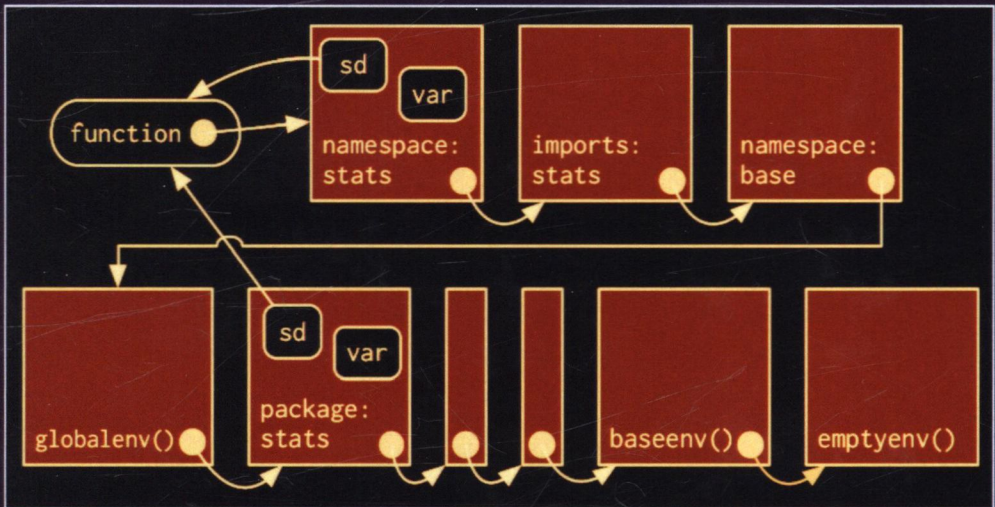


The R Series

Advanced R



Hadley Wickham

 CRC Press
Taylor & Francis Group

A CHAPMAN & HALL BOOK

“The author has become one of the foremost authorities on this topic and is well known and appreciated throughout the entire R community. This is the great strength of the book and the primary reason it deserves to be published. It addresses a topic where there is already a growing number of books, but few have the depth, the technical accuracy, and the authority of this one.”

—Bill Venables, CSIRO

Advanced R presents useful tools and techniques for attacking many types of R programming problems, helping you avoid mistakes and dead ends. With more than ten years of experience programming in R, the author illustrates the elegance, beauty, and flexibility at the heart of R.

The book develops the necessary skills to produce quality code that can be used in a variety of circumstances. You will learn:

- The fundamentals of R, including standard data types and functions
- Functional programming as a useful framework for solving wide classes of problems
- The positives and negatives of metaprogramming
- How to write fast, memory-efficient code

This book not only helps current R users become R programmers but also shows existing programmers what’s special about R. Intermediate R programmers can dive deeper into R and learn new strategies for solving diverse problems while programmers from other languages can learn the details of R and understand why R works the way it does.



CRC Press

Taylor & Francis Group
an informa business

www.crcpress.com



Advanced R

Wickham



Advanced R

Hadley Wickham



CRC Press

Taylor & Francis Group

Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

First issued in hardback 2017

© 2014 by Taylor & Francis Group, LLC
CRC Press is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works

ISBN: 978-1-4665-8696-3 (pbk)
ISBN: 978-1-1384-6932-7 (hbk)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access www.copyright.com (<http://www.copyright.com/>) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Visit the Taylor & Francis Web site at
<http://www.taylorandfrancis.com>

and the CRC Press Web site at
<http://www.crcpress.com>

Advanced R

Chapman & Hall/CRC

The R Series

Series Editors

John M. Chambers
Department of Statistics
Stanford University
Stanford, California, USA

Torsten Hothorn
Division of Biostatistics
University of Zurich
Switzerland

Duncan Temple Lang
Department of Statistics
University of California, Davis
Davis, California, USA

Hadley Wickham
RStudio
Boston, Massachusetts, USA

Aims and Scope

This book series reflects the recent rapid growth in the development and application of R, the programming language and software environment for statistical computing and graphics. R is now widely used in academic research, education, and industry. It is constantly growing, with new versions of the core software released regularly and more than 5,000 packages available. It is difficult for the documentation to keep pace with the expansion of the software, and this vital book series provides a forum for the publication of books covering many aspects of the development and application of R.

