Journal of Statistical Software

April 2020, Volume 93, Book Review 1. doi: 10.18637/jss.v093.b01

Reviewer: Abdolvahab Khademi
University of Massachusetts

Flexible Imputation of Missing Data (2nd Edition)

Stef van Buuren
Chapman & Hall/CRC, Boca Raton, 2018.
ISBN 9781138588318. xxvii+416 pp. USD 91.95 (H).
https://www.crcpress.com/9781138588318

Occurrence of missing data can cause serious issues, including decreased sample size, biased estimates, and algorithmic problems. Therefore, proper treatment of missing data is a significant part of data analysis in statistics, especially in clinical and experimental studies.

Treatment of missing data is usually included in a section of its own in most textbooks, presenting best or most convenient practices, based on the methods presented and the statistical sophistication of the audience. However, the contexts, complexity, and severity of missing data are so diverse and complicated that its treatment warrants a volume of its own. There are currently several books on missing data, ranging from practical to more theoretical. *Flexible Imputation of Missing Data (2nd Edition)* is an updated addition to the literature on missing data, which combines practice, theory, and applications using the R programming language.

The twelve chapters of the book are grouped in four sections: the basics, advanced techniques, case studies, and extensions. Exposition in each chapter is accompanied by plenty of graphs, code, examples, and exercises. The code and the entire book are available online at the author's personal website. The programming language of the book is R and the treatment of missing data is performed by the package **mice**, which was created by the author.

Chapter 1, *introduction*, provides motivation for dealing with missing data, how missing data occur (e.g. by nonresponse or attrition), current practice in the literature, categories of missing data (MCAR, MAR, MNAR), and common fixes and their advantages and drawbacks. Common practices, such as listwise and pairwise deletion methods, mean imputation, (stochastic) regression regression, last/baseline observation carried forward (LOCF and BOCF), and indicator method (popular in public health) are discussed. The largest portion of the chapter is dedicated to multiple imputation. According to the author, the emphasis of the entire book is on multiple imputation because of its efficiency and statistical properties compared to other methods.

The entire Chapter 2, *multiple imputation*, is devoted to the method of multiple imputation (MI), starting with a historical sketch of the concept and practice of MI. The historical

section is not only anecdotal, but also very informative due to its reference to some works that shaped the foundation of the study of missing data. In the following sections, topics in incomplete data are elaborated, including incomplete-data perspective (which in my opinion is an excellent point by the author), causes of missing data, notation in the book, rigorous definitions of MCAR, MAR, MNAR (and how to simulate them), and ignorable and nonignorable missing data models. Other topics presented in this chapter include the goal of MI, sources of variation in MI (sampling, missing vales, simulation), characteristics of a proper imputations (unbiased population estimand, unbiased sampling variance, confidence valid estimate of variance due to missing data), variance ratio, and degrees of freedom for testing MI. Once an MI is performed, procedures are needed to evaluate the performance of the MI. These procedures are explained next in this chapter along with an example and R code. This chapter ends with discussion on imputation versus prediction, when not to use MI, and how to determine how many MI's are needed in a given missing data analysis scenario.

Applications of MI in common statistical methods are introduced in Chapter 3, *univariate missing data*. A motivating example is introduced at the beginning of the chapter, upon which the author conceptually builds five methods of imputing the missing data. These foundation methods are then taken up and implemented in the rest of the chapter on different inferential methods, including normal linear regression, t-distribution based methods, classification and regression trees, general linear models, count data, semi-continuous data, censored data, and data with nonignorable missingness. Imputation methods are clearly explained in each section with the algorithm and the R code.

In situations where there is more than one variable with missing data, which is very common in real world research and data analysis, different methods and tools are needed. Chapter 4, *multivariate missing data*, introduces imputation techniques for such cases. The author begins with different patterns of missing data in multivariate research, followed by some illustrative plots and R code and output on a dataset. In addition, some useful indices are presented that show the relationship between the variables in the dataset and the potential contribution of those variables in imputing the missing data (inbound and outbound indices, and influx and outflux coefficients). The rest of the chapter focuses on more nuanced issues, theory, and algorithms. Imputation of monotone missing data patterns, joint modeling of continuous and discrete data, and fully conditional specification (FCS) are discussed. Numerical issues in the FCS modeling are discussed in technical depth. Familiarity with stochastic processes and numerical iterative methods are assumed in this section. The chapter ends with an example using the MICE algorithm. The theoretical discussion in this chapter will help the audience when setting the parameters in the **mice** package. Therefore, understanding of the technical aspects, though probably daunting for the beginner, is recommended.

