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 Mu~noz Avila
University Medical Center Utrecht

Valentijn de Jong
University Medical Center Utrecht

Gerko Vink
Utrecht University

Thomas Debray
University Medical Center Utrecht

Abstract

Multilevel data is not spared the ubiquitous problem of missing information. This is a tutorial paper on imputing incomplete multilevel data with **mice**. Including methods for ignorable and non-ignorable missingness. Footnotes in the current version show work in progress/under construction. The last section is not part of the manuscript, but purely for reminders.

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

1. Introduction

In many contemporary data analysis efforts, some form of hierarchical or clustered structure is recorded. For example, students clustered within schools, or patients clustered within studies in individual patient data meta-analyses. Analyzing such multilevel data requires specialized techniques that take the clustered structure into account, since ignoring it may be harmful to the statistical inferences and can yield biased results. Imagine a case where cross-level interactions between unit-level variables and cluster-level variables are present. The cluster to which a unit belongs may then influence the unit-level observations–and vice versa for each of the units that make up the cluster (Hox, Moerbeek, and van de Schoot 2017). These relations can and should be taken into account when developing analysis models for multilevel data for the simple reason that groups of observations share some common variance.

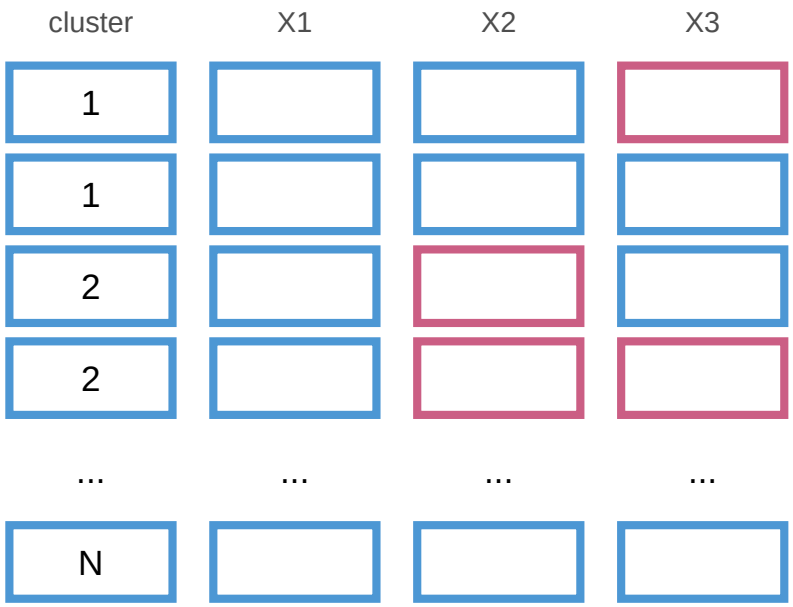


Figure 1: Missingness in multilevel data

The variability due to clustering is often measured by means of the intraclass coefficient (ICC). The ICC can be seen as the percentage of variance that can be attributed to the cluster-level, where a high ICC would indicate that a lot of variability is due to the cluster structure. Multilevel models typically accommodate for this variability by including separate intercepts for each cluster. Such fixed effects relieve the restriction imposed by single-level models: equal group means across clusters. Additionally, models may include random effects of predictors across clusters, and random error terms (heterogeneous residual error variances; see e.g. Hox et al. (2017) and de Jong, Moons, Eijkemans, Riley, and Debray (2021)). 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'.

1.1. Incomplete multilevel data

The process of analyzing multilevel data is further complicated when not all data entries are observed. Just as with single level data, missingness may occur at the unit level. But with multiple levels of data comes the potential for clustered missingness. Missingness in multilevel data can therefore be categorized into two general patterns: systematic missingness and sporadic missingness (Resche-Rigon, White, Bartlett, Peters, and Thompson 2013). We have visualized the difference between these types of missingness in Figure 1. The figure shows an $n \times p$ set $\mathbf{X} = X_1, \ldots, X_p$, with n units distributed over N clusters, and $p = 3$ columns. Column `X1` is completely observed. Column `X2` is systematically missing and column `X3` is sporadically missing. Systematic missingness can be further subdivided into unobserved constants (i.e., the same value within clusters) and non-measured random variables (which may differ per unit within clusters). In Figure 1, the former would imply that the unobserved

values for units 3 and 4 on column `X2` are identical. With the latter, the values would differ. The optimal strategy for dealing with the missingness may therefore depend on the observed missing data pattern.

Another characteristic of the missing data to take into account in analyses is the mechanism behind the missingness. Although the essence of the true non-response mechanism may not be known, it can be inferred or assumed to be one of the following:

- Missing Completely At Random (MCAR), where the probability to be missing is equal across all data entries;
- Missing At Random (MAR), where the probability to be missing depends on observed information;
- Missing Not At Random (MNAR), where the probability to be missing depends on unrecorded information, making the missingness non-ignorable (Rubin 1976; Meng 1994). Depending on the assumed missingness mechanism, missing data handling strategies may be more or less suitable, see e.g., Yucel (2008) and Hox, van Buuren, and Jolani (2015).

