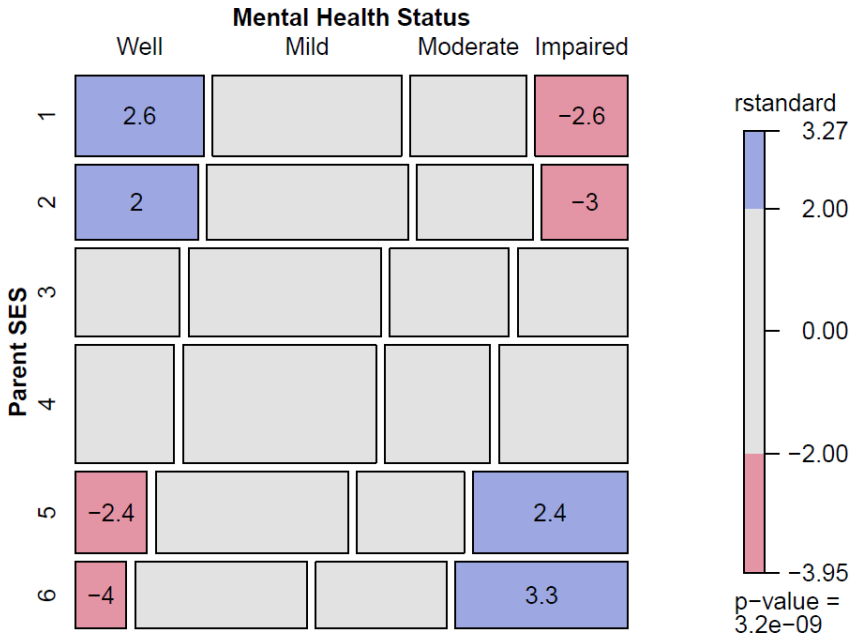
Discrete Data Analysis with R

Visualization and Modeling Techniques for Categorical and Count Data

Michael Friendly
York University

David Meyer
UAS Technikum Wien

March 12, 2015



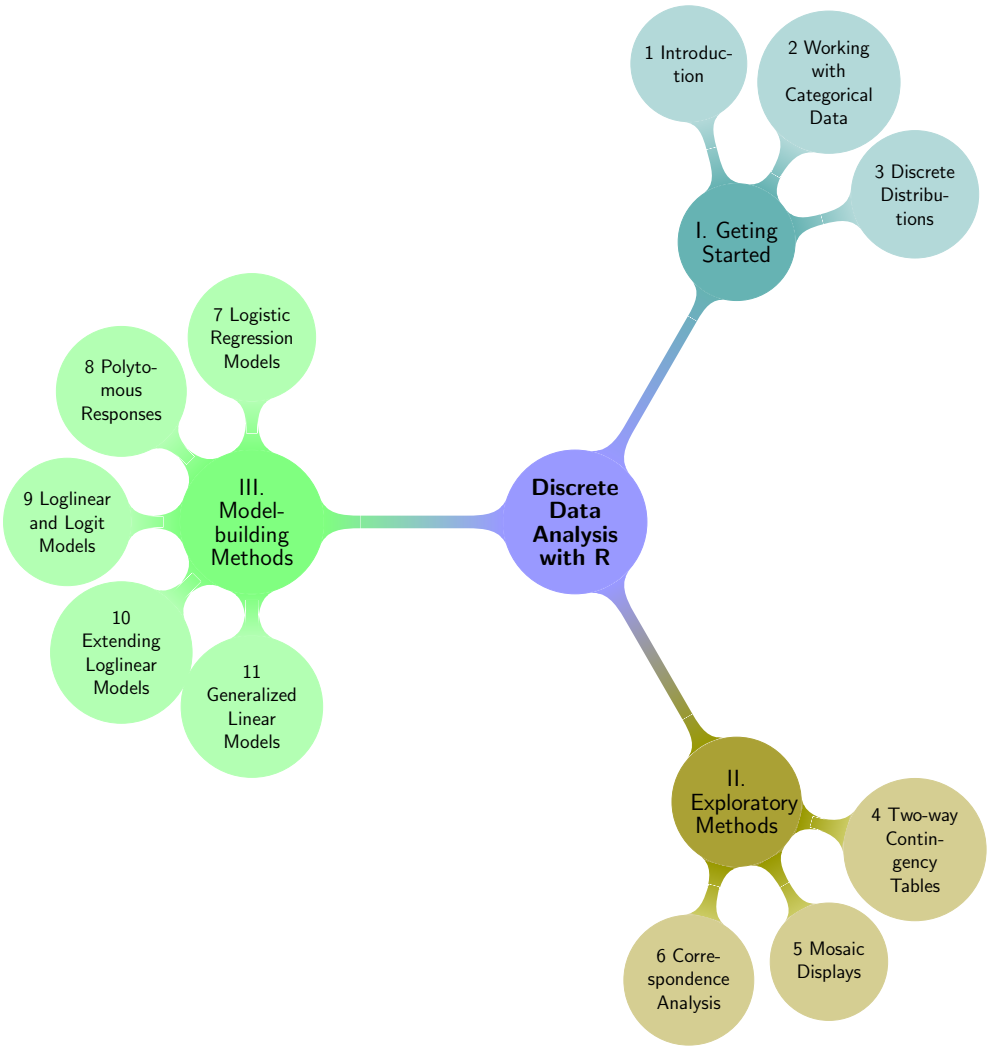
Discrete Data Analysis with R

Visualization and Modeling Techniques for Categorical and Count Data

Michael Friendly
York University

David Meyer
UAS Technikum Wien

with contributions by
Achim Zeilleis
Universität Insbruck



Contents

Preface

> The greatest value of a picture is when it forces us to notice what we never expected to see.
>
> Tukey (1977, p. vi)

Data analysis and graphics

This book stems from the conviction that data analysis and statistical graphics should go hand-in-hand in the process of understanding and communicating statistical data. Statistical summaries compress a data set into a few numbers, the result of an hypothesis test, or coefficients in a fitted statistical model, while graphical methods help to explore patterns and trends, see the unexpected, identify problems in an analysis and communicate results and conclusions in principled and effective ways.

This interplay between analysis and visualization has long been a part of statistical practice for *quantitative data*. Indeed, the origin of correlation, regression and linear models (regression, ANOVA) can arguably be traced to Francis Galton's (1886) visual insight from a scatterplot of heights of children and their parents on which he overlaid smoothed contour curves of roughly equal bivariate frequencies and lines for the means of $Y \mid X$ and $X \mid Y$ (described in Friendly and Denis (2005), Friendly *et al.* (2013)).

The analysis of discrete data is a much more recent arrival, beginning in the 1960s and giving rise to a few seminal books in the 1970s (Bishop *et al.*, 1975, Haberman, 1974, Goodman, 1978, Fienberg, 1980). Agresti (2013, Chapter 17) presents a brief historical overview of the development of these methods from their early roots around the beginning of the 20^{th} Century.

