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.

Introduction

The negative binomial is traditionally derived from a Poisson–gamma mixture model. However, the negative binomial may also be thought of as a member of the single parameter exponential family of distributions. This family of distributions admits a characterization known as generalized linear models (GLMs), which summarizes each member of the family. Most importantly, the characterization is applicable to the negative binomial. Such interpretation allows statisticians to apply to the negative binomial model the various goodness-of-fit tests and residual analyses that have been developed for GLMs.

Poisson regression is the standard method used to model count response data. However, the Poisson distribution assumes the equality of its mean and variance – a property that is rarely found in real data. Data that have greater variance than the mean are termed *Poisson overdispersed*, but are more commonly designated as simply *overdispersed*. Negative binomial regression is a standard method used to model overdispersed Poisson data.

When the negative binomial is used to model overdispersed Poisson count data, the distribution can be thought of as an extension to the Poisson model. Certainly, when the negative binomial is derived as a Poisson–gamma mixture, thinking of it in this way makes perfect sense. The original derivation of the negative binomial regression model stems from this manner of understanding it, and has continued to characterize the model to the present time.

As mentioned above, the negative binomial has recently been thought of as having an origin other than as a Poisson–gamma mixture. It may be derived as a generalized linear model, but only if its ancillary or heterogeneity parameter is entered into the distribution as a constant. The straightforward derivation of the model from the negative binomial probability distribution function (PDF) does not, however, equate with the Poisson–gamma mixture-based version of the negative binomial. Rather, one must convert the canonical link and inverse canonical link to log form. So doing produces a GLM-based negative binomial that yields identical parameter estimates to those calculated by the mixture-based model. As a non-canonical linked model, however, the standard errors will differ slightly from the mixture model, which is typically estimated using a full maximum likelihood procedure. The latter uses by default the observed information matrix to produce standard errors. The standard GLM algorithm uses Fisher scoring to produce standard errors based on the expected information matrix – hence the difference in standard errors between the two versions of negative binomial. The GLM negative binomial algorithm may be amended though to allow production of standard errors based on observed information. When this is done, the amended GLM-based negative binomial produces identical estimates and standard errors to that of the mixture-based negative binomial. This form of negative binimoal was called the *log-negative binomial* by Hilbe (1993a), and was the basis of a well-used SAS negative binomial macro (Hilbe, 1994b). It is also the form of the negative binomial found in Stata's *glm* command as well as in the SAS/STAT GENMOD procedure in SPSS's GLZ command, and in GENSTAT's GLM program.

Regardless of the manner in which the negative binomial is estimated, it is nevertheless nearly always used to model Poisson overdispersion. The advantage of the GLM approach rests in its ability to utilize the specialized GLM fit and residual statistics that come with the majority of GLM software. This gives the analyst the means to quantitatively test different modeling strategies with tools built into the GLM algorithm. This capability is rarely available with models estimated using full maximum likelihood or full quasi-likelihood methods.

In this book we shall discuss in greater depth the two methods of estimating negative binomial data that have been outlined above. The complete derivation of both methods will be given, together with discussion of how the algorithms may be altered to deal with count data that should not be modeled using simple Poisson or standard negative binomial methods. In fact, we shall devote considerable space to describing the base Poisson regression model, and the manner in which its assumptions may be violated. In addition, we shall find that just as Poisson models can be overdispersed, negative binomial models can as well. Following an examination of estimating methods and overviews of both the Poisson and negative binomial models, the remainder of the book is devoted to a discussion of how to understand and deal with various enhancements to both the Poisson and traditional negative binomial models.

Extensions to the respective Poisson and negative binomials are made depending on the type of underlying problem that is being addressed. Extended models include, among others, those for handling excessive response zeros – zero-inflated Poisson, zero-inflated negative binomial, and hurdle models; for handling responses having no possibility of zero counts – zero-truncated Poisson and zero-truncated negative binomial; having responses with structurally absent values – truncated and censored Poisson and negative binomial; and having longitudinal or clustered data – fixed, random, and mixed effects negative binomial as well as negative binomial GEE. Models may also have to be devised for situations when the data can be split into two or more distributional subsets. In fact, both Poisson and negative binomial models have been extended to account for a great many count response modeling situations. We shall attempt to give an overview of each of the major varieties mentioned here, which should provide the researcher with a map or guideline of how to handle a wide variety of count modeling situations.

Typically, extensions to the Poisson model precede analogous extensions to the negative binomial. For example, statisticians have recently created random parameter and random intercept count models to deal with certain types of correlated data. The first implementations were based on the Poisson distribution. Nearly all literature dealing with random parameter count models relates to the Poisson. Negative binomial versions have only surfaced within the past couple of years, primarily as a result of the work of William Greene. The only software available for modeling negative binomial random parameter and intercept models is LIMDEP, and even at that, it has not yet been made part of its menu system procedures.

Of the two general count regression models, the negative binomial has greater generality. In fact, as will be discussed at greater length later in the text, the Poisson can be considered as a negative binomial with an ancillary or heterogeneity parameter value of zero. It seems clear that having an understanding of the various negative binomial models, basic as well as complex, is essential for anyone considering serious research dealing with count models.

It is important to realize that the negative binomial has been derived and presented with different parameterizations. Some authors employ a variance function that clearly reflects a Poisson–gamma mixture. With the Poisson variance defined as μ and the gamma as μ^2/α, the negative binomial variance is then characterized as $\mu + \mu^2/\alpha$. The Poisson–gamma mixture is clear. This parameterization is the same as that originally derived by Greenwood and Yule (1920). An inverse relationship between μ and α was also used to define the negative binomial variance in McCullagh and Nelder (1989), to which some authors refer when continuing this manner of representation.

However, shortly after the publication of that text, Nelder developed his KK system (1992), a user-defined negative binomial macro written for use with Genstat software. In this system he favored the direct relationship between α and μ^2 – resulting in a negative binomial variance function of $\mu + \alpha\mu^2$.

Nelder has continued to prefer the direct relationship in his subsequent writings (1994). Still, relying on the 1989 work, a few authors have continued to use the originally defined relationship, even as recently as Faraway (2006).

The direct parameterization of the negative binomial variance function was favored by Breslow (1984) and Lawless (1987) in their highly influential seminal articles on the negative binomial. In the decade of the nineties, the direct relationship was used in the major software implementations of the negative binomial: Hilbe (1993b, 1994a) – XploRe and Stata, Greene (2006) – LIMDEP, and Johnston (1997) – SAS. The direct parameterization was also specified in Hilbe (1994a), Long (1997), Cameron and Trivedi (1998), and most articles and books dealing with the subject. Recently Long and Freese (2003, 2006), Hardin and Hilbe (2001, 2007), and a number of other recent authors have employed the direct relationship as the preferred variance function. It is rare now to find current applications using the older inverse parameterization.

The reason for preferring the direct relationship stems from the use of the negative binomial in modeling overdispersed Poisson count data. Considered in this manner, α is directly related to the amount of overdispersion in the data. If the data are not overdispersed, i.e. the data are Poisson, then $\alpha = 0$. Increasing values of α indicate increasing amounts of overdispersion. Values for data seen in practice typically range from 0 to about 4.

Interestingly, two books have been recently published, Hoffmann (2004) and Faraway (2006), asserting that the negative binomial is not a true generalized linear model. However, the GLM status of the negative binomial depends on whether it is a member of the single-parameter exponential family of distributions. If we assume that the overdispersion parameter, α, is known and is ancillary, resulting in what has been called a LIMQL (limited information maximum quasi-likelihood) model (see Greene, 2003), then the negative binomial is a GLM. On the other hand, if α is considered to be a parameter to be estimated, then the model may be estimated as FIMQL (full information maximum quasi-likelihood), but it is not a GLM.

In this text, the negative binomial is estimated as both a GLM and as a full maximum (quasi-)likelihood model. As a GLM, the model has associated fit and residual statistics, which can be of substantial use during the modeling process. However, in order to obtain a value of α, i.e. to make α known, it must be estimated. The traditional, and most reasonable, method of estimating α is by a non-GLM maximum likelihood algorithm. Extensions to the negative binomial model, e.g. zero-inflated, zero-truncated, and censored models, are nearly all based on FIMQL methods. I shall be using both methods of estimation when modeling basic Poisson and negative binomial data. How these methods are used together will become apparent as we progress through the text.

The first chapter provides a brief overview of count response regression models. Incorporated in this discussion is an outline of the variety of negative binomial models that have been constructed from its basic parameterization. Each extension from the base model is considered as a response to a violation of model assumptions. We list seven types of violation to the standard negative binomial model. Enhanced negative binomial models are identified as solutions to the respective violations.

Chapter 2 examines the two major methods of parameter estimation relevant to modeling Poisson and negative binomial data. We begin by illustrating the construction of distribution-based statistical models. That is, starting from a probability distribution, we follow the logic of establishing the estimating equations that serve as the focus of the fitting algorithms. Given that the Poisson and traditional negative binomial, also referred to as NB-2, are members of the exponential family of distributions, we define the exponential family and its constituent terms. In so doing we derive the iteratively reweighted least squares (IRLS) algorithm and the form of the algorithm required to estimate the model parameters. Secondly, we define maximum likelihood estimation and show how the modified Newton–Raphson algorithm works in comparison to IRLS. We shall discuss the reason for differences in output between the two estimation methods, and explain when and why differences occur.

Chapter 3 is devoted to the derivation of the Poisson log-likelihood and estimating equations. The Poisson traditionally serves as the basis for deriving the negative binomial – at least for one variety of negative binomial. Regardless, Poisson regression is the fundamental method used to model counts. We identify how overdispersion is indicated from Poisson model output, and some of the methods used to deal with it. We also discuss the rate parameterization of the count models. We find that rates can be thought of in a somewhat analogous manner to the denominators in binomial models. There are important differences though – which we discuss. The subject relates to the topic of offsets.

Chapter 4 details the difference in real versus apparent overdispersion. Criteria are specified which can be used to distinguish real from apparent overdispersion. Simulated examples are constructed that show how apparent overdispersion can be eliminated. We show how overdispersion affects otherwise equi-dispersed data. Finally, scaling of standard errors, application of robust variance estimators, jackknifing, and bootstrapping of standard errors are all evaluated in terms of their effect on inference. An additional section related to negative binomial overdispersion is provided, showing that overdispersion is a problem for all count models, not simply for Poission models. This chapter is vital to the development of the negative binomial model.

In Chapter 5 we define the negative binomial probability distribution function (PDF) and proceed to derive the various statistics required to model the canonical and traditional form of the distribution. Additionally, we derive the Poisson–gamma mixture parameterization that is used in maximum likelihood algorithms. In this chapter it becomes clear that the negative binomial is a full member of the exponential family of generalized linear models. We discuss the nature of the canonical form, and the problems that have been claimed to emanate when applying it to real data. We then re-parameterize the canonical form of the model to derive the traditional log-linked form (NB-2).

In Chapter 6 we discuss the development and interpretation of the NB-2 model. Examples are provided that demonstrate how the negative binomial is used to accommodate overdispersed Poisson data. Goodness-of-fit statistics are examined, in particular methods used to determine whether the negative binomial fit is statistically different from a Poisson. Residuals appropriate to evaluation of a negative binomial analysis are derived and explained.

Chapter 7 addresses alternative parameterizations of the negative binomial. We begin with a discussion of the geometric model, a simplification of the negative binomial where the overdispersion parameter has a value of one. When the value of the overdispersion parameter is zero, NB-2 reduces to a Poisson model. The geometric distribution is the discrete correlate of the negative exponential distribution. We then address the interpretation of the canonical link derived in Chapter 5. We thereupon derive and discuss how the linear negative binomial, or NB-1, is best interpreted. Finally, the NB-2 model is generalized in the sense that the ancillary or overdispersion parameter itself is parameterized by user-defined predictors for generalization from scalar to observation-specific interpretation. NB-2 can also be generalized to parameterize the negative binomial exponent. This model is called the NB-P model.

Chapter 8 deals with a common problem faced by researchers handling real data. In many situations the data at hand exclude a zero count. Other data situations have an excessive number of zeros – far more than defined by usual count distributions.

Zero-truncated and zero-inflated Poisson (ZIP) and negative binomial (ZINB) models, as well as hurdle models, have been developed to accommodate these two types of data situations. Hurdle models are typically used when the data have excessive zero counts, much like zero-inflated models. There are differences, however. Detailed are logit, probit, and complementary loglog negative binomial hurdle models. Finally, we examine negative binomial models having endogenous stratification.

Chapter 9 discusses truncated and censored data and how they are modeled using appropriately adjusted Poisson and negative binomial models. Two types of parameterizations are delineated for censored count models: econometric or dataset-based censored and survival, or observation-based censored, parameterizations.

The final chapter addresses the subject of negative binomial panel models. These models are used when the data are either clustered or when they are in the form of longitudinal panels. We derive and examine unconditional and conditional fixed effects and random effects Poisson and negative binomial regression models. Population averaged panel models, also referred to as generalized estimating equations (GEE) are also examined as are random intercept and random coefficient multilevel negative binomial models.

Several appendices are associated with the text. The titles are listed in the Contents.

1
Overview of count response models

Count response models are a subset of discrete response regression models. Discrete models address non-negative integer responses. Examples of discrete models include:

RESPONSE

Binary: binary logistic and probit regression
Proportional: grouped logistic, grouped complementary loglog
Ordered: ordinal logistic and ordered probit regression
Multinomial: discrete choice logistic regression
Count: Poisson and negative binomial regression

A count response consists of any discrete response of counts, e.g. the number of hits recorded by a Geiger counter, patient days in the hospital, and goals scored at major contests. All count models aim to explain the number of occurrences, or counts, of an event. The counts themselves are intrinsically heteroskedastic, right skewed, and have a variance that increases with the mean of the distribution.

1.1 Varieties of count response model

Poisson regression is the basic count model upon which a variety of other count models are based. The Poisson distribution may be characterized as

$$f_y(y;\mu) = e^{-\mu}\mu^y/y!, \qquad y = 0, 1, 2, \ldots; \qquad \mu > 0 \tag{1.1}$$

where the random variable y is the count response and parameter μ is the mean. Often, μ is also called the rate or intensity parameter. Unlike most other distributions, the Poisson does not have a distinct scale parameter. Rather, the scale is assumed equal to the location parameter μ.

The Poisson distribution may also include an exposure variable associated with μ. The variable, t, is considered to be the length of time or exposure during which events or counts occur. If $t = 1$, then the Poisson probability distribution reduces to the standard form. If t is a constant, or varies between events, then the distribution can be parameterized as

$$f_y(y; \mu) = e^{-t\mu}(t\mu)^y/y! \tag{1.2}$$

When included in the data, modelers enter the natural log of t as an offset in the model estimation. Playing an important role in estimating both Poisson and negative binomial models, offsets are discussed at greater length in Chapter 3.

A unique feature of the Poisson distribution is the relationship of its mean to the variance – they are equal. This relationship is termed equidispersion. The fact that it is rarely found in real data has driven the development of more general count models, which do not assume such a relationship.

The Poisson regression model derives from the Poisson distribution. The relationship between μ, β, and x, the fitted mean of the model, parameters, and model covariates or predictors respectively, is parameterized such that $\mu = \exp(x\beta)$. So doing guarantees that μ is positive for all values of η, the linear predictor, and for all parameter estimates. By attaching the subscript, ι, to μ, y, and x, the parameterization can be extended to all observations in the model. The subscript can also be used when modeling non-iid observations.

As shall be described in greater detail later in this book, the Poisson model carries with it various assumptions. Violations of Poisson assumptions usually result in overdispersion, where the variance of the model exceeds the value of the mean. Violations of equidispersion indicate correlation in the data, which affect standard errors of the parameter estimates. Model fit is also affected. Chapter 4 is devoted to this discussion.

A simple example of how distributional assumptions may be violated will likely be instructional at this point. We begin with the base count model – the Poisson. The Poisson distribution defines a probability distribution function for non-negative counts or outcomes. For example, given a Poisson distribution having a mean of 2, some 39% of the outcomes are predicted to be zero. If, in fact, we are given an otherwise Poisson distribution having a mean of 2, but with 50% zeros, it is clear that the Poisson distribution may not adequately describe the data at hand. When such a situation arises, modifications are made to the Poisson model to account for discrepancies in the goodness of fit of the underlying distribution. Models such as zero-inflated Poisson and zero-truncated Poisson directly address such problems.

The above discussion regarding distributional assumptions applies equally to the negative binomial. A traditional negative binomial distribution having a mean of 2 and an ancillary parameter of 1.5 yields a probability of approximately 40% for an outcome of zero. When the observed number of zeros substantially differs from the theoretically imposed number of zeros, the base negative binomial model can be adjusted in a manner similar to the adjustments mentioned for the Poisson.

Early on, researchers developed enhancements to the Poisson model, which involved adjusting the standard errors in such a manner that the presumed overdispersion would be dampened. Scaling of the standard errors was the first method developed to deal with overdispersion from within the GLM framework. It is a particularly easy tactic to take when the Poisson model is estimated as a generalized linear model. We shall describe scaling in more detail later in the text. Nonetheless, most count models required more sophisticated adjustments than simple scaling.

Again, the negative binomial is normally used to model overdispersed Poisson data, which spawns our notion of the negative binomial as an extension of the Poisson. However, distributional problems affect both models, and negative binomial models themselves may be overdispersed. Both models can be extended in similar manners to accommodate any extra correlation or dispersion in the data that result in a violation of the distributional properties of each respective distribution (Table 1.1). The enhanced or advanced Poisson or negative binomial model can be regarded as a solution to a violation of the distributional assumptions of the primary model.

The following list enumerates the types of extensions that are made to both Poisson and negative binomial regression. Thereafter, we provide a bit more detail as to the nature of the assumption being violated and how it is addressed by each type of extension. Later chapters are devoted to a more detailed examination of each of these model types.

Earlier in this chapter we described violations of Poisson and negative binomial distributions as related to excessive zero counts. Each distribution has an expected numbers of counts for each value of the mean parameter; we saw how for a given mean, an excess – or deficiency – of zero counts result in overdispersion. However, it must be understood that the negative binomial has an additional ancillary or heterogeniety parameter, which, in concert with the value of the mean parameter, defines (in a probabilistic sense) specific expected values of counts. Substantial discrepancies in the number of counts, i.e. how many zeros, how many ones, how many twos, and so forth, observed in the data from the expected frequencies defined by the given mean and ancillary parameter (NB model), result in correlated data and hence overdispersion. The first two items in Table 1.1 directly address this problem.

Table 1.1. *Violations of distributional assumptions*

1	No zeros in data
2	Excess zeros in data
3	Data separable into two distributions
4	Censored observations
5	Truncated data
6	Data structured as panels: clustered and longitudinal data
7	Some responses occur based on the value of another variable

Violation 1: The Poisson and negative binomial distributions assume that zero counts are a possibility. When the data to be modeled originate from a generating mechanism that structurally excludes zero counts, then the Poisson or negative binomial distribution must be adjusted to account for the missing zeros. Such model adjustment is not used when the data can have zero counts, but simply do not. Rather, an adjustment is made only when the data must be such that it is not possible to have zero counts. Hospital length of stay data are a good example. When a patient enters the hospital, a count of one is given. There are no lengths of stay recorded as zero days. The possible values for data begin with a count of one. Zero-truncated Poisson and zero-truncated negative binomial models are normally used for such situations.

Violation 2: The Poisson and negative binomial distributions define an expected number of zero counts for a given value of the mean. The greater the mean, the fewer zero counts expected. Some data, however, come with a high percentage of zero counts – far more than are accounted for by the Poisson or negative binomial distribution. When this occurs statisticians have developed regression models called zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB). The data are assumed to come from a mixture of two distributions where the structural zeros from a binary distribution are mixed with the non-negative integer outcomes (including zeros) from a count distribution. Logistic or probit regression is typically used to model the structural zeros, and Poisson or negative binomial regression is used for the count outcomes. If we were to apply a count model to the data without explicitly addressing the mixture, it would be strongly affected by the presence of the excess zeros. This inflation of the probability of a zero outcome is the genesis of the zero-inflated name.

Violation 3: When the zero counts of a Poisson or negative binomial model do not appear to be generated from their respective distributions, one may separate the model into two parts, somewhat like the ZIP and ZINB models above. However, in the case of hurdle models, the assumption is that a threshold must be crossed from zero counts to actually entering the counting process. For example, when modeling insurance claims, clients may have a year without claims – zero counts. But when one or more accidents occur, counts of claims follow a count distribution, e.g. Poisson or negative binomial. The logic of the severability in hurdle models differs from that of zero-inflated models. Hurdle models are sometimes called zero-altered models, giving us model acronyms of ZAP and ZANB.

Like zero-inflated models, hurdle or zero-altered algorithms separate the data into zero versus positive counts: modeling zero (1) versus positive count (0) as a logit, probit, or complementary loglog model, and counts from one to the upper range of counts as Poisson or negative binomial. Zero-inflated likelihood functions differ from the likelihood functions of similar hurdle models. We shall address these differences in more detail in Chapter 9. Good references to these discussions in particular can be found in Greene (1994) and Cameron and Trivedi (1998).

Violation 4: At times certain observations are censored from the rest of the model. With respect to count response models, censoring takes two forms. In either case a censored observation is one that contributes to the model, but for which exact information is missing.

The traditional form, which I call the econometric parameterization, revalues censored observations as the value of the lower or upper valued non-censored observation. Left-censored data take the value of the lowest non-censored count; right-censored data take the value of the highest non-censored count. Another parameterization, which can be referred to as the survival parameterization, considers censoring in the same manner as is employed with survival models. That is, an observation is left censored to when events are known to enter into the data; they are right censored when events are lost to the data due to withdrawal from the study, loss of information, and so forth. The log-likelihood functions of the two parameterizations differ, but the parameter estimates calculated are usually not too different.

Violation 5: Truncated observations consist of those that are entirely excluded from the model, at least for the period of truncation. Unlike the econometric parameterization of censoring described in Violation 4, truncated data are excluded, not revalued, from the model. Data can be left censored from below, or right censored from above.

Violation 6: Longitudinal data come in the form of panels. For example, in health studies, patients given a drug may be followed for a period of time to ascertain effects occurring during the duration of taking the drug. Each patient may have one or more follow-up tests. Each set of patient observations is considered a panel. The data consist of a number of panels. However, observations within each panel cannot be considered independent – a central assumption of maximum likelihood theory. Within-panel correlation result in overdispersed data. Clustered data result in similar difficulties. In either case, methods have been developed to accommodate extra correlation in the data due to the within-panel correlation of observations. Such models, however, do require that the panels themselves are independent of one another, even though the observations within the panels are not. Generalized estimating equations (GEE), fixed-effects models, and random-effects models have been widely used for such data.

Violation 7: Data sometimes come to us in such a manner that an event does not begin until a specified value of another variable has reached a certain threshold. One may use a selection model to estimate parameters of this type of data. Greene (1994) summarizes problems related to selection models. Selection models for count response data are in their infancy, and more theoretical work needs to be done. At this point only LIMDEP incorporates a negative binomial selection model among its offerings.

Table 1.2 provides a schema of the major types of negative binomial regression model. A similar schema may also be presented characterizing varieties of the Poisson. Some exceptions exist, however. Little development work has been committed to the exact statistical estimation of negative binomial standard errors. However, substantial work has been done on Poisson models of this type – particularly by Cytel Corp, manufacturers of LogXact software. Additionally, models such as heterogeneous negative binomial and NB-P have no correlative Poisson model.

1.2 Estimation

There are two basic approaches to estimating models of count data. The first is by full maximum likelihood (ML or FML); the second is by an iteratively re-weighted least squares (IRLS) algorithm, which is based on a simplification of the full maximum likelihood method. IRLS is intrinsic to the estimation of generalized linear models (GLMs), as well as to certain extensions to the generalized linear model algorithm. We examine the details of estimation in the following chapter. However, we can mention here that both methods are commonly used for the analysis of Poisson and negative binomial data. The negative

Table 1.2. *Varieties of negative binomial model*

```
1  Negative binomial (NB)
       NB2
       NB1
       NB-C (canonical)
       Geometric
       NB-H (Heterogeneous negative binomial)
       NB-P
2  Zero-adjusting models
       Zero-truncated NB
       Zero-inflated NB
       NB with endogenous stratification (G)
       Hurdle NB models
           NB-logit hurdle    / geometric-logit hurdle
           NB-probit hurdle   / geometric-probit hurdle
           NB-cloglog hurdle  / geometric-cloglog hurdle
3  Censored NB
       Censored NB-E: econometric parameterization
       Censored NB-S: survival parameterization
4  Sample selection NB models
5  Panel NB models
       Unconditional Fixed Effects NB
       Conditional Fixed Effects NB
       Generalized Estimating Equations
       Linear Mixed NB Models
           Random intercept NB
           Random parameter NB
           Latent Class NB models
6  Exact NB model
```

binomial, conceived as a Poisson–gamma mixture model, is usually estimated by maximum likelihood using a variety of the Newton–Raphson estimating algorithm. A commonly used variety is the Marquardt (1963) modification, which has itself been modified for use in leading commercial packages. Such a method allows the estimation of the negative binomial ancillary or overdispersion parameter, which we refer to as α.

The negative binomial was first supported in the generalized linear models software by Hilbe (1994), who incorporated it into the GLM procedures of both XploRe and Stata software. So doing allowed use of the full range of GLM fit and residual capabilities. However, since the standard IRLS GLM algorithm allows estimation of only the mean parameter, μ, or exp(βx), the ancillary or heterogeneity parameter must be entered into the GLM algorithm as a known constant. It is not itself estimated. The scale parameter for all GLM count models is defined as one, and does not enter into the estimation process.

There are methods to estimate α (Breslow, 1984; Hilbe, 1993a) based on an iterative covering algorithm with an embedded IRLS that forces the deviance-dispersion statistic to a value of 1.0. But the resulting estimated value of α typically differs from the value estimated using maximum likelihood. Moreover, no standard error is obtained for the estimated α using the cited IRLS approaches. Typically though, the difference in estimated α between the two methods is not very different. Both methods will be examined in detail in the next chapter, together with recommendations of how they can be used together for a given research project.

1.3 Fit considerations

Fit statistics usually take two forms: so-called goodness-of-fit statistics and residual statistics. With respect to the negative binomial, goodness-of-fit generally relates to the relationship of the negative binomial model to that of a Poisson model on the same data. Assuming that the purpose of negative binomial modeling is to model otherwise Poisson overdispersed data, it makes sense to derive fit statistics aimed at determining whether a negative binomial model is statistically successful in that regard. Two of the commonly used fit statistics include a Score test, or Lagrange multiplier test, and a Vuong test (Vuong, 1989). These, as well as others, will later be examined in more detail.

Residual analysis of negative binomial models generally follows the model varieties constructed for members of the family of generalized linear models. These include unstandardized and standardized Pearson and deviance residuals, as well as the studentized forms for both. Additionally, Anscombe residuals have been developed for the NB-2 negative binomial (Hilbe, 1994a; Hardin and Hilbe, 2001). In application, however, values of the Anscombe are similar to those of the standardized deviance residuals, with the latter being substantially less complicated to calculate. Most commercial GLM software packages provide standardized deviance residuals as options. At the time of this writing, only Stata provides Anscombe residuals as a standard option. A commonly used residual fit analysis for negative binomial models plots standardized deviance residuals against the fitted values, μ.

1.4 Brief history of the negative binomial

Negative binomial regression can be viewed as a nonlinear regression model, estimated by maximum likelihood, or as a member of the family of generalized linear models. The basic or traditional form of negative binomial – NB-2 – employs a natural log link function. So doing forces the fitted values to be positive, as is appropriate for counts. In this respect, it is similar to the Poisson. Again, as stated earlier, when the negative binomial ancillary parameter has a value of zero, the model is Poisson. However, the canonical link of the negative binomial, unlike the Poisson, is not the log link. It is simply called the canonical link, taking the form: $\ln(\alpha\mu/(1+\alpha\mu))$.

When considered as a generalized linear model, the negative binomial follows in the tradition originating from Gauss (1823), who developed the theory of normal or Gaussian regression, also called ordinary least squares regression (OLS). Likelihood theory was developed by Ronald Fisher in 1922, the same year as he constructed the first complimentary loglog model. Two years earlier, Greenwood and Yule had derived the negative binomial probability distribution function as the probability of observing y failures before the rth success in a series of Bernoulli trials. The negative binomial probability may be construed differently though, as reflected in the structure of the function. Regardless, the negative binomial has been derived from the binomial, as above, as well as from the Poisson as a Poisson–gamma mixture. The contagion or mixture concept of the negative binomial originated with Eggenberger and Polya (1923).

George Beall (1942) and F. J. Anscombe (1949) followed on the efforts's of Bartlett (1947) and his analysis of square root transforms on Poisson data, by examining variance stabilizing transformations for overdispersed data. Anscombe's work entailed the construction of the first negative binomial regression model, but as a one-parameter nonlinear regression. A maximum likelihood Poisson regression model was not fully developed until Birch in 1963, which he used to analyze tables of counts. The rate parameterization of the Poisson took another 11 years to develop. John Nelder, of London's Imperial College of Science and Technology, originated offsets as an admitted afterthought, not realizing until later the extent to which they could be used.

The theory of generalized linear models was developed by John Nelder and R. W. M. Wedderburn in 1972 (Nelder and Wedderburn, 1972), then expanded by Nelder thereafter. Following Wedderburns's untimely death, Nelder wrote the first version of GLIM software, which allowed users to estimate GLMs for a limited set of exponential family members. Although GLIM did not have a specific option for negative binomial models, one could use the *open* option to craft such a model. In 1982 Nelder joined with Peter McCullagh to write the first edition of *generalized linear models*, in which the negative binomial regression model was described. Negative binomial regression software was only available as a user defined macro in GLIM or in Genstat, via Nelder's KK system extension until 1993. In that year Hilbe incorporated the negative binomial into both XploRe and Stata GLM software. He also wrote a GLM-based negative binomial macro in SAS, which was the only SAS program for the model until Gordon Johnston made it part of the SAS/STAT GENMOD procedure in 1998. In 1994, Venables posted a GLM-based negative binomial to Statlib using S-Plus.

Maximum likelihood estimation of the negative binomial began with Plackett in 1981, while working with categorical data which he could not fit using a Poisson approach.

Geometric hurdle models were developed by Mullahy (1986), with a later enhancement to negative binomial hurdle models. William Greene's LIMDEP was the first commercial package to offer negative binomial regression models to its users (1987 [2006]). Stata was next with a maximum likelihood negative binomial (1994). Called *nbreg*, Stata's negative binomial command was later enhanced to allow modeling of both NB-1 as well as NB-2 parameterizations. In 1998, Stata offered a generalized negative binomial, *gnbreg*, in which the heterogeniety parameter itself could be parameterized. It should be emphasized that this command does not address the generalized negative binomial distribution, but rather it allows a generalization of the scalar overdispersion parameter such that parameter estimates can be calculated showing how model predictors comparatively influence overdispersion. Following LIMDEP, I have referred to this model as a heterogeneous negative binomial, or NB-H, since the model extends NB-2 to permit observed sources of heterogeneity in the overdispersion parameter. In the meantime, LIMDEP has continuously added to its initial negative binomial offerings. It currently estimates nearly every negative binomial-related model that shall be discussed in this monograph. In 2006 Greene developed a new parameterization of the negative binomial, called NB-P, which estimates both the traditional negative binomial ancillary parameter, as well as the exponent of the second term of the variance function.

As mentioned in the Preface, R can be used to model negative binomial data using the MASS package available from the R package library. The first R implementation was written by Ihaka and Gentleman in 1996. The software is under continuous development.

1.5 Summary

Negative binomial models have been derived from two different origins. First, and initially, the negative binomial can be thought of as a Poisson–gamma mixture designed to model overdispersed Poisson count data. Conceived of in this manner, estimation usually takes the form of a maximum likelihood

Newton–Raphson type algorithm. This parameterization estimates both the mean parameter, μ, as well as the ancillary or heterogeneity parameter, α. Extensions to this approach allow, for example, α itself to be parameterized (NB-H), as well as the negative binomial exponent (NB-P). Violations of distributional assumptions are addressed by various adjustments to the base negative binomial (NB-2) model. Examples include models such as ZINB, zero-truncated negative binomial, and censored negative binomial regression.

Secondly, the negative binomial can be derived as a full member of the single parameter exponential family of distributions, and hence be considered as one of the generalized linear models. The value of this approach rests on the ability to evaluate the model using well-tested GLM goodness-of-fit statistics as well as to employ the host of associated GLM-defined residuals. Estimation in this case takes the form of Fisher scoring, or iteratively re-weighted least squares. Since the traditional GLM algorithm only allows estimation of parameter θ, which gives us the value of μ, ancillary parameter α must be specified directly into the estimating algorithm as a known constant – it is not itself estimated. Although this is a drawback for its usefulness in modeling, the ability to assess fit in part offsets this problem.

Many statisticians use both methods in concert when engaging in a modeling task. Initial estimation is performed using a full maximum likelihood procedure, with the resultant estimated value of α then inserted into a GLM algorithm. The respective value of each method thereby contributes to the overall modeling task.

We shall next examine the foremost methods of estimation.

2
Methods of estimation

Two general methods are used to estimate count response models: (1) iteratively re-weighted least squares algorithm based on the method of Fisher scoring, and (2) a maximum likelihood Newton–Raphson type algorithm. Although the maximum likelihood approach was first used with both the Poisson and negative binomial, we shall discuss it following our examination of IRLS. We do this for strictly pedagogical purposes, as will become evident as we progress.

2.1 Derivation of the IRLS algorithm

The traditional generalized linear models (GLM) algorithm, from the time it was implemented in GLIM (generalized linear interactive modeling) through its current implementations in Stata, S-Plus, and other GLM software, uses some version of an IRLS estimating algorithm. This method arises from Fisher scoring, which substitutes the expected Hessian matrix for the observed Hessian matrix in a Taylor series defined updating step for a solution of the estimating equation. The resulting Newton–Raphson or updating equation for the regression coefficients may be written in terms of ordinary least squares (OLS) due to the simplification afforded by Fisher scoring. The reason for its initial development had much to do with the difficulty of modeling individual GLM models using full maximum likelihood algorithms. In the late 1960s and early 1970s, statistical software was limited to mainframe batch runs. That is, one wrote an estimating algorithm in a higher programming language such as FORTRAN, tied it together with data stored on cards, and submitted it to a mainframe, which usually resided at a remote site. If one desired to make changes, then the entire batch file required rewriting. Each submission to the mainframe had a cost applied to it, normally charged to the department of the user. Problems with initial values, difficulties with convergence, and other such difficulties resulted in a modeling project taking substantial time and money.

When Wedderburn proposed an IRLS algorithm for estimating regression models based on the exponential family of distributions, specific models still had to be submitted on cards and to a mainframe. The difference was, however, that there were substantially fewer difficulties with convergence, and one algorithm could be used for all members of the class. All that required alteration from one model type to another, e.g. a logit model compared with a probit model, or a gamma compared with a Poisson, was a change in the specification of the link and variance functions. The algorithm took care of the rest. Time savings transferred to money savings.

As previously mentioned, GLIM was the first commercial software implementation of GLM. When desktop computing became available on PCs starting in August 1981, Numerical Algorithms Group in the UK, the manufacturer of GLIM, quickly implemented a desktop version of the GLIM software. For the first time models such as logit, probit, Poisson, gamma, and so forth could be modeled in an interactive and inexpensive manner. An international GLIM user group emerged whose members designed macros to extend the basic offerings. The negative binomial was one of these macros, but was not published to the general user base.

The important point in this discussion is that the IRLS algorithm was designed to be a single covering algorithm for a number of related models. Also important to understand is that the IRLS algorithm is a simplified maximum likelihood algorithm and that its derivation is similar to that of the derivation of general Newton–Raphson type models. We turn to this demonstration next.