Ignoring the missingness in analyses can be extremely harmful to inferences. 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 and is widely known as imputation. Imputing (i.e., filling in) the missing values separates the missing data problem from the scientific problem: missing data are replaced by plausible values whereafter the completed data is analysed as if it were completely observed. The R package **mice** has become the de-facto standard for imputation by chained equations, which iteratively solves the missingness on a variable-by-variable basis. **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.

1.2. Aim of this paper

This papers serves as a tutorial for imputing incomplete multilevel data with **mice**. We provide practical guidelines and code snippets for different missing data situations. For reasons of brevity, we focus on imputation by chained equations exclusively (although the alternative, joint modeling imputation for multilevel data or **jomo** Quartagno, Grund, and Carpenter (2019), has been implemented in **mice** as well). Other useful resources for the analysis of incomplete multilevel data include the R packages **mitml**, **miceadds**, and **mdmb**, and the empirical work by Audigier, White, Jolani, Debray, Quartagno, Carpenter, van Buuren, and Resche-Rigon (2018) and Grund, Lüdtke, and Robitzsch (2018). Note that this tutorial paper assumes a basic level of knowledge on multilevel models. We're providing an overview of implementations. It's up-to the reader to decide which multilevel strategy suits their data. We won't go into detail for the different methods (and equations). This paper is just a software tutorial, so we'll keep it practical. Assumed knowledge also includes the use of the 'piping operator', `%>%`, adopted from the `magrittr` package and avidly used among `tidy` workflows, and the `lme4` notation for multilevel models.

To illustrate how to impute incomplete multilevel data, we structure this tutorial around three case studies:

- `mice::popmis` (simulated data on perceived popularity, $n = 2,000$ pupils across $N = 100$ schools, with univariate MAR missingness);
- `GJRM::hiv` (simulated data on HIV diagnoses, $n = 6,416$ patients across $N = 9$ regions, with univariate MNAR missingness)
- `metamisc::impact` (empirical data on traumatic brain injuries, $n = 11,022$ patients across $N = 15$ studies, without `NA`s).

In each case study we highlight one aspect of the imputation workflow. With the `mice::popmis` data, we show different ways of designing an imputation model and what happens if you misspecify this model. With the `GJRM::hiv` data we extend the imputation model to include Heckman-type selection-inclusion methods. And with the `metamisc::impact` data we provide a example of multivariate missingness in real-world data. For all case studies we discuss the nature of the incomplete data, the imputation model(s), and evaluation of the imputed data: A. Choose an analysis model (so the imputation model will be compatible with the analyses); B. Evaluate the incomplete data; C. Develop the imputation model(s); D. Impute the missingness; E. Evaluate the imputations.

2. How (not) to impute (Case Study I: Popularity)

In this section we'll go over the different steps involved with imputing incomplete multilevel data. The data we're using is the `popmis` dataset from the `mice` package. This is a simulated dataset with pupils ($n = 2000$) clustered within schools ($N = 100$). In this tutorial we'll use the following variables:

- `school`, school identification number (clustering variable);
- `popular`, pupil popularity (self-rating between 0 and 10; unit-level);
- `sex`, pupil sex (0=boy, 1=girl; unit-level);
- `texp`, teacher experience (in years; cluster-level).

The analysis model corresponding to this dataset is multilevel regression with random intercepts, random slopes and a cross-level interaction. The outcome variable is `popular`, which is predicted from the unit-level variable `sex` and the cluster-level variable `texp`. The regression equation[1] and `lme4` notation for this model are

$$\text{popular}_{ij} = \gamma_{00} + \gamma_{10}\ \text{sex}_{ij} + \gamma_{01}\ \text{texp}_j + \gamma_{11}\ \text{texp}_j \times \text{sex}_{ij} + u_{0j} + u_{1j}\ \text{sex}_{ij} + e_{ij}$$

```
popular ~ 1 + sex + texp + sex:texp + (1 + sex | school)
```

TODO: add all estimimates, seed, and version nr., this also shows the workflow for pooling etc.

[1] add the 'level notation' (Bryk and Raudenbush, 1992) and/or matrix notation ('linear mixed effects model'; Laird and Ware, 1982) too?

