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

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, postgraduate 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.

Newburgh Alain F. Zuur
Newburgh Elena N. Ieno
Den Burg 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 Hernádez-Milian for commenting on earlier drafts and Kathleen Hills (The Lucidus Consultancy) for editing the text.

Contents