Negative Binomial Regression

Second Edition

This second edition of *Negative Binomial Regression* provides a comprehensive discussion of count models and the problem of overdispersion, focusing attention on the many varieties of negative binomal regression. A substantial enhancement from the first edition, the text provides the theoretical background as well as fully worked out examples using Stata and R for most every model having commercial and R software support. Examples using SAS and LIMDEP are given as well. This new edition is an ideal handbook for any researcher needing advice on the selection, construction, interpretation, and comparative evaluation of count models in general, and of negative binomial models in particular.

Following an overview of the nature of risk and risk ratio and the nature of the estimating algorithms used in the modeling of count data, the book provides an exhaustive analysis of the basic Poisson model, followed by a thorough analysis of the meanings and scope of overdispersion. Simulations and real data using both Stata and R are provided throughout the text in order to clarify the essentials of the models being discussed. The negative binomial distribution and its various parameterizations and models are then examined with the aim of explaining how each type of model addresses extra-dispersion. New to this edition are chapters on dealing with endogeny and latent class models, finite mixture and quantile count models, and a full chapter on Bayesian negative binomial models. This new edition is clearly the most comprehensive applied text on count models available.

JOSEPH M. HILBE is a Solar System Ambassador with NASA's Jet Propulsion Laboratory at the California Institute of Technology, an Adjunct Professor of statistics at Arizona State University, and an Emeritus Professor at the University of Hawaii. Professor Hilbe is an elected Fellow of the American Statistical Association and elected Member of the International Statistical Institute (ISI), for which he is the founding Chair of the ISI astrostatistics committee and Network. He is the author of *Logistic Regression Models*, a leading text on the subject, co-author of *R for Stata Users* (with R. Muenchen), and of both *Generalized Estimating Equations* and *Generalized Linear Models and Extensions* (with J. Hardin).

Negative Binomial Regression

Second Edition

JOSEPH M. HILBE

Jet Propulsion Laboratory,
California Institute of Technology and
Arizona State University

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town,
Singapore, São Paulo, Delhi, Mexico City

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521198158

First published 2007
Reprinted with corrections 2008
Second edition 2011
Reprinted with corrections 2012

Printed in the United Kingdom at the University Press, Cambridge

A catalogue record for this publication is available from the British Library

Library of Congress Cataloguing in Publication data
Hilbe, Joseph.
Negative binomial regression / Joseph M. Hilbe. – 2nd ed.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-521-19815-8 (hardback)
1. Negative binomial distribution. 2. Poisson algebras. I. Title.
QA161.B5H55 2011
519.2′4 – dc22 2010051121

ISBN 978-0521-19815-8 Hardback

Additional resources for this publication at http://works.bepress.com/joseph_hilbe/

Contents

Preface to the second edition

The aim of this book is to present a detailed, but thoroughly clear and understandable, analysis of the nature and scope of the varieties of negative binomial model that are currently available for use in research. Modeling count data using the standard negative binomial model, termed NB2, has recently become a foremost method of analyzing count response models, yet relatively few researchers or applied statisticians are familiar with the varieties of available negative binomial models, or how best to incorporate them into a research plan.

Note that the Poisson regression model, traditionally considered as the basic count model, is in fact an instance of NB2 – it is an NB2 with a heterogeneity parameter of value 0. We shall discuss the implications of this in the book, as well as other negative binomial models that differ from the NB2. Since Poisson is a variety of the NB2 negative binomial, we may regard the latter as more general and perhaps as even more representative of the majority of count models used in everyday research.

I began writing this second edition of the text in mid-2009, some two years after the first edition of the text was published. Most of the first edition was authored in 2006. In just this short time – from 2006 to 2009/2010 – a number of advancements have been made to the modeling of count data. The advances, however, have not been as much in terms of new theoretical developments, as in the availability of statistical software related to the modeling of counts. Stata commands have now become available for modeling finite mixture models, quantile count models, and a variety of models to accommodate endogenous predictors, e.g. selection models and generalized method of moments. These commands were all authored by users, but, owing to the nature of Stata, the commands can be regarded as part of the *Stata* repertoire of capabilities.

R has substantially expanded its range of count models since 2006 with many new functions added to its resources; e.g. zero-inflated models, truncated, censored, and hurdle models, finite-mixture models, and bivariate count models, etc. Moreover, R functions now exist that allow non-parametric features to be added to the count models being estimated. These can assist in further adjusting for overdispersion identified in the data.

SAS has also enhanced its count modeling capabilities. SAS now provides the ability of estimating zero-inflated count models as well as the NB1 parameterization of the negative binomial. Several macros exist that provide even more modeling opportunities, but at this time they are still under construction. When the first edition of this book was written, only the Poisson and two links of the negative binomial as found in SAS/STAT **GENMOD** were available in SAS.

