The Stata Journal

The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

The *Stata Journal* is indexed and abstracted by *CompuMath Citation Index*, *Current Contents/Social and Behavioral Sciences*, *RePEc: Research Papers in Economics*, *Science Citation Index Expanded* (also known as *SciSearch*), *Scopus*, and *Social Sciences Citation Index*.

For more information on the *Stata Journal*, including information for authors, see the webpage

http://www.stata-journal.com

Subscriptions are available from SAGE Publishing via telephone 805-499-9774 (U.S. customers) or 44-(0)20-73248701 (International), email journals@sagepub.com, or online at

https://journals.sagepub.com/home/stj

The *Stata Journal* (ISSN 1536-867X) is a quarterly journal of Stata Press and is published by SAGE Publishing in association with StataCorp LLC. Stata, stata, Stata Press, Mata, mata, and NetCourse are registered trademarks of StataCorp LLC.

The Stata Journal (2021)
21, Number 3, pp. 814–817 DOI: 10.1177/1536867X211045581

Review of Michael N. Mitchell's Data Management Using Stata: A Practical Handbook, Second Edition

William D. Dupont
Vanderbilt University School of Medicine
Nashville, TN
william.dupont@vumc.org

Abstract. In this article, I review *Data Management Using Stata: A Practical Handbook, Second Edition*, by Michael N. Mitchell (2020, Stata Press).

Keywords: gn0087, book review, data management

1 Introduction

Data management is a critical component of any scientific study. First and foremost is the need for reproducible results. In a lengthy review process, it is all too easy to lose track of how the original analyses were performed. More importantly, it is vital that authors can, if challenged, reproduce their published results or perform sensitivity analyses that are germane to criticisms of their work. Being able to reproduce our own work is the first step in enabling others to validate our findings (Lithgow, Driscoll, and Phillips 2017). Accurate data dictionaries must be created, and audit trails of data edits must be maintained. Complex studies require designing databases that will enable flexible analyses. Working with data obtained from multiple sources requires harmonization (Fortier et al. 2011; Hamilton et al. 2011), while analyzing administrative medical databases often requires expert adjudication of whether patients meet entry criteria, experience exposures of interest, or suffer study outcomes (Harpe 2009).

Mitchell (2020) is a well-written, comprehensive introduction to those aspects of data management that can be facilitated by using Stata. One of Stata's strengths is the ease with which one can obtain help in performing specific tasks. However, there is value in reading about a topic when one is not focused in finding a specific command for the task at hand, in that it can lead to finding aspects of Stata that one never knew one needed. For example, one can write sophisticated data management code without knowing about the existence of frames. However, once one becomes aware of their existence, they can make many tasks easier to perform. Mitchell's text provides an excellent resource for a reader who seeks a comprehensive review of many topics associated with data management.

2 Content

Chapter 1 gives an overview of the text and provides links to other online resources. Chapter 2 deals with reading and importing data in different formats. Chapter 3 explains how to save and export data files. Chapters 4 and 5 are concerned with data cleaning, labeling variables, and creating data dictionaries of value labels. Chapter 6 describes the different types of Stata variables and explains how to create and recode them. I found Mitchell's discussion of string expressions and how they can be manipulated to be particularly helpful. Chapter 7 deals with appending and merging datasets. Chapter 8 is concerned with processing observations across subgroups, while chapter 9 explains Stata's powerful `reshape` and `collapse` commands. Chapters 10 and 11 give an introduction to programming in Stata, including a helpful discussion of Stata macros and Stata's looping commands. An appendix gives an overview of Stata syntax, logical expressions, functions, missing values, variable lists, frames, and other topics.

Throughout the book, concepts and commands are illustrated with simple examples. The datasets used in this text are all available online.

3 Strengths and limitations

The strengths of this text are the clarity of its explanations and the many simple examples that are given throughout the book. What this book covers, it covers well.