Since the data is simulated and the missingness is induced, we can compare our inferences after imputation to the true complete data. The data is created in such a way that the clustering variable `school` explains quite some variance in the outcome variable `popular`. We express this using the intraclass correlation, ICC = 0.58. We'll evaluate the ICC after each missing data strategy, and compare the estimated fixed effects:

```
# A tibble: 4 x 2
  term        estimate
  <chr>       <chr>
1 (Intercept) " 3.314 [ 2.998,  3.629]"
2 sex         " 1.330 [ 1.069,  1.590]"
3 texp        " 0.110 [ 0.090,  0.130]"
4 sex:texp    "-0.034 [-0.051, -0.017]"

# A tibble: 4 x 2
  term                 estimate
  <chr>                <chr>
1 sd__(Intercept)      0.642
2 cor__(Intercept).sex 0.077
3 sd__sex              0.476
4 sd__Observation      0.626
```

Incomplete data

Load the data into the environment and select the relevant variables:

```
R> popmis <- mice::popmis[, c("school", "popular", "sex", "texp")]
```

The missingness is univariate and sporadic, which is illustrated in the missing data pattern in Figure 2. The ICC in the incomplete data is 0.56.

To develop the best imputation model for the incomplete variable `popular`, we need to know whether the missingness depends on the observed values of other variables. We'll highlight only one other variable to illustrate, but ideally one would inspect all relations. The questions we'll ask are: 'Does the missing data of pupil popularity (`popular`) depend on observed teacher popularity (`texp`)?'. This can be evaluated statistically, but visual inspection usually suffices. We'll make a histogram of `texp` separately for the pupils with known popularity and missing popularity.

In Figure 3 we see that [**update this part**] 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.) We've made it observable by examining the relations between the missingness in popular and the observed data in `texp`.

Complete case analysis (not recommended)

Complete case analysis ignores the observations with missingness altogether, which may even introduce bias in MCAR situations.