Yet curiously, associated graphical methods for categorical data were much slower to develop. This began to change as it was recognized that counts, frequencies and discrete variables required different schemes for mapping numbers into useful visual representations (Friendly, 1995, 1997), some quite novel. The special nature of discrete variables and frequency data vis-a-vis statistical graphics is now more widely accepted, and many of these new graphical methods (e.g., mosaic displays, fourfold plots, diagnostic plots for generalized linear models) have become, if not main stream, then at least more widely used in research, teaching and communication.

Much of what had been developed through the 1990s for graphical methods for discrete data was described in the book *Visualizing Categorical Data* (Friendly, 2000) and was implemented in SAS® software. Since that time, there has been considerable growth in both statistical methods for the analysis of categorical data (e.g., generalized linear models, zero-inflation models, mixed models for hierarchical and longitudinal data with discrete outcomes), along with some new graphical methods for visualizing and interpreting the results (3D mosaic plots, effect plots, diagnostic plots, etc.) The bulk of these developments have been implemented in R, and the time is right for an in-depth treatment of modern graphical methods for the analysis of categorical data, to which you are now invited.

Goals

This book aims to provide an applied, practically-oriented treatment of modern methods for the analysis of categorical data—discrete response data and frequency data—with a main focus on graphical methods for exploring data, spotting unusual features, visualizing fitted models and presenting or explaining results.

We describe the necessary statistical theory (sometimes in abbreviated form) and illustrate the practical application of these techniques to a large number of substantive problems: how to organize the data, conduct an analysis, produce informative graphs, and understand what they have to say about the data at hand.

Overview and organization of this book

This book is divided into three parts. Part I, Chapters 1–3, contains introductory material on graphical methods for discrete data, basic R skills needed for the book and methods for fitting and visualizing one-way discrete distributions.

Part II, Chapters 4–6, is concerned largely with simple, traditional non-parametric tests and exploratory methods for visualizing patterns of association in two-way and larger frequency tables. Some of the discussion here introduces ideas and notation for loglinear models that are treated more generally in Part III.