SPSS has added the **Genlin** procedure to its functions. **Genlin** is the SPSS equivalent of Stata's **glm** command and SAS's **GENMOD** procedure. **Genlin** provides the now standard GLM count models of Poisson and three parameterizations of negative binomial: log, identity, and canonical linked models. At this writing, SPSS supports no other count models. LIMDEP, perhaps one of the most well-respected econometric applications, has more capabilities for modeling count data than the others mentioned above. However, there are models we discuss in this text that are unavailable in LIMDEP.

This new edition is aimed to update the reader with a presentation of these new advances and to address other issues and methodologies regarding the modeling of negative binomial count data that were not discussed in the first edition. The book has been written for the practicing applied statistician who needs to use one or more of these models in their research. The book seeks to explain the types of count models that are appropriate for given data situations, and to help guide the researcher in constructing, interpreting, and fitting the various models. Understanding model assumptions and how to adjust for their violations is a key theme throughout the text.

In the first edition I gave Stata examples for nearly every model discussed in the text. LIMDEP was used for examples discussing sample selection and mixed-effects negative binomial models. Stata code was displayed to allow readers to replicate the many examples used throughout the text. In this second edition I do the same, but also add R code that can be used to emulate Stata output. Nearly all output is from Stata, in part because Stata output is nicely presented and compact. Stata commands also generally come with a variety of post-estimation commands that can be used to easily assess fit. We shall discuss these commands in considerable detail as we progress through the book. However, R users can generally replicate Stata output insofar as possible by pasting the source code available on the book's website into the R script editor and running it. Code is also available in tables within the appropriate area of discussion. R output is given for modeling situations where Stata does not have the associated command. Together the two programming languages provide the researcher with the ability to model almost every count model discussed in the literature.

I should perhaps mention that, although this text focuses on understanding and using the wide variety of available negative binomial models, we also address several other count models. We do this for the purpose of clarifying a corresponding or associated negative binomial model. For example, the Poisson model is examined in considerable detail because, as previously mentioned, it is in fact a special case of the negative binomial model. Distributional violations of the Poisson model are what has generally motivated the creation and implementation of other count models, and of negative binomial models in particular. Therefore, a solid understanding of the Poisson model is essential to the understanding of negative binomial models.

I believe that this book will demonstrate that negative binomial models are core to the modeling of count data. Unfortunately they are poorly understood by many researchers and members of the statistical community. A central aim of writing this text is to help remedy this situation, and to provide the reader with both a conceptual understanding of these models, and practical guidance on the use of software for the appropriate modeling of count data.

New subjects discussed in the second edition

In this edition I present an added examination of the nature and meaning of risk and risk ratio, and how they differ from odds and odds ratios. Using 2×2 and $2\times k$ tables, we define and interpret risk, risk ratio, and relative risk, as well as related standard errors and confidence intervals. We provide two forms of coefficient interpretation, and some detail about how they are related. Also emphasized is how to test for model dispersion. We consider at length the meaning of extra-dispersion, including both under- and overdispersion. For example, a model may be both Poisson overdispersed and negative binomial under-dispersed. The nature of overdispersion and how it can be identified and accommodated is central to our discussion. Also of prime importance is an understanding of the consequences that follow when these assumptions are violated.

I also give additional space to the NB-C, or canonical negative binomial. NB-C is unlike all other parameterizations of the negative binomial, but is the only one that directly derives from the negative binomial probability mass function, or PMF. It will be discovered that certain types of data are better modeled using NB-C; guidelines are provided on its applicability in research.

In the first edition I provided the reader with Stata code to create Poisson, NB2 and NB1 synthetic count models. We now examine the nature of these types of synthetic models and describe how they can be used to understand the relationship of models to data. In addition, synthetic models are provided to estimate two-part synthetic hurdle models. This code can be useful as a paradigm for the optimization or maximum likelihood estimation of more complex count models. Also discussed in this edition are marginal effects and discrete change, which have a prominent place in econometrics, but which can also be used with considerable value in other disciplines. I provide details on how to construct and interpret both marginal effects and discrete change for the major models discussed in the text.

New models added to the text from the first edition include: finite mixture models, quantile count models, bivariate negative binomial, and various methods used to model endogeneity. We previously examined negative binomial sample selection models and negative binomial models with endogenous stratification, but we now add instrumental variables, generalized method of moments, and methods of dealing with predictors having missing values. Stata now supports these models.

