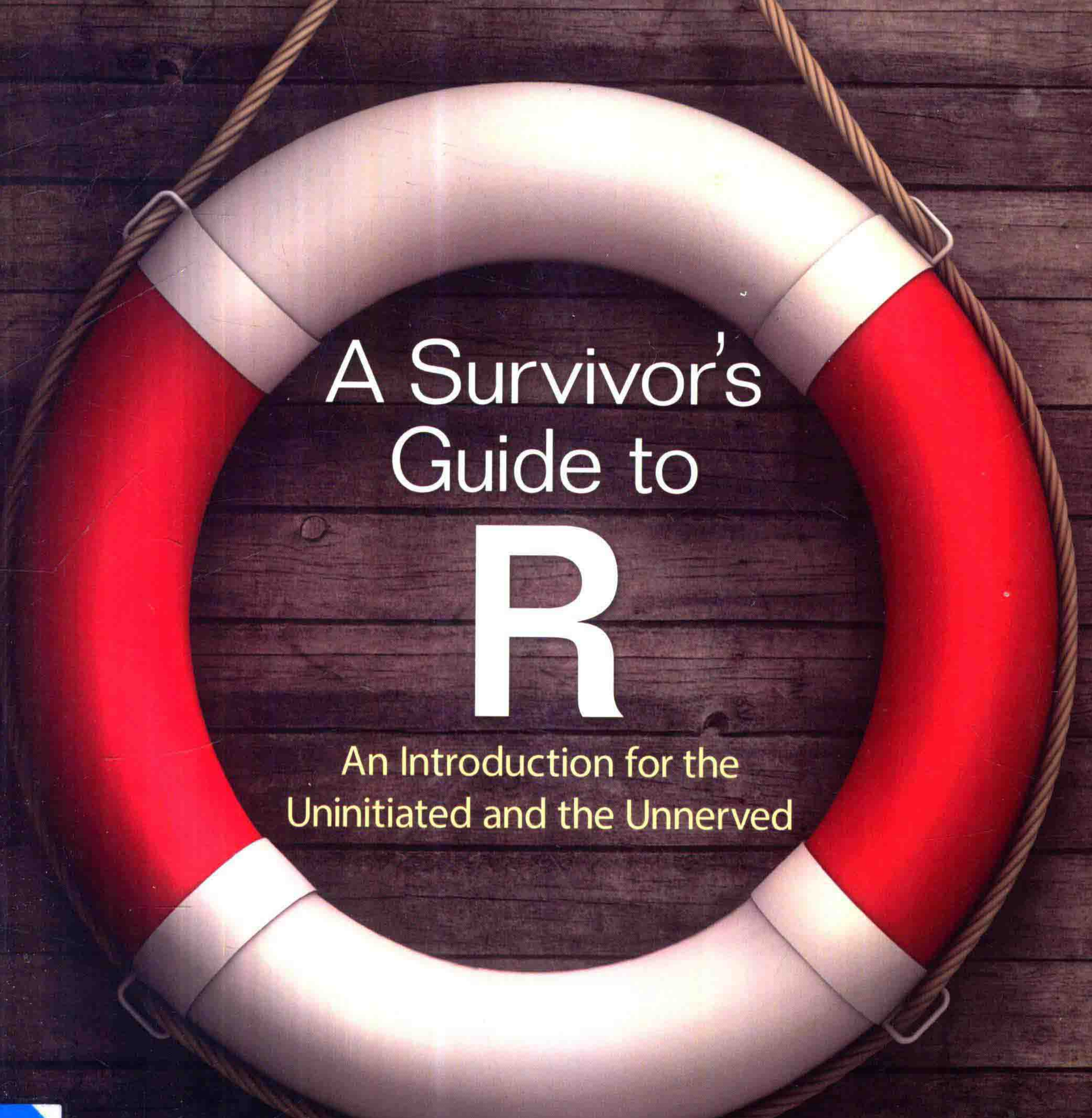

A Survivor's Guide to R

An Introduction for the Uninitiated and the Unnerved

Kurt Taylor Gaubatz

A Survivor's Guide to R

An Introduction for the Uninitiated and the Unnerved

Kurt Taylor Gaubatz
Old Dominion University

SAGE

Los Angeles | London | New Delhi
Singapore | Washington DC

SAGE

Los Angeles | London | New Delhi
Singapore | Washington DC

FOR INFORMATION:

SAGE Publications, Inc.
2455 Teller Road
Thousand Oaks, California 91320
E-mail: order@sagepub.com

SAGE Publications Ltd.
1 Oliver's Yard
55 City Road
London, EC1Y 1SP
United Kingdom

