

Cambridge University Press

978-1-107-57646-9 - A First Course in Statistical Programming with R: Second Edition

W. John Braun and Duncan J. Murdoch

Frontmatter

[More information](#)

A First Course in Statistical Programming with R

This new, color edition of Braun and Murdoch's bestselling textbook integrates use of the RStudio platform and adds discussion of newer graphics systems, extensive exploration of Markov chain Monte Carlo, expert advice on common error messages, motivating applications of matrix decompositions, and numerous new examples and exercises.

This is the only introduction you'll need to start programming in R, the computing standard for analyzing data. Co-written by an R Core Team member and an established R author, this book comes with real R code that complies with the standards of the language. Unlike other introductory books on the R system, this book emphasizes programming, including the principles that apply to most computing languages, and techniques used to develop more complex projects. Solutions, datasets, and any errata are available from the book's website. The many examples, all from real applications, make it particularly useful for anyone working in practical data analysis.

W. John Braun is Deputy Director of the Canadian Statistical Sciences Institute. He is also Professor and Head of the Departments of Computer Science, Physics, Mathematics and Statistics at the University of British Columbia Okanagan. His research interests are in the modeling of environmental phenomena, such as wildfire, as well as statistical education, particularly as it relates to the R programming language.

Duncan J. Murdoch is a member of the R Core Team of developers, and is co-president of the R Foundation. He is one of the developers of the `rgl` package for 3D visualization in R, and has also developed numerous other R packages. He is also a professor in the Department of Statistical and Actuarial Sciences at the University of Western Ontario.

Cambridge University Press

978-1-107-57646-9 - A First Course in Statistical Programming with R: Second Edition

W. John Braun and Duncan J. Murdoch

Frontmatter

[More information](#)

Cambridge University Press

978-1-107-57646-9 - A First Course in Statistical Programming with R: Second Edition

W. John Braun and Duncan J. Murdoch

Frontmatter

[More information](#)

A First Course in Statistical Programming with R

Second Edition

W. John Braun and Duncan J. Murdoch



CAMBRIDGE
UNIVERSITY PRESS

Cambridge University Press
978-1-107-57646-9 - A First Course in Statistical Programming with R: Second Edition
W. John Braun and Duncan J. Murdoch
Frontmatter
[More information](#)

CAMBRIDGE
UNIVERSITY PRESS

32 Avenue of the Americas, New York NY 10013

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org

Information on this title: www.cambridge.org/9781107576469

© W. John Braun and Duncan J. Murdoch 2007, 2016

This publication is in copyright. Subject to statutory exception and to the provisions of relevant collective licensing agreements, no reproduction of any part may take place without the written permission of Cambridge University Press.

First published 2007

Second edition 2016

Printed in the United States of America by Sheridan Books, Inc.

A catalogue record for this publication is available from the British Library.

ISBN 978-1-107-57646-9 Hardback

Additional resources for this publication at www.cambridge.org/9781107576469.

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for external or third-party Internet Web sites referred to in this publication and does not guarantee that any content on such Web sites is, or will remain, accurate or appropriate.

Contents

	Preface to the second edition	<i>page xi</i>
	Preface to the first edition	xiii
1	Getting started	1
1.1	What is statistical programming?	1
1.2	Outline of this book	2
1.3	The R package	3
1.4	Why use a command line?	3
1.5	Font conventions	4
1.6	Installation of R and RStudio	4
1.7	Getting started in RStudio	5
1.8	Going further	6
2	Introduction to the R language	7
2.1	First steps	7
2.2	Basic features of R	11
2.3	Vectors in R	13
2.4	Data storage in R	22
2.5	Packages, libraries, and repositories	27
2.6	Getting help	28
2.7	Logical vectors and relational operators	34
2.8	Data frames and lists	37
2.9	Data input and output	43
3	Programming statistical graphics	49
3.1	High level plots	50
3.2	Choosing a high level graphic	62
3.3	Low level graphics functions	63
3.4	Other graphics systems	70