Finally, Stata 11 appeared in late July 2009, offering several capabilities related to our discussion which were largely unavailable in the version used with the first edition of this text. In particular, Stata's **glm** command now allows maximum likelihood estimation of the negative binomial heterogeneity parameter, which it did not in earlier versions. R's **glm** function and SAS's STAT/**GENMOD** procedure also provide the same capability. This option allows easier estimation and comparative evaluation of NB2 models.

I created the code and provided explanations for the development and use of a variety of synthetic regression models used in the text. Most were originally written using pseudo-random number generators I published in 1995, but a few were developed a decade later. When Stata offered their own suite of pseudo-random number generators in 2009, I re-wrote the code for constructing these synthetic models using Stata's code. This was largely done in the first two months of 2009. The synthetic models appearing in the first edition of this book, which are employed in this edition, now use the new code, as do the other synthetic models. In fact, Stata code for most of the synthetic models appearing in this text was published in the *Stata Journal* (Volume 10:1, pages 104–124). Readers are referred to that source for additional explanation on constructing synthetic models, including many binomial models that are not in this book. It should be noted that the synthetic Stata models discussed in the text have corresponding R scripts provided to duplicate model results.

R data files, functions, and scripts written for this text are available in the COUNT package that can be downloaded on CRAN. I recommend Muenchen and Hilbe (2010), *R for Stata Users* (Springer) for Stata users who wish to understand the logic of R functions. For those who wish to learn more about Stata, and how it can be used for data management and programming, I refer you to the books published by Stata Press. See *www.stata-press.com*. The websites which readers can access to download data files and user-authored commands and functions are:

Cambridge University Press: www.cambridge.org/9780521198158

Data sets, software code, and electronic *Extensions* to the text can be downloaded from: http://works.bepress.com/joseph_hilbe/

Errata and a post-publication *Negative Binomial Regression Extensions* can be found on the above sites, as well as at: http://www.statistics.com/hilbe/nbr.php. Additional code, text, graphs and tables for the book are available in this electronic document.

I should mention that practicing researchers rarely read this type of text from beginning straight through to the end. Rather, they tend to view it as a type of handbook where a model they wish to use can be reviewed for modeling details. The text can indeed be used in such a manner. However, most of the chapters are based on information contained in earlier chapters. Material explained in Chapters 2, 5, 6, and 7 is particularly important for mastery of the discussion of negative binomial models presented in Chapters 8 and 9. Knowledge of the material given in Chapters 7 and 8 is fundamental to understanding extended Poisson and negative binomial models. Because of this, I encourage you to make an exception and to read sequentially though the text, at least through Chapter 9. For those who have no interest at all in the derivation of count models, or of the algorithms by which these models are estimated, I suggest that you only skim through Chapter 4. Analogical to driving a car and understanding automotive mechanics, one can employ a sophisticated negative binomial model without understanding exactly how estimation is achieved. But, again as with a car, if you wish to re-parameterize a model or to construct an entirely new count model, understanding the discussion in Chapter 4 is essential. However, I have reiterated certain important concepts and relationships throughout the text in order to provide background for those who have not been able to read earlier chapters, and to emphasize them for those who may find reminding to be helpful.

To those who made this a better text

I wish to acknowledge several individuals who have contributed specifically to the second edition of *Negative Binomial Regression*. Chris Dorger of Intel kindly read through the manuscript, identifying typos and errors, and re-derived the derivations and equations presented in the text in order to make certain that they are expressed correctly. He primarily checked Chapters 2 through 10. Dr. Dorger provided assistance as the book was just forming, and at the end, prior to submission of the manuscript to the publisher. He spent a substantial amount of time working on this project. Dr. Tad Hogg, a physicist with the Institute of Molecular Manufacturing, spent many hours doing the same, and provided excellent advice regarding the clarification of particular subjects. James Hardin, with whom I have co-authored two texts on statistical modeling and whom I consider a very good friend, read over the manuscript close to the end, providing valuable insight and help, particularly with respect to Chapters 13 and 14. Andrew Robinson of the University of Melbourne provided invaluable help in setting up the COUNT package on CRAN, which has all of the R functions and scripts for the book, as well as data in the form of data frames. He also helped in re-designing some of the functions I developed to run more efficiently. We are now writing a book together on the *Methods of Statistical Model Estimation* (Chapman & Hall/CRC). Gordon Johnston of SAS Institute, author of the GENMOD procedure and longtime friend, provided support on SAS capabilities related to GEE and Bayesian models. He is responsible for developing the SAS programs used for the examples in Section 15.3, and read over Sections 15.1 and 15.2, offering insightful advice and making certain that the concepts expressed were correct. Allan Medwick, who recently earned a Ph.D. in statistics at the University of Pennsylvania read carefully through the entire manuscript near the end, checking for readability and typos; he also helped running the SAS models in Chapter 14. The help of these statisticians in fine-tuning the text has been invaluable, and I deeply appreciate their assistance. I am alone responsible for any remaining errors, but readers should be aware that every attempt has been made to assure accuracy.

