Modeling Count Data

This definitive entry-level text, authored by a leading statistician in the field, offers clear and concise guidelines on how to select, construct, interpret, and evaluate count data. Written for researchers with little or no background in advanced statistics, the book presents treatments of all major models, using numerous tables, insets, and detailed modeling suggestions. It begins by demonstrating the fundamentals of modeling count data, including a thorough presentation of the Poisson model. It then works up to an analysis of the problem of overdispersion and of the negative binomial model, and finally to the many variations that can be made to the base count models. Examples in Stata, R, and SAS code enable readers to adapt models for their own purposes, making the text an ideal resource for researchers working in health, ecology, econometrics, transportation, and other fields.

Joseph M. Hilbe is a solar system ambassador with NASA's Jet Propulsion Laboratory, California Institute of Technology; an adjunct professor of statistics at Arizona State University; an emeritus professor at the University of Hawaii; and an instructor for Statistics.com, a web-based continuing-education program in statistics. He is currently president of the International Astrostatistics Association, and he is an elected Fellow of the American Statistical Association, for which he is the current chair of the section on Statistics in Sports. Author of several leading texts on statistical modeling, Hilbe also serves as the coordinating editor for the Cambridge University Press series Predictive Analytics in Action.

Other Statistics Books by Joseph M. Hilbe

Generalized Linear Models and Extensions (2001, 2007, 2013 – with J. Hardin)

Generalized Estimating Equations (2002, 2013 – with J. Hardin)

Negative Binomial Regression (2007, 2011)

Logistic Regression Models (2009)

Solutions Manual for Logistic Regression Models (2009)

R for Stata Users (2010 – with R. Muenchen)

Methods of Statistical Model Estimation (2013 – with A. Robinson)

A Beginner's Guide to GLM and GLMM with R: A Frequentist and Bayesian Perspective for Ecologists (2013 – with A. Zuur and E. Ieno)

Quasi–Least Squares Regression (2014 – with J. Shults)

Practical Predictive Analytics and Decisioning Systems for Medicine (2014 – with L. Miner, P. Bolding, M. Goldstein, T. Hill, R. Nisbit, N. Walton, and G. Miner)

Modeling Count Data

Joseph M. Hilbe
Arizona State University
and
Jet Propulsion Laboratory,
California Institute of Technology

Cambridge
University Press

CAMBRIDGE
UNIVERSITY PRESS

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

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/9781107611252

First published 2014

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

ISBN 978-1-107-02833-3 Hardback
ISBN 978-1-107-61125-2 Paperback

Additional resources for this publication at www.cambridge.org/9781107611252

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

Cambridge University Press
978-1-107-02833-3 — Modeling Count Data
Joseph M. Hilbe
Frontmatter
More Information

Preface

Modeling Count Data is written for the practicing researcher who has a reason to analyze and draw sound conclusions from modeling count data. More specifically, it is written for an analyst who needs to construct a count response model but is not sure how to proceed.

A count response model is a statistical model for which the dependent, or response, variable is a count. A count is understood as a nonnegative discrete integer ranging from zero to some specified greater number. This book aims to be a clear and understandable guide to the following points:

- How to recognize the characteristics of count data
- Understanding the assumptions on which a count model is based
- Determining whether data violate these assumptions (e.g., overdispersion), why this is so, and what can be done about it
- Selecting the most appropriate model for the data to be analyzed
- Constructing a well-fitted model
- Interpreting model parameters and associated statistics
- Predicting counts, rate ratios, and probabilities based on a model
- Evaluating the goodness-of-fit for each model discussed

There is indeed a lot to consider when selecting the best-fitted model for your data. I will do my best in these pages to clarify the foremost concepts and problems unique to modeling counts. If you follow along carefully, you should have a good overview of the subject and a basic working knowledge needed for constructing an appropriate model for your study data. I focus on understanding the nature of the most commonly used count models and on the problem of dealing with both over- and underdispersion, as well as on Poisson and negative binomial regression and their many variations. However, I also introduce several other count models that have not had much use in research because of the unavailability of commercial software for their estimation. In particular, I also discuss models such as the Poisson inverse Gaussian, generalized Poisson, varieties of three-parameter negative binomial, exact Poisson, and several other count models that will provide analysts with an expanded ability to better model the data at hand. Stata and/or R software and guidelines are provided for all of the models discussed in the text.