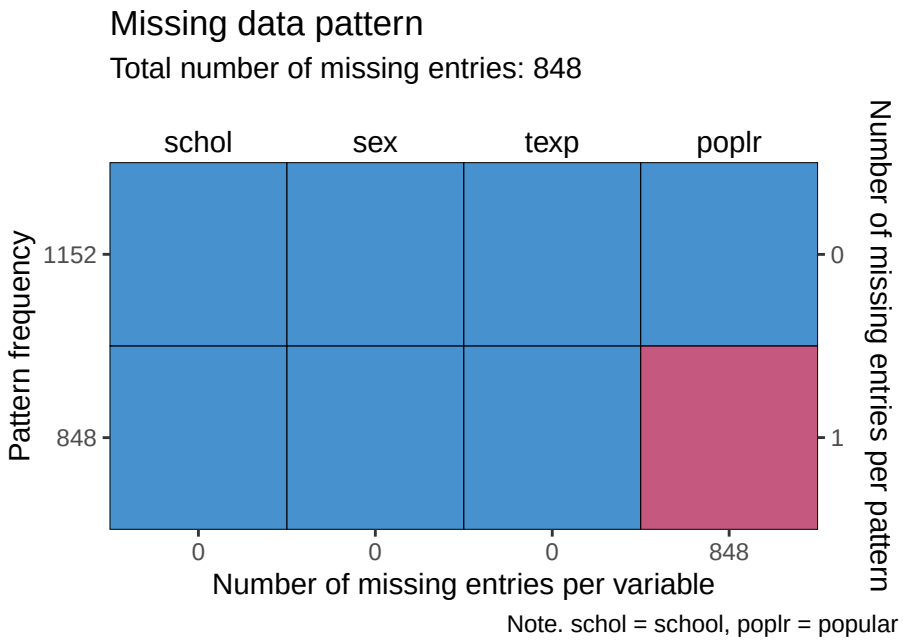


Figure 2: Missing data pattern in the popularity data

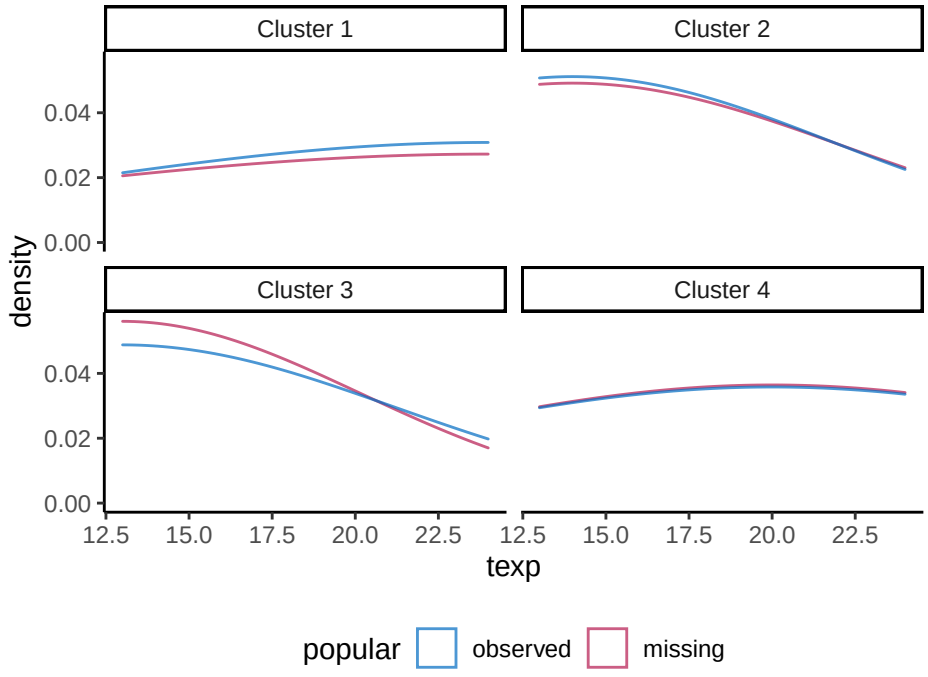


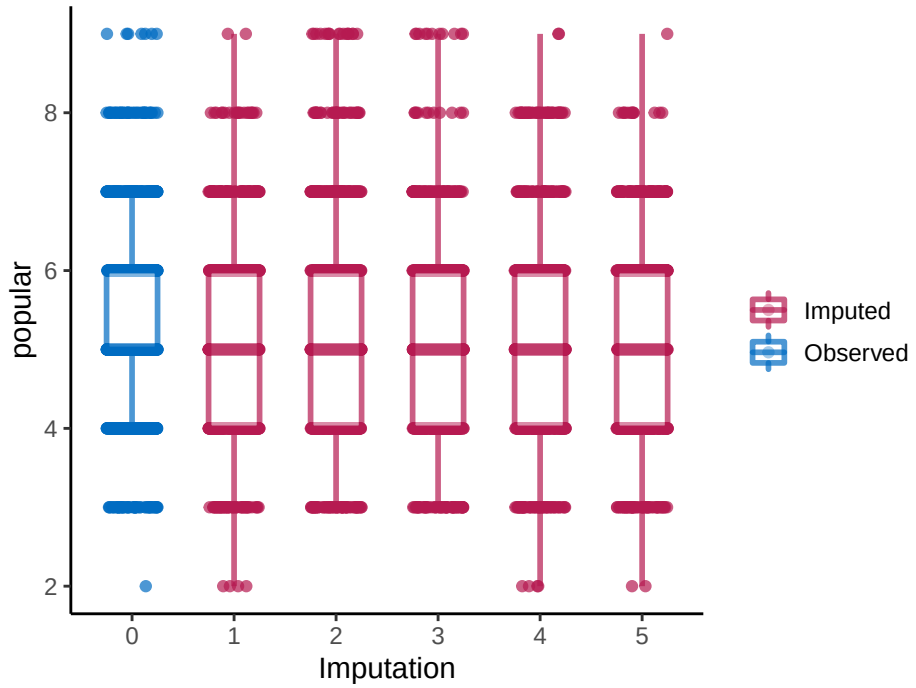
Figure 3: Conditional distributions in the popularity data

Imputation ignoring the cluster variable (not recommended)

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 identifier `school` as predictors to impute other variables.

```
R> # dry run to get imputation parameters
R> ini <- mice(popmis, maxit = 0)
R>
R> # extract predictor matrix and adjust
R> pred <- ini$pred
R> pred[, c("school")] <- 0
R>
R> # impute the data, ignoring the cluster structure
R> imp_ignored <- mice(popmis, maxit = 1, pred = pred, print = FALSE)
```



As the original ICCs show, 100% of the variance in `texp` can be attributed to the clustering variable `school`. 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 school has teacher

experience recorded, we can deductively impute the correct (i.e. true) value for every pupil in the school.

Add: Assumes exchangability between units.

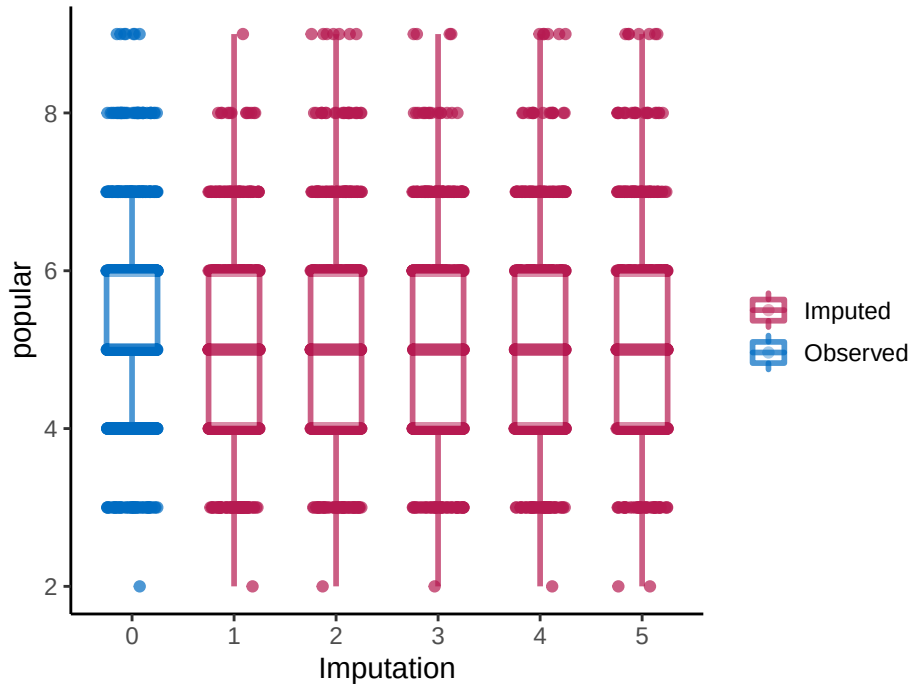
Imputation with the cluster variable as predictor (not recommended)