Multiple imputation work flow comprises three phases: imputation of the missing data m times, analysis of the m imputed datasets, and pooling of the parameters across m analyses. Chapter 5, *analysis of imputed data*, exclusively focuses on phase two of the MI work flow. The work flow of choice by the author is introduced first (using the **mice** package), followed by some common practices that the author does not recommend, such as parameter averaging and data stacking. Next, pooling methods based on normal distribution are presented. In the multiple-parameter cases (such as when a categorical variable is encoded as dummy variables), different statistics are introduced, such as multivariate Wald test (D_1), combined test statistics (D_2), and likelihood ratio test (D_3). Stepwise model selection is specifically treated for the potential problems it raises in imputation. To eliminate sampling error, the bootstrap method

in the context of MI is introduced, followed by a short discussion on parallel processing implementation of each imputation by a separate thread or core.

In section II of the book (*advanced techniques*), the author discusses MI techniques in multilevel modeling and causal inference. This section comprises three chapters. In Chapter 6, *imputation in practice*, issues arising in real world data analyses are discussed, primarily in relation to the algorithm and package **mice**. This chapter provides practical solutions to several key points the practitioner needs to address before and during the execution of MI. The points addressed include type of missingness, form of imputation model, set of predictors to include in the imputation process, imputing variables that are function of other variables, setting up imputation and the number of iterations, and the number of MI datasets. Technical considerations, such as visit sequence, convergence of algorithm, and diagnostics are clearly discussed. The author provides practical advice from experience and literature, accompanied by examples, code, and illustrative graphs. For the practitioner, this chapter provides a wealth of practical information that will be immediately applicable.

Imputation of missing data in more complex statistical models, such as random effects, mixed effects, and hierarchical models, are discussed in Chapter 7, *multilevel multiple imputation*. A brief introduction to multilevel modeling is given, contrasted with single level statistics, the notation used in the chapter, and complications involved in the imputation of missing data with multilevel models. Joint modeling and fully conditional specification methods are revisited in this chapter for multilevel data. In the rest of the chapter, the author walks through an example using the **mice** package on different variations of multilevel data, such as intercept-only model, random intercepts, random intercepts with interaction, and random slopes. Emphasis is on level 2 missing data because they are more challenging to impute. Overall, those who use this book to apply the techniques in their own work and research will find this chapter very practical and straightforward with excellent extended examples, plots, code, and helpful interpretation of the results.

Causal inference in experimental studies, such as in medical, psychological, and educational research, is usually discussed at aggregate level. However, in more specific studies, causal effects are of interest at the experimental unit level, such as the individual level. Chapter 8, *individual causal effects*, discusses methods and issues in imputation of missing data in the framework of causal effects when the effect is studied at the unit level (*heterogeneity in treatment effect*). The first few pages of this chapter is devoted to introducing the concepts of individual causal effects (ICE). Next, the author presents imputation methods in the FCS framework (naive FCS and FCS with a prior). Extensions to the framework, especially the use of control variates is also discussed in this brief chapter. Most of the contents of the chapter comprise examples, plots, and code in R, together with extensive discussion of the output. The rich plots in this chapter show how informative plots can be in communication of statistical results.

The third section of the book presents case studies in which real research and data are presented. In Chapter 9, *measurement issues*, real data are used to illustrate how to deal with too many independent variables, sensitivity analysis, self-reported data, and dependence in observations. Chapter 10, *selection issues*, presents MI challenges when rows (cases) are missing, such as in drop-out situations in panel designs or non-response. Chapter 11, *longitudinal data*, presents example data in studies where measurements are repeated over longer time.

Section four of the book (*extensions*), includes only Chapter 12, *conclusion*. In this chapter,

the author provides tips and advice regarding practical issues such as reporting, file management, and ideas for future research on missing data analysis.

Flexible Imputation of Missing Data (2nd Edition) will definitely appeal to practitioners who analyze real world data with missing values, particularly clinical and health data. The book covers all types of missing data and missing data patterns. There are several aspects of the book that make it accessible and distinguished from other volumes currently available. The most prominent feature of this book is the clarity of exposition achieved by presenting clear description, examples, plots, and code. In addition, the example data sets used in the book are very familiar to researchers and practitioners in clinical and health data analysis, creating a tangible connection between the text and the practice. Students can use the example datasets to understand very clearly the process of missing data management and analysis. Another great feature of the book is the use of the R programming language and focus on one package. This structure gives the book coherence in terms of methods and tools used. In addition, on the theory side there is enough information, challenging questions, and reference to literature that make this book a rich resource for theoretical researchers. The intended audience of this book are practitioners in data analysis (especially biostatisticians), advanced graduate students, and theoretical researchers.

Reviewer:

Abdolvahab Khademi
University of Massachusetts
Department of Mathematics and Statistics
Amherst MA 01002, United States of America
E-mail: khademi@math.umass.edu

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
April 2020, Volume 93, Book Review 1 *Published:* 2020-04-18
doi:10.18637/jss.v093.b01