SAGE Publications India Pvt. Ltd.
B 1/I 1 Mohan Cooperative Industrial Area
Mathura Road, New Delhi 110 044
India

SAGE Publications Asia-Pacific Pte. Ltd.
3 Church Street
#10-04 Samsung Hub
Singapore 049483

Acquisitions Editor: Vicki Knight
Assistant Editor: Katie Guarino
Editorial Assistant: Yvonne McDuffee
Production Editors: Kelly DeRosa, Stephanie Palermini
Copy Editor: QuADS Prepress (P) Ltd.
Typesetter: C&M Digitals (P) Ltd.
Proofreader: Wendy Jo Dymond
Indexer: Kurt Taylor Gaubatz
Cover Designer: Anupama Krishnan
Marketing Manager: Nicole Elliott

Printed in the United States of America

Library of Congress Cataloging-in-Publication Data

Gaubatz, Kurt Taylor, author.
A survivor's guide to R : an introduction for the uninitiated and the unnerved / Kurt Taylor Gaubatz.

pages cm
Includes bibliographical references and index.

ISBN 978-1-4833-4673-1 (pbk. : acid-free paper)
ISBN 978-1-4833-4688-5 (web PDF)

1. Statistics—Data processing. 2. R (Computer program language) I. Title.

QA276.45.R3G38 2015
005.26′2—dc23 2013045399

This book is printed on acid-free paper.

SUSTAINABLE FORESTRY INITIATIVE
Certified Chain of Custody
Promoting Sustainable Forestry
www.sfiprogram.org
SFI-01268
SFI label applies to text stock

14 15 16 17 18 10 9 8 7 6 5 4 3 2 1

211

A Survivor’s Guide to R

For Kathy, of course.

List of Tables

List of Figures

About the Author

Kurt Taylor Gaubatz is an associate professor in the Department of Political Science and Geography and in the Graduate Program in International Studies at Old Dominion University (ODU). He teaches a range of courses in international relations, international law, and research methods. Before coming to ODU in 2000, he was the Visiting John G. Winant Lecturer in American Foreign Policy at Oxford University (Nuffield College) and was on the political science and international relations faculty at Stanford University. He has served as the Susan Louise Dyer Peace Fellow in the National Fellows program at the Hoover Institution and was a Pew Faculty Fellow in International Affairs with the Kennedy School of Government at Harvard University. He did his undergraduate work in economics at the University of California, Berkeley. He holds master's degrees in international law from the Fletcher School of Law and Diplomacy and in theology from Princeton Theological Seminary. He earned his PhD in political science from Stanford University. He is the author of *Elections and War* (Stanford University Press, 1999) as well as a number of prominent articles mostly focused on international law and on the relationship between domestic politics and international relations. More information can be found at www.sagepub.com/gaubatz.

Preface

A few years ago, I was at a conference chatting with one of the most distinguished and technologically capable political scientists I know. This is someone who came to political science with an undergraduate degree in math from Caltech and is the author of a major text on game theory as well as a number of prominent articles using sophisticated statistical analysis. He recounted the experience of sitting in on an advanced seminar on Bayesian statistics. The statistics were pretty straightforward, he said. The real challenge was coming to grips with R for the first time. When I mentioned that I intended to switch to R for one of my introductory statistics classes, he shuddered.

This story might come as a revelation to many in the community of advanced R users, who view R syntax as essentially second nature. Having worked with R on a daily basis for many years, they have little trouble making it sit, lie down, and roll over. They are somewhat surprised when others think that the only trick R knows is playing dead.

I did start using R for teaching, and I, and every one of the students in those classes, survived. At its core, this book is a step-by-step guide to how we did that. In fact, although R does have a steep learning curve—on first encounter, it is often intimidating and unnerving—it has proven to have a number of significant advantages for teaching and learning statistics.

R is powerful and inexpensive (free!). It is rapidly becoming the package of choice for advanced statistical analysis across a number of fields. Moreover, it has probably been assigned to you, so you just have to buckle down and learn it. The purpose of this book is to help you survive and even to thrive in that process. The approach I take is to focus primarily on the challenges of using R to manage, manipulate, and visualize your data, rather than the usual approach of jumping right into conducting statistical analysis with R.

I take this alternative approach for three reasons. In the first place, data management is the foundation for all statistical analysis. Getting your data into the right form for analysis is a critical skill. Yet data management issues are rarely taught in statistics classes, where appropriate and well-groomed data sets appear to float down directly from heaven. This book provides the opportunity to get a handle on some of those essential background skills. Second, once you have learned the basic structure and rules of R in this context, you will find it much easier to follow up with learning the statistical procedures, which you will most likely do in the context of a statistics class and text. Finally, separating the statistics from the teaching of R allows the book to serve both as a tutorial and as a reference in which you can quickly find the commands and procedures that otherwise are mixed in and hidden among the statistical content of traditional texts.

