Missing data and imputation 102

E-mail distributed on 28-11-2022

Dear all,

First a note that the stat support webspace has been updated with the materials of the most recent workshop on logistic regression. Second, I would like to continue the topic of the last support mail, and elaborate on what **multiple stochastic imputation** is (usually abbreviated MI), and how it can be run correctly.

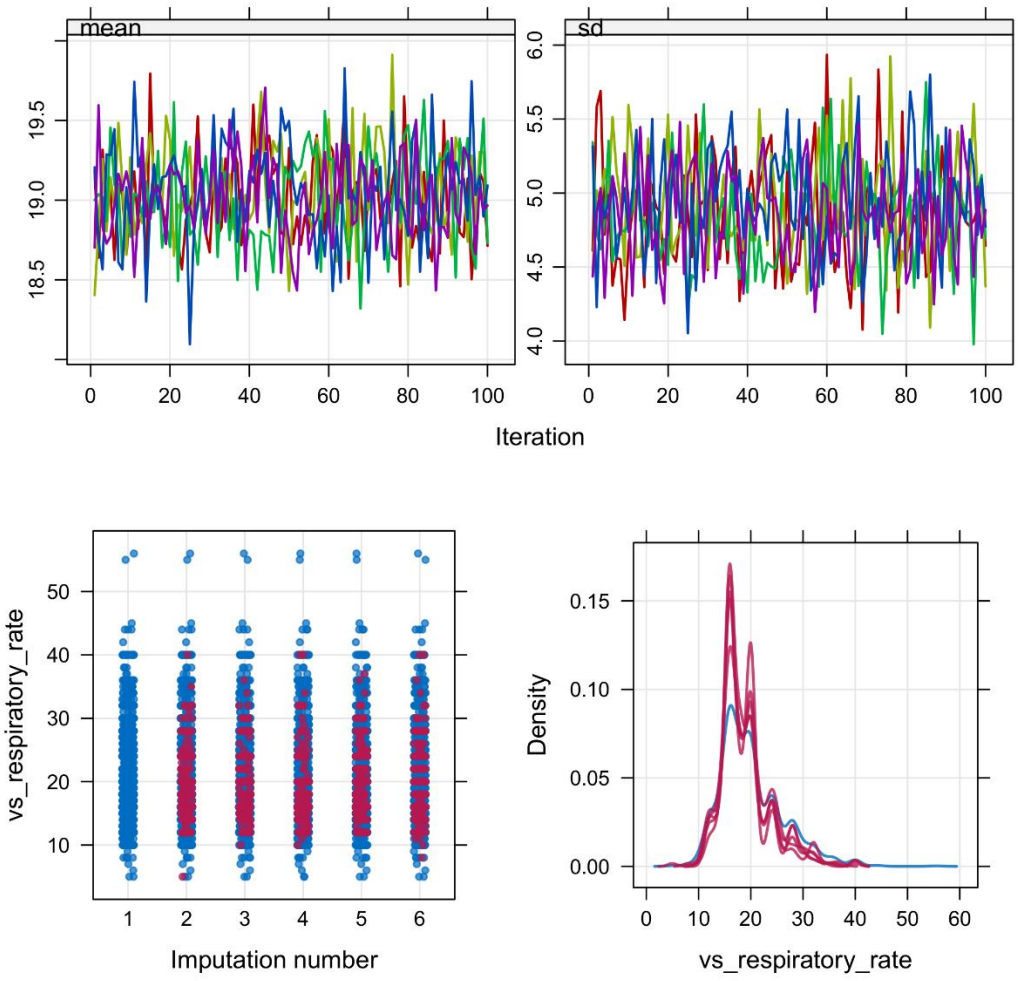
MI is a method for imputing missing data that combines the idea of model-based imputations with the addition of (random) noise, drawn from the model's distributional part (=the "stochastic" part). By doing this, we simulate a value that could have been observed as a real measurement, as opposed to imputing only the model's mean guess, which is likely too perfect. Moreover, we repeat this process *M* times (5-10 are typical choices) to create *M* imputed data sets (=the "multiple" part). The target analysis (e.g., an ANOVA) is then conducted on all imputed data sets, and results are pooled using pooling rules. This procedure simulates natural variability in the imputed data and allows to quantify the contribution of missingness on the final result.

In R, MI can be executed with the R package **mice**, which uses Multiple Imputation by Chained Equations (MICE; van Buuren & Groothuis-Oudshoorn, 2011). Using this package, running MI still requires a few practical choices, as well as cautions (van Buuren (2012):

- **Imputation sequence:** When more than 1 variable has missing values, it is best to start by imputing the variable with the least missingness, so that its imputed values can be used to predict missingness in the variable with the second least missingness, and so on. This is known as *monotone imputation*, although mice allows customized imputation sequences.
- **Imputation model:** A wide range of imputation models is now available, tailored to specific variable types (e.g., continuous, ordinal, categorical). These include standard regression models, Bayesian models, and machine learners. More complex models may be preferred here due to being more flexible in generating plausible imputed values (e.g., random forest).
- **Predictive mean matching:** Often, the imputed values generated by the imputation model are transformed to match the nearest observed value (e.g., 5.3 is altered to 5). This procedure is known as *predictive mean matching* (PMM) and ensures that imputed values are always strictly within the distribution of the observed values. This is especially attractive for discreet/integer variables, where fractional values may otherwise not even exist

- **Derived variables:** Derived variables such as interactions or transformations should never be imputed directly. E.g., we impute weight and height, and then calculate BMI, rather than imputing BMI directly.
- **Dependent variable:** Contrary to intuition, it is recommended to use the DV of the target analysis for imputation (Little, 1992; Moons et al., 2006), as well as variables related to missingness that may not be used in the target analysis.
- **Convergence:** Because MICE relies on computational MCMC sampling, a number of iterations need to be set. This number should be at least 50 or 100.

Once imputations are generated, their quality needs to be checked with diagnostic plots (see attached example). This includes **(1)** checking convergence of the MCMC chains (no visible trends) and **(2)** checking the distribution of the imputed values (red) against the observed values (blue). These distributions should be similar, with no imputed values outside of the observed data range. If you detect problems here it means you may need to change the imputation model or increase the number of MCMC iterations.



When imputation quality is judged to be adequate, we can proceed to the final target analysis. This analysis is run each of the *M* imputed data sets, with results aggregated according to correct pooling rules. The mice package facilitates this part with automated functions.

When reporting an analysis on imputed data, make sure to always fully explain the imputation procedure in the methods section, to report pooled results of the analysis, and to quantify the impact of the missing data on the final results (e.g., the standard deviation on your test statistics).

Best,
Ben

Reference: Van Buuren (2012). *Flexible Imputation of Missing Data*. CRC Press.

Ben Meuleman, Ph.D.
Statistician
Swiss Center for Affective Sciences
University of Geneva | Campus Biotech
Chemin des Mines 9 | CH-1202 Genève
ben.meuleman@unige.ch | +41 (0)22 379 09 79