Negative Binomial Regression

At last – a book entirely devoted to the negative binomial model and its many variations. Every model currently offered in a commercial statistical software package is discussed in detail – how each is derived, how each resolves a distributional problem, and numerous examples of their application. Many of these models have never before been thoroughly examined in a text on count response models: the canonical negative binomial; the NB-P model, where the negative binomial exponent is itself parameterized; and negative binomial mixed models. Written for practicing researchers and statisticians who need to update their knowledge of Poisson and negative binomial models, the book provides a comprehensive overview of estimating methods and algorithms used to model counts, as well as specific modeling guidelines, model selection techniques, methods of interpretation, and assessment of model goodness of fit. Data sets and modeling code are provided on a companion website.

JOSEPH M. HILBE is an Emeritus Professor at the University of Hawaii and Adjunct Professor of Statistics in the School of Social and Family Dynamics at Arizona State University. He has served as both the Software Reviews Editor and overall Associate Editor for *The American Statistician* since 1997 and is currently on the editorial boards of five academic journals in statistics. An elected Fellow of both the American Statistical Association and Royal Statistical Society, Hilbe is author, with James Hardin, of *Generalized Estimating Equations* and two editions of *Generalized Linear Models and Extensions*.

Negative Binomial Regression

JOSEPH M. HILBE
Arizona State University

CAMBRIDGE
UNIVERSITY PRESS

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

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/9780521857727

First published 2007

Printed in the United Kingdom at the University Press, Cambridge

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

ISBN 978-0-521-85772-7 hardback

Contents

Preface

This is the first text devoted specifically to the negative binomial regression model. Important to researchers desiring to model count response data, the procedure has only recently been added to the capabilities of leading commercial statistical software. However, it is now one of the most common methods used by statisticians to accommodate extra correlation – or overdispersion – when modeling counts. Since most real count data modeling situations appear to involve overdispersion, the negative binomial has been finding increased use among statisticians, econometricians, and researchers who commonly analyze count response data.

This volume will explore both the theory and varieties of the negative binomial. It will also provide the reader with examples using each type of major variation it has undergone. However, of prime importance, the text will also attempt to clarify discrepancies regarding the negative binomial that often appear in the statistical literature. What exactly is a negative binomial model? How does it relate to other models? How is its variance function to be defined? Is it a member of the family of generalized linear models? What is the most appropriate manner by which to estimate parameters? How are parameters to be interpreted, and evaluated as to their worth? What are the limits of its applicability? How has it been extended to form more complex models? These are important questions that have at times found differing answers depending on the author. By examining how the negative binomial model arises from the negative binomial probability mass function, and by considering how major estimating methods relate to the estimation of its parameters, we should be able to clearly define each variety of negative binomial as well as the logic underlying the respective extensions.

The goal of this text is to serve as a handbook of negative binomial regression models, providing the reader with guidelines of how to best implement the model into their research. Although we shall provide the mathematics of how

the varieties of negative binomial model are derived, the emphasis will be on clarity and application. The text has been written to be understandable to anyone having a general background in maximum likelihood theory and generalized linear models. To gain full benefit of the theoretical aspects of the discussion, the reader should also have a working knowledge of elementary calculus.

The Stata statistical package (http://www.stata.com) is used throughout the text to display example model output. Although many of the statistical models discussed in the text are offered as a standard part of the commercial package, I have written a number of more advanced negative binomial models using Stata's proprietary higher programming language. These programs, called *ado* files by Stata, display results that appear identical to official Stata procedures. Some 25 of these Stata programs have been posted to the Boston College School of Economics SSC archive, accessed at: http://ideas.repec.org/s/boc/bocode.html. Programs are ordered by year, with the most recent posted at the bottom of the respective year of submission. Most statistical procedures written for this text can be found in the 2004 files.