4	Programming with R	76
4.1	Flow control	76
4.2	Managing complexity through functions	91
4.3	The <code>replicate()</code> function	97
4.4	Miscellaneous programming tips	97
4.5	Some general programming guidelines	100
4.6	Debugging and maintenance	107
4.7	Efficient programming	113
5	Simulation	120
5.1	Monte Carlo simulation	120
5.2	Generation of pseudorandom numbers	121
5.3	Simulation of other random variables	126
5.4	Multivariate random number generation	142
5.5	Markov chain simulation	143
5.6	Monte Carlo integration	147
5.7	Advanced simulation methods	149
6	Computational linear algebra	158
6.1	Vectors and matrices in R	159
6.2	Matrix multiplication and inversion	166
6.3	Eigenvalues and eigenvectors	171
6.4	Other matrix decompositions	172
6.5	Other matrix operations	178
7	Numerical optimization	182
7.1	The golden section search method	182
7.2	Newton–Raphson	185
7.3	The Nelder–Mead simplex method	188
7.4	Built-in functions	191
7.5	Linear programming	192
Appendix	Review of random variables and distributions	209
	Index	212

Expanded contents

	Preface to the second edition	<i>page xi</i>
	Preface to the first edition	xiii
I	Getting started	1
1.1	What is statistical programming?	1
1.2	Outline of this book	2
1.3	The R package	3
1.4	Why use a command line?	3
1.5	Font conventions	4
1.6	Installation of R and RStudio	4
1.7	Getting started in RStudio	5
1.8	Going further	6
2	Introduction to the R language	7
2.1	First steps	7
2.1.1	R can be used as a calculator	7
2.1.2	Named storage	9
2.1.3	Quitting R	10
2.2	Basic features of R	11
2.2.1	Functions	11
2.2.2	R is case-sensitive	12
2.2.3	Listing the objects in the workspace	13
2.3	Vectors in R	13
2.3.1	Numeric vectors	13
2.3.2	Extracting elements from vectors	14
2.3.3	Vector arithmetic	15
2.3.4	Simple patterned vectors	16
2.3.5	Vectors with random patterns	17
2.3.6	Character vectors	17
2.3.7	Factors	18
2.3.8	More on extracting elements from vectors	19

2.3.9	Matrices and arrays	19
2.4	Data storage in R	22
2.4.1	Approximate storage of numbers	22
2.4.2	Exact storage of numbers	24
2.4.3	Dates and times	25
2.4.4	Missing values and other special values	25
2.5	Packages, libraries, and repositories	27
2.6	Getting help	28
2.6.1	Built-in help pages	28
2.6.2	Built-in examples	29
2.6.3	Finding help when you don't know the function name	30
2.6.4	Some built-in graphics functions	31
2.6.5	Some elementary built-in functions	33
2.7	Logical vectors and relational operators	34
2.7.1	Boolean algebra	34
2.7.2	Logical operations in R	34
2.7.3	Relational operators	36
2.8	Data frames and lists	37
2.8.1	Extracting data frame elements and subsets	39
2.8.2	Taking random samples from populations	40
2.8.3	Constructing data frames	40
2.8.4	Data frames can have non-numeric columns	40
2.8.5	Lists	41
2.9	Data input and output	43
2.9.1	Changing directories	43
2.9.2	<code>dump()</code> and <code>source()</code>	43
2.9.3	Redirecting R output	44
2.9.4	Saving and retrieving image files	45
2.9.5	The <code>read.table</code> function	45

3 | Programming statistical graphics 49

3.1	High level plots	50
3.1.1	Bar charts and dot charts	50
3.1.2	Pie charts	53
3.1.3	Histograms	54
3.1.4	Box plots	55
3.1.5	Scatterplots	57
3.1.6	Plotting data from data frames	57
3.1.7	QQ plots	60
3.2	Choosing a high level graphic	62
3.3	Low level graphics functions	63
3.3.1	The plotting region and margins	63
3.3.2	Adding to plots	64
3.3.3	Adjusting axis tick labels	66
3.3.4	Setting graphical parameters	68
3.4	Other graphics systems	70
3.4.1	The <code>ggplot2</code> package	70
3.4.2	The <code>lattice</code> package	72

