

# Use R!

*Advisors:*

Robert Gentleman • Kurt Hornik • Giovanni Parmigiani

# Use R!

---

*Series Editors:* Robert Gentleman, Kurt Hornik, and Giovanni Parmigiani

*Albert:* Bayesian Computation with R

*Bivand/Pebesma/Gómez-Rubio:* Applied Spatial Data Analysis with R

*Claude:* Morphometrics with R

*Cook/Swayne:* Interactive and Dynamic Graphics for Data Analysis: With R and GGobi

*Hahne/Huber/Gentleman/Falcon:* Bioconductor Case Studies

*Kleiber/Zeileis,* Applied Econometrics with R

*Nason:* Wavelet Methods in Statistics with R

*Paradis:* Analysis of Phylogenetics and Evolution with R

*Peng/Dominici:* Statistical Methods for Environmental Epidemiology with R:  
A Case Study in Air Pollution and Health

*Pfaff:* Analysis of Integrated and Cointegrated Time Series with R, 2nd edition

*Sarkar:* Lattice: Multivariate Data Visualization with R

*Spector:* Data Manipulation with R

Alain F. Zuur • Elena N. Ieno •  
Erik H.W.G. Meesters

# A Beginner's Guide to R

 Springer

Alain F. Zuur  
Highland Statistics Ltd.  
6 Laverock Road  
Newburgh  
United Kingdom AB41 6FN  
highstat@highstat.com

Elena N. Ieno  
Highland Statistics Ltd.  
6 Laverock Road  
Newburgh  
United Kingdom AB41 6FN  
bio@highstat.com

Erik H.W.G. Meesters  
IMARES, Institute for Marine  
Resources & Ecosystem Studies  
1797 SH 't Horntje  
The Netherlands  
erik.meesters@wur.nl

ISBN 978-0-387-93836-3 e-ISBN 978-0-387-93837-0  
DOI 10.1007/978-0-387-93837-0  
Springer Dordrecht Heidelberg London New York

Library of Congress Control Number: 2009929643

© Springer Science+Business Media, LLC 2009

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden. The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

Printed on acid-free paper

Springer is part of Springer Science + Business Media ([www.springer.com](http://www.springer.com))

*To my future niece (who will undoubtedly  
cost me a lot of money)*

Alain F. Zuur

*To Juan Carlos and Norma*

Elena N. Ieno

*For Leontine and Ava, Rick, and Merel*

Erik H.W.G. Meesters

# Preface

## The Absolute R Beginner

For whom was this book written?

Since 2000, we have taught statistics to over 5000 life scientists. This sounds a lot, and indeed it is, but with some classes of 200 undergraduate students, numbers accumulate rapidly (although some courses have involved as few as 6 students). Most of our teaching has been done in Europe, but we have also conducted courses in South America, Central America, the Middle East, and New Zealand. Of course teaching at universities and research organisations means that our students may be from almost anywhere in the world. Participants have included undergraduates, but most have been MSc students, post-graduate students, post-docs, or senior scientists, along with some consultants and nonacademics.

This experience has given us an informed awareness of the typical life scientist's knowledge of statistics. The word "typical" may be misleading, as those scientists enrolling in a statistics course are likely to be those who are unfamiliar with the topic or have become rusty. In general, we have worked with people who, at some stage in their education or career, have completed a statistics course covering such topics as mean, variance,  $t$ -test, Chi-square test, and hypothesis testing, and perhaps including half an hour devoted to linear regression.

There are many books available on doing statistics with R. But this book does not deal with statistics, as, in our experience, teaching statistics and R at the same time means two steep learning curves, one for the statistical methodology and one for the R code. This is more than many students are prepared to undertake. This book is intended for people seeking an elementary introduction to R. Obviously, the term "elementary" is vague; elementary in one person's view may be advanced in another's.

R contains a high "you need to know what you are doing" content, and its application requires a considerable amount of logical thinking. As statisticians, it is easy to sit in an ivory tower and expect the life scientist to knock on our door and ask to learn our language. This book aims to make that language as simple

as possible. If the phrase “absolute beginner” offends, we apologize, but it answers the question: For whom is this book intended?