I also thank Professor James Albert of Bowling Green University for his insightful assistance in helping develop R code for Monte Carlo simulations. Robert Muenchen, co-author with me of *R for Stata Users* (2010), assisted me in various places to streamline R code, and at times to make sure my code did exactly as I had intended. The efforts of these two individuals also contributed much to the book's value.

Again, as in the first edition, I express my gratitude and friendship to the late John Nelder, with whom I have discussed these topics for some 20 years. In fact,

I first came to the idea of including the negative binomial family in standard generalized linear models software in 1992 while John and I hiked down and back up the Grand Canyon Angel Trail together. We had been discussing his newly developed *kk* add-ons to GenStat software, which included rudimentary negative binomial code, when it became clear to me that the negative binomial, with its heterogeneity parameter as a constant, is a full-fledged member of the single parameter exponential family of distributions, and as such warranted inclusion as a GLM family member. It was easy to then parameterize the negative binomial probability distribution into exponential family form, and abstract the requisite GLM link, mean, variance, and inverse link functions – as any other GLM family member. It was also clear that this was not the traditional negative binomial, but that the latter could be developed by simple transformations. His visits to Arizona over the years, and our long discussions, have led to many of the approaches I now take to GLM–related models, even where a few of my findings have run counter to some of the positions he initially maintained.

I must also acknowledge the expertise of Robert Rigby and Mikis Stasinopoulos, authors of R’s **gamlss** suite of functions, who upon my request rewrote part of the software in such a manner that it can now be used to estimate censored and truncated count models, which were not previously available using R. These capabilities are now part of **gamlss** on CRAN. Also to be acknowledged is Masakazu Iwasaki, Clinical Statistics Group Manager for Research & Development at Schering-Plough K.K., Tokyo, who re-worked code he had earlier developed into a viable R function for estimating bivariate negative binomial regression. He amended it expressely for this text; it is currently the only bivariate negative binomial function of which I am aware, and is available in the COUNT package. Malcolm Faddy and David Smith provided code and direction regarding the implementation of Faddy’s namesake distribution for the development of a new count model for both under- and overdispersion. Robert LaBudde, President of Least Cost Formulations, Ltd., and adjunct professor of statistics at Old Dominion University, provided very useful advice related to R programming, for which I am most grateful. The above statisticians have helped add new capabilities to R, and have made this text a more valuable resource as a result.

The various participants in my **Statistics.com** web course, *Modeling Count Data*, should be recognized. Nearly all of them are professors teaching statistics courses in various disciplines, or researchers working in the corporate world or for a governmental agency. Their many questions and comments have helped shape this second edition. Elizabeth Kelly of the Los Alamos National Laboratories in New Mexico and Kim Dietrich, University of Washington, provided very helpful assistance in reviewing the manuscript at its final stage, catching several items that had escaped others, including myself. They also offered suggestions for clarification at certain points in the discussion, many of which were taken. Denis Shah, Department of Plant Pathology at Kansas State University, in particular, asked many insightful questions whose answers found their way into the text. I also wish to again express my appreciation to Diana Gillooly, statistics editor at Cambridge University Press, for her advice and for her continued confidence in the value of my work. Clare Dennison, assistant editor (Math/Computer science) at Cambridge, is also to be thanked for her fine help with the text, as are Joanna Endell-Cooper (production editor), Mairi Sutherland (freelance editor), Simon Crump (production manager), and Nicola Philps (Production Editor Manufacturing). They very kindly accommodated my many amendments. Ms Sutherland provided excellent editorial advice when required, reviewing every word of the manuscript. I alone am responsible for any remaining errors. In addition, I acknowledge the longtime assistance and friendship of Pat Branton of *Stata* Corp, who has helped me in innumerable ways over the past 20 years. Without her support over this period, this book would very likely not have been written.

Finally, I wish to thank the members of my family, who again had to lose time I would have otherwise spent with them. To my wife, Cheryl, our children, Heather, Michael, and Mitchell, and to my constant companion, Sirr, a white Maltese, I express my deepest appreciation. I dedicate this text to my late parents, Rader John and Nadyne Anderson Hilbe. My father, an engineering supervisor with Douglas Aircraft during the Second World War, and a UCLA mathematics professor during the early 1940s, would have very much appreciated this volume.