3.4.3	The <code>grid</code> package	73
3.4.4	Interactive graphics	74
4	Programming with R	76
4.1	Flow control	76
4.1.1	The <code>for()</code> loop	76
4.1.2	The <code>if()</code> statement	82
4.1.3	The <code>while()</code> loop	86
4.1.4	Newton's method for root finding	87
4.1.5	The <code>repeat</code> loop, and the <code>break</code> and <code>next</code> statements	89
4.2	Managing complexity through functions	91
4.2.1	What are functions?	91
4.2.2	Scope of variables	94
4.2.3	Returning multiple objects	95
4.2.4	Using S3 classes to control printing	95
4.3	The <code>replicate()</code> function	97
4.4	Miscellaneous programming tips	97
4.4.1	Always edit code in the editor, not in the console	97
4.4.2	Documentation using <code>#</code>	98
4.4.3	Neatness counts!	98
4.5	Some general programming guidelines	100
4.5.1	Top-down design	103
4.6	Debugging and maintenance	107
4.6.1	Recognizing that a bug exists	108
4.6.2	Make the bug reproducible	108
4.6.3	Identify the cause of the bug	109
4.6.4	Fixing errors and testing	111
4.6.5	Look for similar errors elsewhere	111
4.6.6	Debugging in RStudio	111
4.6.7	The <code>browser()</code> , <code>debug()</code> , and <code>debugonce()</code> functions	112
4.7	Efficient programming	113
4.7.1	Learn your tools	114
4.7.2	Use efficient algorithms	114
4.7.3	Measure the time your program takes	116
4.7.4	Be willing to use different tools	117
4.7.5	Optimize with care	117
5	Simulation	120
5.1	Monte Carlo simulation	120
5.2	Generation of pseudorandom numbers	121
5.3	Simulation of other random variables	126
5.3.1	Bernoulli random variables	126
5.3.2	Binomial random variables	128
5.3.3	Poisson random variables	132
5.3.4	Exponential random numbers	136
5.3.5	Normal random variables	138
5.3.6	All built-in distributions	140

x | EXPANDED CONTENTS

5.4	Multivariate random number generation	142
5.5	Markov chain simulation	143
5.6	Monte Carlo integration	147
5.7	Advanced simulation methods	149
5.7.1	Rejection sampling	150
5.7.2	Importance sampling	152
6	Computational linear algebra	158
6.1	Vectors and matrices in R	159
6.1.1	Constructing matrix objects	159
6.1.2	Accessing matrix elements; row and column names	161
6.1.3	Matrix properties	163
6.1.4	Triangular matrices	164
6.1.5	Matrix arithmetic	165
6.2	Matrix multiplication and inversion	166
6.2.1	Matrix inversion	167
6.2.2	The <i>LU</i> decomposition	168
6.2.3	Matrix inversion in R	169
6.2.4	Solving linear systems	170
6.3	Eigenvalues and eigenvectors	171
6.4	Other matrix decompositions	172
6.4.1	The singular value decomposition of a matrix	172
6.4.2	The Choleski decomposition of a positive definite matrix	173
6.4.3	The QR decomposition of a matrix	174
6.5	Other matrix operations	178
6.5.1	Kronecker products	179
6.5.2	<code>apply()</code>	179
7	Numerical optimization	182
7.1	The golden section search method	182
7.2	Newton–Raphson	185
7.3	The Nelder–Mead simplex method	188
7.4	Built-in functions	191
7.5	Linear programming	192
7.5.1	Solving linear programming problems in R	195
7.5.2	Maximization and other kinds of constraints	195
7.5.3	Special situations	196
7.5.4	Unrestricted variables	199
7.5.5	Integer programming	200
7.5.6	Alternatives to <code>lp()</code>	201
7.5.7	Quadratic programming	202
Appendix	Review of random variables and distributions	209
	Index	212

Preface to the second edition

A lot of things have happened in the R community since we wrote the first edition of this text. Millions of new users have started to use R, and it is now the premier platform for data analytics. (In fact, the term “data analytics” hardly existed when we wrote the first edition.)