IRLS methodology, like maximum likelihood methodology in general, is ultimately based on a probability distribution or probability mass function. Generalized linear models software typically offers easy specification of variance functions defined by eight families, each of which is a probability function. These include:

GLM DISTRIBUTIONAL FAMILIES

```
Gaussian              Bernoulli
Binomial              Gamma
Inverse Gaussian      Poisson
Geometric             Negative Binomial
```

We may express the GLM probability function as

$$f(y;\ \theta,\ \varphi) \tag{2.1}$$

where y is the response, θ is the location or mean parameter, and φ is the scale parameter. Count models by definition set the scale to a value of one. The outcome y, of course, has the distributional properties appropriate to the family used in estimation. Probability functions determine properties of the response, or data, given values of the mean and scale parameters. Maximum likelihood, on the other hand, bases estimation on the likelihood.

The likelihood function is the reverse of the probability function. Rather than data being determined on the bases of mean and scale values, the mean, and possibly scale, parameters are estimated on the basis of the given data. The underlying goal of likelihood is to determine which parameters make the given data most likely. This parameterization can be characterized as

$$L(\theta, \phi; y) \tag{2.2}$$

Statisticians normally employ the natural log of the likelihood function in order to facilitate estimation. The prime reason is the observations and their respective parameter estimates enter the likelihood function in a multiplicative manner. However, it is much easier to estimate parameters if their relationship is instead additive. In fact, for many modeling situations, using a likelihood function rather than a log-likelihood function would not allow the estimation process to get off the ground. An excellent discussion of this topic, together with numeric examples, can be found in Gould, Pitblado, and Sribney (2006). Also see Edwards (1972).

The log-likelihood function can be written as

$$\mathcal{L}(\theta, \phi; y) \tag{2.3}$$

The first derivative of the log-likelihood function is called the gradient; the second derivative is the Hessian. These functions play an essential role in the estimating process for both IRLS and traditional Newton–Raphson type algorithms.

Gradient – first derivative of $\mathcal{L}$
Hessian – second derivative of $\mathcal{L}$

Derivation of the iteratively re-weighted least squares algorithm is based on a modification of a two-term Taylor expansion of the log-likelihood function. In its original form, Taylor expansion appears as

$$\begin{aligned} 0 &= f(X_0) + (X_1 - X_0)f'(X_0) \\ &\quad + \frac{(X_1 - X_0)^2}{2!}f''(X_0) + \frac{(X_1 - X_0)^3}{3!}f'''(X_0) + \cdots \end{aligned} \tag{2.4}$$

The first two terms reduce to

$$0 = f(X_0) + (X_1 - X_0)f'(X_0) \tag{2.5}$$

which can be recast to

$$X_0 = X_1 - \frac{f(X_0)}{f'(X_0)} \tag{2.6}$$

The Newton–Raphson method of estimation adopts the above by using the score or gradient of the log-likelihood function as the basis of parameter estimation. The form is

$$\beta_r = \beta_{r-1} - \frac{\partial\mathcal{L}(\beta_{r-1})}{\partial^2\mathcal{L}(\beta_{r-1})} \tag{2.7}$$

where

$$\partial\mathcal{L} = \frac{\partial\mathcal{L}}{\partial\beta} \quad \text{and} \quad \partial^2\mathcal{L}\frac{\partial^2\mathcal{L}}{\partial\beta\partial\beta'} \tag{2.8}$$

In traditional nomenclature, we let

$$U = \partial\mathcal{L} \quad \text{and} \quad H = \partial^2\mathcal{L} \tag{2.9}$$

Then

$$\beta_r = \beta_{r-1} - H^{-1}U \tag{2.10}$$

where

$$H = H_{r-1} \quad \text{and} \quad U = U_{r-1} \tag{2.11}$$

Newton–Raphson estimates β_r, the model parameter estimates, by iteratively finding solutions for *H* and *U*, which define β_r. β_r resets itself to β_{r-1} in each subsequent iteration until some predefined threshold is reached. We shall see, however, that the matrix *H* used by the Newton–Raphson is the observed information matrix. IRLS, on the other hand, defines *H* as the expected information matrix. We next turn to how both methods define *U*.

2.1.1 Solving for $\partial\mathcal{L}$ or U – the gradient

In exponential family form, the log-likelihood function is expressed as

$$\frac{\partial\mathcal{L}}{\partial\beta_j} = \sum_{i=1}^{n_i} \frac{y_i\theta_i - b(\theta_i)}{\alpha_i(\phi)} + \sum_{i=1}^{n_i} C(y_i, \phi) \tag{2.12}$$

Solving for $\mathcal{L}$, with respect to β, can be performed using the chain rule

$$\frac{\partial\mathcal{L}}{\partial\beta_j} = \sum_{i=1}^{n_i} = \left(\frac{\partial\mathcal{L}}{\partial\theta}\right)_i \left(\frac{\partial\theta}{\partial\mu}\right)_i \left(\frac{\partial\mu}{\partial\eta}\right)_i \left(\frac{\partial\eta}{\partial\beta_j}\right) \tag{2.13}$$

Solving for each term yields

$$\frac{\partial \mathcal{L}}{\partial \theta_i} = \sum_{i=1}^{n_i} \frac{y_i\theta_i - b'(\theta_i)}{\alpha_i(\phi)} = \sum_{i=1}^{n_i} \frac{y_i - \mu_i}{\alpha(\phi)} \tag{2.14}$$

We obtain the above formula by solving each of the terms of the chain.

We have $b'(\theta_i) = \mu_i$

$$\frac{\partial \mu_i}{\partial \theta_i} = \frac{\partial b'(\theta_i)}{\partial \theta_i} = b''(\theta_i) = V(\mu_i); \quad \frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{V(\mu_i)} \tag{2.15}$$

and

$$\frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial (x_i\beta_j)}{\partial \beta_j} = x_{ij}, \qquad \text{since} \quad \eta_i = x_i\beta_j \tag{2.16}$$

and

$$\frac{\partial \mu_i}{\partial \eta_i} = [g^{-1}(\eta_i)]' = \frac{1}{\partial \eta_i / \partial \mu_i} = \frac{1}{g'(\mu_i)} \tag{2.17}$$

which is the derivative of the link function with respect to μ.

Substitutions of expressions specify that the maximum likelihood estimator of β is the solution of the vector-based estimating equation

$$\sum_{i=1}^{n_i} \frac{(y_i - \mu_i)x_i}{\alpha_i(\phi)V(\mu_i)g'(\mu_i)} = \frac{(y_i - \mu_i)x_i}{\alpha_i(\phi)V(\mu_i)} \left(\frac{\partial \mu}{\partial \eta}\right)_i = 0 \tag{2.18}$$

where y and μ are scalars, x is a $1 \times p$ row vector, and the resulting sum is a $p \times 1$ column vector.

The next step in the derivation takes two turns, based on a decision to use the observed or expected information matrix. Again, Newton–Raphson type maximum likelihood estimation uses the observed matrix; IRLS, or Fisher scoring, uses the expected. We first address the latter.

2.1.2 Solving for $d^2\mathcal{L}$

The traditional GLM algorithm substitutes I for H, the Hessian matrix of observed second derivatives. I is the second of two equivalent forms of Fisher information given by

$$I = -E\left[\frac{\partial^2 \mathcal{L}}{\partial \beta_j \partial \beta_k'}\right] = E\left[\frac{\partial \mathcal{L}}{\partial \beta_j}\frac{\partial \mathcal{L}}{\partial \beta_k'}\right] \tag{2.19}$$

Solving the above yields

$$I = \frac{\partial}{\partial \beta_j}\left[\frac{(y_i - \mu_i)x_j}{\alpha_i(\phi)V(\mu_i)}\left(\frac{\partial}{\partial \eta}\right)_i\right] * \frac{\partial \mathcal{L}}{\partial \beta_k}\left[\frac{(y_i - \mu_i)x_k}{\alpha_i(\phi)V(\mu_i)}\left(\frac{\partial \mu}{\partial \eta}\right)_i\right] \tag{2.20}$$

$$I = \frac{(y_i - \mu_i)^2 x_j x_k}{\alpha_i(\phi)V(\mu_i)^2}\left(\frac{\partial \mu}{\partial \eta}\right)_i^2 \tag{2.21}$$

Since

$$(y_i - \mu_i)^2 = \alpha_i(\phi)V(\mu_i) \tag{2.22}$$

and letting

$$V(y_i) = \alpha_i(\phi)V(\mu_i) = (y_i - \mu_i)^2 \tag{2.23}$$

I therefore becomes formulated as

$$I = \frac{x_j x_k}{V(y_i)}\left(\frac{\partial \mu}{\partial \eta}\right)_i^2 = \frac{x_j x_k}{V(y_i)g'^2} \tag{2.24}$$

Putting the various equations together we have

$$\beta_r = \beta_{r-1} - \left[\frac{x_j x_k}{V(y_i)}\left(\frac{\partial \mu}{\partial \eta}\right)_i^2\right]^{-1}\left[\frac{x_k(y_i - \mu_i)}{V(y_i)}\left(\frac{\partial \mu}{\partial \eta}\right)_i\right] \tag{2.25}$$

Multiply both sides by *I* yields

$$\left[\frac{x_j x_k}{V(y_i)}\left(\frac{\partial \mu}{\partial \eta}\right)_i^2\right]\beta_r = \left[\frac{x_j x_k}{V(y_i)}\left(\frac{\partial \mu}{\partial \eta}\right)_i^2\right]\beta_{r-1} + \left[\frac{x_k(y_i - \mu_i)}{V(y_i)}\left(\frac{\partial \mu}{\partial \eta}\right)_i\right] \tag{2.26}$$

We next let weight *W* equal

$$W(\text{weight}) = \frac{1}{V(y_i)}\left(\frac{\partial \mu}{\partial \eta}\right)_i^2 \tag{2.27}$$

with the linear predictor, η, given, as

$$\eta_i = x_k \beta_{r-1} \tag{2.28}$$

We next convert the above algebraic representation to matrix form. This can be done in parts. First, given the definition of *W* above, the following substitution may be made

$$\left[\frac{x_j x_k}{V(y_i)}\left(\frac{\partial \mu}{\partial \eta}\right)_i^2\right]\beta_r = [X'WX]\beta_r \tag{2.29}$$

Secondly, recalling the definition of $V(y)$ and W

$$\frac{x_k(y_i - \mu_i)}{V(y_i)}\left(\frac{\partial\mu}{\partial\eta}\right)_i = \frac{x_k(y_i - \mu_i)}{\frac{1}{W}\left(\frac{\partial\mu}{\partial\eta}\right)_i^2}\left(\frac{\partial\mu}{\partial\eta}\right)_i \tag{2.30}$$

Thirdly, since $\eta = x_k\,\beta_{r-1}$, we have, in matrix form

$$\left[\frac{x_j x_k}{V(y_i)}\left(\frac{\partial\mu}{\partial\eta}\right)_i^2\right]\beta_{r-1} = X'W\eta_i \tag{2.31}$$

Combining the terms involved, we have

$$[X'WX]\beta_r = X'W\eta_i + \left[\frac{x_k(y_i - \mu_i)}{\frac{1}{W}\left(\frac{\partial\mu}{\partial\eta}\right)_i^2}\left(\frac{\partial\mu}{\partial\eta}\right)_i\right] \tag{2.32}$$

$$[X'WX]\beta_r = X'W\eta_i + \left[X_k W(y_i - \mu_i)\left(\frac{\partial\eta}{\partial\mu}\right)_i\right] \tag{2.33}$$

Finally, letting z, the model working response, be defined as

$$z_i = \eta_i + (y_i - \mu_i)\left(\frac{\partial\eta}{\partial\mu}\right)_i \tag{2.34}$$

we have

$$[X'WX]\beta_r = X'Wz_i \tag{2.35}$$

so that, by repositioning terms, β_r is equal to

$$\beta_r = [X'WX]^{-1}XWz_i \tag{2.36}$$

which is a weighted regression matrix used to iteratively update estimates of parameter vector β_r, as well as values for μ, η, and the deviance function. Iteration typically culminates when the difference in deviance values between two iterations is minimal, usually 10^{-6}. Some software uses the minimization of differences in the log-likelihood function as the basis of iteration. Others use differences in parameter estimates. In any case, the results are statistically identical. However, since the deviance is itself used to assess the fit of a GLM model, as well as being a term in the goodness-of-fit BIC statistic, it has enjoyed more use in commercial software implementations of GLM. The log-likelihood function is also used to assess fit, and is a term in the AIC goodness-of-fit statistic. The use of deviance over log-likelihood is a matter of preference and tradition. We generally calculate both statistics. Some software iterates based on the deviance statistic, then calculates the log-likelihood function at the end of the iteration process from the final values of μ and η, thereby providing a wider range of post-estimation fit statistics.

Remember that the matrix form of the estimation relationship between parameter and data for ordinary least squares regression (OLS) is

$$\beta_r = [X'X]^{-1}Xy \tag{2.37}$$

The formula we derived is simply a weighted version of the OLS algorithm. Since the IRLS algorithm is iterative, and cannot be solved in one step for models other than the basic Gaussian model, the response is redefined as a function of the linear predictor – hence the value of z rather than y. A consideration of the iteration or updating process is our next concern.

2.1.3 The IRLS fitting algorithm

The IRLS algorithm, using an expected information matrix, may take one of several forms. Not using subscripts, a standard schema is:

1 Initialize the expected response, μ, and the linear predictor, η, or $g(\mu)$.

2 Compute the weights as

$$W^{-1} = Vg'(\mu)^2 \tag{2.38}$$

where $g'(\mu)$ is the derivative of the link function and V is the variance, defined as the second derivative of the cumulant, $b''(\theta)$.

3 Compute a working response, a one term Taylor linearization of the log-likelihood function, with a standard form of (using no subscripts)

$$z = \eta + (y - \mu)g'(\mu) \tag{2.39}$$

4 Regress z on predictors $X_1 \ldots X_n$ with weights, W, to obtain updates on the vector of parameter estimates, β.

5 Compute η, or $X\beta$, based on the regression estimates.

6 Compute μ, or $E(y)$, as $g^{-1}(\mu)$.

7 Compute the deviance or log-likelihood function.

8 Iterate until the change in deviance or log-likelihood between two iterations is below a specified level of tolerance, or threshold.

Again, there are many modifications to the above scheme. However, most traditional GLM software implementations use methods similar to the above.

The GLM IRLS algorithm for the general case is presented in Table 2.1. The algorithm can be used for any member of the GLM family. We later demonstrate how substitution of specific functions into the general form for link, $g(\mu)$,

Table 2.1. *Standard GLM estimating algorithm (expected information matrix)*

```
Dev = 0
$\mu$ = (y + 0.5)/(m + 1)      /* binomial */
$\mu$ = (y + mean(y))/2        /* non − binomial */
$\eta$ = g($\mu$)                   /* linear predictor */
WHILE (abs($\Delta$Dev) > tolerance){
    w = 1/(V$g'^2$)
    z = $\eta$ + (y − $\mu$)$g'$ − offset
    $\beta$ = $(X'wX)^{-1}X'wz$
    $\eta$ = $X'\beta$ + offset
    $\mu$ = $g^{-1}(\eta)$
    Dev0 = Dev
    Dev = Deviance function
    $\Delta$Dev = Dev − Dev0
}
Chi2 = $\Sigma$ (y − $\mu$)$^2$/V($\mu$)
AIC = (−2LL + 2p)/n/* AIC is sometimes defined w/o $\eta$*/
BIC = Dev − (dof)ln(n)/* alternative definitions exist*/

Where p = number of model predictors + const
n = number of observations in model
dof = degrees of freedom(n − p)
```

inverse link, $g^{-1}(\eta)$, variance, *V*, and deviance or log-likelihood function create different GLM models. All other aspects of the algorithm remain the same, hence allowing the user to easily change models. Typically, with parameter estimates being of equal significance, the preferred model is the one with the lowest deviance, or highest log-likelihood, as well as the lowest AIC or BIC statistic. AIC is the acronym for the Aikake Information Criterion, which is based on the log-likelihood function; BIC represents the Baysean Information Criterion, which is usually based on the deviance value. These will be discussed at greater length later in the text. For count response models, statistics reflecting overdispersion need to be considered as well.

Table 2.1 provides a schematic view of the IRLS estimating algorithm, employing the traditional GLM expected information matrix for the calculation of standard errors. The algorithm relies on change in the deviance (*Dev*) value as the criterion of convergence. We have also added formulae for calculating the Pearson χ^2, the AIC, and BIC statistics.

Other terms needing explanation are *m*, the binomial denominator, g', the first derivative of the link, and the two variables: *Dev*0, the value of the deviance in the previous iteration, and Δ*Dev*, the difference in deviances between iterations. When the difference reaches a small value, or tolerance, somewhere in the range of 10^{-6}, iterations cease and the resultant parameter estimates, standard errors, and so forth are displayed on the screen.

There are of course variations of the algorithm found in the table. But the variations largely deal with whether iteration takes the form of a WHILE-DO loop, an IF-THEN loop, or some other looping technique. The form of Table 2.1 is that used for Stata's original GLM algorithm.

2.2 Newton–Raphson algorithms

In this section we discuss the derivation of the major estimating algorithms that have been used for the negative binomial and its many variations. We first address the Newton–Raphson algorithm, followed by the technique known as Fisher scoring, which is used with iteratively re-weighted least squares algorithms. The latter is the standard estimation method used for generalized linear models (GLM). We conclude by showing how the parameterization of the GLM mean, μ, can be converted to $X'\beta$.

2.2.1 Derivation of the Newton–Raphson

There are a variety of Newton–Raphson type algorithms. Few software programs use the base version, which is a simple root-finding procedure. Rather, they use one of several types of modified Newton–Raphson algorithm to produce maximum likelihood estimates of the parameters. A type of modified Marquardt algorithm is perhaps the most popular commercial software implementation. Moreover, complex models require methods other than Newton–Raphson or Marquardt. Quadrature methods are commonly used with random effects and mixed models; simulation-based methods are employed when no other estimating algorithm appears to work, or when estimation takes an inordinate amount of time.

In any event, a marked feature of the Newton–Raphson approach is that standard errors of the parameter estimates are based on the observed information matrix. However, the IRLS algorithm can be amended to allow calculation of observed information-based standard errors. The theoretical rationale and calculational changes required for being able to do this is detailed in Hardin and Hilbe (2001). GLM software such as SAS, Stata, XploRe, and LIMDEP allow the user to employ either the expected or observed information matrix. We note, though, that the more complex observed matrix reduces to the expected when the canonical link is used for the model. The log link is canonical for the Poisson, but not for the negative binomial. Therefore, which algorithm, and which modification, is used will have a bearing on negative binomial standard errors, and consequently on the displayed significance of parameter estimates. Differences between the two matrices are particularly notable when modeling small numbers of observations.

Derivation of terms for the estimating algorithm begin as a Taylor linearization and continue through the calculaton of the gradient, or first derivative of the likelihood function. Fisher scoring, used as the basis of the GLM estimating algorithm, calculates the matrix of second derivatives based on the expected information matrix

$$I = -E\left[\frac{\partial^2 \mathcal{L}}{\partial\beta_j \partial\beta_k'}\right] = E\left[\frac{\partial\mathcal{L}}{\partial\beta_j}\frac{\partial\mathcal{L}}{\partial\beta_k'}\right] \tag{2.40}$$

Newton–Raphson methodology, on the other hand, calculates the second derivatives of the likelihood on the basis of the observed information matrix, which allows estimation of likelihood-based algorithms other than the more limited exponential family form

$$H = \left[\frac{\partial^2 \mathcal{L}}{\partial\beta_j \partial\beta_k}\right] = \sum_{i=1}^{n}\frac{1}{\alpha_i(\phi)}\left[\frac{\partial}{\partial\beta_k}\right]\left\{\frac{y_i-\mu_i}{V(\mu_i)}\left(\frac{\partial\mu}{\partial\eta}\right)_i x_j x_i\right\} \tag{2.41}$$

Solved, the above becomes

$$\begin{aligned} &-\sum_{i=1}^{n}\underline{\frac{1}{\alpha\,(\phi)}}\;\underset{\text{EIM}}{\underline{\Bigg[\frac{1}{V(\mu_i)}\left(\frac{\partial\mu}{\partial\eta}\right)_i^2 - (\mu_i - y_i)}} \\ &\times \underset{\text{OIM}}{\underline{\underline{\left\{\frac{1}{V\,(\mu_i)^2}\left(\frac{\partial\mu}{\partial\eta}\right)_i^2\frac{\partial V\,(\mu_i)}{\partial\mu} - \frac{1}{V\,(\mu_i)}\left(\frac{\partial^2\mu}{\partial\eta^2}\right)_i\right\}\Bigg]}}}\;\underline{x_{ji}x_{ki}} \end{aligned} \tag{2.42}$$

Note that in the case of the canonical link terms including and following the term, $-(\mu - y)$, in the above formula cancel, reducing to the value of the expected information. Compare (2.42) with (2.21). A single line is drawn under the formula for the expected information matrix. The double line rests under the added terms required for the observed information.

Table 2.2 provides a schema for the modified Newton–Raphson algorithm used for the SAS/STAT GENMOD procedure.

We note here that this algorithm uses a convergence method based on the elementwise absolute differences between the vector of parameter estimates. βn represents the new β, βc represents the previously calculated, but current β. The intercept, sometimes represented as α_0, is included in comparison vectors. Elements of the entire parameter vector, $\alpha_0 + \beta_1 + \beta_2 + \cdots + \beta_n$, must

Table 2.2. *A Newton–Raphson algorithm*

```
$g = g(\mu) =$ link
$g' = g'(\mu) = 1^{st}$ derivative of link,wrt $\mu$
$g'' = g''(\mu) = 2^{nd}$ derivative of link, wrt $\mu$
$V = V(\mu) =$ variance
$V' = V(\mu) =$ derivative of variance
$m =$ binomial denominator
$y =$ response
$p =$ prior weight
$\phi =$ phi, a constant scale parameter
off = offset

$\mu = (y + mean(y))/2$ : binomial $= (y + 0.5)/(m + 1)$
$\eta = g$
$\beta n = 0$
while MAX(ABS($\beta n - \beta c$)) > tolerance {
     $\beta c = \beta n$
   $z = p(y - \mu)/(Vg'\phi)$                          < a column vector >
    $s = X'z$                                          < gradient >
     $We = p/(\phi V g'^2)$                            < weight : expected IM >
     $Wo = We + p(y - \mu)\{(Vg'' + V'g')/(V^2 g'^3 \phi)\}$   < observedIM >
     $Wo = diag(Wo)$                                   < diagonalize Wo >
    $H = -X'WoX$                                       < Hessian >
     $\beta n = \beta c - H^{-1}s$  :==:  $\beta c + (X'WoX)^{-1}\ X'(p(y - \mu))$
                    :==:  $(X'WoX)^{-1}\ X'W[\eta + (y - \mu)g']$
                    :==:  $(X'WoX)^{-1}\ X'Wz$ < if $z = \eta + (y - \mu)g'$ >
     $\eta = X'\beta n + off$                          < linear predictor >
     $\mu = g^{-1}(\eta)$                              < inverse link >
}
```

not change (much) from one iteration to another in order for the algorithm to converge.

The Newton–Raphson type algorithm takes the general form of Table 2.3. Initial values must be provided to the algorithm at the outset. Some software sets initial values to all zeros or all ones. Others calculate a simpler model, perhaps an OLS regression, to obtain initializing parameter estimates. Negative binomial algorithms typically use the parameter estimates from a Poisson model on the same data when initializing parameters.

The algorithm employs a maximum difference in log-likelihood functions as well as a maximum difference in parameter estimates as the criterion of convergence. The first terms in the algorithm are calculated by the means shown in our previous discussion. The observed information matrix is used to calculate *H*. Maximum likelihood parameter estimates are calculated in line four of the loop. The remaining terms deal with the updating process. One can observe the similarity in the algorithms presented in Table 2.2 and Table 2.3.

Table 2.3. *Maximum likelihood: Newton–Raphson*

```
Initialize $\beta$
WHILE (ABS($\beta$n − $\beta$o) > tol & ABS($\Lambda$n − $\Lambda$o) > tol) {
    g = $\partial$L/.$\partial\beta$
    H = $\partial^2$L/.$\partial\beta^2$
    $\beta$o = $\beta$n
    $\beta$n = $\beta$o − H$^{-1}$g
    Lo = Ln
    Ln
}
V($\mu$)
```

2.2.2 GLM with OIM

The GLM algorithm can be modified to accommodate the calculation of standard errors based on the observed information matrix, which is the inverse of the negative Hessian. The main feature of the alteration is an extension of the w term in the standard GLM. w is defined as

$$w = 1/\{V(\mu)g'(\mu)^2\} \tag{2.43}$$

where $V(\mu)$ is the variance function and $g'(\mu)^2$ is the square of the derivative of the link function. GLM terminology calls this a model weight. Weighting in terms of frequency weights are termed prior weights, and are entered into the algorithm in a different manner.

Returning to the modification of w necessary to effect an observed information matrix, w is amended and entered into the GLM IRLS algorithm as

$$w_0 = w + (y - \mu)\{V(\mu)g''(\mu) + V'(\mu)g'(\mu)\}/\{V(\mu)^2 g'(\mu)^3\} \tag{2.44}$$

so that it reflects (2.42). The full working algorithm is presented in Table 2.4. In terms of the Hessian, w_0 is defined as

$$H = \frac{\partial^2 \mathcal{L}}{\partial\beta_j \partial\beta_k'} \tag{2.45}$$

2.2.3 Parameterizing from μ to $X'\beta$

One finds parameterization of the GLM probability and log-likelihood functions, as well as other related formulae, in terms of both μ and $X'\beta$. μ is defined as the fitted value, or estimated mean, $E(y)$, whereas $X'\beta$ is the linear predictor. GLM terminology also defines the linear predictor as η. Hence $X'\beta = \eta$.

Transforming a log-likelihood function from a parameterization with respect to μ to that of $X'\beta$ is fairly simple. Making such a transformation is at times

Table 2.4. *Standard GLM estimating algorithm (observed information matrix)*

```
 Dev = 0
   $\mu$ = (y + 0.5)/(m + 1)              /* binomial */
   $\mu$ = (y + mean(y))/2                /* non - binomial */
   $\eta$ = g($\mu$)                           /* g; linear predictor */
WHILE (abs($\Delta$Dev) > tolerance)  {
   V = V($\mu$)
   V' = $1^{st}$ derivative of V
   g' = $1^{st}$ derivative of g
  g'' = $2^{nd}$ derivative of g
   w = 1/(V$g'^2$)
   z = $\eta$ + (y - $\mu$)g' - offset
  Wo = w + (y - $\mu$)(Vg'' + V'g')/($V^2 g'^3$)
   $\beta$ = (X'WoX)$^{-1}$X'Woz
   $\eta$ = X'$\beta$ + offset
   $\mu$ = $g^{-1}$($\eta$)
Dev0 = Dev
 Dev = Deviance function
$\Delta$Dev = Dev - Dev0
}
Chi2 = $\Sigma$ (y - $\mu$)$^2$/V($\mu$)
AIC  = (-2LL + 2p)/n
BIC  = Dev - (dof)ln(n)

Where p = number of model predictors + const
      n = number of observations in model
    dof = degrees of freedom (n - p)
```

required when one needs to estimate more-complex count models, such as zero-truncated or zero-inflated models.

The method involves substituting $X'\beta$ for η and substituting the inverse link function of μ at every instance of μ in the formula. For an example, we use the Poisson model. Shown without subscripts, the probability distribution function may be expressed as Equation (1.1)

$$f_y(y;\,\mu) = e^{-\mu}\mu^y/y!$$

The Poisson log-likelihood function may then be derived as

$$\mathcal{L}(\mu; y) = \Sigma\{y\ln(\mu) - \mu - \ln(y!)\} \tag{2.46}$$

Since the Poisson has a link defined as $\ln(\mu)$, the inverse link is

$$\mu = \exp(X'\beta) \tag{2.47}$$

Substituting into (2.46) yields

$$\begin{aligned}\mathcal{L}(\mu; y) &= \Sigma\{y\ln(\mu) - \mu - \ln(y!)\}\\ \mathcal{L}(\beta; y) &= \Sigma\{yX'\beta - \exp(X'\beta) - \ln(y!)\}\end{aligned} \tag{2.48}$$

We also see the above expressed as

$$\mathcal{L}(\beta; y) = \Sigma\{y(x\beta) - \exp(x\beta) - \ln\Gamma(y + 1)\} \tag{2.49}$$

where $y!$ can be calculated in terms of the log-gamma function, $\ln\Gamma(y + 1)$.

The first derivative of the Poisson log-likelihood function, in terms of $x\beta$, is

$$\Sigma\{yx - x\exp(x\beta)\} \tag{2.50}$$

or

$$\Sigma\{(y - \exp(x\beta))x\} \tag{2.51}$$

Solving for parameter estimates, β, entails setting the above to zero and solving

$$\Sigma\{(y - \exp(x\beta))x\} = 0 \tag{2.52}$$

All Newton–Raphson maximum likelihood algorithms use the $x\beta$ parameterization. μ is normally used with GLM-based estimation models. Both parameterizations produce identical parameter estimates and standard errors when the observed information matrix is used in the IRLS GLM algorithm. A similar transformation can be performed with negative binomial models. Except for the base NB2 model, all estimating algorithms use the $x\beta$ parameterization.

2.3 The exponential family

The probability function for the exponential family of distributions is commonly expressed as

$$f(y_i; \mu_i\phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\alpha(\phi)} + C(y_i; \phi)\right\} \tag{2.53}$$

where

θ_i is the canonical parameter or link function

$b(\theta_i)$ is the cumulant

$\alpha(\phi)$ is the scale parameter, set to one in discrete and count models

$C(y_i; \phi)$ is the normalization term, guaranteeing that the probability function sums to unity.

The exponential family form is unique in that the first and second derivatives of the cumulant, with respect to θ, respectively produce the mean and variance functions. The important point to remember is that if one can convert a probability function into exponential family form, its unique properties can be easily used to calculate the mean and variance, as well as facilitate estimation of parameter estimates based on the distribution. All members of the class of generalized linear models can be converted to the exponential form

$$b'(\theta_i)/d\theta = \text{mean}$$
$$b''(\theta_i)/d\theta = \text{variance}$$

2.4 Residuals for count response models

When modeling, using either full Newton–Raphson maximum likelihood or IRLS, it is quite simple to calculate the linear predictor as

$$x\beta = \eta = \alpha_0 + \beta_1 + \beta_2 + \cdots + \beta_n$$

Each observation in the model has a linear predictor value. For members of the exponential family, an easy relationship can be specified for the fitted value based on the linear predictor. The normal or Gaussian regression model has an identity canonical link, i.e. $\eta = \mu$. The canonical Poisson has a natural log link, hence $\eta = \ln(\mu) = \ln(\exp(x\beta)) = x\beta$. The traditional form of the negative binomial also has a log link, although it is not the canonical form. The linear predictor and fit are essential components of all residuals.

The basic or raw residual is defined as the difference between the observed response and the predicted or fitted response. When y is used to identify the response, $\hat{y}$ or μ is commonly used to characterize the fit. Hence

$$\text{Raw residual} = y - \hat{y} \quad \text{or} \quad y - \mu \quad \text{or} \quad y - E(y)$$

Other standard residuals used in the analysis of count response models include:

Pearson: $R^p = (y - \mu)/\text{sqrt}(V)$

Deviance: $R^d = \text{sgn}(y - \mu)\text{sqrt(deviance)}$
Note: $\Sigma(R^d)^2 =$ model deviance statistic

Standardized residuals: Divide residual by sqrt(1-hat), which aims to make its variance constant. hat = stdp^2*V

Studentized residuals: Divide residual by scale, ϕ. (See McCullagh and Nelder (1989), p. 396.)

Standardized–studentized: Divide by both standardized and studentized adjustments; e.g. $R^p : (y - \mu)/\{\phi V(\mu)^* \text{sqrt}(1 - h)\}$

In the above formulae we indicated the model distribution variance function as V, the hat matrix diagonal as *hat* and the standard error of the prediction as *stdp*.

A scale value, ϕ, is user defined, and employed based on the type of data being modeled. See McCullagh and Nelder (1989, p. 396), for a detailed account of these residuals.

We mentioned earlier that the Anscombe residual (Anscombe, 1972) has values close to those of the standardized deviance. There are times, however, when this is not the case, and the Anscombe residual performs better than R^d. Anscombe residuals attempt to normalize the residual so that heterogeneity in the data, as well as outliers, become easily identifiable.

Anscombe residuals use the model variance function. The variance functions for the three primary count models are

Poisson:	$V = \mu$
Geometric:	$V = \mu(1+\mu)$
NB2:	$V = \mu + \alpha\mu^2 \quad \text{or} \quad \mu(1+\alpha\mu)$

Anscombe defined the residual which later became known under his name as

$$R^A = \frac{A(y_i) - A(\mu_i)}{A'(\mu_i)\text{sqrt}(V(\mu_i))} \tag{2.54}$$

where

$$A(.) = \int d\mu_i / V^{1/3}(\mu_i) \tag{2.55}$$

The calculated Anscombe residuals for the three basic count models are, without showing subscripts,

Poisson:

$$3(y^{2/3} - \mu^{2/3})/(2\mu^{1/6}) \tag{2.56}$$

Geometric:

$$\frac{\{3\{(1+y)^{2/3} - (1+\mu)^{2/3}\} + 3(y^{2/3} - \mu^{2/3})\}}{2(\mu^2+\mu)^{1/6}} \tag{2.57}$$

Negative binomial:

$$\frac{\{3/a\{(1+\alpha y)^{2/3} - (1+\alpha\mu)^{2/3}\} + 3(y^{2/3} - \mu^{2/3})\}}{2(\alpha\mu^2+\mu)^{1/6}} \tag{2.58}$$

The negative binomial Anscombe has also been calculated in terms of the hypergeometrix2F1 function. See Hilbe (1993a) and Hardin and Hilbe (2001) for a complete discussion.

$$y^{2/3}H(2/3, 1/3, 5/3, y/\alpha) - \mu^{2/3}H(2/3, 1/3, 5/3, \mu/\alpha) \tag{2.59}$$

$$= 2/3B(2/3, 2/3)\{y - B_{\text{I}}(2/3, 2/3, \mu/\alpha)\} \tag{2.60}$$

where H is the hypergeometrix2F1 function, B is the beta function, and B_I is the incomplete beta function. Hilbe (1994) and Hardin and Hilbe (2001) show that the two-term beta function has the constant value of 2.05339. It is also noted that the value of α is the negative binomial ancillary parameter.

We are ready to proceed to the more detailed derivation of the Poisson model, which is considered to be the base or standard count response regression model.

2.5 Summary

We have discussed the two foremost methods used to estimate count response models – Newton–Raphson and Fisher scoring. Both are maximum likelihood methods, using the likelihood function, or its derived deviance function, as the basis for estimation. That is, both methods involve the maximization of the likelihood score function in order to estimate parameter values. Standard errors are obtained from the Hessian, or, more correctly, from the information matrix, which is calculated as the second derivative of the likelihood function.

We also mentioned that most software applications use some variation of the traditional or basic Newton–Raphson algorithm. For our purposes, we refer to these methods collectively as full maximum likelihood methods of estimation. These methods produce standard errors based on the observed information matrix.

Fisher scoring is typically based on an iteratively re-weighted least squares (IRLS) algorithm. It is the algorithm traditionally used for estimation of generalized linear models (GLM). Fisher scoring is a simplification of the full maximum likelihood method that is allowed due to the unique properties of the exponential family of distributions, of which all GLM members are instances. Standard errors produced by this method are generally based on the expected information matrix. In the case of canonically linked GLMs, the observed information matrix reduces to the expected (see Equation (2.42)), resulting in standard errors of the same value. For example, a log-linked Poisson, which is the canonical link, can be estimated using a form of the Newton–Raphson algorithm employing the observed information matrix, or by considering it as a member of the family of generalized linear models, using the expected information matrix. In either case the calculated standard errors will be identical, except for perhaps very small rounding errors. Non-canonical linked GLMs will produce standard errors that are different from those produced using full maximum likelihood methods.