I am supposing that most people who will use this book start with little to no background in modeling count response data, although readers are expected to have a working knowledge of a major statistical software package, as well as a basic understanding of statistical regression. I provide an overview of maximum likelihood and iterative reweighted least squares (IRLS) regression in Sections 1.4.2 and 1.4.3, which assume an elementary understanding of calculus, but I consider these two sections as optional to our discussion. They are provided for those who are interested in how the majority of models we discuss are estimated. I recommend that you read these sections, even if you do not have the requisite mathematical background. I have attempted to present the material so that it will still be understood. Various terms are explained in these sections that will be used throughout the text.

Seasoned statisticians can also learn new material from the text, but I have specifically written it for researchers or analysts, as well as students at the upper-division to graduate levels, who want an entry-level book that focuses on the practical aspects of count modeling. The book is also addressed to statistical and predictive analytics consultants who find themselves faced with a project involving the modeling of count data, as well as to anyone with an interest in this class of statistical models. It is written in guidebook form, with lots of bullet points, tables, and complete statistical programming code for all examples discussed in the book.

Many readers of this book may be acquainted with my text *Negative Binomial Regression* (Cambridge University Press), which was first published in 2007. A substantially enhanced second edition was published in 2011. That text addresses nearly every count model for which there existed major statistical software support at the time of the book's publication. *Negative Binomial Regression* was primarily written for those who wish to understand the mathematics behind the models as well as the specifics and applications of each model. I recommend it for those who wish to go beyond the discussions found in *Modeling Count Data.*

I primarily use two statistical software packages to demonstrate examples of the count models discussed in the book. First, the Stata 13 statistical package (http://www.stata.com) is used throughout the text to display example model output. I show both Stata code and output for most of the modeling examples. I also provide R code (www.r-project.org) in the text that replicates, as far as possible, the Stata output. R output is also given when helpful. There are also times when no current Stata code exists for the modeling of a particular procedure. In such cases, R is used. SAS code for a number of the models discussed in the book is provided in the Appendix. SAS does not support many of the statistical functions and tests discussed later in the book, but its count-modeling capability is growing each year. I will advise readers on the book's web site as software for count models is developed for these packages. I should mention that I have used Stat/Transfer 12 (2013, Circle Systems) when converting data between statistical software packages. The user is able to convert between 37 different file formats, including those used in this book. It is a very helpful tool for those who must use more than one statistical or spreadsheet file.

Many of the Stata statistical models discussed in the text are offered as a standard part of the commercial package. Users have also contributed count model "commands" for the use of the greater Stata community. Developers of the user-authored commands used in the book are acknowledged at the first use of the software. James Hardin and I have both authored and coauthored a number of the more advanced count models found in the book. Many derive from our 2012 text *Generalized Linear Models and Extensions, 3rd edition* (Stata Press; Chapman & Hall/CRC). Several others in the book are based on commands we developed in 2013 for journal article publications. I should also mention that we also coauthored the current version of Stata's **`glm`** command (2001), although Stata has subsequently enhanced various options over the past 12 years as new versions of Stata were released. Several of the R functions and scripts used in the book were coauthored by Andrew Robinson and me for use in our book (Hilbe and Robinson 2013). Data sets and functions for this book, as well as for Hilbe (2011), are available in the **`COUNT`** package, which may be downloaded from any **`CRAN`** mirror site. I also recommend installing **`msme`** (Hilbe and Robinson), also available on **`CRAN`**. I have also posted all of my user-authored Stata commands and functions, as well as all data sets used in the book, on the book's web site at the following address: http://works.bepress.com/joseph_hilbe/. The book's page with Cambridge University Press is at www.cambridge.org/9781107611252.

