Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Flexible Imputation of Missing Data, 2nd ed.

Boca Raton, FL: Chapman & Hall/CRC Press, 2018, xxvii + 415 pp., $91.95(H), ISBN: 978-1-13-858831-8.

Shu Yang

To cite this article: Shu Yang (2019) Flexible Imputation of Missing Data, 2nd ed., Journal of the American Statistical Association, 114:527, 1421-1421, DOI: 10.1080/01621459.2019.1662249

To link to this article: https://doi.org/10.1080/01621459.2019.1662249

worthwhile to mention that these approaches are derived from the literature of pure mathematics, which is often considered not directly applicable to real world problems. I will recommend the book to data scientists who want to extend their field of expertise, and to gain a different yet powerful view on complex systems.

Kaixian Yu
AI Labs, DiDi Chuxing
Beijing, China
yukaixian@didiglobal.com

Flexible Imputation of Missing Data, 2nd ed. Stef van Buuren. Boca Raton, FL: Chapman & Hall/CRC Press, 2018, xxvii+415 pp., $91.95(H), ISBN: 978-1-13-858831-8.

Missing data are frequently encountered in practice. A broader class of missing data is called incomplete data, which includes data with measurement error, multilevel data with latent variables, and potential outcomes in causal inference. Due to its intuitive appeal, multiple imputation has been the most popular method for handling incomplete data: it multiply fills in the missing values and pools analyses based on straightforward rules.

The book focuses mainly on the multiple imputation approach to incomplete data but not on alternative approaches, such as weighting procedures and likelihood-based approaches. The main objective of the book is to provide a tool kit for practitioners to execute multiple imputation. The first edition of this book (van Buuren 2012) has been popular and well-received in the statistics and applied research communities. The main text explains the basic and key ideas underpinning multiple imputation, how to implement it in practice, and how to report and interpret the results. Moreover, it uses a reader-friendly style with lots of worked-out examples for illustration based on the MICE package. Therefore, the book can also be viewed as an extended MICE tutorial. Drawing from the author's own work and from the most recent developments in the field, the new edition expands discussions on important topics or incorporates new topics. This edition features new chapters on multiple imputation of multilevel data, causal inference by multiple imputation of the potential outcomes, and new data-adaptive algorithms for imputation such as predictive mean matching, trees, and many other machine learning tools.

The aspect I like most about this book is that it is well-balanced between practicality and technicality. On the one hand, it avoids too much mathematical and technical details and uses graphical tools and visual displays to aid understanding. Practitioners are able to understand the big picture and follow the code provided in the book. On the other hand, theoreticians are able to find technical materials and references to journal articles for in-depth investigation.

Divided into three parts, this book begins by an introduction of useful background material and an overview. Chapter 2 reviews the history of multiple imputation and introduces the general taxonomy of missing data mechanisms to set the stage for the whole book. Chapters 3 and 4 cover imputation methods for univariate missing data and multivariate missing data, respectively. Chapter 5 reviews post-imputation statistical analyses of the multiply imputed data. Dedicated statistical tests and model selection techniques are also included. The second part covers various advances of multiple imputation. Chapter 6 covers practical issues such as model specification and diagnosis for multivariate missing data. Chapter 7 is a new chapter discussing multiple imputation for nested data accounting for the multilevel data structure. Chapter 8 is also a new chapter, discussing multiple imputation for causal inference of individualized treatment effects. Chapters 9–11 use case studies to illustrate applications of multiple imputation with item non-response, measurement error, and drop out. The last part discusses limitations, reporting steps, and various extensions.

Multiple imputation was originated for survey data, which often contain design weights (or sample weights) to account for sample selection. However, there is no coverage of such topics in the current book. It would be useful to include a new chapter devoted to multiple imputation with survey data. There are many important topics that are worth discussion, such as how to incorporate design weights in the imputation model, weighted or unweighted estimation, and valid inference procedures.

In summary, the author has succeeded in achieving his objective of providing a practical introduction to multiple imputation for incomplete data. The book is particularly appropriate for substantive researchers and graduate students. For educators, the book would also make a reference for a course on the introduction and practical implementation of multiple imputation. Anyone who often utilizes complete-case analyses for handling missing data will find this book a good alternative to solve their problems.

Reference

van Buuren, S. (2012), *Flexible Imputation of Missing Data*, Boca Raton, FL: Chapman and Hall/CRC Press. [1421]

Shu Yang
North Carolina State University
Raleigh, NC
syang24@ncsu.edu

Missing and Modified Data in Nonparametric Estimation: With R Examples. Sam Efromovich. Boca Raton, FL: Chapman & Hall/CRC Press, 2018, xv+448 pp., $105.00(H), ISBN: 978-1-13-805488-2.

Missing and modified data are common complications in statistical analysis. In this book, missing data are considered as those elements in the data matrix that are unavailable, while modified data arise from a process in which the data values are modified in some way. Some examples of modified data are biased data, censored data, truncated data, and data