It is possible, though, to adjust the GLM IRLS algorithm so that the observed information matrix is used to calculate standard errors rather than the expected. We showed how this can be accomplished, and how it results in standard errors that are the same no matter which of the two major methods of estimation are used.

Finally, we addressed the type of residuals that are used to evaluate the worth of count response models. Residuals based on GLM methodology provide the statistician with a rather wide range of evaluative capability. It is recommended that standardized deviance and Anscombe residuals be used with the fitted values, μ, to determine if the data are appropriate for the model used in estimation.

Exercises

1 What type of models are more appropriately estimated using a Newton–Raphson type of algorithm rather than IRLS?

2 State the essential difference between the expected information matrix and the observed information matrix? Under what conditions do they reduce to the same formula?

3 Using a higher language such as R, Stata, SAS-IML, or similar facility, develop a functional IRLS algorithm based on Table 2.1 with the following values

$$g(\mu) = \ln(\mu) \qquad g^{-1}(\eta) = \exp(\eta) \qquad V = \mu$$
$$\text{Deviance} = 2\Sigma\{y \ln(y/\mu) - (y - \mu)\}$$

The expression, g', in Table 2.1 is the same as $g'(\mu)$ or the first derivative of $g(\mu)$ wrt μ. Use the non-binomial initialization for μ and include the Chi2, AIC, and BIC statistics only as a bonus.

4 Using the data *ex*2_4 given below, together with the algorithm developed in question 3, model y on x_1 and x_2. Determine the parameter estimates and standard errors for x_1 and x_2.

Data	*y*	*x*1	*x*2
	1	1	61
	1	1	65
	2	1	59
	3	1	52
	4	1	56
	4	1	67
	5	1	63
	5	1	58
	8	1	56
	8	0	58

5 Amend the algorithm developed in question 3 so that the observed information matrix is used as in Table 2.4. Run the model using the same data as in question 4. Why is there no difference in results?

6 Show why the observed information matrix collapses to the expected information matrix of a canonically linked GLM.

7 AIC statistics can be calculated for all maximum likelihood models. Why is it the case that not all maximum likelihood models can be evaluated using a BIC statistic?

3
Poisson regression

Poisson regression is the standard or base count response regression model. We have seen in previous discussion that other count models deal with data that violate the assumptions carried by the Poisson model. Since the model does play such a central role in count response modeling, we begin with an examination of its derivation and structure, as well as how it can be parametermized to model rates. The concept of overdispersion is introduced in this chapter, together with two tests that have been used to assess its existence and strength.

3.1 Derivation of the Poisson model

A primary assumption is that of equidispersion, or the equality of the mean and variance functions. When the value of the variance exceeds that of the mean, we have what is termed overdispersion. Negative binomial regression is a standard way to deal with certain types of Poisson overdispersion; we shall find that there are a variety of negative binomial based models, each of which address the manner in which overdispersion has arisen in the data. However, to fully appreciate the negative binomial model and its variations, it is important to have a basic understanding of the derivation of the Poisson as well as an understanding of the logic of its interpretation.

Maximum likelihood models, as well as the canonical form members of generalized linear models, are ultimately based on an estimating equation derived from a probability distribution. In the case of the Poisson, the probability function can be expressed as

$$f_y(y; \mu) = e^{-\mu}\mu^y/y! \tag{3.1}$$

for $y = \{0, 1, \dots\}$ and $\mu > 0$.

The conversion of the Poisson PDF to log-likelihood form is accomplished by casting it as a member of the exponential family of distributions given, without subscripts, as

$$f(y;\mu,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{\alpha(\phi)} + C(y;\phi)\right\} \tag{2.53}$$

Equation (3.1) may be caste into exponential family form for a calculated random sample as

$$f_y(y;\mu) = \Sigma\{\exp(y\ln(\mu) - \mu - \ln(y!)\} \tag{3.2}$$

The log-likelihood function is a transformation of the probability function in which the parameters are estimated to make the given data most likely

$$\mathcal{L}(\mu;y) = \Sigma\{y\ln(\mu) - \mu - \ln(y!)\} \tag{3.3}$$

With subscripts to indicate the individual observation contribution to the overall log-likelihood function, we have

$$\mathcal{L}(\mu_i;y_i) = \Sigma\{y_i\ln(\mu_i) - \mu_i - \ln(y_i!)\} \tag{3.4}$$

The canonical link and cumulant terms can then be abstracted from the above, without showing subscripts, as

$$\text{LINK:}\ \ \theta = \ln(\mu) = \eta \tag{3.5}$$

$$\text{CUMULANT:}\ \ b(\theta) = \mu \tag{3.6}$$

The inverse link is a re-interpretation of μ with respect to η, the linear predictor. Transformation yields

$$\text{INVERSE LINK:}\ \eta = \exp(\mu) \tag{3.7}$$

Recalling that the exponential family mean is defined as the first derivative of the cumulant with respect to θ, and the variance as the second derivative with respect to θ, we calculate the Poisson mean and variance as

$$\text{MEAN:}\ b(\theta) = \frac{\partial b}{\partial\mu}\frac{\partial\mu}{\partial\theta} = (1)(\mu) = \mu \tag{3.8}$$

VARIANCE:

$$b''(\theta) = \frac{\partial^2 b}{\partial\mu^2}\left(\frac{\partial\mu}{\partial\theta}\right)^2 + \frac{\partial b}{\partial\mu}\frac{\partial^2\mu}{\partial\theta^2} = (0)(1) + (\mu)(1) = \mu \tag{3.9}$$

We see the equality of the Poisson mean and variance functions.

Since the derivative of the link is important to the estimating algorithm, we have

$$\text{DERIVATIVE OF LINK:} \quad \frac{\partial\theta}{\partial\mu} = \frac{\partial\{\ln(\mu)\}}{\partial\mu} = \frac{1}{\mu} \tag{3.10}$$

Recall that Equation (2.47) in the previous chapter specified the Poisson mean, μ, as equal to $\exp(x'\beta)$. We can therefore make the following translation such that μ takes the value $\exp(x'\beta)$: Note that the $x'\beta$ parameterization is used in all full maximum likelihood estimating algorithms

$$\frac{1}{\mu} = \frac{1}{\exp(x'\beta)} \tag{3.11}$$

The Poisson log-likelihood, parameterized in terms of $x'\beta$, is therefore given, with subscripts, as

$$\mathcal{L}(\beta_i; y_i) = \Sigma\{y_i(x_i\beta) - \exp(x_i\beta) - \ln(y_i!)\} \tag{3.12}$$

or

$$\mathcal{L}(\beta_i; y_i) = \Sigma\{y_i(x_i\beta) - \exp(x_i\beta) - \ln\Gamma(y_i + 1)\} \tag{3.13}$$

When the response has a value of zero, the log-likelihood function reduces to

$$\mathcal{L}(\beta_i; y_i = 0) = -\exp(x_i\beta) \tag{3.14}$$

Returning to the traditional GLM parameterization of the mean as μ, the GLM deviance function is defined as

$$\text{DEV} = 2\Sigma\{\mathcal{L}(y_i; y_i) - \mathcal{L}(\mu_i; y_i)\} \tag{3.15}$$

The deviance is a measure of the difference between the full, or saturated, and model likelihoods. It is a likelihood ratio test of the full to the model likelihoods. Traditionally, the deviance statistic has been used as the basis of convergence for GLM algorithms (see Table 3.1). It has also been used as a goodness-of-fit statistic, with lower positive values representing a better fitted model.

Substituting the appropriate Poisson terms into the saturated likelihood parameterization entails substituting the value of y for each instance of μ. This gives us

$$\begin{aligned} \text{DEV} &= 2\Sigma\{y_i \ln(y_i) - y_i - y_i \ln(\mu_i) + \mu_i\} && (3.16)\\ &= 2\Sigma\{y_i \ln(y_i/\mu_i) - (y_i - \mu_i)\} && (3.17) \end{aligned}$$

We may recall Table 2.1, which schematized the generic IRLS algorithm. We can now substitute the above Poisson statistics into Table 2.1 to develop a paradigm IRLS-type Poisson regression.

Table 3.1. *Poisson regression algorithm*

```
$\mu$ = (y + mean(y))/2
$\eta$ = ln($\mu$)
WHILE(abs($\Delta$dev)tolerance)   {
$u$= (y − $\mu$)/$\mu$
w = $\mu$
z = $\eta$ + $u$ − offset
$\beta$= (X'wX)$^{-1}$X'wz
$\eta$ = X'$\beta$ + offset
$\mu$ = exp($\eta$)
oldDev = dev
dev = 2$\Sigma${yln(y/$\mu$) − (y−$\mu$)}
$\Delta$dev = dev − oldDev
}
    Chi2 =$\Sigma$(y−$\mu$)$^2$/$\mu$
    AIC = (−2*$\Sigma$(yln($\mu$)−$\mu$−lngamma(y + 1)) + 2*p)/n
    BIC = 2$\Sigma$(y*ln(y/$\mu$) − (y−$\mu$)) − df*ln(n)
    /* n = number of model observations */
    /* p = number of model predictors */
    /* df = model degrees of freedom */
```

The Poisson regression model is also considered as a nonlinear regression to be estimated using maximum likelihood methods. But in order to do so, we must calculate the derivatives of the log-likelihood function, which define the gradient and observed Hessian matrix.

The gradient vector, or first derivative of the Poisson log-likelihood function with respect to parameters β, is calculated as

$$\frac{\partial(\mathcal{L}(\beta; y_i))}{\partial\beta} = \Sigma(y_i - \exp(x_i\beta))x_i \tag{3.18}$$

Setting Equation (3.18) to zero, such that

$$\Sigma(y_i - \exp(x_i\beta))x_i = 0 \tag{3.19}$$

provides for the solution of the parameter estimates.

The Hessian is calculated as the negative inverse of the second derivative of the log-likelihood function

$$\frac{\partial(\mathcal{L}(\beta; y_i))}{\partial\beta\partial\beta'} = [-\Sigma(\exp(x_i\beta))x_i x_j]^{-1} \tag{3.20}$$

The square-roots of the respective terms on the diagonal of the negative inverse Hessian are the values of parameter standard errors. A Newton–Raphson algorithm can be used for maximum likelihood estimation of parameters

$$\beta_{i+1} = \beta_i - H^{-1}g \tag{3.21}$$

Poisson models are typically used to either summarize predicted counts based on a set of explanatory predictors, or are used for interpretation of exponentiated estimated slopes, indicating the expected change or difference in the incidence rate ratio of the outcome based on changes in one or more explanatory predictors. An example will demonstrate how each of these modeling concerns appears in fact.

This example comes from Medicare hospital length of stay data from the state of Arizona. The data are limited to only one diagnostic group. In addition, patient data have been randomly selected to be part of this data set [data = *medpar*].

The model response is *los*, length of stay, a count of the days each patient spent in the hospital. Predictors include *hmo*, whether the patient belonged to an HMO (1/0), *white*, if the patient identifies themselves as primarily Caucasian (1/0), and a three level factor predictor, *type**, related to the type of admission: 1 = elective, 2 = urgent, and 3 = emergency. *type* = 1 is specified as the referent.

```
Generalized linear models           No. of obs       =       1495
Optimization     :  ML              Residual df      =       1490
                                    Scale parameter  =          1
Deviance         =  8142.666001     (1/df) Deviance  =   5.464877
Pearson          =  9327.983215     (1/df) Pearson   =  6.260391<
                                    AIC              =   9.276131
Log likelihood   =  −6928.907786    BIC              =  −2749.057
=================================================================
-----------------------------------------------------------------
             |            OIM
         los |    Coef.   Std.Err.     z     P>|z|    [95% Conf.Interval]
-------------+---------------------------------------------------
         hmo | −.0715493   .023944   −2.99   0.003   −.1184786    −.02462
       white |  −.153871  .0274128   −5.61   0.000   −.2075991   −.100143
       type2 |  .2216518  .0210519   10.53   0.000    .1803908   .2629127
       type3 |  .7094767   .026136   27.15   0.000    .6582512   .7607022
        cons |  2.332933  .0272082   85.74   0.000    2.279606    2.38626
-----------------------------------------------------------------
=================================================================
```

In the parameterization that follows, the coefficients are exponentiated to assess the relationship between the response and predictors as incidence rate ratios. I have altered the formatting of descriptive and fit statistics above the table of parameter estimates, standard errors, and so forth. No statistical values have been deleted; empty lines of output have been merged. I follow this method throughout the text.

```
        |            OIM
    los |     IRR   Std.Err.      z    P>|z|   [95% Conf.Interval]
--------+----------------------------------------------------------
    hmo | .9309504  .0222906   −2.99   0.003   .8882708    .9756806
  white | .8573826  .0235032   −5.61   0.000   .8125327     .904708
  type2 | 1.248137  .0262756   10.53   0.000   1.197685    1.300713
  type3 | 2.032927  .0531325   27.15   0.000   1.931412    2.139778
```

We observe that the number of days a patient is in the hospital is increased by 24% if the patient entered as an urgent rather than as an elective admission. A patient stays twice as long as an elective if they entered as an emergency admit.

We also see that HMO patients stay in the hospital slightly less time than do non-HMO patients – 7% less time; non-white patients stay in the hospital about 14% longer than do white patients.

On the surface the model appears acceptable. However, the Pearson dispersion value is 6.26, far exceeding unity. The dispersion statistic has been indicated with a "<" to the immediate right of the statistic. Thus it appears that the model is overdispersed. The Lagrange multiplier and Z tests also indicate overdispersion (not shown).

Expected counts (of days of stay) can be calculated for a user defined set of predictor values. We can predict on the basis of the model that an HMO non-white patient entering the hospital as an urgent admission has a predicted length of stay of 12 days.

This is particularly easy to calculate since all predictors are binary. Thus

$$\text{\_constant} + \beta_1{}^*1 - \beta_2{}^*1 = \text{linear predicator}$$
$$2.3329 + .2217 - .07155 = 2.48305$$

which is the value of the linear predictor. Next we apply the inverse link to determine the fitted value, μ, which in this case is a predicted count

$$\exp(2.48305) = 11.977741 \quad \text{or} \quad 12$$

Unfortunately the data are not well fitted, as we shall determine. This gives rise to the following observed values of *los*, given the above criteria:

```
los | hmo    white     type1     type2     type3     mu
----+--------------------------------------------------------
  1 | 1      0         0         1         0         11.97757
 14 | 1      0         0         1         0         11.97757
  3 | 1      0         0         1         0         11.97757
 19 | 1      0         0         1         0         11.97757
```

The four observed values are 1, 3, 14, and 19. The mean of these counts is

$$(1 + 3 + 14 + 19)/4 = 9.25$$

A quick assessment of the fit of this model tells us that it is overdispersed.

An easy way to check for possible overdispersion in a Poisson model is to look at the Pearson-based dispersion statistic that is typically displayed in model output. The dispersion is defined as the ratio of the Pearson statistic to the degrees of freedom, or the number of observations less predictors. In this case we have $9327.983/1490 = 6.260$. Such a value for a model consisting of some 1500 observations is clearly excessive. Ideally, if there is no overdispersion in the data, the dispersion statistic will have a value of 1.0.

A value of near 6.25 indicates overdispersion, but only additional investigation will inform us if it is real, or only apparent. We shall address this subject in depth in the following chapter.

3.2 Parameterization as a rate model

We briefly addressed the rate parameterization of the Poisson model in Chapter 1. Although μ is sometimes said to be an intensity or rate parameter, it is such only when thought of in conjunction with a constant coefficient, t. The rate parameterization of the Poisson PDF can be expressed as

$$f(y_i; \mu_i) = e^{-t_i \mu_i}(t_i \mu_i)^{y_i}/y_i! \tag{1.2}$$

t represents the length of time, or exposure, during which events or counts occur. t can also be thought of as an area in which events occur, each associated with a specific count. For instance, when using a Poisson model with disease data, $t_i \mu_i$ can be considered as the rate of disease incidence in specified geographic areas, each of which may differ from other areas in the population. Again, the incidence rate of hospitalized bacterial pneumonia patients can be compared across counties within the state. A count of such hospitalizations divided by the population size of the county, or by the number of total hospitalizations for all diseases, results in the incidence rate ratio (IRR) for that county. When $t = 1$, the model is understood as applying to individual counts without a consideration of size. Many commercial software applications indicate exponentiated Poisson coefficients as incidence rate ratios. IRR is also used with exponentiated negative binomial coefficients.

When employing a rate parameter to a Poisson model, statisticians enter the natural log of t as an offset into the estimating algorithm. The fitted value is expressed as

$$\mu_i = \exp(x_i \beta + \ln(t_i)) \tag{3.22}$$

Tables 2.1 and 2.4 show how $\ln(t)$ is entered into the estimating algorithm. Called an offset, $\ln(t)$ is entered into the algorithm as a constant.

An example of a Poisson model parameterized as a rate is provided below. The data are from the Canadian National Cardiovascular Disease registry called, FASTRAK. They have been grouped by covariate patterns from individual observations. The response is *die*, which is a count of the number of deaths of patients having a specific pattern of predictors. Predictors are *anterior*, which indicates if the patient has had a previous anterior myocardial infarction; *hcabg*, if the patient has a history of having had a CABG procedure; and *killip* class, a summary indicator of the health of the patient, with increasing values indicating increased disability. The number of observations sharing the same pattern of covariates is recorded in the variable *case*. This value is log-transformed and entered into the model as an offset.

```
. glm die anterior hcabg kk2-kk4, fam(poi) eform lnoffset(cases)

Generalized linear models          No. of obs         =          15
Optimization     :   ML            Residual df        =           9
                                   Scale parameter    =           1
Deviance         =   10.93195914   (1/df) Deviance    =    1.214662
Pearson          =   12.60791065   (1/df) Pearson     =    1.400879
                                   AIC                =     4.93278
Log likelihood   =   -30.99584752  BIC                =   -13.44049

------------------------------------------------------------------------
            |              OIM
        die |       IRR    Std. Err.      z    P>|z|   [95% Conf. Interval]
------------+-----------------------------------------------------------
   anterior | 1.963766    .3133595    4.23   0.000   1.436359    2.684828
      hcabg | 1.937465    .6329708    2.02   0.043   1.021282    3.675546
        kk2 | 2.464633    .4247842    5.23   0.000    1.75811    3.455083
        kk3 | 3.044349    .7651196    4.43   0.000    1.86023    4.982213
        kk4 | 12.33746    3.384215    9.16   0.000   7.206717    21.12096
      cases | (exposure)
------------------------------------------------------------------------
```

The Pearson dispersion is relatively low at 1.40, but given a total observation base of 5388, the added 40% overdispersion may represent a lack of model fit. We shall delay this discussion until the next section where we deal specifically with models for overdispersed data.

3.3 Testing overdispersion

The concept of overdispersion is central to the understanding of negative binomial models. Nearly every application of the negative binomial is in response to perceived overdispersion in a Poisson model. We shall address the problem of ascertaining whether indicators of overdispersion represent real overdispersion in the data, or only apparent. Apparent overdispersion can usually be accommodated by various means in order to eradicate it from the model. However, real overdispersion is a problem affecting the reliability of both the model parameter estimates and fit in general.

We showed one manner in which overdispersion could be detected in a Poisson model. We will address other methods in the next chapter. However, two related, yet well-used, tests are at times provided in commercial software applications. These are the Z and Lagrange multiplier tests.

A score test to evaluate whether the amount of overdispersion in a Poisson model is sufficient to violate the basic assumptions of the model is defined as:

$$\text{Z TEST}: \quad Z_i \frac{(y_i - \mu_i)^2 - y_i}{\mu_i\ \text{sqrt}(2)} \tag{3.23}$$

The test is *post-hoc*, i.e. performed subsequent to modeling the data. Using the *medpar* data set as earlier delineated, we first model the data using maximum likelihood Poisson regression:

```
Poisson Regression                  Number of obs   =    1495
                                    Wald chi2(4)    =  866.32
Log likelihood  =   −6928.9078      Prob > chi2     =  0.0000

------------------------------------------------------------------
   los |    IRR     Std. Err.      z     P>|z|   [95% Conf.  Interval]
------------------------------------------------------------------
   hmo | .9309504   .0222906    −2.99   0.003     .8882708    .9756806
 white | .8573825   .0235032    −5.61   0.000     .8125327     .904708
 type2 | 1.248137   .0262756    10.53   0.000     1.197685    1.300713
 type3 | 2.032927   .0531325    27.15   0.000     1.931412    2.139778
------------------------------------------------------------------

AIC Statistic   =      9.276      BIC Statistic   = −2749.057
Deviance        =   8142.666      Dispersion      =      5.465
LM Value        = 62987.860 <     LM Chi2(1)      =      0.000
```

Note: I have included additional fit statistics under the table of parameter estimates that are not in the commercial software output. The software to produce this output is available on the website for this book.

Using Stata, the statistic may be calculated using:

```
predict xb
gen mu = exp(xb)
gen double z=((los-mu)^2-los)/ (mu*sqrt(2))
regress z
```

```
  Source |       SS      df    MS             Number of obs   =    1495
---------+------------------------------------------------------------
         |                                    F(0, 1494)      =    0.00
   Model |         0      0                   Prob > F        =
Residual | 348013.82   1494   232.940977      R-squared       =  0.0000
---------+------------------------------      Adj R-squared   =  0.0000
   Total | 348013.82   1494   232.940977      Root MSE        =  15.262
---------+------------------------------------------------------------
       z |  Coef.     Std. Err.    t      P>|t|    [95% Conf. Interval]
---------+------------------------------------------------------------
   _cons | 3.704561   .394732     9.39   0.000    2.930273   4.478849
```

The Z score test is 3.7, with a t-probability of <0.0005. Z tests the hypothesis that the Poisson model is overdispersed. In practice, it tests whether the data should be modeled as Poisson or negative binomial. This example indicates that the hypothesis of no overdispersion is rejected, i.e. it is likely that real overdispersion does exist in the data.

The Lagrange multiplier test is given as:

$$\text{LAGRANGE MULTIPLIER TEST:}\quad \chi^2 = \frac{(\Sigma_i \mu_i^2 - ny)^2}{2\Sigma_i \mu_i^2}, \text{ with 1 dof} \tag{3.24}$$

Again, using Stata commands to calculate the statistic, we have:

```
. summ los, meanonly    /* solving for Lagrange
. Multplier */
. scalar nybar = r(sum)
. gen double musq = mu*mu
. summ musq, meanonly
. scalar mu2 = r(sum)
. scalar chival = (mu2-nybar)^2/(2*mu2)
. display ″LM value =″ chival n ″P-value =″
  chiprob(1,chival)
LM value = 62987.861
P-value = 0
```

With one degree of freedom, the test appears to be significant – the hypothesis of no overdispersion is again rejected. See model output above.

3.4 Summary

The Poisson model is the paradigm or basic count response model. We discussed the derivation of the model and how the basic Poisson algorithm can be amended to allow estimation of rate models, i.e. how many counts are in a certain defined area or over various time periods. The rate parameterization of the Poisson model is also appropriate for modeling counts that are weighted by person years.

A central distributional assumption of the Poisson model is the equivalence of the Poisson mean and variance. This assumption is rarely met with real data. Usually the variance exceeds the mean, resulting in what is termed as *overdispersion*. Underdispersion occurs when the variance is less than the nominal mean, but this rarely occurs in practice. Overdispersion is, in fact, the norm, and gives rise to a variety of other models that are extensions of the basic Poisson model.

Negative binomial regression is nearly always thought of as the model that is to be used instead of Poisson when overdispersion is present in the data. Because overdispersion is so central to the modeling of counts, we next address it, and investigate how we determine if it is real or only apparent.

Exercises

1 How are the constants between two Poisson models related in which the response of one has values five times greater than the other? How are the Pearson dispersion statistics related? Both models have identical predictors. Formulate a general principal for these two relationships.

2 Amend the GLM-based Poisson regression algorithm as shown in Table 3.1 so that the canonical natural log link is changed to the identity link. (a) Amend it so that standard errors are based on the expected information matrix. (b) Amend it so that standard errors are based on the observed information matrix [difficult].

3 How does the offset in a Poisson model relate to the binomial response denominator in a grouped logistic regression model?

4 What is the relationship between the Pearson χ^2 statistic and the Lagrange multiplier test for Poisson overdispersion?

5 Model the HIV data below with a Poisson model. The response is *infec*, the number of patients infected. The natural log of *cases* is the offset, with *cd4* and *cd8* as two explanatory predictors. Since the values given for the predictors represent ranges of marker values, they should be factored into three levels each. Prepare a well-fitted Poisson model.

infec	cases	cd4	cd8
1	1	0	2
2	2	1	2
4	7	0	0
4	12	1	1
1	3	2	2
2	7	1	0
0	2	2	0
0	13	2	1

6 Show how the Poisson PDF is the same as the negative binomial with an ancillary parameter of zero.

7 The following data *horsekick* are the famous “horse-kick” data set collected by Bortkewitsch for the period 1875–1894. His final data included the frequency of horse kicks for ten corps of Prussian soldiers over the 20-year period. Determine if a Poisson model is an appropriate measure of the data.

Deaths	0	1	2	3	4	>=5
Frequency	109	65	22	3	1	0

4
Overdispersion

This chapter can be considered as a continuation of the former. Few real-life Poisson data sets are truly equi-dispersed. Overdispersion to some degree is inherent to the vast majority of Poisson data. Thus, the real question deals with the amount of overdispersion in a particular model – is it statistically sufficient to require a model other than Poisson? This is a foremost question we address in this chapter, together with how we differentiate between real and apparent overdispersion.

4.1 What is overdispersion?

Not all overdispersion is real; apparent overdispersion can sometimes be identified and the model amended to eliminate it. We first address the difference between real and apparent overdispersion, and what can be done about the latter.

1 What is overdispersion?
Overdispersion in Poisson models occurs when the response variance is greater than the mean.

2 What causes overdispersion?
Overdispersion is caused by positive correlation between responses or by an excess variation between response probabilities or counts. Overdispersion also arises when there are violations in the distributional assumptions of the data.

3 Why is overdispersion a problem?
Overdispersion may cause standard errors of the estimates to be underestimated; i.e., a variable may appear to be a significant predictor when it is in fact not significant.

4 How is overdispersion recognized?

A model may be overdispersed if the value of the Pearson (or χ^2) statistic divided by the degrees of freedom (dof) is greater than 1.0. The quotient of either is called the dispersion. Small amounts of overdispersion are of little concern; however, if the dispersion statistic is greater than 1.25 for moderate sized models, then a correction may be warranted. Models with large numbers of observations may be overdispersed with a dispersion statistic of 1.05.

5 What is apparent overdispersion; how may it be corrected?

Apparent overdispersion occurs when:

(a) the model omits important explanatory predictors;
(b) the data include outliers;
(c) the model fails to include a sufficient number of interaction terms;
(d) a predictor needs to be transformed to another scale; or when
(e) the assumed linear relationship between the response and the link function and predictors is mistaken, i.e. the link is misspecified.

4.2 Handling apparent overdispersion

We can show the impact of the various causes of apparent overdispersion, delineated in 5(a)–(e) above, by creating simulated data sets. Each constructed data set will entail a specific cause for the overdispersion observed in the display of model output.

We shall first create a base Poisson data set consisting of three normally distributed predictors.

CREATION OF A SIMULATED BASE POISSON MODEL

Construct a data set with the following constructed predictors:

```
Constant == 1.00        x1 == 0.50
x2 == −0.75             x3 == 0.25
```

Stata code to create the simulated data consists of the following:

```
. set obs 10000
. gen x1 = invnorm(uniform())
. gen x2 = invnorm(uniform())
. gen x3 = invnorm(uniform())
. gen xb=1 +.5*x1 +.75*x2 +.25*x3
. genpoisson y, xbeta(xb)
```

Created response variable, *y*, has an observed count distribution appearing as (excluding counts greater than 25):

y	Freq.	Percent	Cum.
0	1458	14.58	14.58
1	1808	18.08	32.66
2	1507	15.07	47.73
3	1160	11.60	59.33
4	899	8.99	68.32
5	735	7.35	75.67
6	509	5.09	80.76
7	397	3.97	84.73
8	301	3.01	87.74
9	208	2.08	89.82
10	174	1.74	91.56
11	151	1.51	93.07
12	104	1.04	94.11
13	86	0.86	94.97
14	63	0.63	95.60
15	68	0.68	96.28
16	53	0.53	96.81
17	50	0.50	97.31
18	40	0.40	97.71
19	35	0.35	98.06
20	30	0.30	98.36
21	20	0.20	98.56
22	15	0.15	98.71
23	18	0.18	98.89
24	11	0.11	99.00
25	12	0.12	99.12

y is next modeled on the three randomly generated predictors:

```
. glm y x1 x2 x3, nolog fam(poi)

Generalized linear models          No. of obs       =     10000
Optimization  :  ML                Residual df      =      9996
                                   Scale parameter  =         1
Deviance      =  10640.17865       (1/df) Deviance  =  1.064444
Pearson       =  9871.053244   =>  (1/df) Pearson   =  .9875003
                                   AIC              =  3.718926
Log           =  −18590.62923      BIC              = −81426.38
likelihood

                    OIM
    y |   Coef.    Std. Err.      z      P>|z|    [95% Conf. Interval]
------+-----------------------------------------------------------------
   x1 |  .5005922   .0049171   101.81   0.000     .4909548    .5102296
   x2 | −.745506     .00495   −150.61   0.000    −.7552079   −.7358042
   x3 |  .2496633   .0048491    51.49   0.000     .2401593    .2591673
_cons |  1.002132   .0066982   149.61   0.000     .9890042    1.015261
```

Since the data are randomly distributed, other simulated data sets will have slightly different values. If we ran several hundred simulated models, however, we would find that the parameter estimates would equal the values we assigned them, and that the Pearson dispersion statistic, defined as the Pearson statistic divided by the model degrees of freedom, would equal 1.0. Note the Pearson dispersion statistic in the above model is 0.9875, with the parameter estimates approximating the values we specified.

DELETE PARAMETER X1

We now omit predictor X1, and again model the data.

```
. glm y x2 x3, nolog fam(poi)

Generalized linear models     No. of obs       =      10000
Optimization  : ML            Residual df      =       9997
                              Scale parameter  =          1
Deviance       = 21047.48811  (1/df) Deviance  =    2.10538
Pearson        = 21655.83485  (1/df) Pearson   =   2.166233
                              AIC              =   4.759457
Log            = −23794.28397 BIC              =  −71028.28
likelihood
```

```
─────────────────────────────────────────────────────────────────────
             │              OIM
           y │    Coef.    Std. Err.      z  P>|z|  [95% Conf. Interval]
─────────────┼───────────────────────────────────────────────────────
          x2 │ −.7382381    .0049227  −149.97  0.000  −.7478865  −.7285898
          x3 │  .2374987    .0048487    48.98  0.000   .2279954   .2470019
       _cons │  1.131906    .0061875   182.93  0.000   1.119779   1.144034
─────────────────────────────────────────────────────────────────────
```

Parameter estimates deviate from those defined in the base data set – but not substantially. What has notably changed is the dispersion statistic. It has nearly doubled to a value of 2.1. Given a data set of 10 000 observations, the dispersion statistic correctly indicates that the data are overdispersed. The AIC and BIC statistics are also inflated. These fit statistics are commonly used when comparing models; those with lower AIC and BIC statistics are better-fitted.

OUTLIERS IN DATA

We create ten outliers out of the 10 000 values of the response, *y*. This represents 1/10 of 1 percent of the observations. The synthetic values of *y* we created range from 0 to 101, although the 9,997th largest number is 58. The mean is 4.19 and median 3.0.

Two sets of outliers will be generated. One set will add 10 to the first 10 values of *y* in the data, which have been randomized. The second will add 20 rather than 10. Code showing the creation of the outliers, together with a listing of the first 15 values of *y* and outlier values is given below:

```
. gen y10_10 = y
. replace y10_10 = y10_10 + 10 in 1/10
. gen y20_10 = y
. replace y20_10 = y20_10 + 20 in 1/10
. l y y10_10 y20_10 in 1/100
```

```
                y             y10_10              y20_10
 1              0             10                  20
 2              2             12                  22
 3              7             17                  27
 4              6             16                  26
 5              4             14                  24
 6              3             13                  23
 7              7             17                  27
 8              0             10                  20
 9              4             14                  24
10              1             11                  21
11              2              2                   2
12              0              0                   0
13              2              2                   2
14              4              4                   4
15              1              1                   1
```

Modeling the *y*-plus-10 response on the same set of predictors yields:

```
. glm y10_10 x1 x2 x3, fam(poi)

Generalized linear models      No. of obs      =     10000
Optimization    : ML           Residual df     =      9996
                               Scale parameter =         1
Deviance        = 10893.15795  (1/df) Deviance = 1.089752
Pearson         = 10637.66647  (1/df) Pearson  = 1.064192
                               AIC             = 3.746111
Log likelihood  = -18726.55299 BIC             = -81173.4
```

```
-----------------------------------------------------------------
      |            OIM
y10_10|    Coef.   Std. Err.     z     P>|z|  [95% Conf. Interval]
------+----------------------------------------------------------
    x1|  .4990834   .004911   101.63  0.000    .489458   .5087087
    x2| -.7429948  .0049438  -150.29  0.000  -.7526844  -.7333052
    x3|  .2486112  .0048434    51.33  0.000   .2391184   .2581041
 _cons|  1.007355  .0066797   150.81  0.000   .9942634   1.020447
-----------------------------------------------------------------
```

Note that the parameter estimates are nearly identical to the synthetic model having a response of *y*, i.e. *y* with the first ten responses having 10 added to the value *y*. The Pearson dispersion statistic, however, has increased from 0.9875 to 1.0642. Given the large number of observations, a value of 1.06 indicates overdispersion. Of course, we know that the source of the overdispersion results from the ten outliers.