The data files used for examples in the book are real data. The **rwm1984**and **medpar** data sets are used extensively throughout the book. Other data sets used include **titanic**, **heart**, **azcabgptca**, **smoking**, **fishing**, **fasttrakg**, **rwm5yr**, **nuts**, and **azprocedure**. The data are defined where first used. The **medpar**, **rwm5yr**, and **titanic** data are used more than other data in the book. The **medpar** data are from the 1991 Arizona Medicare files for two diagnostic groups related to cardiovascular procedures. I prepared **medpar** in 1993 for use in workshops I gave at the time. The **rwm5yr** data consist of 19,609 observations from the German Health Reform data covering the five-year period of 1984–1988. Not all patients were in the study for all five years. The count response is the number of visits made by a patient to the doctor during that calendar year. The **rwm1984** data were created from **rwm5yr**, with only data from 1984 included – one patient, one observation. The well-known **titanic** data set is from the 1912 *Titanic* ship disaster survival data. It is in grouped format with *survived* as the response. The predictors are *age* (adult vs. child), *gender* (male vs. female), and *class* (1st-, 2nd-, and 3rd-class passengers). Crew members have been excluded.

I advise the reader that there are parts of Chapter 3 that use or adapt text from the first edition of *Negative Binomial Regression* (Hilbe 2007a), which is now out of print, as it was superseded by Hilbe (2011). Chapter 2 incorporates two tables that were also used in the first edition. I received very good feedback regarding these sections and found no reason to change them for this book. Now that the original book is out of print, these sections would be otherwise lost.

I wish to acknowledge five eminent colleagues and friends in the truest sense who in various ways have substantially contributed to this book, either indirectly while working together on other projects or directly: James Hardin, director of the Biostatistics Collaborative Unit and professor, Department of Statistics and Epidemiology, University of South Carolina School of Medicine; Andrew Robinson, director, Australian Centre of Excellence for Risk Analysis (ACERA), Department of Mathematics and Statistics, University of Melbourne, Australia; Alain Zuur, senior statistician and director of Highland Statistics Ltd., UK; Peter Bruce, CEO, Institute for Statistics Education (Statistics.com); and John Nelder, late Emeritus Professor of Statistics, Imperial College, UK. John passed away in 2010, just shy of his eighty-sixth birthday; our many discussions over a 20-year period are sorely missed. He definitely spurred my interest in the negative binomial model. I am fortunate to have known and to have worked with these fine statisticians. Each has enriched my life in different ways.

Others who have contributed to this book's creation include Valerie Troiano and Kuber Dekar of the Institute for Statistics Education; Professor William H. Greene, Department of Economics, New York University, and author of the Limdep econometrics software; Dr. Gordon Johnston, Senior Statistician, SAS Institute, author of the SAS Genmod Procedure; Professor Milan Hejtmanek, Seoul National University, and Dr. Digant Gupta, M.D., director, Outcomes Research, Cancer Treatment Centers of America, both of whom provided long hours reviewing early drafts of the book manuscript. Helen Wheeler, production editor for Cambridge University Press, is also gratefully acknowledged. A special acknowledgment goes to Patricia Branton of Stata Corp., who has provided me with statistical support and friendship for almost a quarter of a century. She has been a part of nearly every text I have written on statistical modeling, including this book.

There have been many others who have contributed to this book as well, but space limits their express acknowledgment. I intend to list all contributors on the book's web site. I invite readers to contact me regarding comments or suggestions about the book. You may email me at hilbe@asu.edu or at the address on my BePress web site listed earlier.

Finally, I must also acknowledge Diana Gillooly, senior editor for mathematical sciences with Cambridge University Press, who first encouraged me to write this monograph. She has provided me with excellent feedback in my attempt to develop a thoroughly applied book on count models. Her help with this book has been invaluable and goes far beyond standard editorial obligations. I also wish to thank my family for yet again supporting my writing of another book. My appreciation goes to my wife, Cheryl L. Hilbe, my children and grandchildren, and our white Maltese dog, Sirr, who sits close by my side for hours while I am typing. I dedicate this book to Cheryl for her support and feedback during the time of this book's preparation.

Joseph M. Hilbe
Florence, Arizona
August 12, 2013