Preface to the second edition

Software systems evolve, and so do the approaches and expertise of statistical analysts.

After the publication of the first edition of *SAS and R: Data Management, Statistical Analysis, and Graphics*, we began a blog in which we explored many new case studies and applications, ranging from generating a Fibonacci series to fitting finite mixture models with concomitant variables. We also discussed some additions to SAS and new or improved R packages. The blog now has hundreds of entries and (according to Google Analytics) has received hundreds of thousands of visits.

The volume you are holding is nearly 50% longer than the first edition, and much of the new material is adapted from these blog entries, while it also includes other improvements and additions which have emerged in the last few years.

We have extensively re-organized the material in the book and created three new chapters. The first *Simulation*, includes examples where data are generated from complex models such as mixed effects models and survival models, and from distributions using the Metropolis–Hastings algorithm. We also explore three interesting statistics and probability examples via simulation. The second is *Special topics*, where we describe some key features, such as processing by group, and detail several important areas of statistics, including Bayesian methods, propensity scores, and bootstrapping. The last is *Case studies*, where we demonstrate examples of some data management tasks, read complex files, make and annotate maps, and show how to "scrape" data from web pages.

We also cover some important new tools, including the use of RStudio, a powerful and easy-to-use front end for R that adds innumerable features to R. In our experience, it at least doubles the productivity of R users, and our SAS-using students find it an extremely comfortable interface that bears some similarity to the SAS GUI.

We have added a separate section and examples that describe "reproducible analysis." This is the notion that code, results, and interpretation should live together in a single place. We used two reproducible analysis systems (SASweave and Sweave) to generate the example code and output in the book. Code extracted from these files is provided on the book web site. In this edition, we provide a detailed discussion of the philosophy and use of these systems. In particular, we feel that the `knitr` and `markdown` packages for R, which are tightly integrated with RStudio, should become a part of every R user's toolbox. We can't imagine working on a project without them.

Finally, we've reorganized much of the material from the first edition into smaller, more focused chapters. Users will now find separate (and enhanced) chapters on data input and output, data management, statistical and mathematical functions, and programming, rather than a single chapter on "data management." Graphics are now discussed in two chapters: one on high-level types of plots such as scatterplots and histograms, and another on customizing the fine details of the plots, such as the number of tick marks and the color of plot symbols.

We're immensely gratified by the positive response the first edition elicited, and hope the current volume will be as useful to you.

On the web

The book website at `http://www.amherst.edu/~nhorton/sasr2` includes the table of contents, the indices, the HELP dataset, example code in SAS and R, a pointer to the blog, and a list of erratum.

Acknowledgments

In addition to those acknowledged in the first edition, we would like to thank Kathryn Aloisio, Gregory Call, J.J. Allaire and the RStudio developers, plus the many individuals who have created and shared R packages or SAS macros. Their contributions to SAS, R, or LaTeX programming efforts, comments, guidance and/or helpful suggestions on drafts of the revision have been extremely helpful. Above all we greatly appreciate Sara and Julia as well as Abby, Alana, Kinari, and Sam, for their patience and support.

Amherst, MA
July, 2014

Preface

XX PULL IN OLD PREFACE!

SAS^{TM} [153] and R [135] are two statistical software packages used in many fields of research. SAS is commercial software developed by SAS Institute; it includes well-validated statistical algorithms. It can be licensed but not purchased. Paying for a license entitles the licensee to professional customer support. However, licensing is expensive and SAS sometimes incorporates new statistical methods only after a significant lag. In contrast, R is free, open-source software, developed by a large group of people, many of whom are volunteers. It has a large and growing user and developer base. Methodologists often release applications for general use in R shortly after they have been introduced into the literature. Professional customer support is not provided, though there are many resources for users. There are settings in which one of these useful tools is needed, and users who have spent many hours gaining expertise in the other often find it frustrating to make the transition.

We have written this book as a reference text for users of SAS and R. Our primary goal is to provide users with an easy way to learn how to perform an analytic task in both systems, without having to navigate through the extensive, idiosyncratic, and sometimes (often?) unwieldy documentation each provides. We expect the book to function in the same way that an English–French dictionary informs users of both the equivalent nouns and verbs in the two languages as well as the differences in grammar. We include many common tasks, including data management, descriptive summaries, inferential procedures, regression analysis, multivariate methods, and the creation of graphics. We also show some more complex applications. In toto, we hope that the text will allow easier mobility between systems for users of any statistical system.