Adding another ten counts to the observations we already made to the first ten observations produce additional overdispersion:

```
. glm y20_10 x1 x2 x3, nolog fam(poi)

Generalized linear models      No. of obs      =      10000
Optimization    : ML           Residual df     =       9996
                               Scale parameter =          1
Deviance        = 11329.94492  (1/df) Deviance =   1.133448
Pearson         = 12982.97046  (1/df) Pearson  =   1.298817
                               AIC             =   3.790353
Log likelihood = -18947.76719 BIC             = -80736.62

-----------------------------------------------------------------------
       |             OIM
y20_10 |   Coef.    Std. Err.      z     P<|z|   [95% Conf. Interval]
-------+---------------------------------------------------------------
    x1 |  .4975823   .0049049   101.45  0.000    .4879689    .5071956
    x2 | -.7404962   .0049376  -149.97  0.000   -.7501736   -.7308187
    x3 |  .2475645   .0048377    51.17  0.000    .2380827    .2570462
 _cons |  1.012549   .0066612   152.01  0.000    .9994931    1.025605
-----------------------------------------------------------------------
```

The Pearson dispersion statistic has increased from an initial value of 0.9875 to 1.2988 – an increase of some 30%. The effect of these outliers, given that they constitute only 1/1000 of the observations in the model, is remarkable. This example provides good evidence of the importance of checking for model outliers in the presence of apparent overdispersion. Once the outliers are corrected the Pearson dispersion reduces to near 1.0. In fact, in some cases where outliers have been identified, but we do not have information as to how they are to be amended, it may be preferable to simply drop them from the model. Dropping the first ten observations results in a model that is nearly identical to the original model:

```
. glm y x1 x2 x3 in 11/10000, nolog fam(poi)

Generalized linear models       No. of obs      =       9990
Optimization    :  ML           Residual df     =       9986
                                Scale parameter =          1
Deviance        =  10630.39818  (1/df) Deviance =    1.06453
Pearson         =  9861.534172  (1/df) Pearson  =    .987536
                                AIC             =   3.719126
Log likelihood  =  -18573.0342  BIC             =  -81334.07

------------------------------------------------------------------------
      |            OIM
    y |  Coef.    Std. Err.      z      P>|z|    [95% Conf. Interval]
------+-----------------------------------------------------------------
   x1 |  .5006113  .0049181    101.79  0.000     .490972     .5102507
   x2 |-.7456277   .0049518   -150.58  0.000   -.755333    -.7359224
   x3 |   .249782  .0048501     51.50  0.000     .240276      .259288
_cons |  1.001902  .0067024    149.48  0.000    .9887655      .015039
------------------------------------------------------------------------
```

CREATION OF INTERACTION

We next consider a Poisson model having an interaction term. We create it in the same manner as we did the base model:

```
. gen x23 = x2*x3
. gen xbi =1+.5*x1−.75*x2+.25*x3+.2*x23
. genpoisson yi, xbeta(xbi)
```

The interaction term, created by `x2` and `x3`, is represented by the predictor `x23`. We furthermore specified a parameter value of 0.2 for the interaction term. The dataset is created with main effects predictors and the interaction:

```
. glm yi x1 x2 x3 x23, nolog fam(poi)

Generalized linear models    No. of obs       =     10000
Optimization :  ML           Residual df      =      9995
                             Scale parameter  =         1
Deviance      = 10816.20559  (1/df) Deviance  =  1.082162
Pearson       = 10047.10403  (1/df) Pearson   =  1.005213
                             AIC              =  3.743824
Log           = −18714.11936 BIC              = −81241.15
likelihood

==================================================================
          |        OIM
       yi |    Coef.   Std. Err.      z    P>|z|   [95% Conf. Interval]
----------+-------------------------------------------------------
       x1 |  .4953458   .004901   101.07  0.000     .48574   .5049516
       x2 | −.7512554  .0049728  −151.07  0.000  −.7610019  −.7415089
       x3 |  .2486223  .0059682    41.66  0.000   .2369248   .2603197
      x23 |  .2014496  .0048848    41.24  0.000   .1918756   .2110236
    _cons |  1.006059  .0067253   149.59  0.000   .9928779   1.019241
==================================================================
```

All parameter estimates appear as expected, including the interaction. Again, the dispersion statistic approximates 1.0. The AIC and BIC statistics are very close to those of the base model.

Now, we model the data without the interaction:

```
. glm yi x1 x2 x3, nolog fam(poi)

Generalized linear models      No. of obs       =     10000
Optimization    : ML           Residual df      =      9996
                               Scale parameter  =         1
Deviance        = 12464.79638  (1/df) Deviance  =  1.246978
Pearson         = 11834.63586  (1/df) Pearson   =  1.183937
                               AIC              =  3.908483
Log likelihood  = −19538.41476 BIC              = −79601.77

==================================================================
          |        OIM
       yi |    Coef.  Std. Err.      z    P>|z|   [95% Conf. Interval]
----------+-------------------------------------------------------
       x1 |  .5008559 .0049115   101.98  0.000   .4912296   .5104823
       x2 | −.7336854 .0049411  −148.49  0.000  −.7433698  −.7240011
       x3 |  .1063038 .0048617    21.87  0.000   .0967751   .1158325
    _cons |  1.035805 .0065672   157.72  0.000   1.022933   1.048676
==================================================================
```

We see that the dispersion rose 18%, AIC higher by .15, and BIC by 1,640. x3 is also quite different from the true model having the interaction term. x3 was one of the two terms from which the interaction was created. This model is apparently overdispersed. Of course, the overdispersion can be accommodated by creating the proper interaction.

TESTING THE LINK

Since we do not normally employ links other than the natural log with count response models, it is perhaps easier to demonstrate testing of the link function by using members of the GLM binomial family. Criterion 5(e) can be evaluated by creating a complementary loglog model using the same data as the base model. This will serve as the true model. When we then model the data using a logistic regression, we find that it is overdispersed. The major difference, however, relates to the parameter estimates.

The command to create a synthetic data set is given as:

```
. genbinomial yc, xbeta(xb) n(50) link(cloglog)
```

The data are modeled, with a binomial denominator of 50, as:

CLOGLOG REGRESSION

```
. glm yc x1 x2 x3, nolog fam(bin 50) link(clog)

Generalized linear models         No. of obs        =     10000
Optimization    :  ML             Residual df       =      9996
                                  Scale parameter   =         1
Deviance        =  8535.227445    (1/df) Deviance   =  .8538643
Pearson         =  10074.44639    (1/df) Pearson    =  1.007848

Variance        :  V(u) =         [Binomial]
function           u*(1-u/50)     [Complementary log-log]
Link function   :  g(u) =
                   ln(-ln(1-u/50))
Log             =  −16215.10735   AIC               =  3.243821
likelihood                        BIC               = −122635.9

------------------------------------------------------------------------
             |            OIM
          yc |   Coef.    Std. Err.      z     P>|z|    [95% Conf. Interval]
-------------+----------------------------------------------------------
          x1 |   .496669  .0026889    184.71  0.000    .4913989   .5019392
          x2 | −.7474651  .0031267   −239.06  0.000   −.7535933   −.741337
          x3 |  .2485862  .0023845    104.25  0.000    .2439127   .2532597
       _cons |  .9962508  .0027719    359.42  0.000     .990818   1.001684
------------------------------------------------------------------------
```

The Pearson dispersion has a value, as expected, of 1.008. Note the values of the AIC and BIC goodness-of-fit statistics.

Analyzing the data as a logistic model yields the following output:

LOGISTIC REGRESSION

```
. glm yc x1 x2 x3, nolog fam(bin 50) link(logit)

Generalized linear models            No. of obs      =     10000
Optimization : ML                    Residual df     =      9996
                                     Scale parameter =         1
Deviance      = 12618.40477          (1/d) Deviance  =  1.262345
Pearson       = 10949.24219          (1/df) Pearson  =  1.095362
Variance      : V(u) = u*(1-u/50)    [Binomial]
function
Link          : g(u) = ln(u/(50-u)) [Logit]
function
                                     AIC             =  3.652139
Log           = −18256.69601         BIC             = −118552.7
likelihood
```

```
------------------------------------------------------------------------
             |            OIM
          yc |  Coef.    Std. Err.      z     P>|z|   [95% Conf. Interval]
-------------+----------------------------------------------------------
          x1 | .9787876  .0053431    183.19  0.000    .9683152      .98926
          x2 |−1.47635   .0060993   −242.05  0.000   −1.488305   −1.464396
          x3 | .4925683  .0048725    101.09  0.000    .4830184    .5021181
       _cons |2.630692   .0070696    372.12  0.000    2.616836    2.644549
------------------------------------------------------------------------
```

Estimates differ greatly from the synthetically created “true” values. The dispersion statistic indicates overdispersion – 1.095; AIC & BIC are higher than with the complementary loglog model, indicating that the latter is preferable to the logistic model.

Again, the Poisson model is rarely used with a non-canonical link; hence comparison of link misspecification for this model is problematic. We shall again look at proper link specification when dealing with negative binomial regression.

TESTING PREDICTOR SCALE

We next construct a Poisson data set where `x1` has been transformed to `x1`-squared with a parameter value of 0.50. Other predictors are given the same parameters as before.

```
.  gen x1sq = x1*x1 /* square x1 */
.  gen xbsq = 1 +.5*x1sq -.75*x2 +.25*x3
.  genpoisson ysq, xbeta(xbsq)

.  glm ysq x1sq x2 x3, nolog fam(poi)
```

```
Generalized linear models     No. of obs       =     10000
Optimization  : ML            Residual df      =      9996
                              Scale parameter  =         1
Deviance      = 10791.22666   (1/d) Deviance   =  1.079554
Pearson       = 10078.38171   (1/df) Pearson   =  1.008241
                              AIC              =  4.274683
Log           = −21369.41684  BIC              = −81275.34
likelihood
==========================================================
           |          OIM
       ysq |   Coef.  Std. Err.     z    P>|z|   [95% Conf. Interval]
-----------+----------------------------------------------------------
      x1sq |  .500438 .0006325   791.20  0.000    .4991983   .5016777
        x2 |−.7523986 .0031063  −242.21  0.000   −.7584869  −.7463103
        x3 | .2486034 .0028204    88.15  0.000    .2430756   .2541312
     _cons | .9996225 .0051985   192.29  0.000    .9894335   1.009811
==========================================================
```

Parameter estimates all approximate the synthetically assigned values, and the dispersion statistic is close to one. We model the data as with the base model, except for the new *y*, which we call *ysq*.

```
. glm ysq x1 x2 x3, nolog fam(poi)
Generalized linear models      No. of obs       =     10000
Optimization  :  ML            Residual df      =      9996
                               Scale parameter  =         1
Deviance      =  353558.2906   (1/d) Deviance   =  35.36998
Pearson       =  5758862.8     (1/df) Pearson   =  576.1167
                               AIC              =  38.55139
Log           =  −192752.9488  BIC              =  261491.7
likelihood
==========================================================
           |          OIM
       ysq |   Coef.  Std. Err.     z    P>|z|   [95% Conf. Interval]
-----------+----------------------------------------------------------
        x1 |−.3259203 .0029165  −111.75  0.000   −.3316366   −.320204
        x2 |−.5801257 .0029266  −198.23  0.000   −.5858617  −.5743898
        x3 |  .327435 .0028917   113.23  0.000    .3217673   .3331027
     _cons |  2.18808 .0036302   602.74  0.000    2.180965   2.195196
==========================================================

. save odtest/* save simulated datasets in one file */
```

Parameter estimates now differ greatly from the true values. Dispersion statistics are both extremely high (Pearson = 576.1), as are the AIC and BIC statistics. The model is highly overdispersed. Note the difference created by not taking into account the quadratic nature of `x1`. Squaring `x1`, of course, results in the correct model.

These examples show how apparent overdispersion may be corrected. The caveat here is that one should never employ another model designed for overdispersed count data until the model is evaluated for apparent overdispersion. A model may in fact be a well-fitted Poisson or negative binomial model once appropriate transformations have taken place. This is not always an easy task, but necessary when faced with indicators of overdispersion. Moreover, until overdispersion has been accommodated either by dealing with the model as above, or by applying alternative models, one may not simply accept seemingly significant p-values. Although it has not been apparent from the examples we have used, overdispersion does many times change the significance with which predictors are thought to contribute to the model. Standard errors may be biased either upwards or downwards.

4.3 Methods of handling real overdispersion

We may summarize the possible remedies that can be made to a model when faced with apparent overdispersion by the following:

OVERDISPERSION ONLY APPARENT

1. Add appropriate predictor
2. Construct required interactions
3. Transform predictor(s)
4. Transform response
5. Adjust for outliers
6. Use correct link function

When faced with indicators of overdispersion, we first check for the possibility of apparent overdispersion. If overdispersion persists, there are a variety of methods that statisticians have used to deal with it – each based on addressing a reason giving rise to overdispersion.

MODELS DEALING WITH POISSON OVERDISPERSION

1. Scale SEs *post hoc*; deviance, chi2 dispersion
2. Scale SEs iteratively; scale term
3. Robust variance estimators
4. Bootstrap or jackknife SE
5. Negative binomial
6. Heterogeneous negative binomial
7. NB-P
8. Generalized Poisson
9. Generalized estimating equations (GEE)
10. Unconditional fixed effects

11 Conditional fixed effects
12 Random effects
13 Latent Class NB
14 Random intercept
15 Random parameter

OVERDISPERSION DUE TO ZERO COUNTS

1 Zero-inflated model
2 Zero-truncated model
3 NB with endogenous stratification
4 Heterogeneous NB w endogenous stratification
5 Hurdle model
6 Selection model

OVERDISPERSION RESULTING FROM CENSORED OBSERVATIONS

1 Censored model E
2 Censored model S

4.3.1 Scaling of standard errors

Scaling of standard errors was the first method used to deal with overdispersion in binomial and count response models. The method replaces the *W*, or model weight, in

$$\beta = (X'WX)^{-1}X'Wz$$

with the inverse square root of the dispersion statistic. Scaling by the deviance entails estimating the model, abstracting the deviance-based dispersion, applying the transformation, then running one additional iteration of the algorithm, but as

$$\beta(X'W_dX)^{-1}X'W_dz \tag{4.1}$$

Scaling in effect adjusts the model standard errors to the value that would have been calculated if the dispersion statistic had originally been 1.0. McCullagh and Nelder (1989) recommend that deviance-based scaling be used with discrete response models, while continuous response models use Pearson-based scaling. Both deviance and Pearson scaling should produce similar standard errors if

Table 4.1. *Poisson algorithm: scaling by chi2 dispersion*

```
$\mu$ = (y + mean(y))/2
$\eta$ = ln($\mu$)
WHILE(abs($\Delta$dev) > tolerance)  {
   u = (y − $\mu$)/$\mu$
   w = $\mu$
   z = $\eta$ + u − offset
   $\beta$ = (X′wX)$^{-1}$X′wz
   $\eta$ = X′$\beta$ + offset
   $\mu$ = exp($\eta$)
   oldDev = dev
   dev = 2$\Sigma${yln(y/$\mu$) − (y − $\mu$)}
   $\Delta$dev = dev − oldDev
}
/*Afterconvergence, calculate*/
   dof = n − pred − 1
   sc = chi2/dof
   w = $\mu$/sqrt
/*Display again with SE′s adjusted by new w*/
```

the model is well fitted. However, simulation studies have demonstrated that Pearson χ^2-based scaling of count models is preferred over deviance-based scaling.

In Table 4.1 we see an IRLS Poisson algorithm showing both offsets and how scaling is calculated.

An example will demonstrate how an overdispersed model can have the standard errors adjusted, providing the user with a more accurate indication of the true standard errors.

A non-scaled model using the *Medpar* data is given as:

```
. glm los hmo white type2 type3, fam(poi) eform

Generalized linear models          No. of obs      =       1495
Optimization     :  ML             Residual df     =       1490
                                   Scale parameter =          1
Deviance         =  8142.666001    (1/df) Deviance =   5.464877
Pearson          =  9327.983215    (1/df) Pearson  =   6.260391
                                   AIC             =   9.276131
Log likelihood   = −6928.907786    BIC             =  −2749.057
```

```
------------------------------------------------------------------------
       |              OIM
   los |     IRR   Std. Err.     z    P>|z|    [95% Conf. Interval]
-------+----------------------------------------------------------------
   hmo | .9309504  .0222906   −2.99  0.003    .8882708     .9756806
 white | .8573826  .0235032   −5.61  0.000    .8125327      .904708
 type2 | 1.248137  .0262756   10.53  0.000    1.197685     1.300713
 type3 | 2.032927  .0531325   27.15  0.000    1.931412     2.139778
------------------------------------------------------------------------
```

The Pearson χ^2 dispersion is an extremely high 6.26, especially considering the relatively large number of observations. For example, based on the original standard error for *hmo* (.0222906), we may calculate a scaled standard error as sqrt(6.260391)* 0.0222906 =.05577281. Note the calculated value in the model output below:

```
. glm los hmo white type2 type3, nolog fam(poi) eform scale(x2)

Generalized linear models          No. of obs       =       1495
Optimization     :  ML             Residual df      =       1490
                                   Scale parameter  =          1
Deviance         =  8142.666001    (1/df) Deviance  =   5.464877
Pearson          =  9327.983215    (1/df) Pearson   =   6.260391
                                   AIC              =   9.276131
Log likelihood   =  −6928.907786   BIC              =  −2749.057

==============================================================
      |             OIM                      [95% Conf.
  los |    IRR    Std. Err.      z   P>|z|   Interval]
------+-------------------------------------------------------
  hmo | .9309504  .0557729  <= −1.19 0.232   .8278113  1.04694
white | .8573826  .0588069     −2.24 0.025   .7495346 .9807484
type2 | 1.248137  .0657437      4.21 0.000    1.12571 1.383878
type3 | 2.032927  .1329416     10.85 0.000   1.788373 2.310923
==============================================================
(Standard errors scaled using square root of Pearson
X2-based dispersion)
```

It needs to be emphasized that the parameter estimates remain unaffected, and only the standard errors are scaled. Apparently model overdispersion biases standard errors such that *hmo* appears to significantly contribute to the model, and our consequent understanding of *los*, when in fact it does not. Scaling by the deviance dispersion produces similar results in this example. (Note: Stata does not include the log-likelihood and AIC statistic for scaled output. Scaling implies misspecification and Stata tries not to include likelihood-based statistics in such cases.)

4.3.2 Quasi-likelihood variance multipliers

Quasi-likelihood (QL) methods were first developed by Wedderbrun (1974). The method is based on GLM principles, but allows parameter estimates to be calculated based only on a specification of the mean and variance of the model observations without regard to those specifications originating from a member of the single-parameter exponential family of distributions. Further generalizations to the quasi-likelihood methodology were advanced by Nelder and Pregibon (1987). Called extended quasi-likelihood (EQL), these methods were designed to evaluate the appropriateness of the QL variance in a model. However, EQL models take us beyond the scope of our discussion. Quasi-likelihood models though are important to understanding extensions to the Poisson and negative binomial models we consider in this text.

Quasi-likelihood methods allow one to model data without explicit specification of an underlying log-likelihood function. Rather, we begin with a mean and variance function, which are not restricted to the collection of functions defined by single-parameter exponential family members, and abstract backward to an implied log-likelihood function. Since this implied log-likelihood is not derived from a probability function, we call it quasi-likelihood or quasi-log-likelihood instead. The quasi-likelihood, or the derived quasi-deviance function, is then used in an IRLS algorithm to estimate parameters just as for GLMs when the mean and variance function are those from a specific member of the single-parameter exponential family.

Derived from Equation (2.18), quasi-likelihood is defined as

$$Q(y_i; \mu_i) = \int_{y_i}^{\mu_i} \frac{y_i - \mu_i}{\phi V(\mu_i)} d\mu_i \tag{4.2}$$

and the quasi-deviance as

$$QD(y_i; \mu_i) = 2\int_{\mu_i}^{y_i} \frac{y_i - \mu_i}{V(\mu_i)} d\mu_i \tag{4.3}$$

In an enlightening analysis of leaf-blotch data, the quasi-deviance was applied by Wedderburn using the logit link and a “squared binomial” variance function $\mu^2(1 - \mu)^2$. However, the same logit could also have been specified with traditional exponential family variance functions. In the case of the Poisson, we see that by taking the integral of $(y - \mu)/\mu$ from μ to y with respect to μ, the resultant equation is the Poisson log-likelihood, but without the final $\ln(y!)$ normalizing term. The normalizing term is what ensures that the sum of the probabilities over the probability space adds to unity. The negative binomial (NB-2) log-likelihood function can be similarly abstracted using the variance function $\mu + \alpha\mu^2$.

The manner in which quasi-likelihood methodology is typically brought to bear on overdispersed Poisson data is to multiply the variance μ by some constant scale value. Indicated as ψ, a quasi-deviance Poisson algorithm is shown in Table 4.2.

The fact that the variance function is multiplied by a constant changes the likelihood, or, in this case, the deviance function, by dividing it by the scale. It is the next stage in amending the Poisson variance function, and

Table 4.2. *Quasi-deviance Poisson regression algorithm variance multiplier*

```
$\mu = (y + mean(y))/2$
$\eta = ln(\mu)$
WHILE(abs($\Delta$dev) > tolerance)  {
   $u = (y - \mu)/\mu$
   $w = \mu^{\star}\psi$
   $z = \eta + u - offset$
   $\beta = (X'wX)^{-1}X'wz$
   $\eta = X'\beta + offset$
   $\mu = exp(\eta)$
   oldDev = dev
   $dev = [2\Sigma\{y ln(y/\mu) - (y - \mu)\}]/\psi$
   $\Delta dev = dev - oldDev$
}
```

log-likelihood/deviance, to accommodate or adjust for overdispersion. We present an example using the same *Medpar* data. In this case we enter the deviance dispersion statistic from the base model as the variance multiplier.

QUASI-LIKELIHOOD: VARIANCE MULTIPLIER

```
. glm los hmo white type2 type3, nolog ef fam(poi) irls
disp(5.464877)

Generalized linear models     No. of obs       =       1495
Optimization  :  MQL Fisher   Residual df      =       1490
                 scoring
                 (IRLS EIM)   Scale parameter  =   5.464877
Deviance      = 1489.999867   (1/df) Deviance  =   .9999999
Pearson       = 1706.897171   (1/df) Pearson   =   1.145569
Variance      :  V(u) = u     [Poisson]
function
Link          :  g(u) = ln(u) [Log]
function

Quasi-likelihood model with   BIC              =  −9401.724
dispersion: 5.464877

==========================================================
     |             EIM
 los |     IRR   Std. Err.    z    P>|z|   [95% Conf. Interval]
----------------------------------------------------------
  hmo|.9309504  .0095353  −6.99 0.000    .912448     .949828
white|.8573826   .010054 −13.12 0.000   .8379019    .8773162
type2|1.248137  .0112399  24.61 0.000     1.2263    1.270362
type3|2.032927  .0227285  63.46 0.000   1.988865    2.077966
==========================================================
```

Extra variation is dampened from the variance by multiplying it by the value of the deviance dispersion, 5.464877. Note that the deviance-dispersion value of this quasi-likelihood model is now 1.0.

Compare the summary statistics of this model with the standard Poisson model applied to the same data. The BIC statistic is substantially less than that of the standard (and scaled) model, indicating a better fit. The deviance statistic is also substantially less than that of the standard models.

```
                   Standard          Quasi-likelihood model
Deviance            8142.67                         1490.00
     BIC           −2749.10                        −9401.72
```

The quasi-likelihood model is not a true likelihood model, and hence the standard errors are not based on a correct model-based Hessian matrix. Leading to the discussion of the next section, we employ a robust or sandwich variance estimator, producing the following adjusted standard errors. Note that the previously indicated statistically significant contribution of both *hmo* and *white* to the model is now called into question.

```
-----------------------------------------------------------------------
      |            Semi-Robust
  los |    IRR     Std. Err.     z    P>|z|    [95% Conf. Interval]
------+----------------------------------------------------------------
  hmo | .9309504  .0481602   −1.38  0.167    .8411858     1.030294
white | .8573826  .0714211   −1.85  0.065    .7282298     1.009441
type2 | 1.248137  .0660044    4.19  0.000    1.125249     1.384445
type3 | 2.032927  .2354717    6.13  0.000    1.620049      2.55103
-----------------------------------------------------------------------
```

4.3.3 Robust variance estimators

Unlike the standard variance estimator, $-H(\beta)^{-1}$, the robust estimator does not need $LL(\beta; x)$ to be based on the correct distribution function for x. Robust variance estimators have also been referred to as sandwich variance estimators. Associated standard errors are sometimes called Huber standard errors or White standard errors. Huber (1967) was the first to discuss this method, which was later independently discussed by White (1980) in the field of econometrics. Robust estimators are implemented in a post-estimation procedure according to the schema outlined in Table 4.3. Readers interested in a more complete exposition can see Hardin (2003).

We shall find that robust variance estimators are quite robust, hence the name, when modeling overdispersion in count response models. They also play an important role when interpreting the Poisson or negative binomial parameter estimates as risk ratios. The robust score equations for the three count response models are listed in Table 4.4.

Table 4.3. *Implementation of robust variance estimators*

```
1/ Estimate the model
2/ calculate the linear predictor, $x\beta$.
3/ calculate score vector: $g' = g(\beta; x) = x\partial LL(x\beta)/\partial x\beta) = ux$
4/ calculate dof adjustment: $n(n-1)$
5/ combine terms: $V(\beta) = V(n/(n-1)\Sigma u^2 x'x)V$
6/ replace model Variance-Covariance matrix with robust
   estimator: an additional iteration with new matrix.
```

Table 4.4. *Robust score equations*

Poisson	:	$y - \exp(x\beta)$
Geometric (log)	:	$(y - \exp(x\beta))/(1 + \exp(x\beta))$
Negative binomial (log)	:	$(y - \exp(x\beta))(1 + \alpha\exp(x\beta))$

An example using the same *Medpar* data is displayed:

POISSON WITH ROBUST VARIANCE ESTIMATOR

```
. glm los hmo white type2 type3, nolog fam(poi) robust eform

Generalized linear models     No. of obs      =       1495
Optimization  : ML            Residual df     =       1490
                              Scale parameter =          1
Deviance      = 8142.666001   (1/df) Deviance =   5.464877
Pearson       = 9327.983215   (1/df) Pearson  =   6.260391
                              AIC             =   9.276131
Log pseudo-   = -6928.907786  BIC             =  -2749.057
likelihood

     |            Robust
  los|    IRR    Std. Err.     z   P>|z|   [95% Conf. Interval]
-----+-----------------------------------------------------------
  hmo|.9309504  .0481602  -1.38  0.167   .8411858     1.030294
white|.8573826  .0714211  -1.85  0.065   .7282298     1.009441
type2|1.248137  .0660044   4.19  0.000   1.125249     1.384445
type3|2.032927  .2354717   6.13  0.000   1.620049      2.55103
```

When robust variance estimators are applied to this type of quasi-likelihood model, we find that the effect of the robust variance overrides the adjustment made to the standard errors by the multiplier. It is as if the initial quasi-likelihood model were not estimated in the first place.

Robust variance estimators can also be applied to models consisting of clustered or longitudinal data. Many data situations take this form. For instance, when gathering treatment data on patients throughout a county, it must be assumed that treatments given by individual providers are more highly correlated within each provider than between providers. Likewise, in longitudinal data, treatment results may be recorded for each patient over a period of time. Again it must be assumed that results are more highly correlated within each patient record than between patients. Data such as these are usually referred to as panel data. Robust variance adjustments of some variety must be applied to the data due to the fact that observations are not independent.

Modified sandwich variance estimators or robust-cluster variance estimators provide standard errors that allow inference that is robust to within group correlation, but assumes that clusters of groups are independent. The procedure to calculate this type of robust estimate begins by summing the scores within each respective cluster. The data set is thereupon collapsed so that there is only one observation per cluster or panel. A robust variance estimator is then determined in the same manner as in the non-cluster case, except *n* is now the number of clusters and *u* consists of cluster sums. Refer to Table 4.3. A complete discussion of robust panel estimators is found in Hardin and Hilbe (2003).

The *Medpar* data provide the hospital provider code with each observation. Called *provnum*, it is entered as an option to obtain the modified sandwich variance estimator. Unlike scaling and variance multipliers, robust estimators may be used with any maximum likelihood algorithm, not only GLM-based algorithms.

POISSON: CLUSTERING BY PROVIDER

```
. glm los hmo white type2 type3, nolog fam(poi)
eform cluster(provnum)

Generalized linear models     No. of obs       =      1495
Optimization  : ML            Residual df      =      1490
                              Scale parameter  =         1
Deviance      = 8142.666001   (1/df) Deviance  =  5.464877
Pearson       = 9327.983215   (1/df) Pearson   =  6.260391
                              AIC              =  9.276131
Log pseudo-   = −6928.907786  BIC              = −2749.057
likelihood
          (Std. Err. adjusted for 54 clusters in provnum)

==========================================================
     |         Robust
 los |   IRR    Std. Err.    z    P>|z|   [95% Conf. Interval]
----------------------------------------------------------
  hmo|.9309504 .0490889  −1.36 0.175.   8395427     1.03231
white|.8573826 .0625888  −2.11 0.035   .7430825    .9892642
type2|1.248137 .0760289   3.64 0.000   1.107674    1.406411
type3|2.032927 .4126821   3.49 0.000   1.365617     3.02632
==========================================================
```

Standard errors are produced by adjusting for the clustering effect on providers – that is, we suppose that the relationship between length of stay (*los*) and predictors is more highly correlated within a provider than between providers. This is a reasonable supposition. Only *hmo* fails to be contributory. Note again that all summary statistics are the same as in the unadjusted model. Also note that the model parameters are not adjusted for clustering. The model is still specified and estimated as if the observations were all, in fact, independent. It is only the standard errors that are adjusted.

4.3.4 Bootstrap and jackknifed standard errors

Bootstrap and jackknife are two additional methods that are used to adjust standard errors when they are perceived to be overdispersed. Non-parametric bootstrapping makes no assumptions about the underlying distribution of the model. Standard errors are calculated based on the data at hand. Samples are repeatedly taken from the data (with replacement), with each sample providing model estimates. The collection of vector estimates for all samples is used to calculate a variance matrix from which reported standard errors are calculated and used as the basis for calculating confidence intervals. Such confidence intervals can be constructed from percentiles in the collection of point estimates or from large sample theory arguments. The example below uses 50 samples of 1495; each sample provides an estimated coefficient vector from which standard errors are calculated. The number of samples may be changed. This method is primarily used with count data when the data are not Poisson or negative binomial, and the model is overdispersed.

```
. glm los hmo white type2 type3, nolog fam(poi) eform vce(boot)
(running glm on estimation sample)

Bootstrap replications (50)
--+-- 1 --+-- 2 --+-- 3 --+-- 4 --+-- 5
. . . . . . . . . . . . . . . . . . . . .
                                            50
Generalized linear models          No. of obs        =        1495
Optimization     :   ML            Residual df       =        1490
                                   Scale parameter   =           1
Deviance         =    8142.666001  (1/df) Deviance   =    5.464877
Pearson          =    9327.983215  (1/df) Pearson    =    6.260391
                                   AIC               =    9.276131
Log likelihood   =   −6928.907786  BIC               =   −2749.057

--------------------------------------------------------------------
        |  Observed   Bootstrap                   Normal-based
    los |  IRR        Std. Err.     z     P>|z|   [95% Conf. Interval]
--------+-----------------------------------------------------------
    hmo |  .9309504   .0432587    −1.54   0.124    .8499111   1.019717
  white |  .8573826   .0690078    −1.91   0.056    .7322583   1.003887
  type2 |  1.248137   .0555599     4.98   0.000    1.143856   1.361924
  type3 |  2.032927   .2612241     5.52   0.000    1.580321    2.61516
--------------------------------------------------------------------
```

Table 4.5. *Comparision of standard errors: Medpar Poisson moel*

	EIM/OIM	D-SCALE	ROBUST	CLUSTER	BOOT	JACK
hmo	.0222906	.0521090	.0481602	.0490889	.0432587	.0484649
white	.0235032	.0549437	.0714211	.0625888	.0690078	.0732884
type2	.0262756	.0614248	.0660044	.0760289	.0555599	.0664530
type3	.0531325	.1242082	.2354717	.4126821	.2612241	.2415363

Standard errors again indicate that *hmo* and *white* are problematic.

Jackknifed standard errors are used for the same purpose as standard errors calculated from bootstrapping. The model is estimated as many times as there are observations in the data – in this case 1495. Each iteration excludes one observation. The collection of estimated coefficient vectors is used to calculate a variance matrix from which the standard errors are reported in the output, together with (large-sample based) confidence intervals.

```
. glm los hmo white type2 type3, nolog fam(poi) vce(jack)
eform Jackknife replications (1495)

Generalized linear models          No. of obs       =       1495
Optimization     :  ML             Residual df      =       1490
                                   Scale parameter  =          1
Deviance         =  8142.666001    (1/df) Deviance  =   5.464877
Pearson          =  9327.983215    (1/df) Pearson   =   6.260391
                                   AIC              =   9.276131
Log likelihood   =  −6928.907786   BIC              =  −2749.057

---------------------------------------------------------------
      |           Jackknife
  los |     IRR   Std. Err.      t   P>|t|  [95% Conf. Interval]
---------------------------------------------------------------
  hmo | .9309504  .0484649   −1.37  0.170  .8405768     1.03104
white | .8573826  .0732884   −1.80  0.072  .7250295    1.013897
type2 | 1.248137   .066453    4.16  0.000  1.124361    1.385538
type3 | 2.032927  .2415363    5.97  0.000    1.6103    2.566474
---------------------------------------------------------------
```

Jackknifing yields standard errors similar to those obtained from bootstrapping. Again, *hmo* and *white* are questionable contributors to the Poisson model.

To compare standard errors between models on the same data, see Table 4.5.

It is clear that the *Medpar* data we modeled as Poisson are overdispersed. The standard model suggests that all predictors are statistically significant. However, when we employ adjustments to the variance in order to accommodate any overdispersion in our inference, we find that *hmo* and *white* are now questionable contributors to the model, and that the adjusted standard errors are fairly much the same for all but *type3* (emergency admissions). Actually, interpretation of coefficients using robust, bootstrapped, or jackknifed standard errors are similar for *type3*, only Pearson-scaling and clustering on provider differ.

4.3.5 Negative binomial overdispersion

Although the subject is rarely discussed, it is implicit in what we have been discussing that negative binomial (NB-2) models may also be overdispersed. We commonly define Poisson overdispersion as occurring when the Poisson variance exceeds the value of the mean. That is because Poisson distributional assumptions equate the two statistics. We have mentioned several reasons that give rise to Poisson overdispersion in count data.

Given a specified, as well as calculated, value of the mean, μ, we may define negative binomial overdispersion as occurring when the calculated model variance exceeds $\mu + \alpha\mu^2$. That is, a count model may be both Poisson and negative binomial overdispersed if the variance produced by the estimated model exceeds the negative binomial variance.

We may simulate Poisson and negative binomial responses by generating random variates for both distributions. Using Stata, we generate a constant value of 3, and use it to calculate a Poisson and a negative binomial random variate.

POISSON

```
. gen cons = 3                 /* constant - mean - of 3.0 n */
. genpoisson yp, xbeta(cons)/* Poisson random number generator*/
. su yp, detail                /* partial output */
     Mean        20.0909   <=
     Std. Dev.   4.508844
     Variance    20.32967  <=

. glm ypc, nolog fam(poi)

Generalized linear models           No. of obs      =      10000
Optimization       : ML             Residual df     =       9999
                                    Scale parameter =          1
Deviance           = 10231.27276    (1/df) Deviance =    1.02323
Pearson            = 10117.83304 => (1/df) Pearson  =   1.011884
Variance function  : V(u) = u       [Poisson]
Link function      : g(u) = ln(u)   [Log]
                                    AIC             =   5.843747
Log likelihood     = -29217.73637   BIC             = -81862.92

-----------------------------------------------------------------
             OIM
  ypc|  Coef.   Std. Err.     z     P>|z|  [95% Conf. Interval]
-----+-----------------------------------------------------------
_cons|3.000267 .002231   1344.81 0.000    2.995894    3.00464
-----------------------------------------------------------------
```

NEGATIVE BINOMIAL

```
. gennbreg ynb, xbeta(cons) alpha(.3)/* NB RNG; alpha=.3 */
. su ynb, detail
    Mean        20.0231  <=
    Std. Dev.   11.72842
    Variance    137.5557  <=
. di 20.0231 + .3*20.02312/* mean + alpha*mean ^ 2 */
. 140.30046 <=

. glm ynbc, nolog fam(nb.2908478)

Generalized linear models             No. of obs      =      10000
Optimization      : ML                Residual df     =       9999
                                      Scale parameter =          1
Deviance          = 10517.23312       (1/df) Deviance =   1.051828
Pearson           = 10066.66474     =>(1/df) Pearson  =   1.006767
Variance function : V(u) =            [Neg. Binomial]
                    u + (.2908478)u^2
Link function     : g(u) = ln(u)      [Log]
                                      AIC             =   7.543505
Log likelihood    = -37716.52489      BIC             = -81576.96

-----------------------------------------------------------------
     |              OIM
ynbc |    Coef.   Std. Err.     z    P>|z|  [95% Conf. Interval]
-----+-----------------------------------------------------------
cons | 2.996887 .0058377   513.37 0.000    2.985445  3.008328
-----------------------------------------------------------------
. save nboverex  /*  save simulated data in one file  */
```

Repeating the simulation will produce slightly different results, but the values for the respective variances will cluster around the appropriate values; the values for the regression statistics will vary as well.

Overdispersion for both the Poisson and negative binomial models is generally indicated by the value of the Pearson Chi2 dispersion statistic, which reflects the underlying variability in the model data, i.e. the variance. The Pearson Chi2, or χ^2, fit statistic is commonly displayed in GLM program output. It may be defined as

$$\text{Pearson } \chi^2 = \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{V(\mu_i)} \tag{4.4}$$

The Pearson statistic is the sum of all model Pearson residuals as defined in Chapter 2.4.