The scope of the series is wide, covering three main threads:

- Applications of R to specific disciplines such as biology, epidemiology, genetics, engineering, finance, and the social sciences.
- Using R for the study of topics of statistical methodology, such as linear and mixed modeling, time series, Bayesian methods, and missing data.
- The development of R, including programming, building packages, and graphics.

The books will appeal to programmers and developers of R software, as well as applied statisticians and data analysts in many fields. The books will feature detailed worked examples and R code fully integrated into the text, ensuring their usefulness to researchers, practitioners and students.

Published Titles

Stated Preference Methods Using R, *Hideo Aizaki, Tomoaki Nakatani, and Kazuo Sato*

Using R for Numerical Analysis in Science and Engineering, *Victor A. Bloomfield*

Event History Analysis with R, *Göran Broström*

Computational Actuarial Science with R, *Arthur Charpentier*

Statistical Computing in C++ and R, *Randall L. Eubank and Ana Kupresanin*

Reproducible Research with R and RStudio, *Christopher Gandrud*

Introduction to Scientific Programming and Simulation Using R, Second Edition, *Owen Jones, Robert Maillardet, and Andrew Robinson*

Nonparametric Statistical Methods Using R, *John Kloeke and Joseph McKean*

Displaying Time Series, Spatial, and Space-Time Data with R, *Oscar Perpiñán Lamigueiro*

Programming Graphical User Interfaces with R, *Michael F. Lawrence and John Verzani*

Analyzing Sensory Data with R, *Sébastien Lê and Theirry Worch*

Analyzing Baseball Data with R, *Max Marchi and Jim Albert*

Growth Curve Analysis and Visualization Using R, *Daniel Mirman*

R Graphics, Second Edition, *Paul Murrell*

Multiple Factor Analysis by Example Using R, *Jérôme Pagès*

Customer and Business Analytics: Applied Data Mining for Business Decision Making Using R, *Daniel S. Putler and Robert E. Krider*

Implementing Reproducible Research, *Victoria Stodden, Friedrich Leisch, and Roger D. Peng*

Using R for Introductory Statistics, Second Edition, *John Verzani*

Advanced R, *Hadley Wickham*

Dynamic Documents with R and knitr, *Yihui Xie*

*To Jeff, who makes me happy, and who made
sure I had a life outside this book.*

Contents

1	Introduction	1
1.1	Who should read this book	3
1.2	What you will get out of this book	3
1.3	Meta-techniques	4
1.4	Recommended reading	5
1.5	Getting help	6
1.6	Acknowledgments	6
1.7	Conventions	8
1.8	Colophon	8
I	Foundations	11
2	Data structures	13
2.1	Vectors	14
2.1.1	Atomic vectors	15
2.1.1.1	Types and tests	15
2.1.1.2	Coercion	16
2.1.2	Lists	17
2.1.3	Exercises	18
2.2	Attributes	19
2.2.0.1	Names	20
2.2.1	Factors	21
2.2.2	Exercises	23
2.3	Matrices and arrays	24

2.3.1	Exercises	26
2.4	Data frames	27
2.4.1	Creation	27
2.4.2	Testing and coercion	28
2.4.3	Combining data frames	28
2.4.4	Special columns	29
2.4.5	Exercises	30
2.5	Answers	31
3	Subsetting	33
3.1	Data types	34
3.1.1	Atomic vectors	34
3.1.2	Lists	37
3.1.3	Matrices and arrays	37
3.1.4	Data frames	38
3.1.5	S3 objects	39
3.1.6	S4 objects	39
3.1.7	Exercises	39
3.2	Subsetting operators	40
3.2.1	Simplifying vs. preserving subsetting	41
3.2.2	\$	43
3.2.3	Missing/out of bounds indices	44
3.2.4	Exercises	45
3.3	Subsetting and assignment	45
3.4	Applications	46
3.4.1	Lookup tables (character subsetting)	46
3.4.2	Matching and merging by hand (integer subsetting)	47
3.4.3	Random samples/bootstrap (integer subsetting)	48
3.4.4	Ordering (integer subsetting)	49