We'll now use `school` 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 (Drechsler 2015; Enders, Mistler, and Keller 2016). But at least, it includes some multilevel aspect. This method is also called 'fixed cluster imputation', and uses N-1 indicator variables representing allocation of N clusters as a fixed factor in the model (Reiter, Raghunathan, and Kinney 2006; Enders et al. 2016). Colloquially, this is 'multilevel imputation for dummies'.

Add: doesn't work with syst missing (only sporadically). There's some pro's and con's. May not differ much if the number of clusters is low.

The more the random effects are of interest, the more you need ml models.

```
R> # adjust the predictor matrix
R> pred <- ini$pred
R> # pred[, "pupil"] <- 0
R>
R> # impute the data, cluster as predictor
R> imp_predictor <- mice(popmis, maxit = 1, pred = pred, print = FALSE)
```



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

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 with random effects

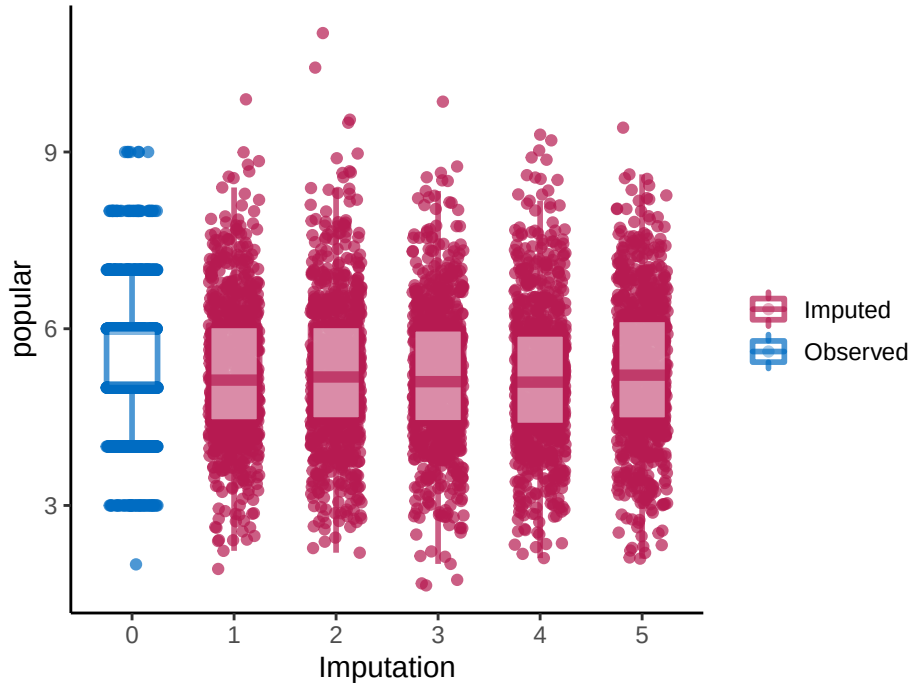
With `2l.norm` we impute the outcome with a multilevel model assuming random slopes for each variable in the imputation model and homogeneous within-cluster variance.

"Van Buuren (2011) considered the homoscedastic linear mixed model as invalid for imputing incomplete predictors, and investigated only the 2l.norm method, which allows for heterogeneous error variances" (Van Buuren 2018).

```
R> pred <- ini$pred
R> pred["popular", ] <- c(-2, 2, 2, 2)
R> #-2 for the cluster variable, 2 for random effects
R> meth <- ini$meth
R> meth <- c("", "2l.norm", "",  "")
R> # meth <- c("", "2l.pmm", "",  "")
R> imp_norm_2l <-
+   mice(
+     popmis %>% mutate(school = as.integer(school)),
+     pred = pred,
+     meth = meth,
+     maxit = 1,
+     print = FALSE
+   )
```

```
Warning: Removed 848 rows containing non-finite values (stat_boxplot).
```

```
Warning: Removed 848 rows containing missing values (geom_point).
```