Moreover, while this book starts with the very basics of installing R and getting it to run simple procedures, it ultimately covers R at a significantly greater depth than you are likely to encounter in a statistics class. This book is designed to carry you beyond the classroom, giving you the opportunity to gain and maintain the kind of facility with R that can make it a functional real-world skill in your analytical toolbox.

Because this book separates the mechanics of working with R from the teaching of statistics, it will be helpful in a wide range of contexts. It is designed to help tackle data problems that arise across a wide range of fields and at different levels of statistical sophistication. Whether you are tackling R in an introductory statistics class or an advanced graduate seminar or are just transitioning to R from another statistics program, you will find this a helpful guide along the way.

For users at the introductory level, Chapter 2 and Appendix B run through most of the procedures that might be encountered in an introductory statistics class. Chapter 3 offers a straightforward approach to understanding object types and their critical role in R. Chapter 5 goes over the basics for summarizing and reviewing data. Chapter 12 is an introduction to R's broad variety of built-in plots.

For those beginning to work on collecting and managing their own data, Chapter 4 goes over the process of getting data into R from a wide range of sources. Chapters 6 and 7 cover sorting, selecting, and transforming data. Chapter 10 teaches the critical skills for merging and aggregating data. Chapter 11 confronts the real-world challenge of dealing with missing data.

For more advanced users, the end of Chapter 7 gets into R programming techniques, including the powerful use of dynamic coding to incorporate variable- and data-driven elements into your R scripts. Chapter 8 deals with the particular issues of textual data and includes a tutorial on the use of regular expressions in R. Chapter 9 does the same thing for the sometimes surprisingly treacherous world of date and time data.

For users at all levels, some of the biggest rewards will come in Chapters 12 to 15, in which I provide a thorough but accessible guide to R's powerful graphics facility. At any level of statistical sophistication, the ability to produce and customize high-quality data visualizations will be a critical 21st-century skill.

There is a book website (http://www.sagepub.com/gaubatz), where I have posted a file with all the R code used in the book. You can go there to see exactly how the code works and to cut and paste for your own projects. You will also find there the example data sets, color versions of many of the plots, and a gallery of additional graphics examples, with the attendant R scripts.

It is likely that you have not chosen to learn R simply for the fun of it. For one reason or another, you have arrived in this somewhat scary place and now have to deal with it. My purpose is to make that as painless as possible. You can survive this. And, at the end of the day, you might just find that it is a little bit fun as well.

ACKNOWLEDGMENTS

I need to start by acknowledging my debt for the extraordinary wealth of materials that the R community has created. The developers of R have done an exceptional job creating a cutting-edge software package that has, in turn, engendered a community of users who are constantly moving R forward. I am occasionally critical in this book of the difficulties new users face in mastering the basics of R and of the difficulties that R experts sometimes have in appreciating the frustrations of new users. But this should not be viewed in any way as diminishing the enormous respect I have for the extraordinary services the R community provides. I have learned much from poring over the archives of the R Help list. It never ceases to amaze me how often it is the giants of the R development team that take time out to answer even the simplest questions from users.

I also need to thank the students in my statistics seminars who gamely served as guinea pigs for this project. Their trail breaking will help spare some of this pain for others who follow in their footsteps. If you find this book of help, you share with me a particularly significant debt of gratitude to several members of the first cohort to use R with me: Melodee Baines, Bill Eliason, Huhe Narisong, Andrew Townsend, and Christopher White. Cody Zimmerman and Scott Duryea also helped move the project forward with their excellent service as research assistants.

Vicki Knight and the editorial team at SAGE Publications have been terrific. Their enthusiasm helped move this project through its final stages. Shamila Swamy (with her team at QuADS Prepress), in particular, applied a careful eye and deft touch to the copyediting. The SAGE process included a very helpful review from a wide range of fields, and I am indebted to Jim Albert, Bowling Green State University; Woody Carter, University of Chicago; Ole J. Forsberg, Oklahoma State University; David Han, University of Texas at San Antonio; Yulan Liang, University of Maryland, Baltimore; A. Dean Monroe, Angelo State University; Charlotte Tate, San Francisco State University; and Toshiyuki Yuasa, University of Houston for their careful comments and insightful suggestions for making this a better and more useful book.

I would also like to thank Kevin Sweeney, who got me started on this whole thing by giving me the opportunity to be involved in some projects that required taking the R plunge.

Finally, of course, there is Kathy, whose constant love and support have been critical even for a sometimes opaque and mysterious project on statistical computing.

Brief Contents