All authors of this book are Windows users and have limited experience with Linux and with Mac OS. R is also available for computers with these operating systems, and all the R code we present should run properly on them. However, there may be small differences with saving graphs. Non-Windows users will also need to find an alternative to the text editor Tinn-R (Chapter 1 discusses where you can find information on this).

## **Datasets used in This book**

This book uses mainly life science data. Nevertheless, whatever your area of study and whatever your data, the procedures presented will apply. Scientists in all fields need to import data, massage data, make graphs, and, finally, perform analyses. The R commands will be very similar in every case. A 200-page book does not offer a great deal of scope for presenting a variety of dataset types, and, in our experience, widely divergent examples confuse the reader. The optimal approach may be to use a single dataset to demonstrate all techniques, but this does not make many people happy. Therefore, we have used ecological datasets (e.g., involving plants, marine benthos, fish, birds) and epidemiological datasets.

All datasets used in this book are downloadable from [www.highstat.com](http://www.highstat.com).

Newburgh  
Newburgh  
Den Burg

Alain F. Zuur  
Elena N. Ieno  
Erik H.W.G. Meesters

# Acknowledgements

We thank Chris Elphick for the sparrow data; Graham Pierce for the squid data; Monty Priede for the ISIT data; Richard Loyn for the Australian bird data; Gerard Janssen for the benthic data; Pam Sikkink for the grassland data; Alexandre Roulin for the barn owl data; Michael Reed and Chris Elphick for the Hawaiian bird data; Robert Cruikshanks, Mary Kelly-Quinn, and John O'Halloran for the Irish river data; Joaquín Vicente and Christian Gortázar for the wild boar and deer data; Ken Mackenzie for the cod data; Sonia Mendes for the whale data; Max Latuhihin and Hanneke Baretta-Bekker for the Dutch salinity and temperature data; and António Mira and Filipe Carvalho for the roadkill data. The full references are given in the text.

This is our third book with Springer, and we thank John Kimmel for giving us the opportunity to write it. We also thank all course participants who commented on the material.

We thank Anatoly Saveliev and Gema Hernández-Milian for commenting on earlier drafts and Kathleen Hills (The Lucidus Consultancy) for editing the text.

# Contents

|                                                                           |     |
|---------------------------------------------------------------------------|-----|
| <b>Preface</b> .....                                                      | vii |
| <b>Acknowledgements</b> .....                                             | ix  |
| <b>1 Introduction</b> .....                                               | 1   |
| 1.1 What Is R? .....                                                      | 1   |
| 1.2 Downloading and Installing R .....                                    | 2   |
| 1.3 An Initial Impression .....                                           | 4   |
| 1.4 Script Code .....                                                     | 7   |
| 1.4.1 The Art of Programming .....                                        | 7   |
| 1.4.2 Documenting Script Code .....                                       | 8   |
| 1.5 Graphing Facilities in R .....                                        | 10  |
| 1.6 Editors .....                                                         | 12  |
| 1.7 Help Files and Newsgroups .....                                       | 13  |
| 1.8 Packages .....                                                        | 16  |
| 1.8.1 Packages Included with the Base Installation .....                  | 16  |
| 1.8.2 Packages Not Included with the Base<br>Installation .....           | 17  |
| 1.9 General Issues in R .....                                             | 19  |
| 1.9.1 Quitting R and Setting the Working Directory .....                  | 21  |
| 1.10 A History and a Literature Overview .....                            | 22  |
| 1.10.1 A Short Historical Overview of R .....                             | 22  |
| 1.10.2 Books on R and Books Using R .....                                 | 22  |
| 1.11 Using This Book .....                                                | 24  |
| 1.11.1 If You Are an Instructor .....                                     | 25  |
| 1.11.2 If You Are an Interested Reader with Limited R<br>Experience ..... | 25  |
| 1.11.3 If You Are an R Expert .....                                       | 25  |
| 1.11.4 If You Are Afraid of R .....                                       | 25  |
| 1.12 Citing R and Citing Packages .....                                   | 26  |
| 1.13 Which R Functions Did We Learn? .....                                | 27  |