Part III, Chapters 7–9, discusses model-based methods for the analysis of discrete data. These are all examples of generalized linear models. However, for our purposes, it has proved more convenient to develp this topic from the specific cases (logistic regression, loglinear models) to the general rather than the reverse.

Chapter ??: ***Introduction*****.** Categorical data require different statistical and graphical methods than commonly used for quantitative data. This chapter outlines the basic orientation of the book toward visualization methods and some key distinctions regarding the analysis and visualization of categorical data.

Chapter ??: ***Working with Categorical Data*****.** Categorical data can be represented in various forms: case form, frequency form, and table form. This chapter describes and illustrates the skills and techniques in R needed to input, create and manipulate R data objects to represent categorical data, and convert these from one form to another for the purposes of statistical analysis and visualization which are the subject of the remainder of the book.

Chapter ??: ***Fitting and Graphing Discrete Distributions*****.** Understanding and visualizing discrete data distributions provides a building block for model-based methods discussed in Part III. This chapter introduces the well-known discrete distributions— the binomial, Poisson, negative-binomial and others— in the simplest case of a one-way frequency table.

Chapter ??: ***Two-way Contingency Tables*****.** The analysis of two-way frequency tables concerns the association between two variables. A variety of specialized graphical displays help to visualize the pattern of association, using area of some region to represent the frequency in a cell. Some of these methods are focused on visualizing an odds ratio (for 2×2 tables), or the general pattern of association, or the agreement between row and column categories in square tables.

Chapter ??: ***Mosaic Displays for* n*-way Tables*****.** This chapter introduces mosaic displays, designed to help to visualize the pattern of associations among variables in two-way and larger tables. Extensions of this technique can reveal partial associations, marginal associations, and shed light on the structure of loglinear models themselves.

Chapter ??: *Correspondence Analysis*. Correspondence analysis provides visualizations of associations in a two-way contingency table in a small number of dimensions. Multiple correspondence analysis extends this technique to n-way tables. Other graphical methods, including mosaic matrices and biplots provide complementary views of loglinear models for two-way and n-way contingency tables.

Chapter ??: *Logistic Regression Models*. This chapter introduces the modeling framework for categorical data in the simple situation where we have a categorical response variable, often binary, and one or more explanatory variables. A fitted model provides both statistical inference and prediction, accompanied by measures of uncertainty. Data visualization methods for discrete response data must often rely on smoothing techniques, including both direct, non-parametric smoothing and the implicit smoothing that results from a fitted parametric model. Diagnostic plots help us to detect influential observations which may distort our results.

Chapter ??: *Loglinear and Logit Models for Contingency Tables*. This chapter extends the model-building approach to loglinear and logit models. These comprise another special case of generalized linear models designed for contingency tables of frequencies. They are most easily interpreted through visualizations, including mosaic displays and effect plots of associated logit models. Special cases arise for ordered categorial variables and square tables that allow more parsimonious models for associations.

Chapter ??: *Generalized Linear Models*. Generalized linear models extend the familiar linear models of regression and ANOVA to include counted data, frequencies, and other data for which the assumptions of independent, normal errors are not reasonable. We rely on the analogies between ordinary and generalized linear models (GLMs) to develop visualization methods to explore the data, display the fitted relationships and check model assumptions. The main focus of this chapter is on models for count data.

Audience

This book has been written to appeal to two broad audiences wishing to learn to apply methods for discrete data analysis:

- Advanced undergraduate, graduate students in the social and health sciences, epidemiology, economics, business and (bio)statistics
- Substantive researchers, methodologists and consultants in various disciplines wanting to be able to use these methods with their own data and analyses.

It assumes the reader has a basic understanding of statistical concepts at least at an intermediate undergraduate level including regression and analysis of variance (for example, at the level of Neter *et al.* (1990), Mendenhall and Sincich (2003)). It is less technically demanding than other modern texts covering categorical data analysis at a graduate level, such as Agresti (2013), *Categorical Data Analysis*, Powers and Xie (2008), *Statistical Methods for Categorical Data Analysis*, and Christensen (1997), *Log-Linear Models and Logistic Regression*. Nevertheless, there are some topics that are a bit more advanced or technical, and these are marked as * or ** sections.

As well, there are also a number of mathematical or statistical topics that we use in passing, but do not describe in these pages (some matrix notation, basic probability theory, maximum likelihood estimation, etc.). Most of these are described in Fox (2015), available online and which serves well as a supplement to this book.

