STATISTICS FOR LINGUISTS

AN INTRODUCTION USING R

BODO WINTER

ROUTLEDGE

Statistics for Linguists: An Introduction Using R

Statistics for Linguists: An Introduction Using R is the first statistics textbook on linear models for linguistics. The book covers simple uses of linear models through generalized models to more advanced approaches, maintaining its focus on conceptual issues and avoiding excessive mathematical details. It contains many applied examples using the R statistical programming environment. Written in an accessible tone and style, this text is the ideal main resource for graduate and advanced undergraduate students of Linguistics statistics courses as well as those in other fields, including Psychology, Cognitive Science, and Data Science.

Bodo Winter is Lecturer in Cognitive Linguistics in the Department of English Language and Applied Linguistics at the University of Birmingham, UK.


Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Statistics for Linguists: An Introduction Using R

Bodo Winter

Routledge
Taylor & Francis Group
NEW YORK AND LONDON

First published 2020
by Routledge
52 Vanderbilt Avenue, New York, NY 10017

and by Routledge
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

Routledge is an imprint of the Taylor & Francis Group, an informa business

Library of Congress Cataloging-in-Publication Data
Names: Winter, Bodo, author.
Title: Statistics for linguists : an introduction using R / by Bodo Winter.
Description: New York, NY : Routledge, 2019. | Includes bibliographical references and index.
Identifiers: LCCN 2019029350 (print) | LCCN 2019029351 (ebook) | ISBN 9781138056084 (hbk) | ISBN 9781138056091 (pbk) | ISBN 9781315165547 (ebk)
Subjects: LCSH: Linguistics—Statistical methods. | R (Computer program language) | Mathematical linguistics.
Classification: LCC P138.5 .W56 2019 (print) | LCC P138.5 (ebook) | DDC 410.1/5195—dc23
LC record available at https://lccn.loc.gov/2019029350
LC ebook record available at https://lccn.loc.gov/2019029351

ISBN: 978-1-138-05608-4 (hbk)
ISBN: 978-1-138-05609-1 (pbk)
ISBN: 978-1-315-16554-7 (ebk)

Typeset in Times New Roman
by Apex CoVantage, LLC

Contents

Acknowledgments

Let me take a few paragraphs to thank the people who have helped with this book. I've been lucky to have been exposed to some excellent statistics teaching. First, I want to thank Benjamin Bergen and Amy Schafer for getting me going with stats while I was a grad student at the University of Hawai'i at Mānoa. Second, I would like to thank Sarah Depaoli and Jack Vevea for teaching excellent graduate-level stats courses at the University of California, Merced. Finally, my biggest thanks go to Roger Mundry. I will never forget your workshops and our pizza nights.

Big thanks also go to Timo Roettger and Márton Sóskuthy for helping me develop the materials for the Birmingham Statistics for Linguists Summer School. I particularly want to thank Timo, a close friend and collaborator during all these years, for continuously challenging me. I hope that at some point I will become the scientist that lives up to his standards.

I want to thank Bruno Nicenboim for providing an excellent review of this book that led to many changes. I also want to thank Kamil Kaźmierski and Keith Wilson for additional suggestions. My student Greg Woodin has read all chapters and the book made a massive jump in quality thanks to his feedback. Another person who has been helping behind the scenes is my father Clive Winter, who has generously proofread early drafts of each chapter.

Special thanks go to the team at Routledge for their patience with me, as well as for their invaluable work on copy-editing and formatting this book (big shout-out for Nikky Twyman for her hard work on the manuscript). Sorry for being perennially late and not following the submission guidelines!

This book would not exist if not for Martine Grice, Anne Hermes, Doris Mücke, and Stefan Baumann. I taught my first course on R and mixed models for the phonetics group at the Institut für Linguistik.

I also want to thank all of the participants of the countless other workshops I have taught. If you were a member of one of my workshops, rest assured that it was your enthusiasm and your questions that allowed me to continuously refine the ways I explain certain concepts. I want to particularly thank the participants of the 2018 stats workshop at the Deafness Cognition and Language Research Centre at UCL, as well the participants of the first Birmingham Statistics for Linguists Summer School (2018).

I also want to thank those countless people who have sent me unsolicited thank-you emails in response to my freely available mixed model tutorial. Thanks for taking the time to reach out!

Finally, I want to thank the people who made life meaningful during the time I was writing this book, or who have supported me in other ways. These include my mum and dad, Mark, Louis, Vincenzo, Maciek, Marcus, Brittany, Jeannette, Matteo, Emily, Suganthi, Tommy, Dan, Jacob, Logan, Brendan, Jim, and Barry. I also want to include the Midlands Out badminton team, my yoga teachers Anna Robottom and Richard George, as well as the team of the Henrietta Street Gym for keeping me in high spirits.