|          |                                                                                                     |    |
|----------|-----------------------------------------------------------------------------------------------------|----|
| <b>2</b> | <b>Getting Data into R</b>                                                                          | 29 |
| 2.1      | First Steps in R                                                                                    | 29 |
| 2.1.1    | Typing in Small Datasets                                                                            | 29 |
| 2.1.2    | Concatenating Data with the <code>c</code> Function                                                 | 31 |
| 2.1.3    | Combining Variables with the <code>c</code> , <code>cbind</code> , and <code>rbind</code> Functions | 34 |
| 2.1.4    | Combining Data with the <code>vector</code> Function*                                               | 39 |
| 2.1.5    | Combining Data Using a <code>Matrix</code> *                                                        | 39 |
| 2.1.6    | Combining Data with the <code>data.frame</code> Function                                            | 42 |
| 2.1.7    | Combining Data Using the <code>list</code> Function*                                                | 43 |
| 2.2      | Importing Data                                                                                      | 46 |
| 2.2.1    | Importing Excel Data                                                                                | 47 |
| 2.2.2    | Accessing Data from Other Statistical Packages**                                                    | 51 |
| 2.2.3    | Accessing a Database***                                                                             | 52 |
| 2.3      | Which R Functions Did We Learn?                                                                     | 54 |
| 2.4      | Exercises                                                                                           | 54 |
| <b>3</b> | <b>Accessing Variables and Managing Subsets of Data</b>                                             | 57 |
| 3.1      | Accessing Variables from a Data Frame                                                               | 57 |
| 3.1.1    | The <code>str</code> Function                                                                       | 59 |
| 3.1.2    | The Data Argument in a Function                                                                     | 60 |
| 3.1.3    | The <code>\$</code> Sign                                                                            | 61 |
| 3.1.4    | The <code>attach</code> Function                                                                    | 62 |
| 3.2      | Accessing Subsets of Data                                                                           | 63 |
| 3.2.1    | Sorting the Data                                                                                    | 66 |
| 3.3      | Combining Two Datasets with a Common Identifier                                                     | 67 |
| 3.4      | Exporting Data                                                                                      | 69 |
| 3.5      | Recoding Categorical Variables                                                                      | 71 |
| 3.6      | Which R Functions Did We Learn?                                                                     | 74 |
| 3.7      | Exercises                                                                                           | 74 |
| <b>4</b> | <b>Simple Functions</b>                                                                             | 77 |
| 4.1      | The <code>tapply</code> Function                                                                    | 77 |
| 4.1.1    | Calculating the Mean Per Transect                                                                   | 78 |
| 4.1.2    | Calculating the Mean Per Transect More Efficiently                                                  | 79 |
| 4.2      | The <code>sapply</code> and <code>lapply</code> Functions                                           | 80 |
| 4.3      | The <code>summary</code> Function                                                                   | 81 |
| 4.4      | The <code>table</code> Function                                                                     | 82 |
| 4.5      | Which R Functions Did We Learn?                                                                     | 84 |
| 4.6      | Exercises                                                                                           | 84 |
| <b>5</b> | <b>An Introduction to Basic Plotting Tools</b>                                                      | 85 |
| 5.1      | The <code>plot</code> Function                                                                      | 85 |
| 5.2      | Symbols, Colours, and Sizes                                                                         | 88 |
| 5.2.1    | Changing Plotting Characters                                                                        | 88 |