In addition, it is not possible to include *all* details of using R effectively for data analysis. It is assumed that the reader has at least basic knowledge of the R language and environment, including interacting with the R console (RGui for Windows, R.app for Mac OS X) or other graphical user

interface (e.g., RStudio), using R functions in packages, getting help for these from R, etc. One introductory chapter (Chapter **??**) is devoted to covering the particular topics most important to categorical data analysis, beyond such basic skills needed in the book.

Textbook use

This book is most directly suitable for one-semester applied advanced undergraduate or graduate course on categorical data analysis with a strong emphasis on the use of graphical methods to understand and explain data and results of analysis. A detailed outline of such a course, together with lecture notes and assignments, is available at the first author's web page, `http://euclid.psych.yorku.ca/www/psy6136/`, using this book as the main text. This course also uses Agresti (2007), *An Introduction to Categorical Data Analysis* for additional readings.

For instructors teaching a more traditional course using one of the books mentioned above as the main text, this book would be a welcomed supplement, because almost all other texts treat graphical methods only perfunctorily, if at all. A few of these contain a brief appendix mentioning software, or have a related web site with some data sets and software examples. Moreover, none actually describe how to do these analyses and graphics with R.

Features

- Provides an accessible introduction to the major methods of categorical data analysis for data exploration, statistical testing and statistical models.
- The emphasis throughout is on computing, visualizing, understanding and communicating the results of these analyses.
- As opposed to more theoretical books, the goal here is to help the reader to translate theory into practical application, by providing skills and software tools for carrying out these methods.
- Includes many examples using real data, often treated from several perspectives.
- The book is supported directly by R packages vcd and vcdExtra, along with numerous other R packages.
- All materials (data sets, R code) will be available online on the web site for the book.
- Each chapter contains a collection of lab exercises, which work through applications of some of the methods presented in that chapter. This makes the book more suitable for both self-study and classroom use.

Acknowledgements

We are grateful to many colleagues, friends, students and internet acquaintences who have contributed to to this book, directly or indirectly.

We thank those who read and commented on various drafts of the book or chapters. In particular, John Fox, Michael Greenacre and several anonymous reviewers gave insightful comments on the organization of the book and made many helpful suggestions. Matthew Sigal used his wizardly skills to turn sketches of conceptual diagrams into final figures.

At a technical level, we were aided by the cooperation of a number of R package authors, who helped to enhance the graphic displays: Achim Zeilleis who served as a guiding hand in the development of the vcd and vcdExtra packages; John Fox and Sandy Weisberg for enhancements to the car and effects packages; Milan Bouchet-Valat for incorporating suggestions dealing with plotting `rc()` solutions into the logmult package; Michael Greenacre and Oleg Nenadic for help to enhance plotting in the ca package; Heather Turner for advice and help with plotting models fit using the gnm package; Jay Emerson for improvements to the gpairs package.

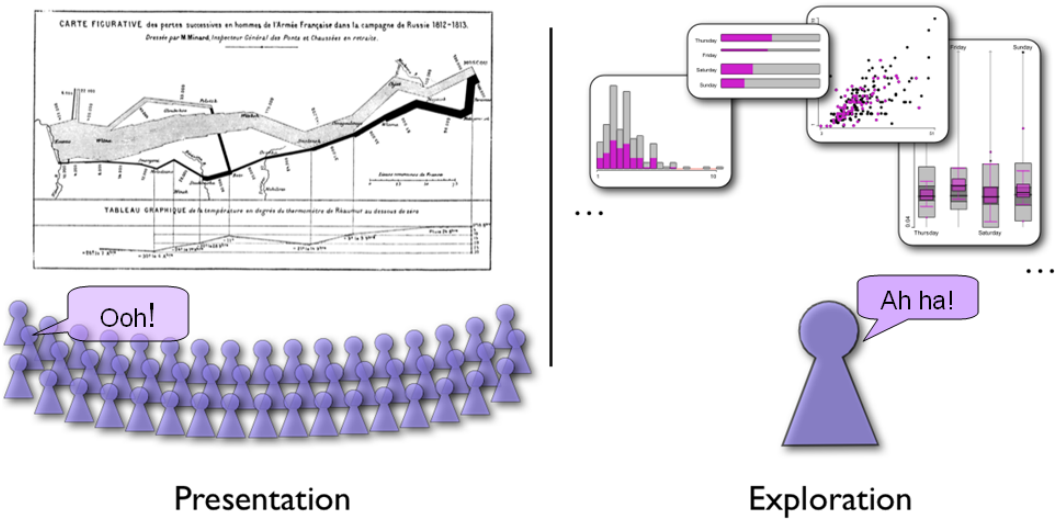
There were also many contributors from the R-Help email list, too many to name them all. Special thanks for generous assistance go to David Carlson, William Dunlap, Bert Gunter, Jim Lemon, Duncan Murdoch, Denis Murphy, Jeff Newmiller, Richard Heiberger, Thierry Onkelinx, Marc Schwartz, David Winsemius, and Ista Zahn.