0 Preface

Approach and How to Use This Book

The language sciences are undergoing a quantitative revolution. There is ever more data, an ever-growing toolkit of statistical methods, and a dedicated push towards incorporating more empirical research into linguistic theorizing. This book is designed to serve as a starting point for researchers from the language sciences and related disciplines to engage with emerging trends in data analysis. The plan is to take the reader from their first steps in R all the way to more advanced techniques such as linear mixed effects models. Along the way, the book aims to foster reproducible research practices.

Although this book is focused on the analysis of linguistic datasets, the examples chosen are accessible to researchers from other fields.

0.1. Strategy of the Book

This book is intended as a full course in basic linear modeling, from descriptive statistics, regression, and multiple regression, over to logistic regression and Poisson regression, all the way up to mixed models. Other books introduce statistics with more 'traditional' methods, such as *t*-tests, Chi-Square tests, ANOVAs, etc. I believe these significance tests are the wrong starting point for learners. When students are introduced to this 'testing framework', they spend most of their time worrying about what test to pick, rather than worrying about how to implement their theoretical understanding of a phenomenon in a statistical model. This book fosters model-based thinking rather than test-based thinking.

I have found out through experience that introducing learners to statistics via the linear model framework is much more engaging than teaching an array of significance tests. When starting one's statistical journey with these traditional methods, statistics seems like a long vocabulary list, and the student is left hanging with a highly compartmentalized view of the field. Moreover, learning statistics via significance tests gives all the wrong incentives. Significance tests such as *t*-tests, Chi-Square tests, and ANOVAs provide 'quick fixes' that encourage the researcher to treat the *p*-value as the ultimate arbiter of truth. Instead, students should be encouraged to think deeply about their theories, and they should be encouraged to spend a lot of time interpreting their models in substantive terms.

This book does not focus on the underlying mathematics, for which more advanced textbooks are available. The tone is deliberately casual, trying to make statistics approachable. Everything is geared towards practical relevance, aimed at researchers and students who want to model their data statistically to address their research questions. Some might argue that it is dangerous to teach linear models and their extensions without delving deeply into the underlying mathematics. Indeed, for every chapter in this book, there are book-length treatments that go into much more detail. So, naturally, I've had to cut corners somewhere in assembling this material. I do not intend to further statistical ignorance. Instead, I believe that there is value in a friendly, practical introduction geared towards usability, with more interested and mathematically capable readers being free to read more advanced texts further down the line.

There are many problems with how statistical methods are applied within the language sciences. For example, people use significance tests and *p*-values without knowing what they mean; people misinterpret main effects in the presence of interactions; and people fit models without being aware of convergence issues. This book is very much written with these problems in mind, based on my experience of teaching workshops to linguists at various institutions. Most of the errors that I see in the context of mixed models, for example, have to do with an insufficient understanding of the underlying regression framework. This is another reason for why this book focuses so much on linear models.

I think that part of the problem the field has with statistics stems from the fact that methods that are too advanced have been given to linguists too early. Baayen's landmark textbook *Analyzing Linguistic Data* (2008) ushered in a new quantitative era in linguistics. However, what many people end up using in practice is a pale reflection of what's discussed in Baayen's book or other texts. This state of affairs results from the fact that there is a scarcity of easy textbooks that provide stepping stones towards more advanced reads. In believe that, as a scientific community, we further bad statistics if we only write textbooks written for mathematically advanced readers.

This book is written in a friendly tone, directly addressing the reader ('you'), and explaining each formula and R function in quite a lot of detail. A lot of the datasets analyzed come from my own work because it allows me to guide the student through some of the reasoning processes that were going on behind the scenes. Altogether, I hope that this is a very relatable and non-threatening introduction to statistics with R.

Finally, this book tries to foster reproducible research practices. For example, the book gives advice about data sharing and reproducible code. Moreover, I will emphasize that a publication without concomitant release of the data and code has to be considered incomplete—unless there are very good reasons for not sharing these materials. Reproducible research takes effort and needs to be trained. This book tries to give the right incentives in these matters.

0.2. Why R?

This book is entirely based on the R programming language. These days, it's safe to say that R is the de facto standard in the language sciences. If you are reading this, chances are that the following paragraphs involve preaching to the choir, since picking up this book likely means that you are convinced of the necessity of using R.

However, since there are still many linguistics and psychology departments that teach their students proprietary software, in particular SPSS, it is worth highlighting why there is absolutely no way around R these days, and why this book *had* to be structured around R.

First, let me make a strong claim to begin with: if you are teaching your students SPSS rather than a programming language, such as R, you are actively disadvantaging their careers, as well as their success as scientists. Not only are there more job posts that ask for R skills than SPSS skills, but R is also much more conducive to open and reproducible research practices, which are required by an increasing number of academic journals and funding bodies. At some point in the near future, it will be difficult for your students to publish if they don't offer their data and code, a practice that the point-and-click structure of SPSS does not actively incentivize. I'd go as far as saying that, at this stage, teaching SPSS to students is unethical, because doing so is inherently directed against the open and community-driven nature of science.