|          |                                                                       |            |
|----------|-----------------------------------------------------------------------|------------|
| 5.2.2    | Changing the Colour of Plotting Symbols . . . . .                     | 92         |
| 5.2.3    | Altering the Size of Plotting Symbols . . . . .                       | 93         |
| 5.3      | Adding a Smoothing Line . . . . .                                     | 95         |
| 5.4      | Which R Functions Did We Learn? . . . . .                             | 97         |
| 5.5      | Exercises . . . . .                                                   | 97         |
| <b>6</b> | <b>Loops and Functions . . . . .</b>                                  | <b>99</b>  |
| 6.1      | Introduction to Loops . . . . .                                       | 99         |
| 6.2      | Loops . . . . .                                                       | 101        |
| 6.2.1    | Be the Architect of Your Code . . . . .                               | 102        |
| 6.2.2    | Step 1: Importing the Data . . . . .                                  | 102        |
| 6.2.3    | Steps 2 and 3: Making the Scatterplot and Adding<br>Labels . . . . .  | 103        |
| 6.2.4    | Step 4: Designing General Code . . . . .                              | 104        |
| 6.2.5    | Step 5: Saving the Graph. . . . .                                     | 105        |
| 6.2.6    | Step 6: Constructing the Loop . . . . .                               | 107        |
| 6.3      | Functions . . . . .                                                   | 108        |
| 6.3.1    | Zeros and NAs . . . . .                                               | 108        |
| 6.3.2    | Technical Information. . . . .                                        | 110        |
| 6.3.3    | A Second Example: Zeros and NAs . . . . .                             | 111        |
| 6.3.4    | A Function with Multiple Arguments. . . . .                           | 113        |
| 6.3.5    | Foolproof Functions . . . . .                                         | 115        |
| 6.4      | More on Functions and the <code>if</code> Statement . . . . .         | 117        |
| 6.4.1    | Playing the Architect Again . . . . .                                 | 118        |
| 6.4.2    | Step 1: Importing and Assessing the Data . . . . .                    | 118        |
| 6.4.3    | Step 2: Total Abundance per Site . . . . .                            | 119        |
| 6.4.4    | Step 3: Richness per Site . . . . .                                   | 120        |
| 6.4.5    | Step 4: Shannon Index per Site . . . . .                              | 121        |
| 6.4.6    | Step 5: Combining Code . . . . .                                      | 122        |
| 6.4.7    | Step 6: Putting the Code into a Function . . . . .                    | 122        |
| 6.5      | Which R Functions Did We Learn? . . . . .                             | 125        |
| 6.6      | Exercises . . . . .                                                   | 125        |
| <b>7</b> | <b>Graphing Tools . . . . .</b>                                       | <b>127</b> |
| 7.1      | The Pie Chart . . . . .                                               | 127        |
| 7.1.1    | Pie Chart Showing Avian Influenza Data. . . . .                       | 127        |
| 7.1.2    | The <code>par</code> Function . . . . .                               | 130        |
| 7.2      | The Bar Chart and Strip Chart . . . . .                               | 131        |
| 7.2.1    | The Bar Chart Using the Avian Influenza Data . . . . .                | 131        |
| 7.2.2    | A Bar Chart Showing Mean Values with Standard<br>Deviations . . . . . | 133        |
| 7.2.3    | The Strip Chart for the Benthic Data . . . . .                        | 135        |
| 7.3      | Boxplot. . . . .                                                      | 137        |
| 7.3.1    | Boxplots Showing the Owl Data . . . . .                               | 137        |
| 7.3.2    | Boxplots Showing the Benthic Data . . . . .                           | 140        |