We do not attempt to exhaustively detail all possible ways available to accomplish a given task in each system. Neither do we claim to provide the most elegant solution. We have tried to provide a simple approach that is easy to understand for a new user, and have supplied several solutions when it seems likely to be helpful. Carrying forward the analogy to an English–French dictionary, we suggest language that will communicate the point effectively, without listing every synonym or providing guidance on native idiom or eloquence.

Who should use this book

Those with an understanding of statistics at the level of multiple-regression analysis will find this book helpful. This group includes professional analysts who use statistical packages almost every day as well as statisticians, epidemiologists, economists, engineers, physicians, sociologists, and others engaged in research or data analysis. We anticipate that this tool will be particularly useful for sophisticated users, those with years of experience in only one system, who need or want to use the other system. However, intermediate-level analysts should reap the same benefit. In addition, the book will bolster the analytic abilities of a relatively new user of either system, by providing a concise reference manual and annotated

examples executed in both packages.

Using the book

The book has three indices, in addition to the comprehensive table of contents. These include: 1) a detailed topic (subject) index in English; 2) a SAS index, organized by SAS syntax; and 3) an R index, describing R syntax. SAS users can use the SAS index to look up a task for which they know the SAS code and turn to a page with that code as well as the associated R code to carry out that task. R users can use the dictionary in an analogous fashion using the R index.

Extensive example analyses are presented; see Table C.1 (p. 405) for a comprehensive list. These employ a single dataset (from the HELP study), described in Appendix C. Readers are encouraged to download the dataset and code from the book website. The examples demonstrate the code in action and facilitate exploration by the reader.

Differences between SAS and R

SAS and R are so fundamentally distinct that an enumeration of their differences would be counter-productive. However, some differences are important for new users to bear in mind.

SAS includes data management tools that are primarily intended to prepare data for analysis. After preparation, analysis is performed in a distinct step, the implementation of which effectively cannot be changed by the user, though often extensive options are available. R is a programming environment tailored for data analysis. Data management and analysis are integrated. This means, for example, that calculating body mass index (BMI) from weight and height can be treated as a function of the data, and as such is as likely to appear within a data analysis as in making a "new" piece of data to keep.

SAS Institute makes decisions about how to change the software or expand the scope of included analyses. These decisions are based on the needs of the user community and on corporate goals for profitability. For example, when changes are made, backwards-compatibility is almost always maintained, and documentation of exceptions is extensive. SAS Institute's corporate conservatism means that techniques are sometimes not included in SAS until they have been discussed in the peer-reviewed literature for many years. While the R Core Team controls base functionality, a very large number of users have developed functions for R. Methodologists often release R functions to implement their work concurrently with publication. While this provides great flexibility, it comes at some cost. A user-contributed function may implement a desired methodology, but code quality may be unknown, documentation scarce, and paid support nonexistent. Sometimes a function which once worked may become defunct due to a lack of backwards-compatibility and/or the author's inability to, or lack of interest in, updating it.

Other differences between SAS and R are worth noting. Data management in SAS is undertaken using row by row (observation-level) operations. R is inherently a vector-based language, where columns (variables) are manipulated. R is case-sensitive, while SAS is generally not.

Where to begin

We do not anticipate that the book will be read cover to cover. Instead, we hope that the extensive indexing, cross-referencing, and worked examples will make it possible for readers to directly find and then implement what they need. A user new to either SAS or R should begin by reading the appropriate Appendix for that software package, which includes a sample session and overview.

On the web

The book website includes the table of contents, the indices, the HELP dataset, example code in SAS and R, and a list of erratum.

Acknowledgments

We would like to thank Rob Calver, Shashi Kumar and Sarah Morris for their support and guidance at Informa CRC/Chapman and Hall, the Department of Statistics at the University of Auckland for graciously hosting NH during a sabbatical leave, and the Office of the Provost at Smith College. We also thank Allyson Abrams, Tanya Hakim, Ross Ihaka, Albyn Jones, Russell Lenth, Brian McArdle, Paul Murrell, Alastair Scott, David Schoenfeld, Duncan Temple Lang, Kristin Tyler, Chris Wild, and Alan Zaslavsky for contributions to SAS, R, or LaTeX programming efforts, comments, guidance and/or helpful suggestions on drafts of the manuscript.

Above all we greatly appreciate Sara and Julia as well as Abby, Alana, Kinari, and Sam, for their patience and support.

Amherst, MA and Northampton, MA
March, 2009