Some statisticians have used the deviance dispersion as the basis for scaling standard errors. However, as will be discussed later in the text, simulation studies indicate that the Pearson dispersion better captures the excess variability, and adjusts standard errors in such a manner as to reflect what the standard errors would be if the excess variability were not present in the data. In any case, Pearson dispersion in excess of 1.0 tends to indicate Poisson or negative binomial overdispersion respectively. Whether the overdispersion is significant depends on: (1) the value of the dispersion statistic, (2) the number of observations in the model, and (3) the structure of the data, e.g. if the data are highly unbalanced.

Enhanced negative binomial models, e.g ZINB, zero-truncated negative binomial, and so forth, attempt to accommodate negative binomial overdispersion just as enhanced Poisson models attempt to accommodate Poisson overdispersion. The difference is that negative binomial regression is itself one of the models used for overdispersed Poisson data. It is important to keep in mind the relationships between the various models, and what it is that each is constructed to do.

The type of overdispersion – Poisson or negative binomial – we deal with when discussing the various models in the text should be evident from the context. Care must be taken though to remember that although both are indicated by the value of the dispersion statistic, the underlying criteria for each differ.

4.4 Summary

We have outlined the methods to determine whether a Poisson model is subject to real or to only apparent overdispersion. Various remedies for apparent overdispersion were detailed, and methods for adjusting the Poisson model standard errors were specified. However, we provided adjustments while still modeling the data as Poisson. Negative binomial regression is a common alternative to enhanced Poisson regression models when dealing with overdispersed data. Additionally, as previously discussed, both Poisson and negative binomial models can themselves be extended to address the specific reasons why overdispersion arises in the data.

Table 4.6 provides a listing of direct adjustments that are commonly applied to the Poisson variance. The first two are discussed in this chapter.

Exercises

1 How can one use simulation to determine whether the Pearson Chi2 dispersion or deviance dispersion better scales standard errors, per the discussion of Section 4.3 above?

Table 4.6. *Methods to directly adjust the variance*

1	Scaling SE's by deviance-dispersion Chi2-dispersion
2	Multiplying V by a constant $\mathrm{QL} = \mathrm{V}\phi$ (with ϕ a constant)
3	Variance multiplied by ancillary parameter $\mathrm{NB{-}1} = \mathrm{V}\phi = \mu(1+\alpha) = \mu + \alpha\mu$
4	Geometric $\mathrm{V}\phi = \mu(1+\mu) = \mu + \mu^2$
5	Negative binomial $\mathrm{NB{-}2} = \mathrm{V}\phi = \mu(1+\alpha\mu) = \mu + \alpha\mu^2$
6	Heterogeneous negative binomial $\mathrm{NB{-}H} = \mu + (\alpha\gamma)\mu^2$ (with α parameterized by γ)
7	Negative binomial P $\mathrm{NB{-}P} = \mu + \alpha\mu^{\nu}$
8	Generalized Estimating Equations (GEE) $\mathrm{GEE} = \mathrm{V}[\text{correlation matrix}]\mathrm{V}'$

2 Compare deviance-dispersion and Pearson χ^2-dispersion as the basis for scaling simulated overdispersed Poisson models. You may use the same type of simulated data sets as were used in Chapter 4.2, or you may create entirely new ones. Provide evidence to support one method over the other in eliminating specific types of overdispersion.

3 Why should a robust variance estimator be used with quasi-likelihood count models?

4 Specify the indicators of overdispersion in Poisson models. Why is it important to test for apparent overdispersion before adjusting a model for real overdispersion?

5 Use the *drg112az* data set found on the text web site. Model length of stay (*los*) on *urgent*, *cabg*, *male*, and *age75*. If the model is overdispersed, adjust the standard errors by employing a robust variance estimator, clustered on hospital code (*hosp*). Discuss the results.

6 Given the variance function $\mu + \alpha\mu^2$, derive the quasi-likelihood and quasi-deviance negative binomial functions.

7 What criteria are used to determine if a robust variance estimator is appropriate for a given count response model?

8 The following data *ticks* are of the count of ticks on sheep. Determine if the distribution is overdispersed? [Data from Fisher (1941), "The negative binomial distribution", *Annals of Eugenics*]

Numb_ticks	Freq	Numb_ticks	Freq
0	4	13	2
1	5	14	2
2	11	15	1
3	10	16	1
4	9	17	0
5	11	18	0
6	3	19	1
7	5	20	0
8	3	21	1
9	2	22	1
10	2	23	1
11	5	24	0
12	0	25	2

5
Negative binomial regression

In this and subsequent chapters we shall discuss the nature and utility of some 25 varieties of negative binomial regression that are useful for modeling count response data. In addition, we also examine certain models that are related to the negative binomial family of models. This chapter will primarily be devoted to an examination of the derivation of the negative binomial model and to the two foremost methods of its estimation. We also consider how the probabilities generated from a negative binomial model differ from the Poisson, as well as how they differ among various negative binomial models based on both mean and ancillary parameters.

5.1 Varieties of negative binomial

I mentioned that the basic negative binomial model can be enhanced to allow for the modeling of a wide range of count response situations. The Poisson can likewise be enhanced to adjust for data that violate, for instance, its essential distributional assumptions. In fact, many of the same distributional problems face both Poisson and negative binomial models. We therefore find similar approaches to the handling of such data for both the Poisson and negative binomial. These include models such as zero-inflated Poisson (ZIP), which is directly related to the zero-inflated negative binomial (ZINB). Other models without a specific negative binomial correlate are also discussed, e.g. generalized Poisson, which has a heterogeneity parameter like the NB-2 model, but which can also be used to model underdispersion. All of the models allow for variance that exceeds the mean – the principal assumption of the Poisson regression model being that of equi-dispersion. With respect to the negative binomial, allowance of extra variation involves: (1) an adjustment to the NB2 variance

Table 5.1. *Negative binomial models*

1	NB-2 $V = \mu + \alpha\mu^2$
2	NB-1 $V = \mu + \alpha\mu$
3	NB-C <Canonical>
4	0-truncated NB
5	0-truncated NB-1
6	0-truncated NB-C
7	0-inflated NB (ZINB)
8	Censored NB
9	NB-logit hurdle
10	NB-cloglog hurdle
11	NB w endogenous stratification
12	NB-H <Heterogeneous>/ w ES
13	Conditional fixed effects NB
14	NB-P
15	GEE NB (population averaged)
16	Beta random effect NB
17	Sample selection NB
18	Geometric
19	0-truncated geometric
20	Canonical geometric
21	Geometric-logit hurdle
22	Geometric-cloglog hurdle
23	Latent class NB
22	Random Intercept NB
22	Random Parameter NB

function, or (2) a modification to the NB-2 probability distribution, resulting in a modified log-likelihood function. The NB-2 variance function, it may be recalled, is expressed as $\mu + \alpha\mu^2$.

Canonical linked models maintain both the likelihood and variance functions of the NB-2 model, but instead modify the link function. It can be argued, however, that it is the NB-2 model that is an alteration of the basic canonical form, which derives directly from the negative binomial probability function. We shall ferret out these complications beginning with this chapter.

Table 5.1 lists the 25 varieties of negative binomial regression for modeling count response data.

Note that we have included the geometric model as a variety of negative binomial. We do this because the geometric is a negative binomial distribution having a value of α equal to one. It may be argued that Poisson should also be included as a variety of negative binomial since a NB-2 model having α equal to zero is a Poisson model. We exclude it, however, since the negative binomial ancillary or heterogeneity parameter, qua negative binomial distribution, can only approach zero, never reach it. On the other hand, as we have observed in practice, a negative binomial model with a value of α close to zero is statistically indistinguishable from a Poisson model. In this sense the Poisson is a variety of negative binomial. However, we have already attended to the model and will be discussing enhanced Poisson models in the context of related negative binomial models. Table 5.1 simply details the specific varieties of negative binomial that will be discussed in this and coming chapters. It does not entail that alternative Poisson-based models will be ignored. We do not need to duplicate that effort.

5.2 Derivation of the negative binomial

We have previously mentioned that the standard negative binomial regression model, which following Cameron and Trivedi (1998) is usually referred to as NB-2, can be derived as either a Poisson–gamma mixture model, or as a member of the exponential family of distributions which serve as the basis of generalized linear models. We shall first address the Poisson–gamma mixture model. For ease of interpretation I shall dispense with subscripts for this chapter. It is understood, therefore, that the terms λ, μ, η, θ, x, y, n, u, r, and p have subscript, i, and xx has $x_i x_i$. Moreover, log-likelihood functions all assume summation across observations; therefore

$$\sum_{i=1}^{n}$$

is assumed to preface log-likelihood functions. The multiplication summation function

$$\prod_{i=1}^{n}$$

is assumed to preface probability functions. However, integration, as a summation, is displayed when required.

5.2.1 Poisson–gamma mixture model

The negative binomial PDF can be derived from the specification of an outcome characterized by

$$f(y;u) = \frac{e^{-\lambda u}(\lambda u)^{y}}{y!} \tag{5.1}$$

which can be thought of as a Poisson model with gamma heterogeneity where the gamma noise has a mean of 1. The gamma mixture accommodates overdispersed or correlated Poisson counts.

The mean of y, conditioned on u, is Poisson with the conditioned mean and variance given by u

$$f\,(y;u) = \int_0^\infty \frac{e^{-\lambda u}\,(\lambda u)^y}{y!} g(u)\partial u \tag{5.2}$$

$$f\,(y;u) = \int_0^\infty \frac{e^{-\lambda u}\,(\lambda u)^y}{y!}\frac{\nu^u}{\Gamma(\nu)} u^{\nu-1} e^{-\nu u}\partial u \tag{5.3}$$

The gamma nature of u is evident in the derivation from Equations (5.1) to (5.2). We carry the derivation further by moving

$$= \frac{\lambda^y}{\Gamma\,(y+1)}\frac{\nu^\nu}{\Gamma\,(\nu)}\frac{\Gamma\,(y+\nu)}{(\lambda+\nu)^{y+\nu}} \tag{5.4}$$

to the left of the integral, with the remaining terms under the integral equaling 1.

We continue as

$$= \frac{\lambda^y}{\Gamma(y+1)}\frac{\nu^\nu}{\Gamma\,(\nu)}\Gamma(y+\nu)\left(\frac{\nu}{\lambda+\nu}\right)^\nu \frac{1}{\nu^\nu}\left(\frac{\lambda}{\lambda+\nu}\right)^y \frac{1}{\lambda^y} \tag{5.5}$$

$$= \frac{\Gamma(y+\nu)}{\Gamma(y+1)\Gamma(\nu)}\left(\frac{\nu}{\lambda+\nu}\right)^\nu \left(\frac{\lambda}{\lambda+\nu}\right)^y \tag{5.6}$$

$$= \frac{\Gamma(y+\nu)}{\Gamma(y+1)\Gamma(\nu)}\left(\frac{1}{1+\lambda/\nu}\right)^\nu \left(1-\frac{1}{1+\lambda/\nu}\right)^y \tag{5.7}$$

Inverting ν, the gamma scale parameter, yields α, the negative binomial ancillary or overdispersion parameter. We also equate λ and μ. Doing so, we then recognize the resulting negative binomial probability function

$$\frac{\Gamma(y+1/\alpha)}{\Gamma(y+1)\Gamma(1/\alpha)}\left(\frac{1}{1+\alpha\mu}\right)^{1/\alpha} \left(1-\frac{1}{1+\alpha\mu}\right)^y \tag{5.8}$$

Equation (5.8) is a commonly observed form of the negative binomial PDF. The last term may be converted to the following, which is another popular expression of the function

$$= \frac{\Gamma(y+1/\alpha)}{\Gamma(y+1)\Gamma(1/\alpha)}\left(\frac{1}{1+\alpha\mu}\right)^{1/\alpha} \left(\frac{\alpha\mu}{1+\alpha\mu}\right)^y \tag{5.9}$$

Important to maximum likelihood, estimating algorithms are the derivatives of the log-likelihood. We saw in Chapter 2 that setting the first derivative of the log-likelihood, with respect to β, to zero and then solving, is the basis of maximum likelihood estimation.

$$\frac{\partial \mathcal{L}}{\partial \beta} = \sum \frac{x\,(y-\mu)}{1+\alpha\mu} = 0 \tag{5.10}$$

$$\frac{\partial \mathcal{L}}{\partial \alpha} = \sum \left\{ \frac{1}{\alpha^2} \left(\ln(1+\alpha\mu) - \ln\left(\frac{\Gamma\,(y+1/\alpha)}{\Gamma\,(1/\alpha)}\right)\right) - \frac{y-\mu}{\alpha\,(1+\alpha\mu)} \right\} = 0 \tag{5.11}$$

$$\frac{-\partial^2 \mathcal{L}}{\partial \beta \partial \beta'} = \sum \frac{\mu}{1+\alpha\mu} xx' = 0 \tag{5.12}$$

$$\frac{\partial^2 \mathcal{L}}{\partial \beta \partial \alpha} = E\left[-\sum \frac{\mu\,(y-\mu)\,xx'}{(1+\alpha\mu)^2} \right] = 0 \tag{5.13}$$

$$\frac{\partial^2 \mathcal{L}}{\partial \alpha^2} = -\sum \frac{1}{\alpha^4} \left(\ln(1+\alpha\mu) - \ln\left(\frac{(\Gamma\,(y+1/\alpha))}{\Gamma\,(1/\alpha)}\right)^2 + \frac{\mu}{\alpha^2\,(1+\alpha\mu)} \right) = 0 \tag{5.14}$$

Negative binomial score functions are provided as

$$\text{SCORE}: (\beta) = (y-\mu)\,/\,(1+\alpha\mu) \tag{5.15}$$

$$\text{SCORE}: (\alpha) = (-1/\alpha)\,\{(\alpha * (\mu - y))\,/\ln(1+\alpha\mu)\} - \ln(1+\alpha\mu) + \Psi_1 - \Psi_2 \tag{5.16}$$

where

$$\Psi_1 = \{\ln\Gamma\,((y+1/\alpha) + .0001) - \ln\Gamma\,((y+1/\alpha) - .0001)\}\,/.0002 \tag{5.17}$$

and

$$\Psi_2 = \{\ln\Gamma\,((1/\alpha) + .0001) - \ln\Gamma\,((1/\alpha) - .0001)\}\,/.0002 \tag{5.18}$$

Note: Ψ represents the digamma function, the derivative of the log-gamma function, $\ln\Gamma()$

5.2.2 Derivation of the GLM negative binomial

Two major forms of the negative binomial may be derived from the negative binomial probability function. Both forms may be considered as members of the exponential family of distributions, and modeled under the framework of generalized linear models. One is the canonical, being derived directly from the PDF, the other is a conversion from the canonical form to the log-linked form. The latter is known as NB-2 or the traditional negative binomial regression model. Utilizing the log-link allows comparison of point estimates to the Poisson model.

We can describe the negative binomial PDF as the probability of observing y failures before the rth success in a series of Bernoulli trials. Under such a description r would be a positive integer. However, there is no compelling mathematical reason to limit this parameter to integers; only to limit r as positive. Although the negative binomial may be parameterized differently, it is always possible to convert terms to produce the final form derived here. Nevertheless, the form with which we begin is expressed as:

NEGATIVE BINOMIAL PDF

$$f(y;r,p) = \binom{y+r-1}{r-1} p^r (1-p)^y \tag{5.19}$$

Converting the negative binomial PDF into exponential family form results in:

EXPONENTIAL FAMILY FORM

$$f(y;r,p) = \exp\left\{ y\ln(1-p) + r(\ln(p)) + \ln\left(\frac{y+r-1}{r-1}\right)\right\} \tag{5.20}$$

From our earlier discussion we found that the canonical link and cumulant can easily be abstracted from a PDF when it is expressed in exponential family form.

LINK, CUMULANT, SCALE

$$\theta = \ln(1-p) => p = 1 - \exp(\theta) \tag{5.21}$$

$$b(\theta) = -r\ln(p) => -r(1-\exp(\theta)) \tag{5.22}$$

$$\alpha(\phi), \text{the scale, is taken as } 1 \tag{5.23}$$

The first and second derivatives, with respect to θ, respectively yield the mean and variance functions.

NEGATIVE BINOMIAL MEAN

$$b'(\theta) = \frac{\partial b}{\partial p}\frac{\partial p}{\partial \theta} = -\frac{r}{p}\{-(1-p)\} = \frac{r(1-p)}{p} = \mu \tag{5.24}$$

NEGATIVE BINOMIAL VARIANCE

$$b''(\theta) = \frac{\partial^2 b}{\partial p^2}\left(\frac{\partial p}{\partial \theta}\right)^2 + \frac{\partial b}{\partial p}\frac{\partial^2 p}{\partial \theta^2} = \frac{r}{p^2}(1-p)^2 + \frac{-r}{p}(1-p) = \frac{r(1-p)}{p^2} \tag{5.25}$$

$V(\mu)$ therefore equals $r(1-p)/p^2$. We now parameterize p and r in terms of μ and α.

$$(1-p)/(\alpha p)=\mu \tag{5.26}$$

$$(1-p)/p=\alpha\mu \tag{5.27}$$

$$p=1/(1+\alpha\mu) \tag{5.28}$$

where $\alpha = 1/r$.

Given the defined values of μ and α, we may re-parameterize the negative binomial PDF such that

$$f(y;\mu,\alpha)=\binom{y+1/\alpha-1}{1/\alpha-1}\left(\frac{1}{1+\alpha\mu}\right)^{1/\alpha}\left(\frac{\alpha\mu}{1+\alpha\mu}\right)^{y} \tag{5.29}$$

Re-expressed in terms of the log-likelihood, Equation (5.29) yields

$$\begin{aligned}\mathcal{L}(\mu;y,\alpha)\;&=\Sigma\exp\{y\ln((\alpha\mu)/(1+\alpha\mu))-(1/\alpha)\ln(1+\alpha\mu)\\&\quad+\ln\Gamma(y+1/\alpha)-\ln\Gamma(y+1)-\ln\Gamma(1/\alpha)\}\end{aligned} \tag{5.30}$$

or

$$\Sigma\exp\{y\ln(\alpha\mu)-(y+1/\alpha)\ln(1+\alpha\mu)+\ln\Gamma\ldots\} \tag{5.31}$$

The GLM deviance function is derived from both the saturated and fitted log-likelihood functions. The saturated function consists of replacing the value of y for each value of μ.

DEVIANCE

$$D=2\sum_{i=1}^{n}\{\mathcal{L}(y;y)-\mathcal{L}(\mu;y)\} \tag{5.32}$$

Substituting the log-likelihood function as specified in either Equation (5.21) or Equation (5.22), we have

$$D_{nb}=2\sum_{i=1}^{n}\{y\ln(y/\mu)-(y+1/\alpha)/\alpha*\ln((1+\alpha y)/(1+\alpha\mu))\} \tag{5.33}$$

Calculating the terms required for the IRLS algorithm, we have

LINK

$$g(\mu)=\theta=\ln((\alpha\mu)/(1+\alpha\mu))=-\ln(1/\alpha\mu+1) \tag{5.34}$$

INVERSE LINK

$$g^{-1}(\mu) = \mu = 1/\{a(e^{-\theta} - 1)\} \tag{5.35}$$

CUMULANT

$$b\,(\theta) = 1/\alpha \ln\left(1/\,(1+\alpha\mu)\right) \tag{5.36}$$

$$= -1/\alpha \ln\,(1+\alpha\mu) \tag{5.37}$$

MEAN, VARIANCE AND DERIVATIVE

$$b'(\theta) = \mu = \frac{\partial b}{\partial \mu}\frac{\partial \mu}{\partial \theta} = \frac{1}{1+\alpha\mu}\mu\,(1+\alpha\mu) = \mu \tag{5.38}$$

$$b''(\theta) = V = \frac{\partial^2 b}{\partial \mu^2}\left(\frac{\partial \mu}{\partial \theta}\right)^2 + \frac{\partial b \partial^2 \mu}{\partial \mu \partial \theta^2} = \mu + \alpha\mu^2 \tag{5.39}$$

$$g'(\theta) = \frac{\partial \theta}{\partial \mu} = \frac{\partial(\ln\,(\alpha\mu/\,(1+\alpha\mu)))}{\partial \mu} = \frac{1}{\mu + \alpha\mu^2} \tag{5.40}$$

IRLS algorithms normally are parameterized in terms of the fit statistic μ rather than $x\beta$, the linear predictor. Maximum likelihood algorithms such as Newton–Raphson or Marquardt are always parameterized as $x\beta$. In Chapter 1 we showed how to convert between the two parameterizations, which is simply substituting the inverse link function for μ. We do that next for the negative binomial log-likelihood function.

$$\begin{aligned}\mathcal{L} = {} & y * \ln(\alpha \exp(xb)/(1+\alpha \exp(xb))) - \ln(1+\alpha \exp(xb))/\alpha \\ & + \ln\Gamma(y+1/\alpha) - \ln\Gamma(y+1) - \ln\Gamma(1/\alpha)\end{aligned} \tag{5.41}$$

5.3 Negative binomial distributions

Figures 5.1–5.11 illustrate negative binomial distributions for various values of both the mean and α. Note that when $\alpha = 1$, all distributions take the form of a geometric distribution, which is the discrete correlate of the continuous negative exponential distribution. Note also that as the mean increases, the probability of a zero decreases.

Range of Mean (0.5, 1, 2, 5, 10) per each Alpha(0, 0.33, 0.67, 1.0, 1.5, 3.0)
Range of Alpha (0, 0.33, 0.67, 1.0, 1.5, 3.0) per each Mean (0.5, 1, 2, 5, 10)

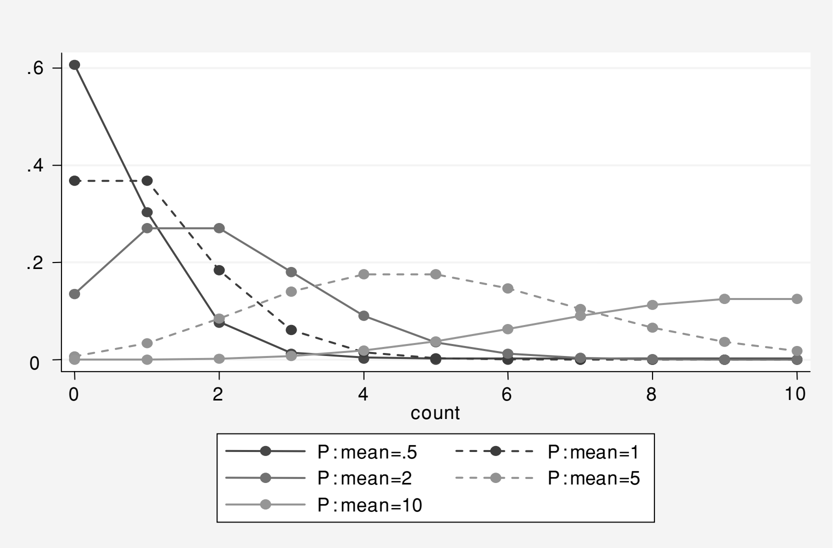

Figure 5.1. Negative binomial distributions: alpha = 0

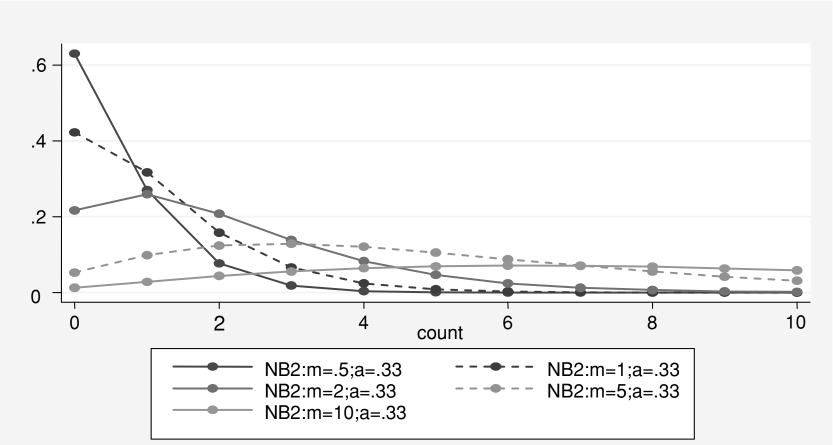

Figure 5.2. Negative binomial distributions: alpha = .33

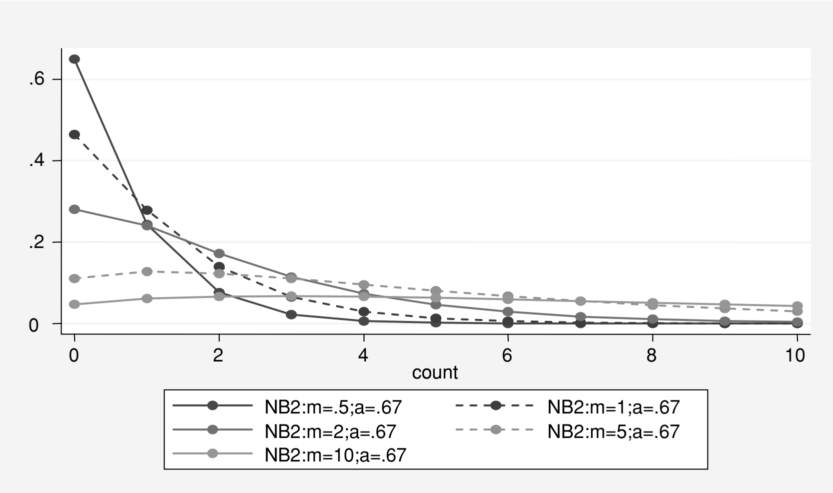

Figure 5.3. Negative binomial distributions: alpha = .67

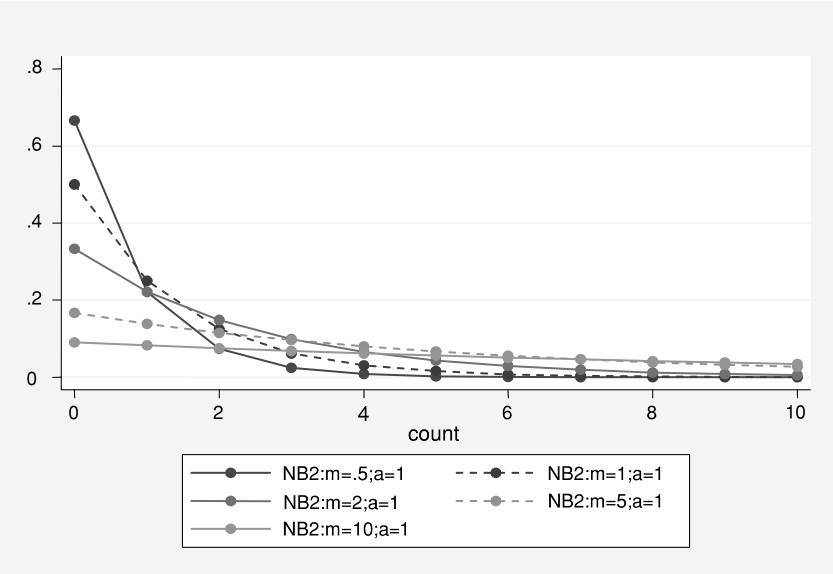

Figure 5.4. Negative binomial distributions: alpha = 1

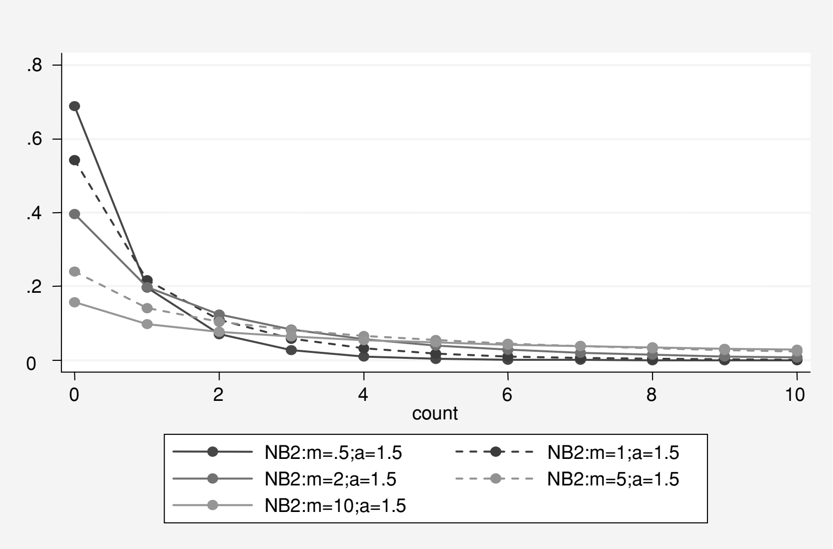

Figure 5.5. Negative binomial distributions: alpha = 1.5

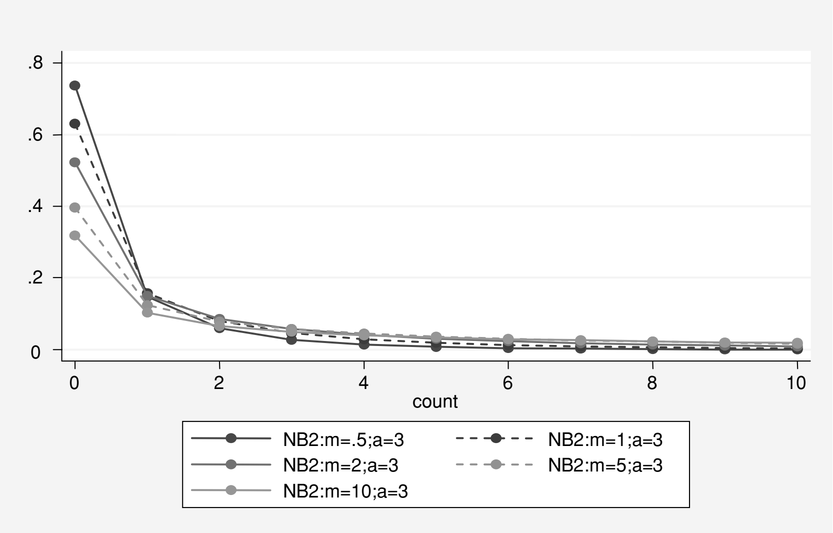

Figure 5.6. Negative binomial distributions: alpha = 3.0

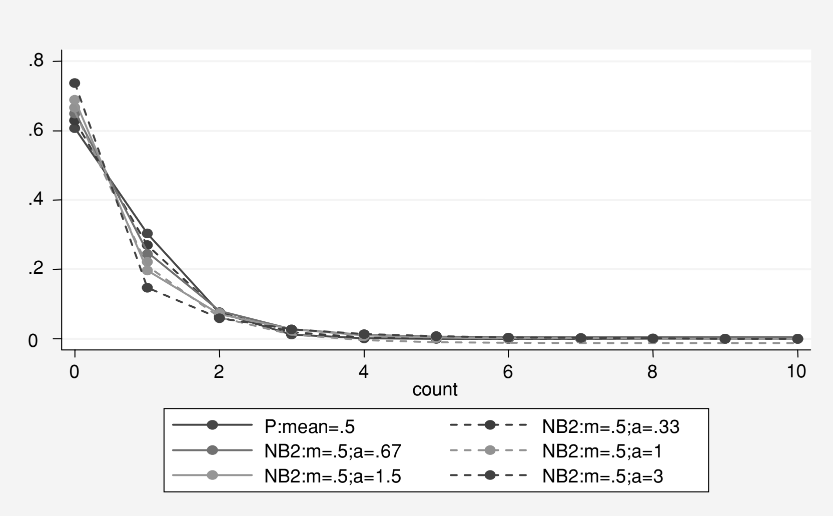

Figure 5.7. Negative binomial distributions: mean = .5

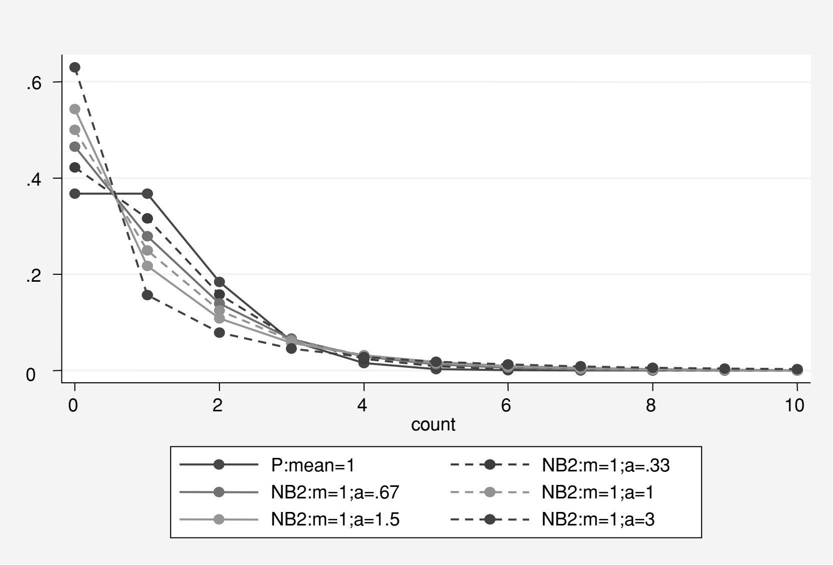

Figure 5.8. Negative binomial distributions: mean = 1

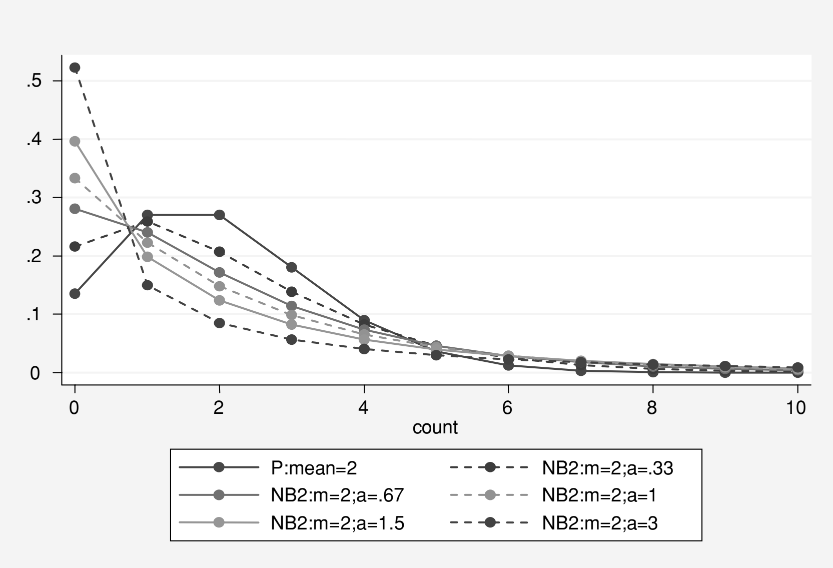

Figure 5.9. Negative binomial distributions: mean = 2

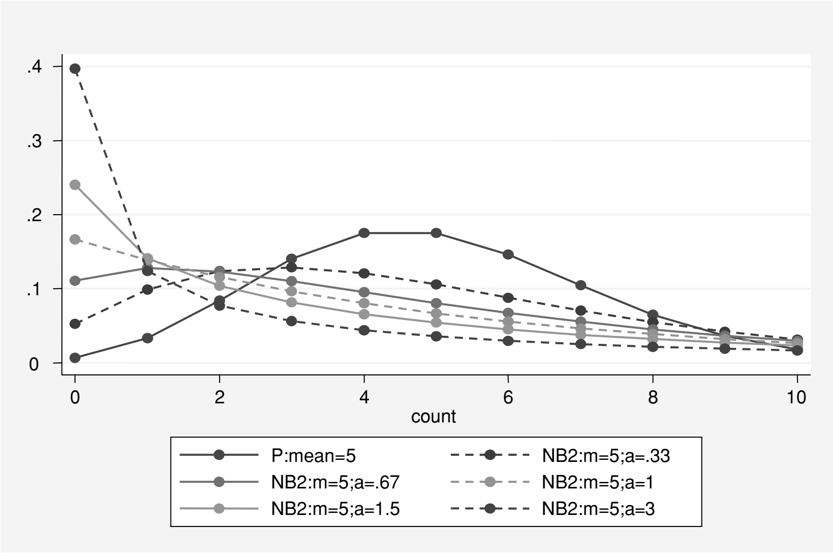

Figure 5.10. Negative binomial distributions: mean = 5

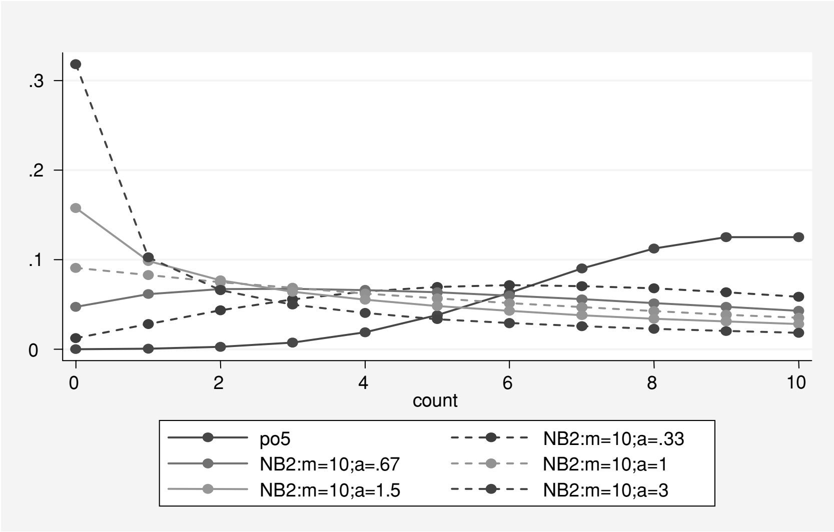

Figure 5.11. Negative binomial distributions: mean = 5

5.4 Algorithms

5.4.1 NB-C: Canonical negative binomial

We have derived the IRLS functions required to construct the canonical form of the negative binomial, which we shall refer to using the acronym, NB-C. When we derived the Poisson–gamma mixture model, the resultant PDF, log-likelihood, cumulant, and so forth, were all appropriate for the traditional negative binomial or NB-2 model. It is for this reason that statisticians have tended to think that this form of the negative binomial is basic. As a mixture model, it is; as a model directly derived from the negative binomial PDF, it is not. This has caused some confusion among those using negative binomial models.