Sometimes I hear the argument that R may be too difficult for students, in particular for undergraduates. In stark contrast to this, I've found that students from all sorts of backgrounds (even without any programming knowledge) can quickly pick up R if it is taught in a friendly manner. Moreover, it helps students that R can be used on their own machines without licensing hassle, and it further helps students that there's by now much more online help for R than for SPSS. Also, the interactive nature of R, as well as the ease with which plots can be created, can be highly engaging to students.

A final point about R is that it allows making 'data wrangling' an integral part of a statistical analysis. Because preprocessing the data and statistical modeling are two sides of the same coin, they should happen within the same software environment. R makes this easier than other software.

0.3. Why the Tidyverse?

These days, there are two 'dialects' or 'styles' of programming in R. One uses mostly 'base R' functions (those that come with the original R distribution). The other dialect uses 'tidyverse' packages. The 'tidyverse' is a label used for the network of packages developed by Hadley Wickham and colleagues, including such widely known packages as `dplyr` and `ggplot2`. Which style should you learn?

Essentially, there's no way around knowing *both* of these styles. A solid foundation in base R is still crucial, even if many tidyverse functions provide easier alternatives. Many web tutorials or discussions in online help forums such as StackOverflow include extensive base R code, but the student will invariably also encounter tidyverse-style code. Given this state of affairs, I think that it is necessary to teach both styles.

That said, the 'tidy' style is much easier to read and I've found that students grasp it more quickly. So I decided that there should be one introductory chapter on base R (Chapter 1), as well as one on the tidyverse (Chapter 2). However, after Chapter 2, the book almost exclusively uses tidyverse-style code from Chapter 2 onwards. Only when base R offers the easier alternative is base R code used.

0.4. R Packages Required for This Book

You need to download and install R and RStudio, which can be downloaded online. The following R packages need to be installed to be able to execute all code in all chapters. The `tidyverse` and `broom` packages should be loaded for every chapter, as they are used throughout the entire book.

```
install.packages('tidyverse')
install.packages('broom')
install.packages('gridExtra')
install.packages('car')
install.packages('MASS')
install.packages('pscl')
install.packages('effsize')
install.packages('lme4')
install.packages('afex')
install.packages('brms')
install.packages('MuMIn')
install.packages('swirl')
install.packages('languageR')
install.packages('emmeans')
```

0.5. What This Book Is Not

To get any false expectations out of the way, let me tell you a few things that this book is *not*.

- This book is not an introduction to the underlying theory and mathematics of regression, or mixed models. For this, there are more advanced materials available. Beware that any introductory text will have to cut corners on some topics, and this one is no exception.
- This book is not an introduction to exploratory techniques, such as exploratory factor analysis, cluster analysis, or classification and regression trees.
- This book is not a 'cookbook' that teaches you a whole range of different techniques. The focus is on regression modeling. Appendix A shows how some basic significance tests (such as *t*-tests) map onto the techniques discussed throughout the book. The concluding chapter of this book, Chapter 16, provides additional arguments why the cookbook approach is limiting students, and why it should be avoided whenever possible.

0.6. How to Use This Book

The materials presented here are intended as a full course. Each chapter combines conceptual introductions to statistical topics with hands-on applications. To maximize learning benefits, it is of utmost importance that you actually execute the R code presented in each chapter. Only by typing in each and every command can you develop the relevant muscle memory to learn the programming language.

All the data that is needed for this book can be accessed via the following Open Science Framework (OSF) repository:

https://osf.io/34mq9/

Some further recommendations:

- Although I do provide script files for all chapters on this webpage, I don't recommend looking at these while reading the book. Only consult these when you get stuck.
- The data can be downloaded from the website. It's possible to work through the entire book in one continued R session. Alternatively, you can also quit R after each chapter and come back to where you finished.
- I highly recommend setting up a folder on your computer where all the materials are saved, and where you create scripts that follow the code presented in each chapter. Annotate the code with your own comments to make it 'your own'.
- There are exercises at the end of each chapter. The solutions to the exercises can also be found on the above-mentioned repository.

0.7. Information for Teachers

This book is intended to be read from front to back. However, Appendix A (on significance tests) and Chapters 9 and 10 can be moved to different points, depending on the needs of a particular course. Likewise, Chapter 16 can be read independently of the other chapters, as well. This book can be used for both undergraduate and postgraduate courses. Chapter 8 on interactions is hard and, if teaching an undergraduate class, I may forestall this chapter at the expense of having more time to discuss inferential statistics (Chapters 9, 10, and 11).

If you've already taught statistics classes, you may be used to teaching a class that is focused on significance tests. In this case, I welcome you to consider the approach adopted in this book. Trust me, it works.

That said, you may want to continue teaching significance tests. In this case, this book could still be a useful textbook for your class, as the issues discussed here also apply to significance tests. Moreover, the methods discussed throughout the chapters have direct correspondences to significance tests, and these correspondences are explained in Appendix A.

All in all, I hope that this book strikes a nice balance between the easy and the advanced, so that readers from all levels will find something useful in it.