|          |                                                               |            |
|----------|---------------------------------------------------------------|------------|
| 7.4      | Cleveland Dotplots . . . . .                                  | 141        |
| 7.4.1    | Adding the Mean to a Cleveland Dotplot . . . . .              | 143        |
| 7.5      | Revisiting the <code>plot</code> Function . . . . .           | 145        |
| 7.5.1    | The Generic <code>plot</code> Function . . . . .              | 145        |
| 7.5.2    | More Options for the <code>plot</code> Function . . . . .     | 146        |
| 7.5.3    | Adding Extra Points, Text, and Lines . . . . .                | 148        |
| 7.5.4    | Using <code>type = "n"</code> . . . . .                       | 149        |
| 7.5.5    | Legends . . . . .                                             | 150        |
| 7.5.6    | Identifying Points . . . . .                                  | 152        |
| 7.5.7    | Changing Fonts and Font Size* . . . . .                       | 153        |
| 7.5.8    | Adding Special Characters . . . . .                           | 153        |
| 7.5.9    | Other Useful Functions . . . . .                              | 154        |
| 7.6      | The Pairplot . . . . .                                        | 155        |
| 7.6.1    | Panel Functions . . . . .                                     | 156        |
| 7.7      | The Coplot . . . . .                                          | 157        |
| 7.7.1    | A Coplot with a Single Conditioning Variable . . . . .        | 157        |
| 7.7.2    | The Coplot with Two Conditioning Variables . . . . .          | 161        |
| 7.7.3    | Jazzing Up the Coplot* . . . . .                              | 162        |
| 7.8      | Combining Types of Plots* . . . . .                           | 164        |
| 7.9      | Which R Functions Did We Learn? . . . . .                     | 166        |
| 7.10     | Exercises . . . . .                                           | 167        |
| <b>8</b> | <b>An Introduction to the Lattice Package . . . . .</b>       | <b>169</b> |
| 8.1      | High-Level Lattice Functions . . . . .                        | 169        |
| 8.2      | Multipanel Scatterplots: <code>xyplot</code> . . . . .        | 170        |
| 8.3      | Multipanel Boxplots: <code>bwplot</code> . . . . .            | 173        |
| 8.4      | Multipanel Cleveland Dotplots: <code>dotplot</code> . . . . . | 174        |
| 8.5      | Multipanel Histograms: <code>histogram</code> . . . . .       | 176        |
| 8.6      | Panel Functions . . . . .                                     | 177        |
| 8.6.1    | First Panel Function Example . . . . .                        | 177        |
| 8.6.2    | Second Panel Function Example . . . . .                       | 179        |
| 8.6.3    | Third Panel Function Example* . . . . .                       | 181        |
| 8.7      | 3-D Scatterplots and Surface and Contour Plots . . . . .      | 184        |
| 8.8      | Frequently Asked Questions . . . . .                          | 185        |
| 8.8.1    | How to Change the Panel Order? . . . . .                      | 186        |
| 8.8.2    | How to Change Axes Limits and Tick Marks? . . . . .           | 188        |
| 8.8.3    | Multiple Graph Lines in a Single Panel . . . . .              | 189        |
| 8.8.4    | Plotting from Within a Loop* . . . . .                        | 190        |
| 8.8.5    | Updating a Plot . . . . .                                     | 191        |
| 8.9      | Where to Go from Here? . . . . .                              | 191        |
| 8.10     | Which R Functions Did We Learn? . . . . .                     | 192        |
| 8.11     | Exercises . . . . .                                           | 192        |

|                                                                                      |     |
|--------------------------------------------------------------------------------------|-----|
| <b>9 Common R Mistakes</b> . . . . .                                                 | 195 |
| 9.1 Problems Importing Data . . . . .                                                | 195 |
| 9.1.1 Errors in the Source File . . . . .                                            | 195 |
| 9.1.2 Decimal Point or Comma Separation . . . . .                                    | 195 |
| 9.1.3 Directory Names . . . . .                                                      | 197 |
| 9.2 Attach Misery . . . . .                                                          | 197 |
| 9.2.1 Entering the Same attach Command Twice. . . . .                                | 197 |
| 9.2.2 Attaching Two Data Frames Containing the Same<br>Variable Names . . . . .      | 198 |
| 9.2.3 Attaching a Data Frame and Demo Data. . . . .                                  | 199 |
| 9.2.4 Making Changes to a Data Frame After Applying the<br>attach Function . . . . . | 200 |
| 9.3 Non-attach Misery . . . . .                                                      | 201 |
| 9.4 The Log of Zero . . . . .                                                        | 202 |
| 9.5 Miscellaneous Errors . . . . .                                                   | 203 |
| 9.5.1 The Difference Between 1 and l. . . . .                                        | 203 |
| 9.5.2 The Colour of 0 . . . . .                                                      | 203 |
| 9.6 Mistakenly Saved the R Workspace. . . . .                                        | 204 |
| <b>References</b> . . . . .                                                          | 207 |
| <b>Index</b> . . . . .                                                               | 211 |