Substituting the canonical and inverse link functions into the GLM algorithm, we produce the canonical form of negative binomial regression as schematized in Table 5.2.

I have used the deviance as the basis for the convergence criterion. The log-likelihood function could have been used as well. The deviance statistic has been the commonly used form since the initial release of the GLIM software package in the early 1980s. A trend is developing, however, to use the log-likelihood instead (when appropriate), calculating the deviance function

Table 5.2. *Negative binomial: canonical*

```
$\mu$ = (y − mean(y))/2            /*initialization of $\mu$ */
$\eta$ = −ln(1/$\alpha\mu$ + 1)          /*NB canonical link */
WHILE(ABS($\Delta$Dev) > tolerance {
    w = $\mu + \alpha\mu^2$                   /*NB variance function */
    z = $\eta$ + (y − $\mu$)/w − offset
    $\beta$ = (X'wX)$^{-1}$X'wz
    $\eta$ = X'$\beta$ + offset             /* calculation of linear predictor */
    $\mu$ = 1/($\alpha$(exp(−$\eta$) − 1))     /*NB inverse link */
    oldDev = Dev
    Dev = 2$\Sigma${ yln(y/$\mu$) − ($y$ + 1/$\alpha$)ln((1 + $\alpha$y)/(1 + $\alpha\mu$))}/* Deviance*/
    $\Delta$Dev = Dev − oldDev
}
```

after the parameters have been estimated. The deviance statistic can be derived directly from the final values of μ and α, and used as a term in the BIC goodness-of-fit statistic. The deviance is also used as a goodness-of-fit statistic in its own right, with lower values indicating a comparatively preferable model. However, most statisticians now prefer the use of the AIC, BIC, and other model-specific fit statistics to the deviance.

Note that α enters the algorithm as a constant. Unlike the traditional NB-2 negative binomial, the canonical linked algorithm incorporates α into the link, inverse link, and variance functions. Having α as a term in the link and inverse link resulted in convergence difficulties with older estimating algorithms. Convergence seemed to be particularly tedious when estimated via Newton-Raphson type maximum likelihood. However, most of the current optimization code is sufficiently sophisticated to handle canonically linked models without difficulty.

The canonical model has not been used for any research project of which I am aware. However, this need not be the case. It is a viable parameterization, and is directly derived from the negative binomial probability and likelihood functions. Using it with various example data has at times resulted in a better fit than modeling with NB-2 or NB-1 models. In addition, exact statistics algorithms can be developed for canonically linked GLM-type models.

Cytel's LogXact has the capability of calculating exact p-values and confidence intervals for logit and Poisson regression models. The logit link is the canonical form derived from the binomial and Bernoulli distributions. The natural log link is canonical for Poisson. Exact statistics developed for negative binomial models can utilize the canonical form.

Table 5.3. *Fisher scoring: expected imformation matrix*

```
$\mu$ = (y − mean(y))/2
$\eta$ = ln($\mu$)
WHILE(ABS($\Delta$Dev) > toleration {
    w = $\mu + \alpha\mu^2$
    z = $\eta$ + (y − $\mu$)/w − offset
    $\beta$ = $(X'wX)^{-1}X'wz$
    $\eta$ = X'$\beta$ + offset
    $\mu$ = exp($\eta$)
    oldDev = Dev
    Dev = 2$\Sigma${ yln(y/$\mu$) − (y + 1/$\alpha$)ln((1 + $\alpha$y)/(1 + $\alpha\mu$))}
    $\Delta$Dev = Dev − oldDev
}
```

5.4.2 NB-2 – expected information matrix

To convert the canonically linked GLM model to a non-canonical natural log link, change the initial values of the link and inverse link.

```
Link:           η = ln(μ)
Inverse link:   μ = exp(η)
```

When we substitute the non-canonical log link into the GLM algorithm, the standard errors change from being generated on the basis of the observed information matrix to the expected information matrix. In smaller and unbalanced data sets, the calculated standard errors will differ from those produced by full maximum likelihood algorithms, which employ observed information matrix-based standard errors. Nevertheless, the differences are not usually enough to change the apparent significance level of the model predictors. However, when those values lie near the edge of a pre-assigned level of significance, e.g. 0.05, apparent significance may change. In medium and large data sets this situation is not usually of concern.

The value of using a Fisher scoring method of estimation, which uses the expected information matrix for production of standard errors, is the simplification of the Newton–Raphson steps to a sequence of weighted ordinary least squares model fits. Furthermore, models can be easily changed from one to another within a distributional family by simply changing the link and inverse link functions. GLMs can be interchanged between families by changing the variance and deviance functions, as well as the link and inverse link functions. Thus, all GLMs are specified through four functions. Creating an overall IRLS algorithm for the estimation of GLMs is thus a relatively simple matter, and it affords a great deal of modeling flexibility.

The IRLS algorithm can be schematized as shown in Table 5.3.

Table 5.4. *Negative binomial regression (log linked) with iterative estimation of α via χ^2 dampening*

```
Poisson y < predictors >
Chi2 = $\Sigma(y - \mu)^2/\mu$
Disp = Chi2/df
$\phi$ = 1/disp
j = 1
WHILE(ABS($\Delta$Disp) > tolerance  {
    oldDisp = Disp
    NB y < predictors >, $\alpha = \phi$
    Chi2 = $\Sigma\{(y - \mu)^2/(\mu + \alpha\mu^2)\}$
    Disp = Chi2/df
    $\phi$ = Disp*$\phi$
    $\Delta$Disp = Disp − oldDisp
    j = j + 1
}
```

The problem of using Fisher scoring for modeling the log-negative binomial is the necessity of entering α into the algorithm as a constant. Alternative values of α result in different parameter estimates and standard errors.

A fairly accurate point estimate of α can be obtained by searching for the value of α which results in the Pearson Chi2 (χ^2) dispersion statistic approximating 1.0. So doing indicates that Poisson overdispersion has been dampened from the model. This value is also close to that produced using maximum likelihood methods, which directly estimate α as a parameter.

Breslow (1984) was the first to develop this idea. Hilbe (1993a) developed an algorithm to iteratively search for the optimal value of α. The algorithm was made into a SAS macro based on the SAS\STAT GENMOD procedure. It was also implemented into Stata (Hilbe, 1993b) and Xplore software (Hilbe and Turlach, 1995).

The logic of the updating algorithm is simple. The algorithm begins by estimating a Poisson model. The inverse of the Pearson χ^2 dispersion statistic is calculated, and is given the value of a constant, ϕ. ϕ is equated with α, and entered into the GLM negative binomial model. After estimation of the negative binomial, another χ^2 dispersion statistic is calculated. This time, however, the value of ϕ is multiplied by the dispersion, resulting in an updated value of ϕ. Convergence is based on minimizing the difference between old and new dispersion statistics. Once convergence is achieved, the value of α is recorded. It is the optimal value of α produced by this dampening algorithm.

Table 5.4 schematizes the algorithm.

Again, the algorithm estimates the negative binomial model, and the ancillary or heterogeneity parameter, α, by iteratively forcing Pearson χ^2 to 1. The deviance may be used in place of χ^2 to determine the dispersion; however, it involves more terms to calculate. On the other hand, as previously mentioned, simulation studies appear to indicate that dampening by the Pearson χ^2 dispersion results in a value of α that is closer to that estimated by full maximum likelihood than is dampening by the deviance dispersion.

5.4.3 NB-2 – observed information matrix

Finally we parameterize the IRLS log-negative binomial (NB-2) by transforming the weight function so that standard errors are based on an observed rather than an expected information matrix. This subject was initially addressed in Chapter 2.

To effect this conversion, various terms need to be calculated, and introduced into the estimating algorithm. Common to both Newton–Raphson and Fisher scoring are:

LINK

$$\eta = g(\mu) = \ln(\mu)$$

hence

$$g'(\mu) = 1/\mu$$

and

$$g''(\mu) = -1/\mu^2$$

VARIANCE

$$V(\mu) = \mu + \alpha\mu^2$$

hence

$$V(\mu) = 1 + 2\alpha\mu$$

and

$$V^2 = (\mu + \alpha\mu^2)^2$$

DEFINING *W*

$$\begin{aligned} u &= (y-\mu)g'(\mu) = (y-\mu)/\mu \\ w^{-1} &= V\{g'(\mu)\}^2 = (\mu+\alpha\mu^2)/\mu^2 = (1+\alpha\mu)/\mu \\ w &= \mu/(\alpha\mu) \end{aligned}$$

Table 5.5. *NB-2 negative binomial with observed information matrix*

```
$\mu = (y - mean(y))/2$
$\eta = ln(\mu)$
WHILE (ABS($\Delta$Dev) > toleration {
    $w = \mu + \alpha\mu) + (y - \mu)\{\alpha\mu/(1 + 2\alpha\mu + \alpha^2 + \mu^2)\}$
    $z = \{\eta + (y - \mu)/(w(1 + \alpha\mu))\} - offset$
    $\beta = (X'wX)^{-1}X'wz$
    $\eta = X'\beta + offset$
    $\mu = exp(\eta)$
    oldDev = Dev
    $Dev = 2\Sigma\{yln(y/\mu) - (y + 1/\alpha)ln((1 + \alpha y)/(1 + \alpha\mu))\}$
    $\Delta Dev = Dev - oldDev$
}
```

DEFINING w_0

The observed information matrix adjusts the weights, w, such that

$$\begin{aligned} w_0 &= w + (y - \mu)\{V(\mu)g''(\mu) + V'(\mu)g'(\mu)\}/\{V^2 g'(\mu)^3\} \\ &= \mu/(1 + \alpha\mu) + (y - \mu)\{[-(\mu + \alpha\mu^2)/\mu^2] \\ &\quad + [(1 + 2\alpha\mu)/\mu]\}/[(\mu + \alpha\mu^2)/\mu^3] \\ &= \mu/(1 + \alpha\mu) + (y - \mu)\{\alpha\mu/(1 + 2\alpha\mu + \alpha^2\mu^2)\} \end{aligned}$$

DEFINING z_0

A revised working variate, z_0, is defined as

$$\begin{aligned} z_0 &= \eta + w_0^{-1} wu \\ &= \eta + (y - \mu)/\{w_0(1 + \alpha\mu)\} \end{aligned}$$

with w and w_0 representing diagonal weight matrices. Substituting w_0 and z_0 into the log-negative binomial algorithm provides the proper adjustment.

Table 5.5 schematizes the IRLS estimating algorithm, which is adjusted such that standard errors are produced from the observed information matrix.

This IRLS algorithm will produce the same estimates and standard errors as that of a full maximum likelihood algorithm, which also estimates α. However, the GLM IRLS algorithm only allows α to be entered as a constant. This is a rather severe limitation of the GLM approach. As we have observed however, an updating algorithm synthesized into the IRLS algorithm can approximate the maximum likelihood value of α. A tactic that many statisticians implement is to first estimate α using a maximum likelihood routine, then substitute that value of α into the a GLM-based algorithm that is adjusted to calculate an observed information matrix as in Table 5.5. Typically the algorithm schematized in Table 5.5 uses a log-likelihood function as the basis of convergence rather than the deviance. Either method produces the same result.

We next address issues related to the development of well-fitted negative binomial models. We shall limit our discussion to the NB-2, or log-negative binomial, model. Subsequent chapters will deal with alternative parameterizations and extensions to the base model.

5.5 Summary

In this chapter we introduced the both the canonical and traditional forms of the negative binomial model. The traditional form of the model is a Poisson–gamma mixture model in which the gamma distribution is used to adjust the Poisson in the presence of overdispersion. The original manner of expressing the negative binomial variance clearly shows this mixture relationship: $\mu + \mu^2/\nu$. μ is the Poisson variance and μ^2/ν the one-parameter gamma distribution variance. We inverted the gamma scale function, ν, to α, the negative binomial ancillary or heterogeneity function: $\mu + \alpha\mu^2$. This provides a direct relationship of α to the amount of overdispersion in the otherwise Poisson model. α is sometimes referred to as the overdispersion parameter.

The traditional negative binomial model, as defined above, is many times referred to as the NB-2 model, with 2 indicating the degree of the exponential term. We also derived the negative binomial model directly from the negative binomial probability fiunction. This is the normal method used to derive any of the members of family of generalized linear models.

When the negative binomial is derived from its probability function, the canonical form is different from the Poisson–gamma mixture model version, of NB-2. We showed the canonical linked algorithm, defining its primary components. We also showed how the canonical form of the model can be amended to the traditional of NB-2 form. We do this by converting the canonical form to a log-linked form, which is the same link as the canonical Poisson. We see then that the GLM log-negative binomial is the same as the NB-2 model that is traditionally estimated using full maximum likelihood methods. As a non-canonical linked model though, it must be further amended to allow estimation of standard errors based on the observed rather than expected information matrix. All of the relevant derivations are shown in this chapter.

By equating the traditional Poisson–gamma mixture model parameterization with the amended GLM log-negative binomial model, we find that either version can be used to estimate parameters and standard errors. The only drawback with the GLM version is that the heterogeneity parameter, α, is not estimated, but rather has to be entered into the GLM model as a constant. Algorithms have been developed using point estimate methods that provide a close approximation of the value of α as compared with the value estimated using full maxiumum likelihood methods. In fact, this was the method used by the author to construct a SAS macro for the negative binomial in 1993.

We argue that a researcher should use a full maximum likelihood algorithm to estimate the negative binomial, but then insert the resultant value of α into a GLM log-negative binomial model, estimating the same parameter estimates and standard errors as in the initial estimation. The value of doing this consists in the use of GLM residuals and fit statistics, which are standard options and output in most commercial GLM applications. This allows the user with a variety of methods to evaluate the fit of the model.

We next turn to the task of modeling with negative binomial regression.

Exercises

1 Refer to question 1 of the Chapter 3 exercises. Is there a general principle for relating the constants of two negative binomial models, one of which has a response with values *x* times that of the other model?

2 What are the essential differences between NB-1 and NB-2 models? Why is NB-2 considered as the standard negative binomial model?

3 Model the data below using both Poisson and negative binomial regression. The natural log of *pyears* is to be entered as an offset. *Deaths* is the response with *age* the explanatory predictor. *Age* represents age groups, and should therefore be entered into the model as leveled. Explain the value of the Pearson-dispersion and what it indicates in this circumstance. Is negative binomial (NB-2) the appropriate model for these data? Is Poisson? Explain.

age	smokes	deaths	pyears
<40	1	32	52407
<40	0	2	18790
41-50	1	104	43248
41-50	0	12	10673
51-60	1	206	28612
51-60	0	28	5710
61-70	1	186	12663
61-70	0	28	2585
>70	1	102	5317
>70	0	31	1462

4. Re-express the formulae from Equations (5.9) to (5.16) from a parameterization in terms of μ to a parameterization in terms of $x\beta$.
5. In the same manner of the graphs shown in 5.3, create a graph showing: (1) a negative binomial with mean of 10 and alpha 0.1, (2) a negative binomial with mean 10 and alpha 0.25, and (3) a Poisson with a mean of 10. What are the reasons for the differences in the three graphs? What type of trend do you see? What do you predict happens to the graph if the mean were changed to 20?
6. Graphically demonstrate how the negative binomial distribution approaches the Poisson as the value of alpha nears zero.
7. For a negative binomial distribution with a mean of 2 and N of 30, calculate the probability of three or fewer events for each of the conditions listed

   $\alpha = 0 \qquad \alpha = 0.5 \qquad \alpha = 1.0 \qquad \alpha = 1.5 \qquad \alpha = 2.0$

8. Show that

$$\begin{aligned}&(1/1 + \alpha\mu)^{1/\alpha}\,(1 - 1/\,(1 + \alpha\mu))^{y}\\ &\quad = ((1/\alpha)\,/\,(1/\alpha + \mu))^{1/\alpha}\,(\mu/(1/\alpha + \mu))^{y}\end{aligned}$$

9. Given a negative multinomial probability mass function of

$$\frac{\Gamma\,(\Sigma y_i + 1/\alpha)}{\Gamma(1/\alpha)\Pi\Gamma\,(y_i + 1)}\left(\frac{1}{1 + \alpha\Sigma\mu_i}\right)^{1/a}\prod\nolimits_{i=1}^{j}\left(1 - \frac{1}{\alpha\Sigma\mu_i}\right)^{y_i}$$

   what value of j reduces the function to a negative binomial as given in Equation (5.8)? Extra credit: Determine the negative multinomial log-likelihood function in terms of Equations (5.30) or (5.31).
10. Discuss how the negative binomial ancillary parameter differs from the scale parameter of models such as gamma and inverse Gaussian.
11. Model the data in exercise 8 of Chapter 4 using negative binomial regression. Compare the parameter estimates and dispersion statistics with a Poisson model of the data. Discuss the reasons for differences, if any.

6
Negative binomial regression: modeling

In this chapter we describe how count response data can be modeled using the NB-2 negative binomial regression. NB-2 is the traditional parameterization of the negative binomial model, and is the one with which most statisticians are familiar. For this chapter, then, any reference to negative binomial regression will be to the NB-2 model unless otherwise indicated.

6.1 Poisson versus negative binomial

When earlier describing how apparent overdispersion may be dealt with in a model, we created simulated data sets to demonstrate the effect of interactions, transformations, and so forth, on the model fit. The same will be done here, showing how the negative binomial model accommodates overdispersion in Poisson data. We begin then by creating simulated negative binomial data, with the value of the ancillary parameter, α, specified as 0.5. To affect the results desired, the number of observations will be set at ten thousand. Table 6.1 shows the steps to create this data set.

Modeling the synthetic data using full maximum likelihood results in the following output:

NEGATIVE BINOMIAL: MAXIMUM LIKELIHOOD

```
. nbreg ynb x1 x2, nolog

Negative binomial regression      Number of obs   =      10000
                                  LR chi2(2)      =   12574.54
Dispersion        =   mean        Prob > chi2     =     0.0000
Log likelihood    =   -19267.411  Pseudo R2       =     0.2460

==========================================================================
     ynb |  Coef.    Std. Err.    z    P>|z|  [95% Conf.    Interval]
---------+----------------------------------------------------------------
      x1 |  1.483868  .0120109 123.54 0.000    1.460327     1.507409
      x2 | -.7434778  .0108178 -68.73 0.000   -.7646803    -.7222752
   _cons |  .5094076  .0126692  40.21 0.000    .4845764     .5342388
---------+----------------------------------------------------------------
/lnalpha | -.7077611  .0266405                -.7599755    -.6555467
---------+----------------------------------------------------------------
   alpha |  .4927462   .013127                 .4676779     .5191582
--------------------------------------------------------------------------
Likelihood-ratio test of alpha = 0:chibar2(01) = 2.2e+04 Prob> =
chibar2 = 0.000
==========================================================================
```

Table 6.1. *Synthetic negative binomial with alpha = 0.5*

```
. set obs 10000
. gen x1 = invnorm(uniform())
. gen x2 = invnorm(uniform())
. gen xb = .5 + 1.5*x1 -.75*x2
. gennbreg ynb, xbeta(xb) alpha(.5) /* Creates NB model */
```

Since we synthesized the data to be modeled using negative binomial regression, it is no surprise that the model fits well. The parameter estimates show close agreement to the deterministic values used to generate the data, and the likelihood ratio test indicates that the data are significantly different than Poisson. Modeled as a GLM, we observe additional statistics:

NEGATIVE BINOMIAL: GLM

```
. glm ynb x1 x2, nolog fam(nb.4927462)

Generalized linear models              No. of obs      =      10000
Optimization     :  ML                 Residual df     =       9997
                                       Scale parameter =          1
Deviance         =  9794.929447        (1/df) Deviance =   .9797869
Pearson          =  10022.89697        (1/df) Pearson  =    1.00259
Variance         :  V(u) =             [Neg. Binomial]
function            u+(.4927462)u^2
Link function    :  g(u) = ln(u)       [Log]
                                       AIC             =   3.854082
Log likelihood   =  -19267.41111       BIC             =  -82280.84

------------------------------------------------------------------------
             |               OIM
         ynb |      Coef.   Std. Err.      z    P>|z|   [95% Conf. Interval]
-------------+----------------------------------------------------------
          x1 |   1.483868   .0120109   123.54   0.000    1.460327    1.507409
          x2 |  -.7434778   .0108177   -68.73   0.000     -.76468   -.7222755
       _cons |   .5094076   .0126692    40.21   0.000    .4845764    .5342387
------------------------------------------------------------------------
```

We now observe (Pearson Chi2 dispersion 1.00) that the data are not overdispersed – as a negative binomial model. That is, the model accommodates for extra correlation that might be in the data. Recall that if we were to re-run the simulation code above, slightly different data, and hence parameters, would result. Setting a random seed value to a specific number allows the same sequence of pseudo-random numbers to be generated so that re-fitting the model produces the same results.

Running a Poisson model on the negative binomial data results in:

POISSON: USING NEGATIVE BINOMIAL DATA

```
. glm ynb x1 x2, nolog fam(poi)

Generalized linear models          No. of obs          =      10000
Optimization      :   ML           Residual df         =       9997
                                   Scale parameter     =          1
Deviance          =   40148.47459  (1/df) Deviance     =   4.016052
Pearson           =   44637.58746  (1/df) Pearson      =   4.465098
                                   AIC                 =   6.090518
Log likelihood    =   −30449.59071 BIC                 =   −51927.3

================================================================
        |                OIM
--------+-------------------------------------------------------
    ynb |     Coef.   Std. Err.      z    P>|z|  [95% Conf. Interval]
--------+-------------------------------------------------------
     x1 |  1.478781  .0041106   359.75  0.000   1.470724   1.486838
     x2 | −.7028355  .0040693  −172.72  0.000  −.7108111  −.6948598
  _cons |  .5378572  .0076955    69.89  0.000   .5227743   .5529401
================================================================
```

The Poisson model indicates that the data are highly overdispersed – with a Pearson dispersion of 4.465. Moreover, the absolute values of the parameter z-values are inflated as compared with the "true" negative binomial values. Inflation of the z statistics result from a deflation of associated standard errors. This is a typical result of overdispersion. Overdispersed Poisson models many times lead us to believe that predictors significantly contribute to the model when in fact they do not.

We next look at the results of scaling Poisson standard errors to see if the model effects of overdispersion are attenuated. Only the table of parameter estimates is shown; the statistics in the header are the same as for the standard model.

POISSON: SCALED STANDARD ERRORS

```
==========================================================================
        |                 EIM
--------+-----------------------------------------------------------------
    ynb |      Coef.     Std. Err.        z      P>|z|     [95% Conf.     Interval]
--------+-----------------------------------------------------------------
     x1 |    1.478781      .008686    170.25    0.000     1.461757     1.495805
     x2 |   −.7028354     .0085987    −81.74    0.000    −.7196886    −.6859823
  _cons |    .5378572     .0162612     33.08    0.000     .5059858     .5697286
--------------------------------------------------------------------------
(Standard errors scaled using square root of Pearson X2-based dispersion)
==========================================================================
```

For this model, scaling roughly doubles the standard errors of each parameter, including the constant. This adjustment is made in the direction of the true negative binomial statistics, but is not nearly enough.

Table 6.2. *Comparison of model z statistics*

		x1	x2	_con
Negative binomial z		123.54,	−68.73	40.21
Poisson z		359.75,	−172.72	69.89
Scaled Poisson z	(Chi2)	170.25,	−81.74	33.08
	(dev)	179.52,	−86.19	34.88
Robust z		50.45	−29.15	17.75
Bootstrapped z		55.08	−25.94	12.66

Standard errors may also be scaled using the deviance dispersion rather than the Pearson Chi2 dispersion statistic, although I do not recommend it. In addition, standard errors may be adjusted using a robust variance estimator or by employing a bootstrap mechanism. Table 6.2 displays the z-statistics for the true negative binomial model and for the alternative Poisson models.

It appears that using robust or bootstrap techniques for estimation of standard errors tend to overadjust standard errors relative to the values of their respective parameter estimates. Still, adjustments to the Poisson do not closely approximate the "true" negative binomial estimates.

It is clear from what has been discussed that when data are referred to as Poisson or count overdispersed, the data may in fact be distributed as negative binomial. That is, overdispersed Poisson data may be the same as negative binomial data. Of course, there are many other types of overdispersed count data – count data situations in which the negative binomial does not account for, or explain, overdispersion in the data. The negative binomial model may have to be itself extended, or other Poisson-type models may have to be evaluated, in order to deal with the data.

Count data have also been analyzed using a gamma model first proposed by Winkelmann (1995), by two-parameter log-gamma or log-inverse Gaussian models (Hilbe, 2000), by a generalized Poisson, by a generalized negative binomial, or by a generalized binomial. The generalized negative binomial was first defined and analyzed by Jain and Consul (1971), but was later amended by Consul and Gupta (1980) following Nelson's (1975) discovery that the distribution did not sum to one when the heterogeneity parameter was less than zero. The generalized binomial, defined by Consul and Gupta (1980) is based on the generalized negative binomial distribution. Famoye (1995) presented an excellent overview of all three generalized models, expanding the generalized binomial to include multiple predictors. Together with the generalized Poisson, the generalized binomial will be discussed at greater length in Chapter 7. All of these methods have primarily been used to handle underdispersed Poisson data, although they can handle overdispersion as well. Regardless, the negative binomial and its many extensions have been used with considerable effectiveness for dealing with a wide variety of count models.

6.2 Binomial versus count models

At times data come to us in the form of individual data that we wish to model as exposures. It is also the case that we may need to decide if data should be modeled as logistic or as a count. To draw out these relationships I shall begin with a noted data set that has traditionally been analyzed using a logistic model. The data come to us from Hosmer and Lemeshow (2003), *Applied Logistic Regression*, 2nd edition. Called the *low birth weight* (*lbw*) data, the response is a binary variable, *low*, which indicates whether the birth weight of a baby is under 2500g (1), or over (0). To simplify the example, only two predictors will be included in the model: *smoke* (1/0), and *race*. *Race* enters the model as a three-level factor variable, with level 1 as the referent. I have also expanded the data threefold, i.e. I multiplied each covariate pattern by three. So doing allows the example to more clearly illustrate the relevant relationships. The initial data set consists of 189 observations; we utilize a subset of 567.

The data are first modeled as a logistic regression. Parameter estimates have been exponentiated to odds ratios. Note also the low value of the AIC and BIC goodness-of-fit statistics.

```
. glm low smoke race2 race3, nolog fam(bin) eform

Generalized linear models      No. of obs      =       567
Optimization    : ML           Residual df     =       563
                               Scale parameter =         1
Deviance        = 659.9241316  (1/df) Deviance =  1.172157
Pearson         = 553.7570661  (1/df) Pearson  =  .9835827
                               AIC             =  1.177997
Log likelihood  = -329.9620658 BIC             = -2909.698

==========================================================
     |           OIM
-----+----------------------------------------------------
 low |Odds Ratio Std. Err.   z   P>|z|  [95% Conf. Interval]
-----+----------------------------------------------------
smoke|3.052631   .6507382  5.24 0.000 2.010116    4.635829
race2|2.956742   .8364414  3.83 0.000 1.698301    5.147689
race3|3.030001   .7002839  4.80 0.000 1.926264    4.766171
==========================================================
```

Observation level data can be converted to grouped format, i.e. one observation per covariate pattern, using the following Stata code. Users of other packages can easily apply the same logic to produce the same result.

Table 6.3. *Low birth weight data*

n = 567	data = lbwch6
variable name	variable label
low	birth weight<2500g
smoke	smoked during pregnancy
race1	race = = white
race2	race = = black
race3	race = = other

```
/* Be certain that data consist of low, smoke, race1
   race2 race3 only */

  . egen grp = group(smoke-race3)
  . egen cases = count(grp), by(grp)
  . egen lowbw = sum(low), by(grp)
  . sort grp
  . by grp: keep if _n == 1 /* discard all but 1st of
                                like CP's */
```

In grouped format, the logit model appears as

```
. glm lowbw smoke race2 race3, nolog fam(bin cases) eform

Generalized linear models      No. of obs      =         6
Optimization    : ML           Residual df     =         2
                               Scale parameter =         1
Deviance        = 9.470810009  (1/df) Deviance =  4.735405
Pearson         = 9.354399596  (1/df) Pearson  =    4.6772
                               AIC             =  7.443406
Log likelihood = -18.33021651 BIC             = -3.209909

--------------------------------------------------------------
                   OIM
lowbw|Odds Ratio Std. Err.   z   P>|z|  [95% Conf. Interval]
-----+--------------------------------------------------------
smoke|3.052631   .6507383  5.24 0.000  2.010117    4.635829
race2|2.956742   .8364415  3.83 0.000  1.698301     5.14769
race3|3.030001   .700284   4.80 0.000  1.926264    4.766172
--------------------------------------------------------------
```

Parameter estimates and associated standard errors and confidence intervals are identical, except for small differences resulting from estimation rounding errors. On the other hand, both AIC and BIC statistics have markedly risen.

Overdispersion makes no sense for a binary response model. Therefore, the dispersion statistic displayed in the output of a binary logistic model does not indicate overdispersion, or, in this instance, equidispersion. For binomial models, overdispersion can only be assessed when the data are formatted as grouped. Here the binomial model has significant overdispersion – 4.677. We could scale the model, estimate robust variance estimators, use bootstrapping and jackknife techniques, or engage in specialized procedures such as the Williams Procedure (Collett, 1989), to handle the overdispersion. We should also look to see if the overdispersion is real or only apparent. In addition, we can model the data using an entirely different GLM family. Once the binary logistic data have been converted to grouped format, the binomial numerator can be modeled as Poisson. The binomial numerator is considered as a count (rather than a success) and the denominator is considered an exposure. Exposures enter the model as a log-transformed offset.

```
. glm lowbw smoke race2 race3, nolog fam(poi) eform
lnoffset(cases)

Generalized linear models      No. of obs      =        6
Optimization    : ML           Residual df     =        2
                               Scale parameter =        1
Deviance        = 9.717852215  (1/df) Deviance = 4.858926
Pearson         = 8.863298559  (1/df) Pearson  = 4.431649
                               AIC             = 7.954159
Log likelihood  = -19.86247694 BIC             = 6.134333

==========================================================
     |             OIM
-----+----------------------------------------------------
lowbw|    IRR     Std. Err.   z   P>|z|  [95% Conf. Interval]
-----+----------------------------------------------------
smoke|  2.020686 .3260025  4.36 0.000   1.472897  2.772205
race2|  1.969159 .4193723  3.18 0.001    1.29718  2.989244
race3|  2.044699 .3655788  4.00 0.000   1.440257   2.90281
cases| (exposure)
==========================================================
```

Not surprisingly, the model is still overdispersed (4.43), which is similar to that of the binomial logistic model (4.68). The primary difference due to reparameterization is in how the estimates are interpreted. The AIC and log-likelihood statistics are similar in each model, but not the BIC. The latter value indicates that the logistic model is preferable.

We know that the model is not finalized. Since the Poisson model is overdispersed, one must check for apparent versus real overdispersion, and take appropriate remedies based on each alternative. Given that the data are truly overdispersed, we can model the data as negative binomial to see if overdispersion is accommodated. A maximum likelihood negative binomial algorithm is initially applied to obtain an estimate of the ancillary parameter, α.

```
. nbreg lowbw smoke race2 race3, nolog irr exposure(cases)

Negative binomial regression        Number of obs    =          6
                                    LR chi2(3)       =       8.37
Dispersion         =     mean       Prob > chi2      =     0.0389
Log likelihood     =    −19.472423  Pseudo R2        =     0.1770

------------------------------------------------------------------
   lowbw |     IRR    Std. Err.  z    P>|z|   [95% Conf.  Interval]
---------+--------------------------------------------------------
   smoke |  2.035876 .4337797   3.34  0.001    1.340873   3.091114
   race2 |  2.072214 .5627557   2.68  0.007    1.216948   3.528558
   race3 |  2.063825 .5048891   2.96  0.003    1.277724   3.333563
   cases | (exposure)
---------+--------------------------------------------------------
/lnalpha | −3.707643 1.557557              −6.760397 −.6548881
---------+--------------------------------------------------------
   alpha |  .0245353 .0382151                .0011588   .5195002
---------+--------------------------------------------------------
Likelihood-ratio test of alpha = 0: chibar2(01) = 0.78
                                    Prob> = chibar2 = 0.189
------------------------------------------------------------------
```

The output indicates that the model is not statistically different from a Poisson model. The likelihood ratio test determining if α is statistically different from zero fails. In effect, the value of α is approximately zero, indicating a preference for the more parsimonious Poisson model.

If considered as a count model, it is unlikely that we can do any better with the data than to consider them as Poisson. If Poisson distributional assumptions are not a concern however, calculating standard errors via bootstrap may be a viable modeling strategy.

```
------------------------------------------------------------------
      |  Observed   Bootstrap                Normal-based
lowbw |    IRR      Std. Err.    z     P>|z|   [95% Conf.  Interval]
------+-----------------------------------------------------------
smoke |  2.020686   .7705485   1.84   0.065   .9569959     4.266656
race2 |  1.969159   .7648926   1.74   0.081   .9196939     4.216171
race3 |  2.044699   1.017267   1.44   0.151   .7711595     5.421438
cases | (exposure)
------------------------------------------------------------------
```

It is clear that the grouped logistic model is preferred over Poisson and negative binomial rate parameterizations. Logistic parameter estimates appear to significantly contribute to understanding the response and both AIC and BIC statistics indicate a better fitted model. Specialized goodness-of-fit tests for logistic models are discussed in texts such as Hosmer and Lemeshow (2003), Collett (1989), and Hardin and Hilbe (2001), but go beyond the scope of our discussion.