RStudio, a cross-platform integrated development environment for R, has had a large influence on the increase in popularity. In this edition we recommend RStudio as the platform for most new users, and have integrated simple RStudio instructions into the text. In fact, we have used RStudio and the `knitr` package in putting together the manuscript.

We have also added numerous examples and exercises, and cleaned up existing ones when they were unclear. Chapter 2 (Introduction to the R language) has had extensive revision and reorganization. We have added short discussions of newer graphics systems to Chapter 3 (Programming statistical graphics). Reference material on some common error messages has been added to Chapter 4 (Programming with R), and a list of pseudorandom number generators as well as a more extensive discussion of Markov chain Monte Carlo is new in Chapter 5 (Simulation). In Chapter 6 (Computational linear algebra), some applications have been added to give students a better idea of why some of the matrix decompositions are so important.

Once again we have a lot of people to thank. Many students have used the first edition, and we are grateful for their comments and criticisms. Some anonymous reviewers also provided some helpful suggestions and pointers so that we could make improvements to the text. We hope our readers find this new edition as interesting and educational as we think it is.

W. John Braun
Duncan Murdoch

November, 2015

Cambridge University Press

978-1-107-57646-9 - A First Course in Statistical Programming with R: Second Edition

W. John Braun and Duncan J. Murdoch

Frontmatter

[More information](#)

Preface to the first edition

This text began as notes for a course in statistical computing for second year actuarial and statistical students at the University of Western Ontario. Both authors are interested in statistical computing, both as support for our other research and for its own sake. However, we have found that our students were not learning the right sort of programming basics before they took our classes. At every level from undergraduate through Ph.D., we found that the students were not able to produce simple, reliable programs; that they didn't understand enough about numerical computation to understand how rounding error could influence their results, and that they didn't know how to begin a difficult computational project.

We looked into service courses from other departments, but we found that they emphasized languages and concepts that our students would not use again. Our students need to be comfortable with simple programming so that they can put together a simulation of a stochastic model; they also need to know enough about numerical analysis so that they can do numerical computations reliably. We were unable to find this mix in an existing course, so we designed our own.

We chose to base this text on R. R is an open source computing package which has seen a huge growth in popularity in the last few years. Being open source, it is easily obtainable by students and economical to install in our computing lab. One of us (Murdoch) is a member of the core R development team, and the other (Braun) is a co-author of a book on data analysis using R. These facts made it easy for us to choose R, but we are both strong believers in the idea that there are certain universals of programming, and in this text we try to emphasize those: it is not a manual about programming in R, it is a course in statistical programming that uses R.

Students starting this course are not assumed to have any programming experience or advanced statistical knowledge. They should be familiar with university-level calculus, and should have had exposure to a course in introductory probability, though that could be taken concurrently: the probabilistic concepts start in Chapter 5. (We include a concise appendix reviewing the probabilistic material.) We include some advanced topics in simulation, linear algebra, and optimization that an instructor may choose to skip in a one-semester course offering.

Cambridge University Press

978-1-107-57646-9 - A First Course in Statistical Programming with R: Second Edition

W. John Braun and Duncan J. Murdoch

Frontmatter

[More information](#)

We have a lot of people to thank for their help in writing this book. The students in Statistical Sciences 259b have provided motivation and feedback, Lutong Zhou drafted several figures, Kristy Alexander, Yiwen Diao, Qiang Fu, and Yu Han went over the exercises and wrote up detailed solutions, and Diana Gillooly of Cambridge University Press, Professor Brian Ripley of Oxford University, and some anonymous reviewers all provided helpful suggestions. And of course, this book could not exist without R, and R would be far less valuable without the contributions of the worldwide R community.

W. John Braun
Duncan Murdoch

February, 2007