Imputation with random effects and heterogeneity

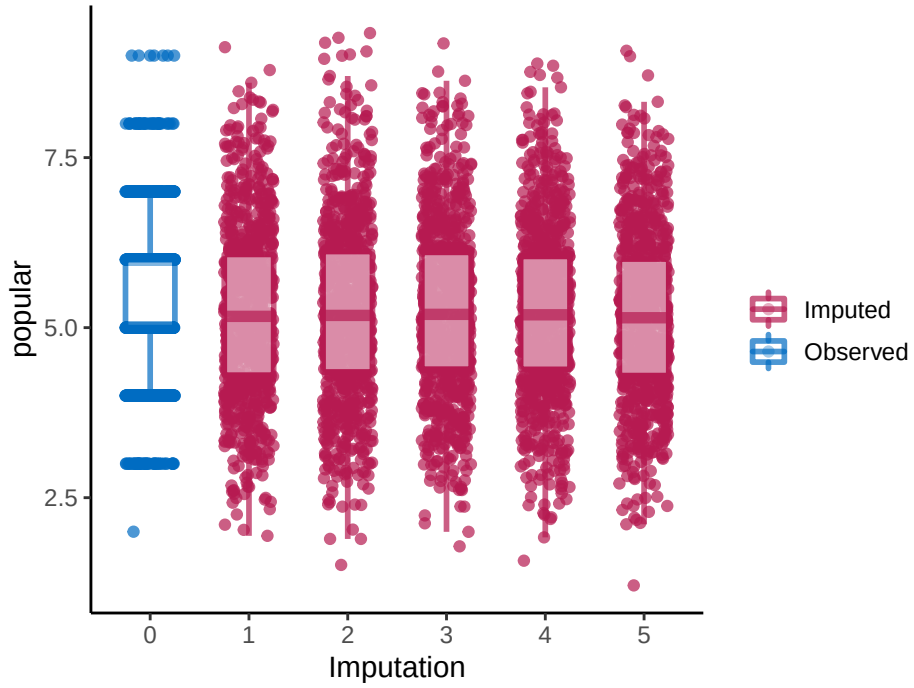
This method assumes random slopes for each variable in the imputation model. In contrast to `2l.norm` this method allows a cluster-specific residual error variance.

Schafer book fortran workflows.

```
R> pred["popular", ] <- c(-2, 2, 1, 2)
R> meth <- c("", "2l.pan", "", "")
R> imp_pan_2l <-
+   mice(
+     popmis %>% mutate(school = as.integer(school)),
+     pred = pred,
+     meth = meth,
+     maxit = 1,
+     print = FALSE
+   )
```

```
Warning: Removed 848 rows containing non-finite values (stat_boxplot).
```

```
Warning: Removed 848 rows containing missing values (geom_point).
```

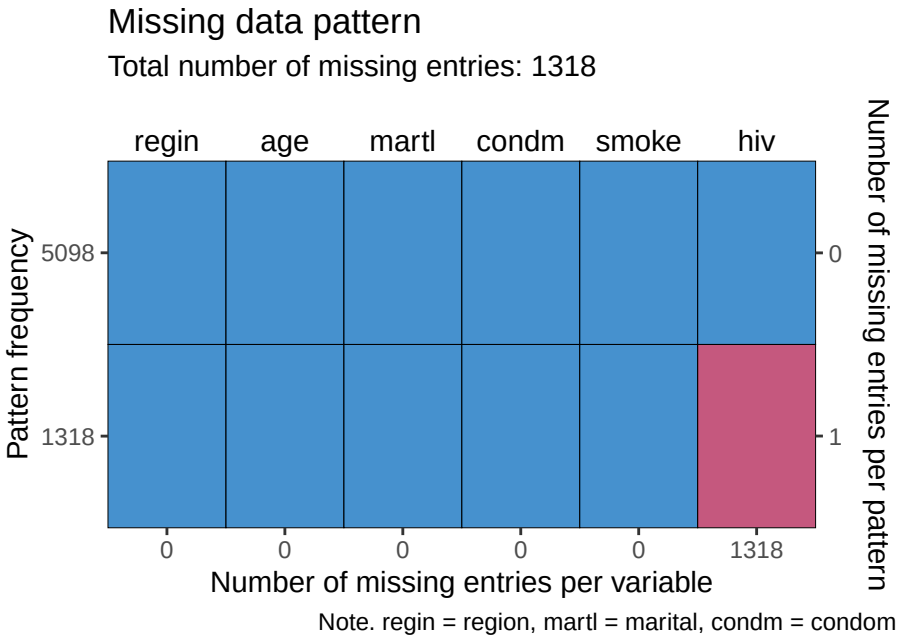


3. How to handle non-random selection (Case study II: HIV)

Data are simulated and included in the `GJRM` package. We will use the following variables:

- `region` Cluster variable,
- `hiv` HIV diagnosis (0=no, 1=yes),
- `age` Age of the patient,
- `marital` Marital status,
- `condom` Condom use during last intercourse,
- `smoke` Smoker (levels; inclusion restriction variable).