An indicator of whether data should be modeled as a grouped logistic or as a rate parameterized count model relates to the ratio of successes to cases

Table 6.4. *Ratio of response values*

	lowbw	cases	%ratio
1	60	165	36.4
2	15	48	31.3
3	12	132	09.1
4	15	36	41.7
5	18	30	60.0
6	57	156	36.5

or counts to exposure respectively. This logic goes back to the derivation of the Poisson distribution from the binomial. Considered in this manner, Poisson models rare binomial events. If the ratio of the Poisson counts to the exposure is small, it is likely that a Poisson or negative binomial model will better fit the data. On the other hand, when the binomial numerator is close to the value of the denominator, it is likely that a logistic, probit, loglog, or complementary loglog model would be preferable. This topic is discussed at length in Hardin and Hilbe (2001, 2007).

The percent ratio of the variables *lowbw* and *cases* are provided in Table 6.4. The values of percent ratio do not indicate a clear preference for a binomial or count model. However, given the rather large mean percent of 36.5, the data appear to lean toward a logistic model. However, this is not always the case. Given a binary response model, we may also have reason to need information concerning risk. When this occurs, the data must be converted to a count response. To do this, however, the binary format must first be converted to grouped, and the binomial denominator must be entered into the count model as an exposure. Example 3 in this chapter will demonstrate a successful use of this method.

6.3 Examples: negative binomial regression

I shall present four examples demonstrating how negative binomial regression can be used to model count response data. These examples will also be used in later chapters when dealing with extensions to the basic form of negative binomial model.

Example 1: Modeling number of marital affairs

For the first example we shall evaluate data from Fair (1978). Although Fair used a tobit model with the data, the outcome measure can be modeled as a count. In fact, Greene (2003) modeled it as Poisson, but given the amount of

Table 6.5. *Example 1: affairs data*

affairs_nb.dta		
obs:	601	
naffairs	number of affairs within last year	
kids	1 = have kids;	0 = no kids
vryunhap	ratemarr = = 1	very unhappily married
unhap	ratemarr = = 2	unhappily married
avgmarr	ratemarr = = 3	avg marriage
hapavg	ratemarr = = 4	happily married
vryhap	ratemarr = = 5	very happily maried
antirel	relig = = 1	anti religious
notrel	relig = = 2	not religious
slghtrel	relig = = 3	slightly religious
smerel	relig = = 4	somewhat religious
vryrel	relig = = 5	very religious
yrsmarr1	yrsmarr = =	0.75 yrs
yrsmarr2	yrsmarr = =	1.5 yrs
yrsmarr3	yrsmarr = =	4.0 yrs
yrsmarr4	yrsmarr = =	7.0 yrs
yrsmarr5	yrsmarr = =	10.0 yrs
yrsmarr6	yrsmarr = =	15.0 yrs

Table 6.6. *Naffair: frequency of counts*

year	Freq.	Percent	Cum.
0	451	75.04	75.04
1	34	5.66	80.70
2	17	2.83	83.53
3	19	3.16	86.69
7	42	6.99	93.68
12	38	6.32	100.00
Total	601	100.00	

overdispersion in the data, employing a negative binomial model is an appropriate strategy (see Table 6.5).

Naffairs is the response variable, indicating the number of affairs reported by the participant in the past year. The classification of counts appears as Table 6.6.

The number of zeros in the data far exceeds the number reasonably expected by the distributional assumptions of both the Poisson and negative binomial. To observe differences between the observed and predicted number of zeros, as well as other values, we first model the data using Poisson regression.

```
. glm naffairs kids avgmarr-vryhap notrel-vryrel
yrsmarr3--yrsmarr6, fam(poi)

Generalized linear models          No. of obs      =        601
Optimization     :  ML             Residual df     =        588
                                   Scale parameter =          1
Deviance         =  2305.835984    (1/df) Deviance =    3.92149
Pearson          =  4088.616155    (1/df) Pearson  =   6.953429
                                   AIC             =   4.701873
Log likelihood   =  −1399.912931   BIC             =  −1456.538

----------------------------------------------------------------------------
         |             OIM
naffairs |     Coef.   Std. Err.      z    P>|z|    [95% Conf.  Interval]
---------+------------------------------------------------------------------
    kids | −.2226308  .1059723   −2.10   0.036   −.4303328  −.0149289
 avgmarr | −.8858196  .1050272   −8.43   0.000   −1.091669  −.6799701
  hapavg | −1.023898  .0859245  −11.92   0.000   −1.192307  −.8554889
  vryhap |  −1.38385  .1009577  −13.71   0.000   −1.581723  −1.185976
  notrel | −.6553382  .1111865   −5.89   0.000   −.8732597  −.4374166
slghtrel | −.5236987  .1113403   −4.70   0.000   −.7419218  −.3054756
  smerel | −1.370688  .1213036  −11.30   0.000   −1.608439  −1.132938
  vryrel | −1.363744  .1589703   −8.58   0.000    −1.67532  −1.052168
yrsmarr3 |  .7578109  .1612081    4.70   0.000    .4418488   1.073773
yrsmarr4 |  1.104536  .1698768    6.50   0.000    .7715832   1.437488
yrsmarr5 |  1.480332  .1648648    8.98   0.000    1.157203   1.803461
yrsmarr6 |  1.480467  .1555978    9.51   0.000    1.175501   1.785433
   _cons |  1.101651  .1648297    6.68   0.000    .7785906   1.424711
----------------------------------------------------------------------------
```

Exponentiating the parameter estimates results in them becoming incidence rate ratios.

```
----------------------------------------------------------------------------
         |            OIM
naffairs |      IRR   Std. Err.      z    P>|z|    [95% Conf.  Interval]
---------+------------------------------------------------------------------
    kids | .8004103  .0848213    −2.10   0.036   .6502927        .985182
 avgmarr |  .412376  .0433107    −8.43   0.000   .3356558       .5066321
  hapavg | .3591922  .0308634   −11.92   0.000   .3035203       .4250753
  vryhap | .2506118  .0253012   −13.71   0.000   .2056204       .3054478
  notrel | .5192664  .0577354    −5.89   0.000   .4175881       .6457023
slghtrel | .5923257  .0659497    −4.70   0.000   .4761979       .7367728
  smerel | .2539321  .0308029   −11.30   0.000   .2001998       .3220856
  vryrel | .2557017   .040649    −8.58   0.000   .1872483         .34918
yrsmarr3 |   2.1336  .3439536     4.70   0.000   1.555581         2.9264
yrsmarr4 | 3.017823  .5126582     6.50   0.000   2.163188       4.210108
yrsmarr5 | 4.394404  .7244825     8.98   0.000   3.181023       6.070621
yrsmarr6 | 4.394996  .6838517     9.51   0.000   3.239764       5.962159
----------------------------------------------------------------------------
```

Observed counts from Table 6.5 can be graphed against predicted counts based on the fitted values, μ_i, from the above model.

Slightly above 75% of observed zero counts clearly differ from the approximate 38% zeros predicted on the basis of the distributional assumption of the Poisson model. From a count of 3 upwards, the empirical and predicted distributions are similar. This type of distributional violation typically results

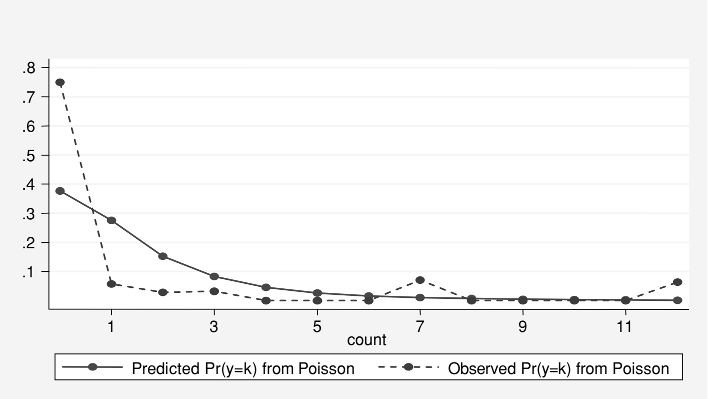

Figure 6.1. Poisson model for number of affairs: observed versus predicted probabilities

in substantial overdispersion, as indicated by the high value for the Pearson Chi2-based dispersion statistic. In this case the dispersion is 6.95.

Several statistical packages provide a Poisson goodness-of-fit test. It is simply the deviance evaluated by a Chi-square distribution with a degree of freedom equal to the number of observations less the number of predictors, including the constant. The statistic here tells us that, given the model, the hypothesis that the data are Poisson is rejected at the <.001 significance level.

```
Goodness-of-fit chi2  =  2305.836
Prob > chi2(588)      =    0.0000
```

Given what we know of the relationship of Pearson dispersion and overdispersion, the above goodness-of-fit test would likely be more effective using the Pearson rather than the deviance dispersion. Regardless, testing indicates that the overdispersion evidenced in both model output and as a result of various comparison tests is in fact real. A Lagrange multiplier test provides a value of 508.85, also indicating overdispersion. Additionally, Figure 6.1 provides excellent visual support for this conclusion.

As previously discussed, two foremost methods used to accommodate Poisson overdispersion are *post-hoc* scaling and application of a modified variance estimator. Typical modifications to the Hessian matrix are (1) White, or Huber, sandwich robust estimator, (2) bootstrapping, and (3) jackknifing. Other modifications are available as well in most of the major statistical packages.

The standard errors that result from an application of a robust variance estimator affects the parameter estimate p-values. Usually overdispersion serves to deflate p-values, perhaps misleading a researcher into believing that a predictor contributes to the model when in fact it does not. Application of a robust variance estimator, scaling of standard errors, or bootstrapping, usually inflates the p-values of an overdispersed count model. At times, however, given the interaction between predictors, such modification appears to work in the opposite direction. But this situation is not the usual.

It is also important to mention that the source of overdispersion in Poisson models may be clearly identifiable. For example, overdispersion arises in longitudinal or clustered data when they are modeled without taking into consideration the extra correlation effected as a result of the similarities of the observations within groups. When an obvious source of overdispersion is identified, we may attempt to find the specifically appropriate remedy for it. For longitudinal and clustered data, a robust variance estimator may be applied by considering each group or cluster to be a single observation, with a summary variance statistic given to each respective cluster. Treated as individuals, each group is then assumed to be independent from one another. Care must be taken when applying this method though. Consider data taken from hospitals within a medium-sized city. If there are only, for instance, four hospitals in the city, and we apply a robust variance estimator on clusters, we effectively reduce the size of the data set to four independent components. The parameter estimates are not affected by this method though – only the Hessian matrix from which the algorithm abstracts model standard errors. Using clustering methods is extremely helpful when dealing with identified overdispersion, but the number of clusters must be sizeable.

Application of a robust variance estimator to the *affairs* data gives us the following output. Notice that *kids* and *yrsmarr3* no longer contribute to the model. However, no source of overdispersion has been identified other than the inflated zeros. We shall later return to this example when we discuss methods of dealing with excessive zeros in the data, e.g. zero-inflated poisson, zero-inflated negative binomial, and hurdle models.

```
=====================================================================
         |             Robust
naffairs |      IRR   Std. Err.     z    P>|z|  [95% Conf.  Interval]
---------+-----------------------------------------------------------
    kids |.8004103 .2437552    −0.73 0.465 .4406457    1.453904
 avgmarr | .412376 .0995915    −3.67 0.000 .2568755    .6620094
  hapavg |.3591922 .0772231    −4.76 0.000 .2356818    .5474287
  vryhap |.2506118 .0655191    −5.29 0.000 .1501296    .4183471
  notrel |.5192664 .1333609    −2.55 0.011 .3138918    .8590145
slghtrel |.5923257 .1459787    −2.12 0.034 .3654112    .9601502
  smerel |.2539321 .0731995    −4.75 0.000 .1443267    .4467745
  vryrel |.2557017 .0953902    −3.66 0.000 .1230809    .5312229
yrsmarr3 |  2.1336 1.039479     1.56 0.120 .8211304    5.543882
yrsmarr4 |3.017823  1.66825     2.00 0.046 1.021293    8.917379
yrsmarr5 |4.394404  2.21657     2.93 0.003 1.635112    11.81007
yrsmarr6 |4.394996 2.230482     2.92 0.004 1.625434    11.88359
=====================================================================
```

Negative binomial regression employs an extra parameter, α, that directly addresses the overdispersion in Poisson models. Generally speaking, there is a direct relationship between the amount of overdispersion in a Poisson model and the value of α in a well-fitted negative binomial model. The relationship is clearly evident in the variance functions of the two models:

$$\begin{aligned}\text{Poisson variance} &= \mu\\ \text{Negative binomial variance} &= \mu + \mu^2\end{aligned}$$

Alternative parameterizations of the negative binomial will be considered in later chapters.

```
. glm naffairs kids avgmarr-vryrel yrsmarr3-yrsmarr6,
eform fam(nb 6.760067)

Generalized linear models              No. of obs      =        601
Optimization      : ML                 Residual df     =        588
                                       Scale parameter =          1
Deviance          = 339.9146951        (1/df) Deviance =   .5780862
Pearson           = 574.2568411        (1/df) Pearson  =   .9766273
Variance          : V(u) =             [Neg. Binomial]
function            u+(6.760067)u^2
Link function     : g(u) = ln (u)      [Log]
                                       AIC             =   2.453377
Log likelihood    = -724.2398359       BIC             = -3422.459


------------------------------------------------------------------
            |             OIM
   naffairs |     IRR  Std. Err.     z   P>|z|   [95% Conf.  Interval]
------------------------------------------------------------------
       kids | 1.091006  .3393095   0.28  0.779   .5930593     2.00704
    avgmarr | .3788406  .1626535  -2.26  0.024   .1633041    .8788527
     hapavg | .3754898  .1370964  -2.68  0.007   .1835748    .7680389
     vryhap | .2491712  .0936138  -3.70  0.000   .1193166    .5203494
    notrel  |  .735144  .3482966  -0.65  0.516   .2904624    1.860608
  slghtrel  | .6610617  .3191906  -0.86  0.391   .2565921    1.703102
     smerel | .2307172  .1071631  -3.16  0.002   .0928358    .5733825
     vryrel | .2202639  .1199509  -2.78  0.005   .0757526    .6404555
   yrsmarr3 |  1.95046  .7752988   1.68  0.093   .8949284    4.250947
   yrsmarr4 | 3.801339  1.695028   2.99  0.003   1.586293    9.109401
   yrsmarr5 | 3.283675  1.471676   2.65  0.008   1.364172    7.904079
   yrsmarr6 | 4.165032   1.61167   3.69  0.000   1.950939    8.891866
------------------------------------------------------------------
```

A non-nested likelihood ratio test of the log-likelihood of the full model against the log-likelihood of the Poisson model ($\alpha = 0$) on the same data informs us of whether the data are Poisson or non-Poisson. A significant p-value is usually

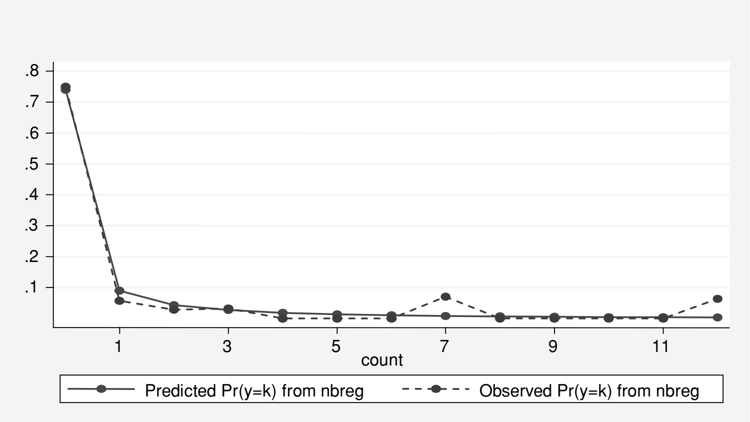

Figure 6.2. Negative binomial model for number of affairs: observed versus predicted probabilities

taken to mean that the model is negative binomial. It may be, but it also may be some variant of the basic negative binomial model.

Is any case, a likelihood ratio test of $\alpha = 0$ (the Poisson model) yields a χ^2, with 1 degree of freedom, of 1351.35. The corresponding p-value is $<.000001$, indicating that the negative binomial model with an α of 6.76 is significantly different from the Poisson. The negative binomial model produces AIC and BIC statistics of 2.45 and -3422.46 respectively. These values compare with the Poisson model statistics of 4.70 and -1456.54. The values for the negative binomial model are clearly and substantially less than those of the corresponding Poisson – indicating again that the data are better modeled as negative binomial rather than Poisson.

Visually comparing the observed and predicted counts for the negative binomial model may help in distinguishing it from the Poisson, as well as assist in determining if it is a preferable model.

Figure 6.2 clearly shows the close association between the observed counts of affairs and the number of affairs predicted on the basis of the negative binomial model. The fit is far superior to that shown for the Poisson model (Figure 6.1). Of particular interest is that the differences between predicted and observed zero counts are now minimal – only 0.01 (Table 6.7). On the other hand, the difference for the Poisson model is some 37.

The model is not yet finalized. There are several predictors that do not contribute to the explanation of the response. Interactions have been checked outside of our discussion and a final model developed, appearing as shown in the output below:

Table 6.7. *Observed vs predicted negative binomial model*

CNT	OBS	PRED
0	.7504	.7394
1	.0566	.0898
2	.0283	.0433
3	.0316	.0265
4	0	.0181

```
. glm naffairs avgmarr-vryhap smerel vryrel yrsmarr4-
yrsmarr6, nolog eform fam(nb 6.908197)

Generalized linear models          No. of obs      =        601
Optimization     : ML              Residual df     =        592
                                   Scale parameter =          1
Deviance         = 339.2113501     (1/df) Deviance =   .5729921
Pearson          = 517.6346425     (1/df) Pearson  =   .8743828
Variance         : V(u) =          [Neg. Binomial]
function           u+(6.908197)u^2
Link function    : g(u) = ln(u)    [Log]
                                   AIC             =   2.446937
Log likelihood = −726.3044443      BIC             = −3448.757

------------------------------------------------------------------
        |            OIM
naffairs|      IRR   Std. Err.     z   P>|z|   [95% Conf. Interval]
--------+---------------------------------------------------------
 avgmarr| .3720682 .1590264   −2.31 0.021    .1609937   .8598769
  hapavg| .3681403 .1352423   −2.72 0.007    .1791888   .7563381
  vryhap| .2514445 .0923022   −3.76 0.000    .1224549    .516307
  smerel| .3047693  .088186   −4.11 0.000    .1728515   .5373648
  vryrel|  .279958 .1149339   −3.10 0.002    .1252105   .6259576
yrsmarr4| 2.824666 1.125855    2.61 0.009    1.293288   6.169344
yrsmarr5| 2.462933 .9828394    2.26 0.024    1.126623   5.384268
yrsmarr6| 3.173011 .9689469    3.78 0.000     1.74397   5.773035
------------------------------------------------------------------
```

All predictors significantly enter the model and both AIC and BIC statistics have been reduced, albeit not significantly. The model can be re-fitted using a robust variance estimator to determine if empirically based standard errors result in a difference in p-value significance. In this case there is minimal difference, with significance moving in the opposite direction, i.e. appearing more contributory than less.

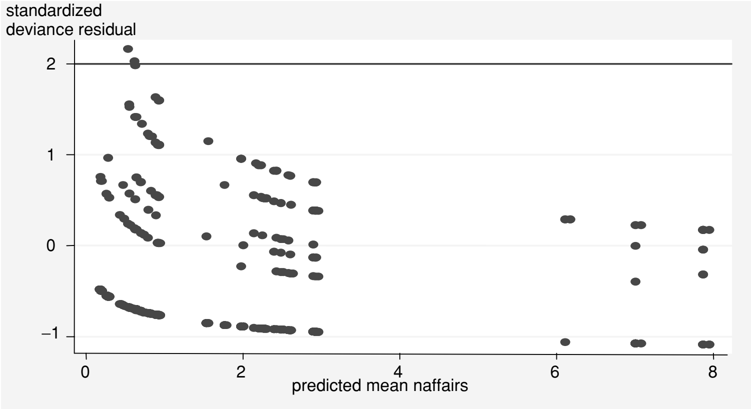

Figure 6.3. Negative binomial model for number of affairs: standardized deviance residuals versus fitted values

```
---------------------------------------------------------------------------
           |               Robust
  naffairs |      IRR    Std. Err.     z    P>|z|   [95% Conf. Interval]
-----------+---------------------------------------------------------------
   avgmarr | .3720682  .1075804   −3.42  0.001    .2111081     .655753
    hapavg | .3681403  .0885893   −4.15  0.000    .2297102    .5899925
    vryhap | .2514445  .0716387   −4.85  0.000    .1438558     .439498
    smerel | .3047693  .0710436   −5.10  0.000    .1929971    .4812731
    vryrel |  .279958  .0883555   −4.03  0.000    .1508174    .5196779
  yrsmarr4 | 2.824666  1.026682    2.86  0.004    1.385417    5.759086
  yrsmarr5 | 2.462933  .7091724    3.13  0.002    1.400746     4.33058
  yrsmarr6 | 3.173011  .7823943    4.68  0.000     1.95697    5.144688
---------------------------------------------------------------------------
```

An analysis of the standardized deviance residuals versus the fitted values shows that only two cases rest outside +/−2.0. These are considered outliers. However, the values are not far above 2.0 −2.029 and 2.165− and represent only two of the 601 cases in the data. Both of these outliers have counts of 12 for the reported number of affairs, being at the extreme high end of the tabulation of counts.

Lastly, it is understood that the levels of factor predictors are evaluated with reference to a referent level. When a contiguous level to the initially assigned referent is not itself significantly different from the referent, we may combine the two levels so that the resultant referent is a combination of the two levels. We do this by simply excluding both levels from the estimation. This was done for marriage status levels 1 and 2, religious status 1–3, and years married groups 1–3. Levels other than combined referent levels significantly contribute to the model (*naffairs*) when compared with the referent.

Table 6.8. *Testing interlevel predictor significance*

```
test avgmarr          =        vryhap
chi2(1)               =        1.19
Prob > chi2           =        0.2751

test smerel           =        vryrel
chi2(1)               =        0.04
Prob > chi2           =        0.8433

test yrsmarr4         =        yrsmarr6
chi2(1)               =        0.09
Prob > chi2           =        0.7623
```

It is also possible to evaluate the levels that are significantly different from the combined referent levels, now simply called the referent. They may not significantly differ from each other. If this turns out to be the case, then significant levels may themselves be combined. All major statistical packages allow levels to be tested for inter-level significance based on a χ^2 or Wald statistic. Stata has a number of ways such can be evaluated. The simplest is to use the commands shown in part in Table 6.8.

Table 6.8 indicates that each non-referent level in the model can be combined with one or more other significant levels. It is likely from an observation of the table that each model predictor can be dichotomized such that each is considered to be binary. For instance, with respect to marital status, a predictor called *marrstatus* can be defined as 1 = levels 3–5; 0 = levels 1–2 (referent). Religious status can be dichotomized as *relstatus* with 1 = levels 4–5; 0 = levels 1–3 (referent) and years married can be likewise be dichotomized as *yrmarr* with 1 = levels 4–6; 0 = levels 1–3 (referent). Such a model may be preferable to the multi-leveled one. I leave it to the reader to determine.

Due to the fact that the response, *naffairs*, consists of some 75% zero counts, we shall later return to these data when considering zero-inflated and hurdle models. Of interest will be a determination if either of these extended models fit the data better than the one we have developed here.

Example 2: Heart procedures

The second example relates to data taken from Arizona cardiovascular patient files in 1991. A subset of the fields was selected to model the differential length of stay for patients entering the hospital to receive one of two standard cardiovascular procedures: CABG and PTCA. CABG is an acronym representing coronary artery bypass surgery; PTCA represents percutaneous transluminal

Table 6.9. *CABG/PTCA: upper frequencies and summary stats*

LOS	PTCA Freq.	Percent	CABG Freq.	Precent
1	147	7.68		
2	399	20.86		
3	292	15.26	1	0.06
4	233	12.18	1	0.06
5	176	9.20	10	0.60
6	149	7.79	48	2.86
7	124	6.48	105	6.26
8	102	5.33	195	11.63
9	66	3.45	200	11.93
10	68	3.55	183	10.92
11	38	1.99	157	9.37
12	23	1.20	129	7.70
13	22	1.15	93	5.55
14	14	0.73	114	6.80
15	14	0.73	81	4.83
16	7	0.37	52	3.10

	Mean	Median	SD
CABG	13.02	11	7.07
PTCA	5.16	4	4.16

coronary angioplasty. Angioplasty is performed by inserting a bulb through the artery to the place containing a blockage near the heart. The bulb is inflated or dilated, clearing the blockage in the affected part of artery. It is substantially safer than a CABG, and, as can be seen from Table 6.9, usually results in an earlier release from the hospital.

Length of stay values are found in the variable *los*; procedure data are found in *procedure*, with 1 = CABG and 0 = PTCA. CABG is considered to be the more difficult procedure of the two. Other controlling or confounding predictors include the following:

```
sex        1 = Male; 0 = Female
admit      1 = Urgent/Emergency; 0 = Elective
age75      1 = age>75; 0 = age <= 75
hospital   encrypted facility code
```

The data, *azprocedure*, consist of 3589 observations. The distribution of counts, together with a listing of the mean, median and standard deviation for each procedure is displayed in Table 6.9

A graphical representation of the differences in length of stay between the two procedures can be found in Figure 6.4.

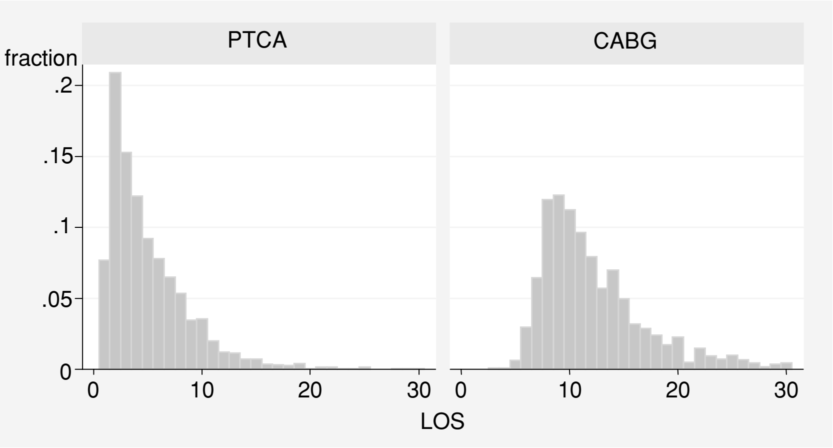

Figure 6.4. Length of stay (LOS) distributions: PTCA versus CABG

It is evident that having a CABG results in a longer hospital stay. The question is whether the difference in stay is statistically significant between the two procedures, controlling for gender, type of admission, and age of patient. Also desired is to determine the probable length of stay given patient profiles.

Modeling the data as Poisson, we have the following output:

```
. glm los procedure sex admit age75, nolog fam(poi) eform

Generalized linear models          No. of obs      =       3589
Optimization    :  ML              Residual df     =       3584
                                   Scale parameter =          1
Deviance        =  8874.147204     (1/df) Deviance =   2.476046
Pearson         =  11499.22422     (1/df) Pearson  =   3.208489
                                   AIC             =   6.238449
Log likelihood  =  −11189.89758    BIC             =  −20463.15

=============================================================
         |              OIM
     los |      IRR    Std. Err.      z    P>|z|  [95% Conf. Interval]
---------+---------------------------------------------------
procedure| 2.612576    .031825    78.84 0.000    2.550939   2.675702
      sex| .8834417  .0104349   −10.49 0.000    .8632245   .9041324
    admit| 1.386239  .0168061    26.94 0.000    1.353688   1.419573
    age75| 1.129999  .0140675     9.82 0.000    1.102761    1.15791
=============================================================
```

Given the large number of cases in the data, a Pearson Chi2 dispersion of 3.21 indicates overdispersion. Possible intrinsic causes may stem from the fact that the data have no zeros, as well as the disparity in the numbers of low counts. There may also be a clustering effect resulting from a higher correlation of procedures being done within providers than being done between providers. First, however, we shall model the data as negative binomial.

```
[Model first using full maximum likelihood negative
binomial to obtain an estimate for α, which is then
included into the GLM algorithm]

. glm los procedure sex admit age75, nolog fam(nb .1601022)
eform

Generalized linear models          No. of obs      =      3589
Optimization    : ML               Residual df     =      3584
                                   Scale parameter =         1
Deviance        = 3525.650017      (1/df) Deviance =  .9837193
Pearson         = 4947.825864      (1/df) Pearson  =  1.380532
Variance        : V(u) =           [Neg. Binomial]
function          u+(.1601022)u^2
Link function   : g(u) = ln(u)     [Log]
                                   AIC             =  5.560626
Log likelihood = -9973.543468      BIC             = -25811.64

-------------------------------------------------------------------------
             |              OIM
         los |      IRR   Std. Err.      z    P>|z|   [95% Conf. Interval]
-------------+-----------------------------------------------------------
   procedure | 2.667403  .0490528    53.35   0.000    2.572973   2.765298
         sex |  .881229  .0168211    -6.62   0.000    .8488693   .9148221
       admit | 1.448736  .0276089    19.45   0.000    1.395621   1.503871
       age75 | 1.127589  .0228369     5.93   0.000    1.083706   1.173249
-------------------------------------------------------------------------
```

The incidence rate ratios between the Poisson and negative binomial models are quite similar. This is not surprising given the proximity of α to zero. On the other hand, the AIC and BIC statistics for the negative binomial model are less – 12% and 26% respectively – than the Poisson. A likelihood ratio χ^2 value of 2432.7, with one degree of freedom, indicates that the model value of α, at 0.16, is nevertheless significantly different from an α of zero.

The fact that hospital length-of-stay data exclude the possibility of having zero counts suggests that the data be modeled using a type of zero-truncated model, e.g. a zero-truncated negative binomial – commonly referred to as a ZINB model. A negative binomial with endogenous stratification model is another possibility. Both of these models will be applied to the data in later chapters.

Example 3: Titanic survival data

These data come from the 1912 Titanic survival data. It consists of 1316 passengers, as well as crew. The crew members have been excluded from the analysis. Only four variables are recorded, with each in binary format. The goal of the study is to assess the risk of surviving. These data have previously been examined using binary logistic regression. Here we shall demonstrate that modeling

Table 6.10. *Titanic survivor dictionary*

n = 1,316		
survived	Survived	1/0
age	Child vs Adult	1/0
sex	Male vs Female	1/0
class1	class = = 1st class	1/0
class2	class = = 2nd class	1/0
class3	class = = 3rd class	1/0

it as logistic is inferior to modeling the data as a count. Converting a binary model to grouped has been shown earlier in this chapter.

The example data are defined in Table 6.10.

Modeled as a binary logistic model, and reporting the exponentiated parameter estimates as odds ratios, we have the following output:

```
. glm survived age sex class2 class3, nolog fam(bin) eform

Generalized linear models          No. of obs      =       1316
Optimization     :  ML             Residual df     =       1311
                                   Scale parameter =          1
Deviance         =  1276.200769    (1/df) Deviance =    .973456
Pearson          =  1356.674662    (1/df) Pearson  =    1.03484
                                   AIC             =   .9773562
Log likelihood   =  -638.1003845   BIC             =  -8139.863

------------------------------------------------------------------
        |Odds      OIM
survived|Ratio     Std. Err.    z     P>|z|  [95% Conf. Interval]
------------------------------------------------------------------
     age|.3479809 .0844397  -4.35   0.000 .2162749     .5598924
     sex|.0935308 .0135855  -16.31  0.000 .0703585     .1243347
  class2|.3640159 .0709594  -5.18   0.000 .2484228     .5333952
  class3|.1709522 .0291845  -10.35  0.000 .1223375     .2388853
------------------------------------------------------------------
```

Recall that goodness-of-fit is evaluated differently for binary binomial models than for count models. It makes no sense to talk about overdispersed binary response data, therefore the Pearson Chi2 dispersion statistic has little value in assessing model worth. AIC and BIC statistics are important, but only when comparing between models. Hosmer and Lemeshow (2003) and Hardin and Hilbe (2001) provide extensive information regarding fit considerations for logistic models.

As seen earlier, an individual or observation level format may be converted to grouped format by using code similar to the following:

Table 6.11. *Titanic data set*

	survive	cases	age	sex	class1	class2	class3
1	14	31	child	women	0	0	1
2	13	13	child	women	0	1	0
3	1	1	child	women	1	0	0
4	13	48	child	man	0	0	1
5	11	11	child	man	0	1	0
6	5	5	child	man	1	0	0
7	76	165	adults	women	0	0	1
8	80	93	adults	women	0	1	0
9	140	144	adults	women	1	0	0
10	75	462	adults	man	0	0	1
11	14	168	adults	man	0	1	0
12	57	175	adults	man	1	0	0

```
egen grp = group(age-class3)
egen cases = count(grp), by(grp)
egen survive = sum(survived), by(grp)
sort grp
by grp: keep if _n == 1
```

The above code groups the 1316 cases into 12 covariate patterns. Table 6.11 consists of the entire re-formatted data set.

The data are modeled as a grouped logistic model. Parameter estimates, standard errors, and so forth are identical to the binary model.

```
. glm survive age sex class2 class3, nolog fam(bin cases)
eform

Generalized linear models          No. of obs       =        12
Optimization     :  ML             Residual df      =         7
                                   Scale parameter  =         1
Deviance         =  110.8437538    (1/df) Deviance  =  15.83482
Pearson          =  100.8828206    (1/df) Pearson   =  14.41183
                                   AIC              =  13.14728
Log likelihood   =  −73.88365169   BIC              =  60.56729

---------------------------------------------------------------------
                         OIM
survive|Odds Ratio  Std. Err.     z    P>|z|  [95% Conf.  Interval]
-------+-------------------------------------------------------------
    age|.3479809    .0844397   −4.35  0.000  .2162749    .5598924
    sex|.0935308    .0135855  −16.31  0.000  .0703585    .1243347
 class2|.3640159    .0709594   −5.18  0.000  .2484228    .5333952
 class3|.1709522    .0291845  −10.35  0.000  .1223375    .2388853
---------------------------------------------------------------------
```

As a grouped logistic model, we find that the data are indeed overdispersed with a Pearson Chi2 based dispersion statistic of 14.41. The AIC and BIC are 13.15 and 60.57 respectively.