The most important limitation of this book is that it does not discuss report writing. A critical aspect of data management is the creation of reproducible reports that are readily updated and that document the key steps in data management and analysis that produce publications. Before Stata 16, an important advantage of R was programs such as `knitr` that enable very flexible report writing. Stata introduced the `putdocx` command in Stata 15 and substantially improved its report-writing capabilities in Stata 16. Stata 16 was released in June of 2019, while Mitchell's text was published in 2020. While this did not provide a great deal of time to incorporate report writing into the text's second edition, a discussion of the `put*`, `dyndoc`, and `markdown` commands would be of great value should Mitchell decide to release a third edition of his text.

I would have liked to see more emphasis on Stata's GUI command interface. No matter how experienced users may be, there are always situations in which they may need to use a command that is new to them or that they have not used in a while. Being able to get the syntax right quickly with the GUI interface and then using the Review window to modify these commands is a real strength of Stata.

The text would have benefited from a discussion of best practices for designing multitable complex databases. Stata does not have structured query language. However, organizing data in tables that are consistent with relational database principles is always a good idea (Hernandez 2020).

4 Conclusion

This is a good text for anyone seeking a comprehensive text on Stata commands that are useful for data management. My only real reservation is that there are so many other help resources available to Stata users that some may find the added value of this text to be limited, given other freely available resources. I find it amazing how often a Google search will get me the information I need. Other valuable resources are Stata's own help files and YouTube tutorials. Users can contact Stata Technical Services, which I have found consistently helpful and friendly. The Stata forum, Statalist, is also helpful, although the volunteer experts on this forum are not always kind to users who ask questions that these experts have already answered.

Thus, I recommend this book to any Stata user who is looking for a comprehensive text on this topic, to users who have ample text budgets, and to departmental libraries that want a comprehensive set of reference works on Stata.

5 References

Fortier, I., D. Doiron, P. Burton, and P. Raina. 2011. Invited commentary: Consolidating data harmonization—How to obtain quality and applicability? *American Journal of Epidemiology* 174: 261–264. https://doi.org/10.1093/aje/kwr194.

Hamilton, C. M., L. C. Strader, J. G. Pratt, D. Maiese, T. Hendershot, R. K. Kwok, J. A. Hammond, W. Huggins, D. Jackman, H. Pan, D. S. Nettles, T. H. Beaty, L. A. Farrer, P. Kraft, M. L. Marazita, J. M. Ordovas, C. N. Pato, M. R. Spitz, D. Wagener, M. Williams, H. A. Junkins, W. R. Harlan, E. M. Ramos, and J. Haines. 2011. Hamilton et al. respond to “Consolidating data harmonization”. *American Journal of Epidemiology* 174: 265–266. https://doi.org/10.1093/aje/kwr191.

Harpe, S. E. 2009. Using secondary data sources for pharmacoepidemiology and outcomes research. *Pharmacotherapy* 29: 138–153. https://doi.org/10.1592/phco.29.2.138.

Hernandez, M. J. 2020. *Database Design for Mere Mortals: A Hands-On Guide to Relational Database Design*. 4th ed. Boston: Addison–Wesley.

Lithgow, G. J., M. Driscoll, and P. Phillips. 2017. A long journey to reproducible results. *Nature* 548: 387–388. https://doi.org/10.1038/548387a.

Mitchell, M. N. 2020. *Data Management Using Stata: A Practical Handbook*. 2nd ed. College Station, TX: Stata Press.

About the author

William D. Dupont is a professor of biostatistics and preventive medicine at Vanderbilt University School of Medicine and is a fellow of the American Statistical Association. He is best known for his research on the relationship between different types of benign breast disease and breast cancer risk. He has also worked on methods and software for power and sample-size calculations. Currently, his research focuses on genetic risk factors for familial prostate cancer, on studies related to opioid abuse, and on the relationship between respiratory syncytial virus infections in infancy and subsequent risk of asthma. He has written a text for teaching multivariable statistical methods to nonstatisticians that use Stata.