The imputation of these date is based on the toy example from IPDMA Heckman Github repo.



From the missing data pattern we see that we can set `maxit` to 1, since there is only one variable with missingness.

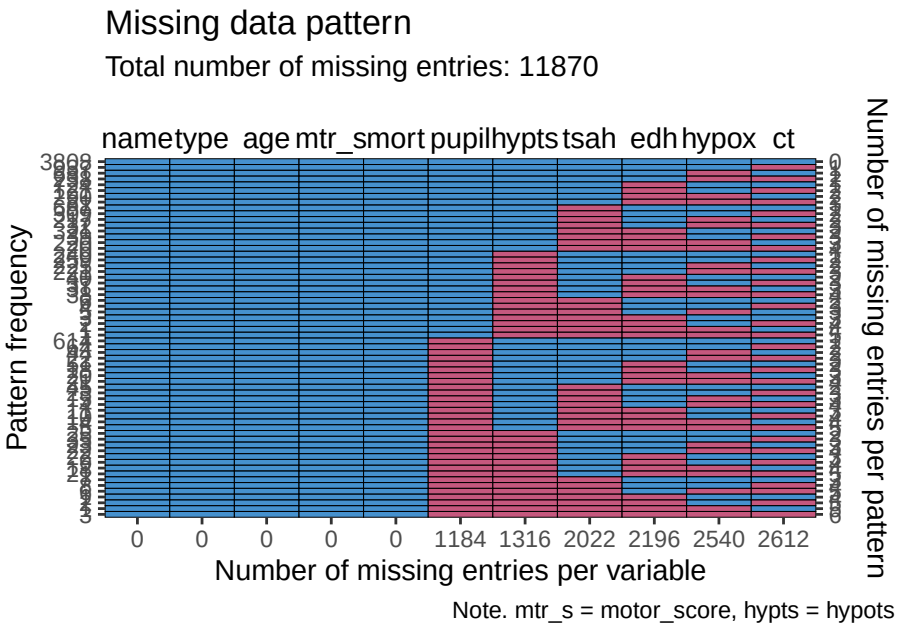
The inclusion restriction variable should be a predictor of the the actual value of the variable of interest, but not of missingness indicator for the variable of interest. In this case, the data were simulated to adhere to this requirement. Namely, β_{smoke} = -0.064, 95% CI [-0.256, 0.126] for the analysis model (`formula = hiv ~ .`), and β_{smoke} = -0.265, 95% CI [-0.422, -0.11] for the selection model (`formula = is.na(hiv) ~ .`). This means the assumptions for the Heckman-type selection model are met.

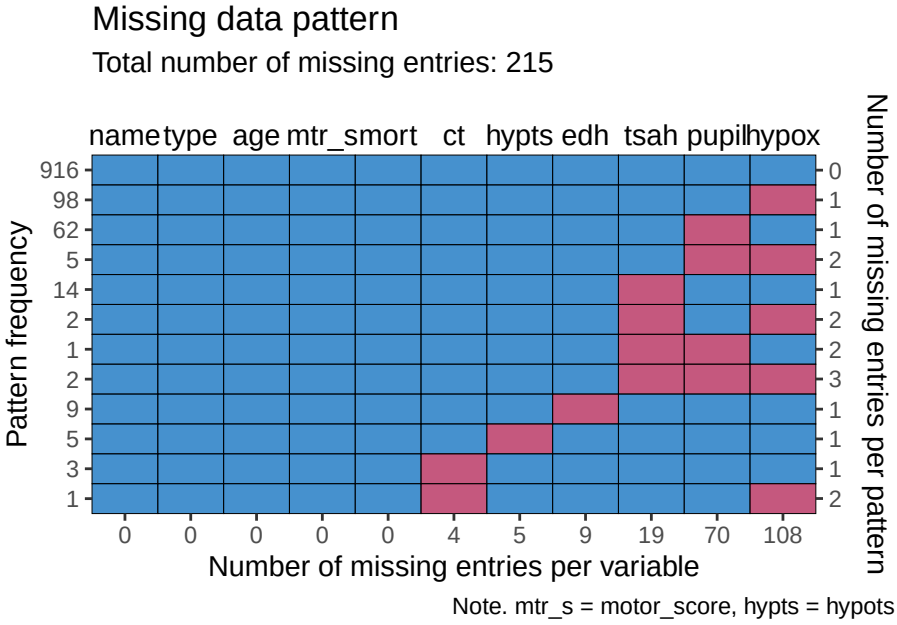
4. How to handle multivariate missingness (Case study III: IMPACT)