LIMDEP software (http://www.limdep.com) is used to display output for examples related to negative binomial mixed models, the NB-P model, negative binomial selection models, and certain types of truncated and censored models. These programs were developed by Prof. William Greene of New York University, author of the LIMDEP package. Stata and LIMDEP statistical software contain more procedures related to negative binomial regression than all other packges combined. I recommend that either of these two packages be obtained if the reader intends to duplicate text examples at their site. A basic NB-2 model in R is provided as part of the MASS package, based in Venables and Ripley (2002). Negative binomial models in R are limited as of this writing, but more advanced models are sure to follow in the near future.

All data sets and Stata ado files related to models used in the text can be downloaded from: www.cambridge.org/XXXXX. Each ado file will have a date of origin associated with it. Occasionally updates or additions will be made to this site; it is recommended that you check it from time to time, updating to the most recent iteration of the procedure of interest. I also intend to post additional materials related to negative binomial modeling at this site.

I shall use the following citation and reference conventions throughout the text. Program commands, variable and data set names, as well as statistical output, are all displayed in Courier New typewriter font. Data sets and command names are in bold, e.g. `medpar, glm`. I shall follow standard conventions with respect to mathematical expressions.

This monograph is based on seminars and classes related to count response models that I have taught over the past 20 years. In particular, the presentation of

the material in this book closely follows the notes used for short courses I taught in November 2005 at the Federal Food and Drug Administration, Rockville, MD, and in Boston as a LearnStat program course sponsored by the American Statistical Association. I learned much from the lively discussions that were associated with the two courses, and have attempted to clarify various issues that seemed murky to several course participants. I have also expanded discussion of areas that were of particular interest to the majority of attendees, with the expectation that these areas will be of like interest to those choosing to read this book.

Note that I reiterate various main statistical points in the early chapters. I have found that there are certain concepts related to count response modeling, as well as to statistical modeling in general, that need to be firmly implanted in a statistician's mind when engaging in the modeling process. I have therefore characterized given concepts from differing points of view as well as reinforced the definitional properties of the statistics by repetition. For those who are approaching generalized linear models, maximum likelihood regression, and count response modeling for the first time, such repetition should prove useful. For those who are already familiar with these concepts, I suggest that you skim over repetitive material and concentrate on the underlying points being made. As we progress through the text, repetitiveness will be kept at a minimum.

Many colleagues have contributed to this work. I owe special appreciation to John Nelder, who spurred my initial interest in negative binomial models in 1992. We spent several hours discussing the relationship of Poisson and generalized linear models (GLMs) in general to negative binomial modeling while hiking a narrow trail from the precipice to the bottom and back of the Grand Canyon. This discussion initiated my desire to include the negative binomial into a GLM algorithm I was developing at the time to use as the basis for evaluating commercial GLM software.

I also wish to acknowledge the valuable influence that James Hardin and William Greene have had on my thinking. Dr Hardin and I collaborated in the writing of two texts on subjects directly related to count models, including the negative binomial. Our frequent discussions and joint projects have shaped many of the opinions I have regarding the negative binomial. He kindly read through the entire manuscript, offering valuable comments and suggestions throughout. Prof. Greene's profound influence can especially be found in the final chapters of this book. As author of LIMDEP, Greene has developed far more software applications relevant to count regression models – and negative binomial models in particular – than any other single individual. He has kindly shared with me his thinking, as well as his writings, on negative binomial models. Additionally, I thank Hyun Kim, University of Massachussetts,

Lowell, who is using the negative binomial in his research. He read through the manuscript and offered many helpful comments, particularly as related to early chapters of the book.

Finally I wish to express appreciation to Diana Gillooly, Statistics Editor, and to Catherine Appleton, Assistant Editor, Science, Technology, and Medicine at Cambridge University Press. Ms Gillooly's encouragement and willingness to extend deadlines when I required more time for research have helped make this book more comprehensive. Ms Appleton provided very useful information related to the technical aspects of the text.

A special thanks go to my wife, Cheryl, and to the two of our children who are living at home, Michael and Mitchell. They were forced to endure many hours without my active attention. Far too often my mind was deep in stat-land, while my body participated in family events. I dedicate this book to them, as well as to my daughter, Heather, and to my late parents, Rader John and NaDyne Anderson Hilbe.