp` can be attributed to the clustering variable `class`. This tells us that the multilevel structure of the data should be taken into account. If we don't, we'll end up with incorrect imputations, biasing the effect of the clusters towards zero.

We can also observe that the teacher experience increases slightly after imputation. This is due to the MNAR missingness in `texp`. Higher values for `texp` have a larger probability to be missing. This may not a problem, however, if at least one pupil in each class has teacher experience recorded, we can deductively impute the correct (i.e. true) value for every pupil in the class.

Imputation model

We'll now use `class` as a predictor to impute all other variables. This is still not recommended practice, since it only works under certain circumstances and results may be biased. But at least, it includes some multilevel aspect. Colloquially, this is 'multilevel imputation for dummies'.

Imputed data




```
      vars incomplete   ignored predictor
1  popular  0.3280070 0.2958571 0.3581435
2 popteach  0.3138658 0.2561198 0.3446899
3     texp  1.0000000 0.4696024 1.0000000
```

Now, we can clearly see that the imputed values of `texp` are higher than the observed values, which is in line with right-tailed MNAR.

The ICCs are way more in line with the ICCs in the incomplete data. But this is a quick and dirty way of imputing multilevel data. We *should* be using a multilevel model.

Imputation model

To include...

2.2. Case study II: IMPACT

`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).

The data is already imputed (Steyerberg et al, 2008), so we'll induce missingness ourselves. For example, MAR missingness varying by cluster.[8]

2.3. Pooling

- Analysis of scientific interest.
- Pooling using `mitml`.
- Pooling 'regular' parameters vs more 'exotic' parameters (SE of residual errors, or autocorrelation)

3. Discussion

- JOMO in **mice** -> on the side for now
- Additional levels of clustering
- More complex data types: timeseries and polynomial relationship in the clustering.

4. Think about

- Adding some kind of help function to mice that suggests a suitable predictor matrix to the user, given a certain analysis model.
- Adding a `multilevel_ampute()` wrapper function in mice.
- Exporting `mids` objects to other packages like `lme4` or `coxme`?

References

de Jong VMT, Moons KGM, Eijkemans MJC, Riley RD, Debray TPA (2021). "Developing More Generalizable Prediction Models from Pooled Studies and Large Clustered Data Sets." *Statistics in Medicine*, **40**(15), 3533–3559. ISSN 1097-0258. doi:10.1002/sim.8981.

[8]Observed data pattern should differ per cluster. So, in cluster 1, the missingness would depend on age, but not in cluster two. Split the dataframe and run `ampute()` on each cluster.

Hox J, van Buuren S, Jolani S (2015). "Incomplete Multilevel Data: Problems and Solutions." In J Harring, L Staplecton, S Beretvas (eds.), *Advances in Multilevel Modeling for Educational Research: Addressing Practical Issues Found in Real-World Applications*, CILVR Series on Latent Variable Methodology, pp. 39–62. Information Age Publishing Inc., Charlotte, NC. ISBN 978-1-68123-328-4.

Hox JJ, Moerbeek M, van de Schoot R (2017). *Multilevel Analysis: Techniques and Applications, Third Edition*. Routledge. ISBN 978-1-317-30868-3.

Meng XL (1994). "Multiple-Imputation Inferences with Uncongenial Sources of Input." *Statistical Science*, **9**(4), 538–558. ISSN 0883-4237, 2168-8745. doi:10.1214/ss/1177010269.

Resche-Rigon M, White IR, Bartlett JW, Peters SA, Thompson SG, Group obotPIS (2013). "Multiple Imputation for Handling Systematically Missing Confounders in Meta-Analysis of Individual Participant Data." *Statistics in Medicine*, **32**(28), 4890–4905. ISSN 1097-0258. doi:10.1002/sim.5894.

Van Buuren S (2018). *Flexible Imputation of Missing Data*. Chapman and Hall/CRC.

Yucel RM (2008). "Multiple Imputation Inference for Multivariate Multilevel Continuous Data with Ignorable Non-Response." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **366**(1874), 2389–2403. doi:10.1098/rsta.2008.0038.

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