`impact` is traumatic brain injury data with patients, $n = 11022$, clustered in studies, $N = 15$. With the following 11 variables:

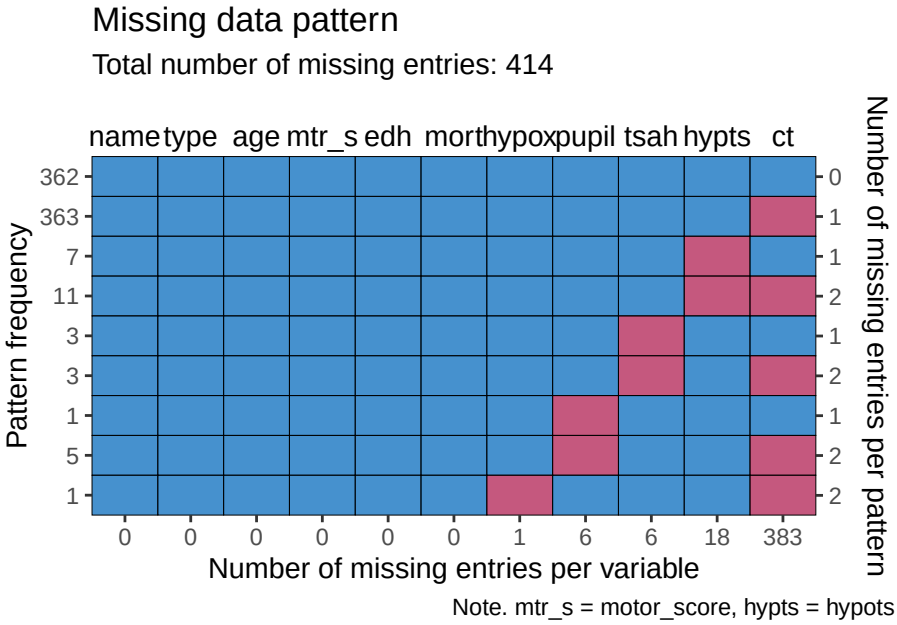
- `name` Name of the study,
- `type` Type of study (RCT: randomized controlled trial, OBS: observational cohort),
- `age` Age of the patient,
- `motor_score` Glasgow Coma Scale motor score,
- `pupil` Pupillary reactivity,
- `ct` Marshall Computerized Tomography classification,
- `hypox` Hypoxia (0=no, 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.[2]





[2]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.



TODO: rotate labels, make missingness % gradient for clusters?? Look at analysis model, maybe copy from GREAT data example e.g., adjusted prognostic effect of ct on unfortunate outcomes.

gradient would be the number of clusters for which this variable is missing

5. Discussion

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

6. 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`?
- Adding a ICC=0 dataset to show that even if there is no clustering it doesn't hurt.
- env dump in repo

References

Audigier V, White IR, Jolani S, Debray TPA, Quartagno M, Carpenter J, van Buuren S, Resche-Rigon M (2018). "Multiple Imputation for Multilevel Data with Continuous and Binary Variables." Statistical Science, **33**(2), 160–183. ISSN 0883-4237, 2168-8745. doi:10.1214/18-STS646. 1702.00971.

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.

Drechsler J (2015). "Multiple Imputation of Multilevel Missing Data—Rigor Versus Simplicity." Journal of Educational and Behavioral Statistics, **40**(1), 69–95. ISSN 1076-9986. doi:10.3102/1076998614563393.

Enders CK, Mistler SA, Keller BT (2016). "Multilevel Multiple Imputation: A Review and Evaluation of Joint Modeling and Chained Equations Imputation." Psychological Methods, **21**(2), 222–240. ISSN 1939-1463. doi:10.1037/met0000063.

Grund S, Lüdtke O, Robitzsch A (2018). "Multiple Imputation of Missing Data for Multilevel Models: Simulations and Recommendations." Organizational Research Methods, **21**(1), 111–149. ISSN 1094-4281. doi:10.1177/1094428117703686.

Hox J, van Buuren S, Jolani S (2015). "Incomplete Multilevel Data: Problems and Solutions." In J Harring, L Stapleton, 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.

Quartagno M, Grund S, Carpenter J (2019). "Jomo: A Flexible Package for Two-level Joint Modelling Multiple Imputation." The R Journal, **11**(2), 205–228. ISSN 2073-4859.

Reiter JP, Raghunathan T, Kinney SK (2006). "The Importance of Modeling the Sampling Design in Multiple Imputation for Missing Data." undefined.

Resche-Rigon M, White IR, Bartlett JW, Peters SAE, Thompson SG (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 0277-6715. doi:10.1002/sim.5894.

Rubin DB (1976). "Inference and Missing Data." Biometrika, **63**(3), 581–592. doi:10.2307/2335739.

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