3.4.5	Expanding aggregated counts (integer subsetting)	50
3.4.6	Removing columns from data frames (character subsetting)	51
3.4.7	Selecting rows based on a condition (logical subsetting)	52
3.4.8	Boolean algebra vs. sets (logical & integer subsetting)	53
3.4.9	Exercises	55
3.5	Answers	55
4	Vocabulary	57
4.1	The basics	57
4.2	Common data structures	59
4.3	Statistics	60
4.4	Working with R	61
4.5	I/O	62
5	Style guide	63
5.1	Notation and naming	63
5.1.1	File names	63
5.1.2	Object names	64
5.2	Syntax	65
5.2.1	Spacing	65
5.2.2	Curly braces	66
5.2.3	Line length	67
5.2.4	Indentation	67
5.2.5	Assignment	67
5.3	Organisation	68
5.3.1	Commenting guidelines	68

6	Functions	69
6.1	Function components	71
6.1.1	Primitive functions	71
6.1.2	Exercises	72
6.2	Lexical scoping	73
6.2.1	Name masking	74
6.2.2	Functions vs. variables	75
6.2.3	A fresh start	76
6.2.4	Dynamic lookup	77
6.2.5	Exercises	78
6.3	Every operation is a function call	79
6.4	Function arguments	81
6.4.1	Calling functions	81
6.4.2	Calling a function given a list of arguments	83
6.4.3	Default and missing arguments	83
6.4.4	Lazy evaluation	84
6.4.5	88
6.4.6	Exercises	89
6.5	Special calls	89
6.5.1	Infix functions	90
6.5.2	Replacement functions	91
6.5.3	Exercises	93
6.6	Return values	94
6.6.1	On exit	97
6.6.2	Exercises	97
6.7	Quiz answers	98

7	OO field guide	99
7.1	Base types	101
7.2	S3	102
7.2.1	Recognising objects, generic functions, and methods	102
7.2.2	Defining classes and creating objects	105
7.2.3	Creating new methods and generics	106
7.2.4	Method dispatch	107
7.2.5	Exercises	109
7.3	S4	111
7.3.1	Recognising objects, generic functions, and methods	111
7.3.2	Defining classes and creating objects	113
7.3.3	Creating new methods and generics	115
7.3.4	Method dispatch	115
7.3.5	Exercises	116
7.4	RC	116
7.4.1	Defining classes and creating objects	117
7.4.2	Recognising objects and methods	119
7.4.3	Method dispatch	119
7.4.4	Exercises	120
7.5	Picking a system	120
7.6	Quiz answers	121
8	Environments	123
8.1	Environment basics	124
8.1.1	Exercises	130
8.2	Recursing over environments	130
8.2.1	Exercises	132
8.3	Function environments	133

8.3.1	The enclosing environment	133
8.3.2	Binding environments	134
8.3.3	Execution environments	136
8.3.4	Calling environments	138
8.3.5	Exercises	140
8.4	Binding names to values	141
8.4.1	Exercises	143
8.5	Explicit environments	144
8.5.1	Avoiding copies	145
8.5.2	Package state	146
8.5.3	As a hashmap	146
8.6	Quiz answers	147
9	Debugging, condition handling, and defensive programming	149
9.1	Debugging techniques	151
9.2	Debugging tools	153
9.2.1	Determining the sequence of calls	154
9.2.2	Browsing on error	155
9.2.3	Browsing arbitrary code	157
9.2.4	The call stack: <code>traceback()</code> , <code>where</code> , and <code>recover()</code>	158
9.2.5	Other types of failure	158
9.3	Condition handling	160
9.3.1	Ignore errors with <code>try</code>	160
9.3.2	Handle conditions with <code>tryCatch()</code>	162
9.3.3	<code>withCallingHandlers()</code>	165
9.3.4	Custom signal classes	166
9.3.5	Exercises	168
9.4	Defensive programming	168
9.4.1	Exercises	169
9.5	Quiz answers	170