The book was written using the knitr package, allowing a relatively seamless integration of LaTeX text, R code, and R output and graphs, so that any changes in the code were automatically incorporated in the book. Thanks are due to Yihui Xie and all the contributors to the `knitr` project for making this possible. We are also grateful to Phil Chalmers and Derek Harnanansingh for assistance in using GitHub to manage our collaboration.

The first author's work on this project was supported by grants from the National Science and Engineering Research Council of Canada (Grant 8150) and a sabbatical leave from York University in 2013–14, during which most of this book was written.

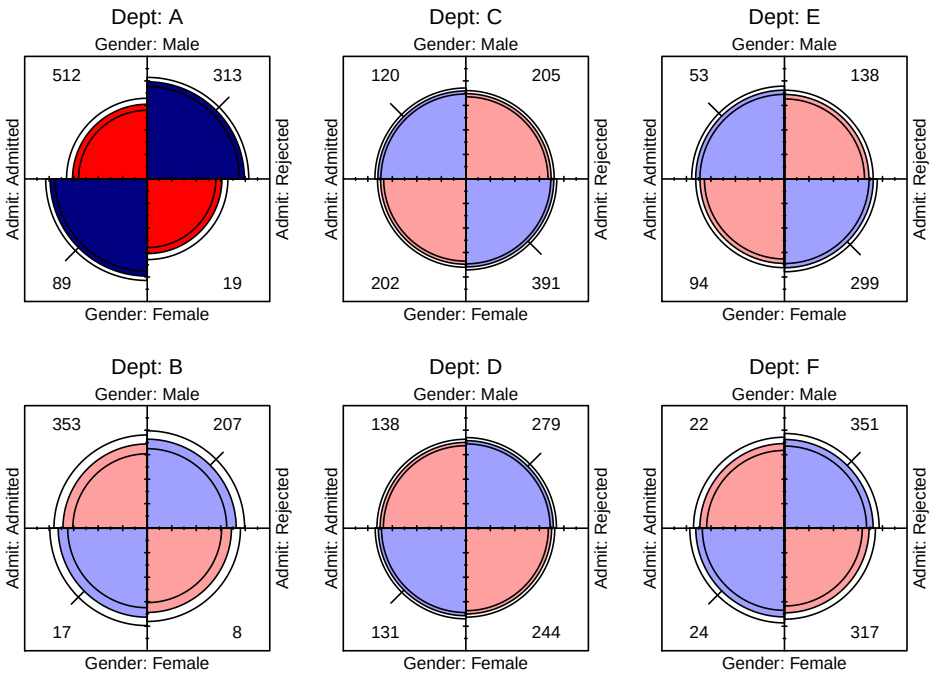
Part I

Introduction



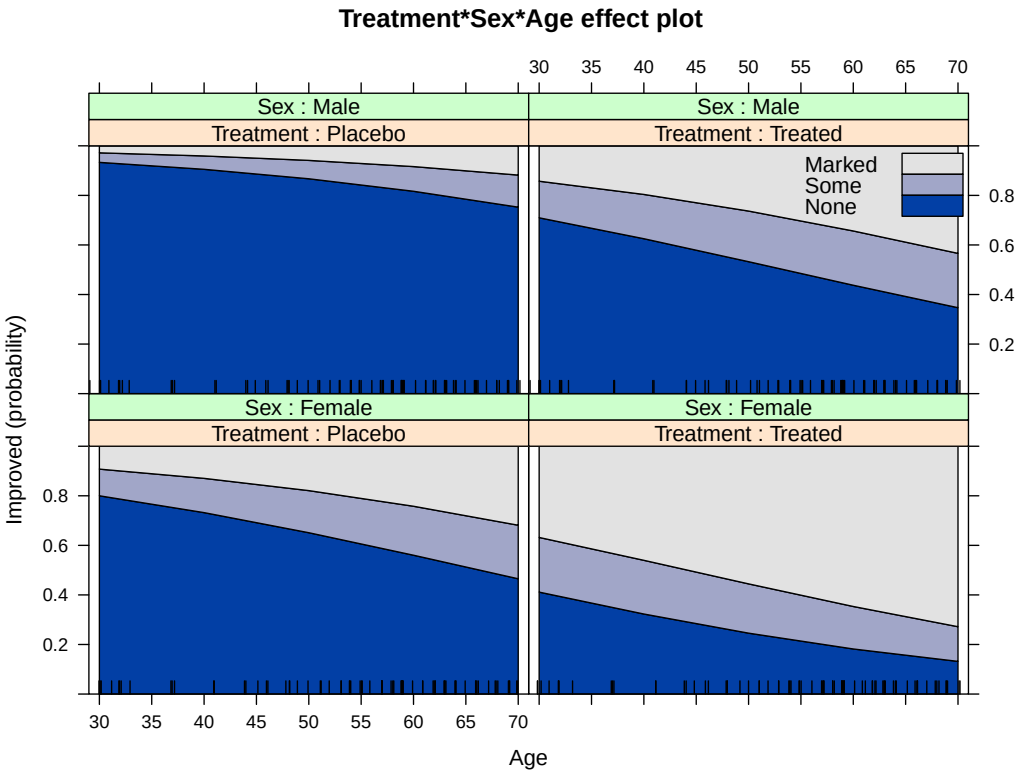
Part II

Exploratory and Hypothesis-testing Methods



Part III

Model-building Methods



References

Agresti, A. (2007). *An Introduction to Categorical Data Analysis*. New York: Wiley, 2nd edn.

Agresti, A. (2013). *Categorical Data Analysis*. Wiley Series in Probability and Statistics. New York: Wiley-Interscience [John Wiley & Sons], 3rd edn.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press.

Christensen, R. (1997). *Log-Linear Models and Logistic Regression*. New York, NY: Springer, 2nd edn.

Fienberg, S. E. (1980). *The Analysis of Cross-Classified Categorical Data*. Cambridge, MA: MIT Press, 2nd edn.

Fox, J. (2015). Appendices to *Applied Regression Analysis, Generalized Linear Models, and Related Methods*, third edition. Online document. Available at `http://socserv.socsci.mcmaster.ca/jfox/Books/Applied-Regression-3E/Appendices.pdf`.

Friendly, M. (1995). Conceptual and visual models for categorical data. *The American Statistician*, 49, 153–160.

Friendly, M. (1997). Conceptual models for visualizing contingency table data. In M. Greenacre and J. Blasius, eds., *Visualization of Categorical Data*, chap. 2, (pp. 17–35). San Diego, CA: Academic Press.

Friendly, M. (2000). *Visualizing Categorical Data*. Cary, NC: SAS Institute.

Friendly, M. and Denis, D. (2005). The early origins and development of the scatterplot. *Journal of the History of the Behavioral Sciences*, 41(2), 103–130.

Friendly, M., Monette, G., and Fox, J. (2013). Elliptical insights: Understanding statistical methods through elliptical geometry. *Statistical Science*, 28(1), 1–39.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute*, 15, 246–263.

Goodman, L. A. (1978). *Analyzing Qualitative Categorical Data: Log-Linear Models and Latent-Strucutre Analysis*. Cambridge, MA: Abt Books.

Haberman, S. J. (1974). *The Analysis of Frequency Data*. Chicago: University of Chicago Press.

Mendenhall, W. and Sincich, T. (2003). *A Second Course in Statistics: Regression Analysis*. Prentice Hall / Pearson Education.

Neter, J., Wasserman, W., and Kutner, M. H. (1990). *Applied Linear Statistical Models : Regression, Analysis of Variance, and Experimental Designs*. Homewood, IL: R. D. Irwin, Inc., 3rd edn.

Powers, D. A. and Xie, Y. (2008). *Statistical Methods for Categorical Data Analysis*. Bingley, UK: Emerald, 2nd edn.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison Wesley.

Author Index

Subject Index

This document was produced using:

```
print(sessionInfo(), locale = FALSE)

## R version 3.1.1 (2014-07-10)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] knitr_1.9
##
## loaded via a namespace (and not attached):
##  [1] colorspace_1.2-5 evaluate_0.5.5   formatR_1.0      grid_3.1.1
##  [5] highr_0.4        lattice_0.20-30  lmtest_0.9-33    MASS_7.3-39
##  [9] stringr_0.6.2    tcltk_3.1.1      tools_3.1.1      vcd_1.3-3
## [13] zoo_1.7-11
```