It is easy to re-parameterize the model as a count response model. Converting the GLM family from binomial to Poisson is all that is required to make the change. Since the canonical link is the default (with most commercial software), it is not necessary to manually change link functions. On the other hand, the total number of observations per group, as indicated in the variable *cases*, is changed from the binomial denominator to the Poisson logged offset.

```
. glm survive age sex class2 class3, nolog fam(poi) eform
lnoffset(cases)

Generalized linear models          No. of obs      =        12
Optimization    :  ML              Residual df     =         7
                                   Scale parameter =         1
Deviance        =  38.30402583     (1/df) Deviance =  5.472004
Pearson         =  39.06072697     (1/df) Pearson  =  5.580104
                                   AIC             =  8.921621
Log likelihood  =  −48.5297265     BIC             =  20.90968

-----------------------------------------------------------------
                    OIM
survive|    IRR     Std. Err.    z     P>|z|  [95% Conf. Interval]
-----------------------------------------------------------------
    age|.6169489   .0898438  −3.32   0.001 .4637587     .8207412
    sex|.3117178   .0296204  −12.27  0.000 .2587484     .3755308
 class2|.6850337   .0805458  −3.22   0.001 .5440367     .8625725
 class3|.463439    .0496088  −7.18   0.000 .3757299     .5716227
  cases|(exposure)
-----------------------------------------------------------------
```

The Poisson model produces an approximate three-fold reduction in dispersion over the grouped logistic model. Additionally, the AIC and BIC statistics are reduced by some 50% and 300% respectively. The deviance statistic has also been reduced from 110.84 to 38.30. Clearly, the Poisson is the preferable model. On the other hand, with a dispersion statistic of 5.58, the model indicates that the data are overdispersed. Modeled as a negative binomial, we have:

```
. glm survive age sex class2 class3, fam(nb .1040345)
eform lnoffset(cases)

Generalized linear models          No. of obs      =        12
Optimization    :  ML              Residual df     =         7
                                   Scale parameter =         1
Deviance        =  12.47948608     (1/df) Deviance =  1.782784
Pearson         =  11.07146766     (1/df) Pearson  =  1.581638
Variance        :  V(u) =          [Neg. Binomial]
function           u+(.1040345)u^2
Link function   :  g(u) = ln(u)    [Log]
                                   AIC             =  8.119471
Log likelihood  =  −43.71682842    BIC             =  −4.91486
```

```
------------------------------------------------------------------------------
         |           OIM
 survive |   IRR     Std. Err.    z    P>|z|  [95% Conf.  Interval]
---------+--------------------------------------------------------------------
     age |.5116907   .1289491  −2.66  0.008  .312248     .8385238
     sex |.3752549   .0887939  −4.14  0.000  .2360003    .5966782
  class2 |.6875551   .2097046  −1.23  0.219  .3781732    1.250041
  class3 |.4037074   .1157954  −3.16  0.002  .2301       .7082993
   cases |(exposure)
------------------------------------------------------------------------------
```

The AIC and BIC statistics are 8.12 (from 8.92) and −4.91 (20.91), with BIC alone indicating better fit. A likelihood ratio test, performed following a previously run (not shown) maximum likelihood negative binomial model – which provided the value of α that was input into the GLM algorithm – indicates that the model is significantly different from Poisson [$\chi^2 = 9.63$; dof = 1; p > $\chi^2 =$ 0.001].

To clean up the model it is necessary to combine class2 with class1 as the referent for *class*. α is slightly increased, with the deviance and AIC lightly decreased. More substantial change is found in the reduced model BIC statistic (−4.9 to −8.2). A final model is presented below.

```
. glm survive age sex class3, nolog fam(nb .1339329)
eform lnoffset(cases)

Generalized linear models          No. of obs      =         12
Optimization    : ML               Residual df     =          8
                                   Scale parameter =          1
Deviance        = 11.71286801      (1/df) Deviance =   1.464109
Pearson         = 8.68621808       (1/df) Pearson  =   1.085777
Variance        : V(u) =           [Neg. Binomial]
function          u+(.1339329)u^2
Link function   : g(u) = ln(u)     [Log]
                                   AIC             =   8.061748
Log likelihood = −44.37048795      BIC             = −8.166385

------------------------------------------------------------------------------
         |           OIM
 survive |    IRR     Std. Err.    z    P>|z|  [95% Conf. Interval]
---------+--------------------------------------------------------------------
     age |  .5410341 .1465874  −2.27  0.023    .318127  .9201288
     sex |  .3996861 .1034007  −3.54  0.000   .2407183  .6636345
  class3 |  .4819995 .1261313  −2.79  0.005   .2886033  .8049927
   cases |(exposure)
------------------------------------------------------------------------------
```

Example 4: Health reform data

A fourth example consists of data from a subset of the 2001 German Socio-Economic Panel (SOEP). The subset was created by Rabe-Hesketh and

Table 6.12. *German health reform data*

Numvisit	Visits to MD office 3 mo prior – response
reform	1 = interview yr after reform: 1998; 0 = pre-reform:1996
badh	1 = bad health; 0 = not bad health
age	Age(yrs 20–60)
educ	Education(yrs 7–18)
loginc	ln(household income DM)
id	Person ID

Skrondal (2005). Only working women are included in these data. Beginning in 1997, German health reform in part entailed a 200% increase in patient co-payment as well as limits in provider reimbursement. Patients were surveyed for the one-year panel (1996) prior to and the one year panel (1998) after reform to assess whether the number of physician visits by patients declined – which was the goal of reform legislation.

The response, or variable to be explained by the model, is *numvisit*, which indicates the number of patient visits to the physicians office during a three-month period. The data set, *mdvisits*, consists of 2227 cases. A tabulation of *numvisit* provides an overview of the observed or empirical count distribution.

The data are first modeled using a Poisson regression. Results appear as:

```
. glm numvisit reform badh age educ loginc, nolog fam(poi)

Generalized linear models          No. of obs      =       2227
Optimization     : ML              Residual df     =       2221
                                   Scale parameter =          1
Deviance         = 7422.124433     (1/df) Deviance =   3.341794
Pearson          = 9681.69202   => (1/df) Pearson  =   4.359159
                                => AIC             =   5.343357
Log likelihood   = -5943.828046    BIC             =  -9698.256

----------------------------------------------------------------------
          |             OIM
 numvisit |     Coef.  Std. Err.     z  P>|z|  [95% Conf.  Interval]
----------+-----------------------------------------------------------
   reform | -.1398842  .0265491  -5.27 0.000  -.1919195  -.0878489
     badh |  1.132628  .0302986  37.38 0.000   1.073244   1.192013
      age |  .0048853  .0012526   3.90 0.000   .0024302   .0073404
     educ | -.0118142  .0059588  -1.98 0.047  -.0234933  -.0001351
   loginc |  .1520247  .0359837   4.22 0.000    .081498   .2225515
    _cons |  -.421508   .268966  -1.57 0.117  -.9486718   .1056558
----------------------------------------------------------------------
```

The model appears prima facie to fit well. Predictors appear significant. However, the dispersion statistic is over 4, indicating substantial overdispersion

Table 6.13. *Tabluation of response: numvisits*

Visits	Freq.	Percent	Cum.	
MD of				
0	665	29.86	29.86	< = large number of 0's
1	447	20.07	49.93	
2	374	16.79	66.73	
3	256	11.50	78.22	
4	117	5.25	83.48	
5	101	4.54	88.01	
6	76	3.41	91.42	
7	21	0.94	92.37	
8	27	1.21	93.58	
9	9	0.40	93.98	
10	61	2.74	96.72	
11	1	0.04	96.77	
12	20	0.90	97.67	
13	5	0.22	97.89	
14	3	0.13	98.02	
15	19	0.85	98.88	
16	2	0.09	98.97	
20	10	0.45	99.42	
24	1	0.04	99.46	
25	3	0.13	99.60	
30	4	0.18	99.78	
36	1	0.04	99.82	
40	2	0.09	99.91	
50	1	0.04	99.96	
60	1	0.04	100.00	
Total	2,227	100.00		

given the large number of observations. The AIC statistic equals 5.343, a value with which we shall later compare alternative models.

Exponentiated coefficients give us incidence rate ratios, and appropriately adjusted standard errors and confidence intervals.

```
                      OIM
numvisit|    IRR    Std. Err.     z    P>|z|   [95% Conf. Interval]
--------+-----------------------------------------------------------
  reform|.8694589 .0230834   −5.27 0.000    .8253733   .9158993
    badh|3.103804 .0940408   37.38 0.000    2.924853   3.293703
     age|1.004897 .0012588    3.90 0.000    1.002433   1.007367
    educ|.9882553 .0058888   −1.98 0.047    .9767805   .9998649
  loginc|1.164189 .0418918    4.22 0.000    1.084911    1.24926
```

Employing a robust sandwich variance estimator gives us the following table of results.

Table 6.14. *Observed proportion and predicted probability for model visits from 0–10*

	Visits	%Visits Obs	%Visits Pred
1	0	0.298608	0.115397
2	1	0.200718	0.233520
3	2	0.167939	0.240503
4	3	0.114953	0.170382
5	4	0.052537	0.096959
6	5	0.045352	0.051449
7	6	0.034127	0.029828
8	7	0.009430	0.020123
9	8	0.012124	0.014646
10	9	0.004041	0.010482
11	10	0.027391	0.007053

```
         |             Robust
numvisit |     IRR    Std. Err.     z    P>|z|   [95% Conf. Interval]
---------+-----------------------------------------------------------
  reform | .8694589  .0515157   −2.36  0.018    .7741322   .9765241
    badh | 3.103804  .2590781   13.57  0.000    2.635382   3.655486
     age | 1.004897  .0029544    1.66  0.097    .9991234   1.010705
    educ | .9882553  .0112303   −1.04  0.299    .9664875   1.010513
  loginc | 1.164189  .0895704    1.98  0.048     1.00123   1.353671
```

There is an approximate 13% reduction in visits to a physician between 1996 and 1998, adjusted for health status, age, education, and the natural log of household income. Age and education are not contributory to model. Recall that the model appears to indicate overdispersion.

The fact that the response consists of 30% zero counts is likely to be a cause of overdispersion. Table 6.14 displays the observed and predicted counts from 0 to 10. Predicted values are derived from the last Poisson model.

This table shows the observed proportions or empirical probabilities for each count up to 10 together with the predicted probability for each count based on the model (given the probability function $f(\mu; y) = e^{-\mu}\mu^{y}/y!$).

Note the near three-fold higher value in the observed frequency for zero counts than for the predicted value. This is reflected in a graph of the table data, with additional counts through 20.

The observed proportions approximate a geometric distribution, which is the discrete correlate of a negative exponential distribution. It is likely that if the model is estimated using a negative binomial regression, the value of α will approximate 1.0. A negative binomial model with $\alpha = 1$ is a geometric model.

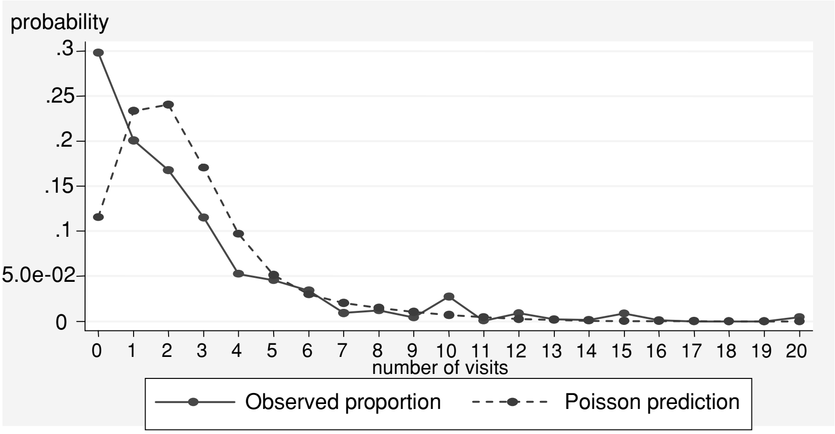

Figure 6.5. Poisson model for number of visits: observed versus predicted probabilities

It is important, though, to first deal with *educ* and *age*. *Educ* does not appear to be contributory to the model and *age* is questionable, particularly when subjected to scaling and adjusted by a robust sandwich variance estimator (not shown here). However, recall that both variables are discrete, with many levels. *Educ* has 16 levels; age has 41, one level for each year from 20 through 60. Both predictors may be considered as continuous, however, each age is measured as a unit, with no decimals. Categorizing each of these predictors may help both understand the differential contribution of education levels and age groups, and also may help in dealing with model overdispersion.

Edu may be left as found, or separated into three separate binary predictors. Commerical statistical software generally prefers one type or another. In any event, one can have the software generate three dummies resulting in predictor levels *educ1*, *educ2*, and *educ3*. Levels are defined by the lowest number of years of education.

```
  edu |        Freq.         Percent           Cum.
------+--------------------------------------------
   7- |        549           24.65            24.65
10.5- |        926           41.58            66.23
  12- |        752           33.77           100.00
------+--------------------------------------------
Total |      2,227          100.00
```

Recalling that study ages range from 20 to 60, *age* may be expanding into four levels: 20–29, 30–39, 40–49, and 50–60.

Modeling with *age1* (20–29) as the referent, it is found that *age2* (30–39) is not statistically significant, implying that there is little difference between the two age divisions. In such a situation the two levels may be combined for an expanded reference group, 20–39. Levels can be labeled and tabulated.

```
    age |        Freq.            Percent           Cum.
--------+-----------------------------------------------
20−39   |        1,352              60.71          60.71
40−49   |          515              23.13          83.83
50−60   |          360              16.17         100.00
--------+-----------------------------------------------
 Total  |        2,227             100.00
```

Re-running the model with levels *educ1–3* and *age1–3*, the model appears as:?

```
.glm numvisit reform badh educ2 educ3 age2 age3 loginc,
nolog fam(poi) eform

Generalized linear models        No. of obs      =      2227
Optimization    :  ML            Residual df     =      2219
                                 Scale parameter =         1
Deviance        =  7398.293267   (1/df) Deviance =  3.334066
Pearson         =  9518.948272   (1/df) Pearson  =  4.289747
                                 AIC             =  5.334452
Log likelihood  =  −5931.912464  BIC             =  −9706.67

            |          OIM
numvisit    |    IRR   Std. Err.     z    P>|z|  [95% Conf. Interval]
------------+--------------------------------------------------------
    reform  |.8706594  .0231258   −5.21  0.000   .8264932   .9171858
      badh  |3.132308  .0938026   38.13  0.000    2.95375   3.321661
     educ2  |1.085224  .0365147    2.43  0.015   1.015965   1.159204
     educ3  |.9221584  .0340499   −2.19  0.028   .8577794   .9913691
      age2  |1.095001  .0352259    2.82  0.005   1.028092   1.166266
      age3  |1.144818  .0410837    3.77  0.000   1.067062    1.22824
    loginc  |1.146791  .0410935    3.82  0.000   1.069012   1.230228
```

Due to overdispersion in the model, and the fact that there may be a clustering effect resulting from multiple visits by the same individual, it is wise to apply a robust cluster variance estimator to the model.

```
. glm numvisit reform badh educ2 educ3 age2 age3 loginc,
fam(poi) eform cluster(id)

Generalized linear models        No. of obs      =     2227
Optimization     : ML            Residual df     =     2219
                                 Scale parameter =        1
Deviance         = 7398.293267   (1/df) Deviance = 3.334066
Pearson          = 9518.948272   (1/df) Pearson  = 4.289747
                                 AIC             = 5.334452
Log              = −5931.912464  BIC             = −9706.67
pseudolikelihood
              (Std. Err. adjusted for 1518 clusters in id)
```

```
---------------------------------------------------------------------
          |            Robust
 numvisit |     IRR    Std. Err.     z   P>|z|   [95% Conf. Interval]
----------+----------------------------------------------------------
   reform | .8706594  .0488115   -2.47  0.013    .7800593   .9717823
     badh | 3.132308  .2628195   13.61  0.000    2.657318   3.692202
    educ2 | 1.085224  .0977545    0.91  0.364    .9095888   1.294773
    educ3 | .9221584  .0774285   -0.97  0.334    .7822307   1.087117
     age2 | 1.095001  .0877094    1.13  0.257     .935909   1.281138
     age3 | 1.144818  .1069266    1.45  0.148    .9533091   1.374798
   loginc | 1.146791  .0946715    1.66  0.097    .9754714   1.348198
---------------------------------------------------------------------
```

Application of a negative binomial model is a reasonable approach to dealing with the excess dispersion in the Poisson model (4.29), particularly when a specific source of the overdispersion has not been identified.

```
. glm numvisit reform badh educ2 educ3 age2 age3 loginc,
fam(nb .9982126) eform

Generalized linear models         No. of obs       =       2227
Optimization     : ML             Residual df      =       2219
                                  Scale parameter  =          1
Deviance         = 2412.452952    (1/df) Deviance  =    1.08718
Pearson          = 2644.673532    (1/df) Pearson   =   1.191831
Variance         : V(u) =         [Neg. Binomial]
function           u+(.9982126)u^2
Link             : g(u) = ln(u)   [Log]
function
                                  AIC              =   4.103197
Log likelihood   = -4560.909631   BIC              =  -14692.51

---------------------------------------------------------------------
          |              OIM
 numvisit |     IRR    Std. Err.     z   P>|z|   [95% Conf. Interval]
----------+----------------------------------------------------------
   reform |  .871369  .0445206   -2.69  0.007    .7883371   .9631464
     badh | 3.134872  .2332147   15.36  0.000    2.709542   3.626969
    educ2 | 1.085687  .0716943    1.24  0.213     .953882   1.235704
    educ3 |  .970105  .0685734   -0.43  0.668    .8445984   1.114262
     age2 | 1.049625  .0666178    0.76  0.445    .9268511   1.188662
     age3 | 1.206782  .0867935    2.61  0.009    1.048115   1.389468
   loginc | 1.134513  .0798461    1.79  0.073    .9883313   1.302316
---------------------------------------------------------------------
```

Negative binomial regression has changed the model statistics. Note at first that the Pearson χ^2 dispersion has been reduced from 4.29 to 1.19. AIC and BIC statistics deflate from 5.33 to 4.10 and from −9706.67 to −14692.51 respectively. The deviance itself has been substantially reduced from 7398 to 2412 – a three-fold reduction. An interesting side is that the value of α is very close to unity. Recall the discussion regarding Figure 6.5, which showed that the shape of the predicted counts from the Poisson model took the form of a geometric distribution. This speculation has been borne out with the value of $\alpha = 0.9982$ in the negative binomial model of the same data.

Adjusting the model by the clustering effect of *id*, and applying robust standard errors to the result, produces the following table of estimates. Notice that the cluster-adjusted standard errors differ very little from base-model standard errors. Moreover, a simple application of robust standard errors to the Hessian matrix without a clustering effect yields little difference to the model with clustering. This mutual lack of effect can be interpreted that there is little if any overdispersion in the resultant model data due to clustering and that any remaining variability in the data comes from a yet to be specified source. Fortunately, the unadjusted negative binomial model accounts for most of the Poisson overdispersion. This conclusion runs concordant to the observed value of the dispersion statistic.

```
         |          (Std. Err. adjusted for 1518 clusters in id)
---------+----------------------------------------------------------
         |              Robust
numvisit |      IRR   Std. Err.      z   P>|z|   [95% Conf.   Interval]
---------+----------------------------------------------------------
  reform |  .871369  .0446676   −2.69  0.007    .7880764     .963465
    badh | 3.134872  .2587493   13.84  0.000    2.666628    3.685337
   educ2 | 1.085687  .0880525    1.01  0.311    .9261247    1.272739
   educ3 |  .970105  .0786037   −0.37  0.708     .827655    1.137072
    age2 | 1.049625  .0772138    0.66  0.510    .9086928    1.212415
    age3 | 1.206782  .1161242    1.95  0.051    .9993571    1.457259
  loginc | 1.134513  .0952187    1.50  0.133    .9624292    1.337365
```

It appears that the negative binomial model – or geometric model – fits the data better than the Poisson model. However, it also appears that the cuts we made in *educ* and *age* have resulted in levels that do not significantly contribute to the model. For those who wish to continue model development, re-classification should be attempted. In addition, since the observed zeros in the data exceed the distributional assumption of the negative binomial, alternative models designed to deal with this specific type of problem should be investigated. Zero-inflated and hurdle models are foremost models that come to mind. Both of these models will be discussed in later chapters and employed on these data.

When modeling a data situation in which two groups are being distinguished to determine differences of some sort between them, it is sometimes instructional to graph the residual data as a whole while allowing the intrinsic groupings to emerge. This may be done by graphing standardized deviance residuals by the predicted mean or μ. A correction value is used with the values of μ, calculated as 2*sqrt(μ).

Two groupings of residuals reflect the reform periods of 1996 and 1998. Standardized deviance residuals greater than 4 can be regarded as possible outliers. Pierce and Schafer (1986) recommend an adjusted standardized deviance using the following formula: $D_{\text{adj}} - d + 1/\{6^{*}\text{sqrt}(\text{mean}(y))\}$. However, I have not found them to be superior to traditionally defined standardized deviance residuals. On the other hand, I have argued for the use of Anscombe residuals (Hilbe, 1994a, 1994b; Hardin and Hilbe, 2001) which are theoretically more appropriate than standardized deviance residuals, although the two values are generally quite similar. When modeling, both should be attempted. If negative binomial Anscombe and standardized deviance residuals differ, then additional tests are required.

Noticeably, the negative binomial fit is preferred over the Poisson. There are many fewer observations with standardized deviance residuals less than −2 or greater than 2. Although the relative shapes of the residuals by fit are similar, their respective variability is substantially different. Those involved in health outcomes research would likely find it interesting to search for the determinants of outlier status.

It is interesting to note that when the predicted number of visits for pre- and post-reform periods are calculated at the average of each predictor versus the average prediction for pre- and post-reform visits, that the distribution lines are fairly similar. The negative binomial model used here is based on *educ* and *age* being separated into their respective levels. Recall that Figure 6.5 is based on *educ* and *age* being considered as continuous variables. Table 6.15 displays mean values of visit for pre- and post-reform periods through ten visits.

Figure 6.7 visually displays the values from 0 through 10 given in Table 6.15. In addition, Figure 6.7 provides average prediction values for pre- and post-reform periods. These distributions are consistent with the previous distributions. (Figure 6.7 and values in Table 6.15 were calculated from a suite of programs created by Prof. Scott Long of the University of Indiana for use with binary and count response models. Aimed at assisting researchers to assess model fit, the programs, termed SPOST, may be downloaded from http://www.indiana.edu/~jslsoc/spost.htm.)

6.4 Summary

The examples used in this chapter demonstrate methods of handling overdispersed Poisson data without going beyond the basic negative binomial model. When identifying a count response model as overdispersed, it is first necessary

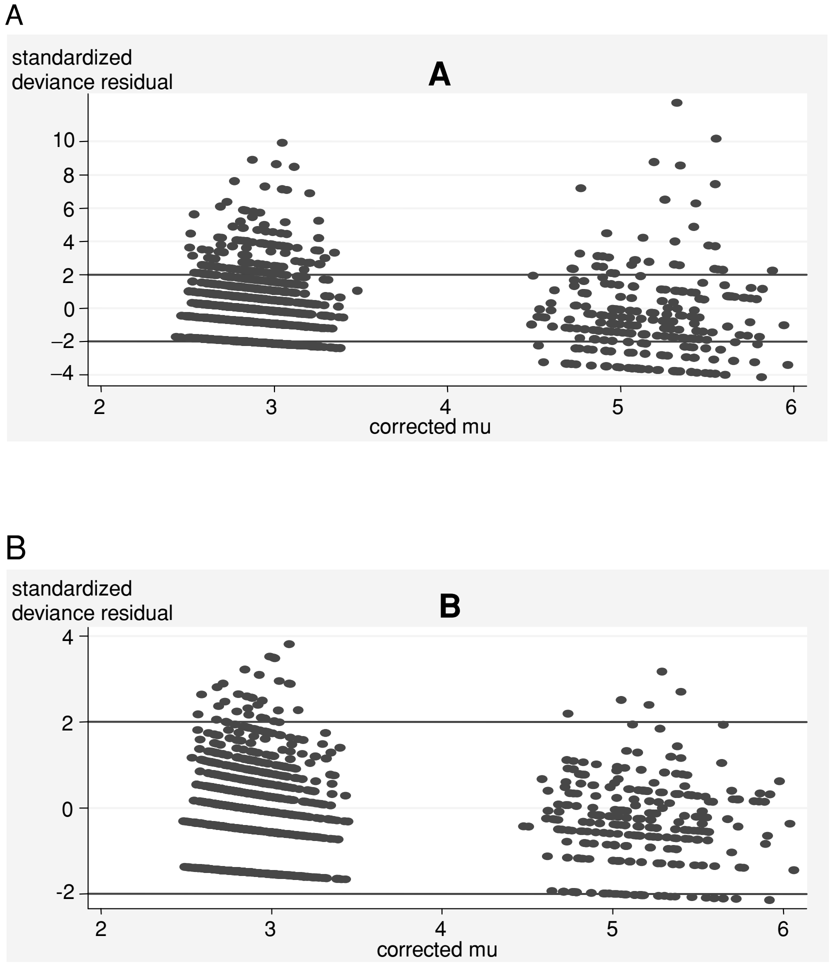

Figure 6.6. German health reform 1996 and 1998: standardized deviance residuals versus corrected predicted mean number of visits. (A) Poisson model; (B) negative binomial (NB-2) model

to determine if the overdispersion is real or only apparent. If it is real, one next determines if it is significant, i.e. whether the variability significantly exceeds Poisson distributional assumptions.

Modeling data using a full maximum likelihood negative binomial allows the statistician to determine if the resultant value of α is statistically different from zero. If not, then the data are to be modeled as Poisson. If there is a statistically significant difference, then a model needs to be found that addresses the genesis of the extra correlation in the data. A basic negative binomial model may be

Table 6.15. *Pre (1996) and post (1998) mean predicted visits with CI's confidence intervals calculated by delta method*

PRE REFORM (1996)			
		95% Conf. Interval	
Rate:	2.5115	[2.3345,	2.6885]
Pr(y = 0|x):	0.2845	[0.2701,	0.2989]
Pr(y = 1|x):	0.2037	[0.1976,	0.2099]
Pr(y = 2|x):	0.1458	[0.1443,	0.1473]
Pr(y = 3|x):	0.1043	[0.1032,	0.1053]
Pr(y = 4|x):	0.0746	[0.0723,	0.0768]
Pr(y = 5|x):	0.0533	[0.0507,	0.0560]
Pr(y = 6|x):	0.0381	[0.0355,	0.0408]
Pr(y = 7|x):	0.0273	[0.0248,	0.0297]
Pr(y = 8|x):	0.0195	[0.0173,	0.0216]
Pr(y = 9|x):	0.0139	[0.0121,	0.0158]
Pr(y = 10|x):	0.0100	[0.0085,	0.0115]
POST REFORM (1998)			
		95% Conf. Interval	
Rate:	2.1884	[2.0329,	2.344]
Pr(y = 0|x):	0.3134	[0.2981,	0.3287]
Pr(y = 1|x):	0.2153	[0.2096,	0.2211]
Pr(y = 2|x):	0.1479	[0.1472,	0.1485]
Pr(y = 3|x):	0.1015	[0.0997,	0.1033]
Pr(y = 4|x):	0.0697	[0.0668,	0.0725]
Pr(y = 5|x):	0.0478	[0.0448,	0.0508]
Pr(y = 6|x):	0.0328	[0.0300,	0.0356]
Pr(y = 7|x):	0.0225	[0.0201,	0.0249]
Pr(y = 8|x):	0.0154	[0.0134,	0.0174]
Pr(y = 9|x):	0.0106	[0.0090,	0.0122]
Pr(y = 10|x):	0.0073	[0.0060,	0.0085]

MEAN VALUES OF REMAINING PREDICTORS					
badh	educ2	educ3	age2	age3	loginc
.11360575	.41580602	.337674	.23125281	.16165245	7.7128263

sufficient to deal with the overdispersion, but if the data violate the distributional assumptions of both Poisson and negative binomial models, then adjustments need to be made to the negative binomial algorithm that directly deal with the source of overdispersion. Some data do not admit for an easy solution.

In the following chapters, the basic or traditional negative binomial algorithm is enhanced to address count data that cannot be modeled using the traditional

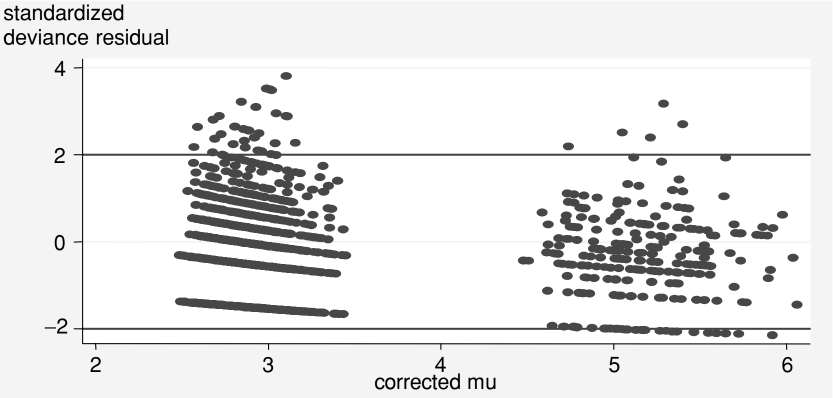

Figure 6.7. Hospital visitations: pre- and post-reform

methods. First to be evaluated are data situations that require an adjustment to the Poisson and negative binomial variance function. Afterwards, data that violate Poisson and negative binomial distributional assumptions are discussed.

Exercises

1 Given the following data, model *kyp* on *start* using a binary response logistic regression model. Determine if the model is overdispersed by converting the data to grouped format. Then determine if the grouped data is equi-dispersed when modeled using Poisson regression.

kyp	start	kyp	start
1	8	0	16
0	9	0	14
0	13	0	12
1	1	0	16
1	8	0	10
0	1	0	15
0	16	0	15
0	16	1	13
0	10	0	13
0	17	0	13
0	13	1	6
0	11	0	13

2 Given the following cross-tabulation, calculate the odds ratio and the relative risk, then model both with *low* as the response, or outcome, and *smoke* the predictor, or exposure. Discuss the relationship between the two models.

birth weight<250	smoked during pregnancy		
0g	0	1	Total
0	86	44	130
1	29	30	59
Total	115	74	189

3 Use the *gss2002_educ* data set to model the highest number of respondent years of education on the basis of a number of explanatory predictors. Construct the best-fitted model using the criteria discussed in the text.

4 Using the Titanic data set, grouped as in the final model output for Example 3, create both negative binomial Anscombe and standard deviance residuals. Graph the two residuals against the fitted negative binomial values. Identify any outliers and discuss the fit of the model on the basis of the two sets of residuals.

5 Model the data in Chapter 3, question 5, using negative binomial regression. Is the overdispersion eliminated? Discuss.

6 Model the data used for Chapter 4, question 5, using a negative binomial model. Does the negative binomial model successfully adjust for the overdispersion found in the Poisson model. Discuss.

7 Model the following data to determine if smoking adds to the risk of developing coronary heart disease. (Data *doll* from Breslow, 1985, pp. 109–143).

	Person-years		Coronary deaths	
Age	Non-smokers	Smokers	Non-smokers	Smokers
35–44	18790	52407	2	32
45–54	10673	43248	12	104
55–64	5710	28612	28	206
65–74	2585	12663	28	186
75–84	1462	5317	31	102

7

Alternative variance parameterizations

Negative binomial regression has traditionally been used to model otherwise overdispersed count or Poisson data. It is now considered to be the general catch-all method used when Poisson data are found to be overdispersed, particularly when the source of overdispersion has not been identified. When we can identify that which gives rise to extra correlation, and hence overdispersion, the basic Poisson and negative binomial algorithms may themselves be further adjusted or enhanced to directly address the identified source of the extra correlation. For example, when overdispersion results from an excess of zero counts in the response, an appropriate strategy is to model the data using either a zero-inflated Poisson (ZIP) or zero-inflated negative binomial (ZINB). Employing a hurdle model may also result in a better fit. On the other hand, if the response is structured such that zero counts are not possible, such as in hospital length of stay data, a zero-truncated Poisson (ZTP) or zero-truncated negative binomial (ZTNB) model may be appropriate.

A variety of alternative models have been developed to address specific facts in the data that give rise to overdispersion. Models dealing with an excess or absence of zeros typically define a mixture that alters the distributional variance of the Poisson distribution. Other models are constructed to alter not the probability and log-likelihood distributions, but rather the Poisson and negative binomial variance functions. We discuss these types of models in this chapter.

Models that address overdispersion by making changes to the Poisson and negative binomial variance function are listed in Table 7.1. Note that the Poisson is regarded as the base count distribution as well as the base variance function. The negative binomial generically deals with overdispersion, and is itself modified given certain data situations.

Table 7.1. *Count model variance functions*

Poisson	:	V = μ
QL Poisson	:	V = $\mu\phi$
Geometric	:	V = μ(1 + μ)
NB-1	:	V = μ(1 + α)
NB-2	:	V = μ(1 + $\alpha\mu$)
NB-H	:	V = μ(1 + ($\alpha\nu$)μ)
NB-P	:	V = $\mu + \alpha\mu^\rho$

NB-2 with $\alpha = 0$ is Poisson
NB-2 with $\alpha = 1$ is Geometric

7.1 Geometric regression

The geometric distribution is a special case of the negative binomial. Using the parameterization developed in Chapter 5, the geometric is the negative binomial with the heterogeneity or overdispersion parameter, α, set to 1.0. GLM software that incorporates the negative binomial as a member family can be used to design geometric models by setting the value of α to a constant value of 1.0; the GLM algorithm should also be set with a log link together with the negative binomial family. Maximum likelihood negative binomial algorithms, on the other hand, generally do not allow estimation of geometric models. Since α is estimated as an additional parameter, it cannot normally be constrained to a user defined value unless the software allows constrained optimization. However, a geometric regression algorithm is simple to design with the appropriate programming language, e.g. SAS's IML, Stata's ML capabilities, or by programming in R.

7.1.1 Derivation of the geometric

The derivation of the geometric follows the same logic as that of the negative binomial, with the exception that the term α is omitted. This greatly simplifies the log-likelihood function, which appears as

$$\mathcal{L}(\mu; y) = \sum_{i=1}^{n} \exp\{y_i \ln(\mu_i/(1+\mu_i)) - \ln(1+\mu_i)\} \tag{7.1}$$

or

$$\mathcal{L}(\mu; y) = \sum_{i=1}^{n} \exp\{y_i \ln(\mu_i) - (y_i+1)^* \ln(1+\mu_i)\} \tag{7.2}$$

Parameterized in terms of the log link function $\ln(\mu) = \beta' x$ or $\ln(\mu) = xb$

$$\mathcal{L}(x_i b; y_i) = \sum_{i=1}^{n} \exp\{y_i \ln(\exp(x_i b)/(1 + \exp(x_i b))) - \ln(1 + \exp(x_i b))\} \tag{7.3}$$

or

$$\mathcal{L}(x_i b; y_i) = \sum_{i=1}^{n} \exp\{y_i^* x_i b - (1 + y_i) \ln(1 + \exp(x_i b))\} \tag{7.4}$$

As a special case of the negative binomial, the geometric mean, variance, and related functions take the same form. Therefore, we have

$$\text{MEAN} = \mu_i \quad \text{or} \quad \exp(x_i b) \quad \text{or} \quad \exp(x_i \beta) \tag{7.5}$$

$$\text{VARIANCE} = \mu_i + \mu_i^2 \quad \text{or} \quad \mu_i(1 + \mu_i) \tag{7.6}$$

$$= \exp(x_i \beta)(1 + \exp(x_i \beta)) \tag{7.7}$$

$$\text{DERIVATIVE OF LINK} = 1/\{\mu_i(1 + \mu_i)\}$$

$$\text{or } 1/\{\exp(x_i \beta)(1 + \exp(x_i \beta))\} \tag{7.8}$$

$$\text{DEVIANCE} = 2 \sum_{i=1}^{n} \{y_i \ln(y_i/\mu_i)$$

$$- (1 + y_i) \ln((1 + y_i)/(1 + \mu_i))\} \tag{7.9}$$

7.1.2 Using the geometric model

Refer to the various graphs presented in Chapter 5 representing the shape of negative binomial data given specified values of the mean and of α. The geometric distribution, with an $\alpha = 1$, produces a shape that is the discrete correlate of the negative exponential distribution. Many types of data fit one of these shapes. If the counts of some item or event deplete in a smooth decreasing manner, then a geometric model will likely fit the data. If the data are modeled using negative binomial regression, the value of α will approximate 1.0. We found this to be the case for Example 4 in the last chapter.

It may be instructional to observe the relationship between the Poisson and geometric models. After all, regardless of the shape of the geometric distribution, the geometric model represents an accommodation of Poisson overdispersion.

We shall first create a simulated geometric data set having the linear predictor being composed of a constant equal to -1, and two parameters with values of

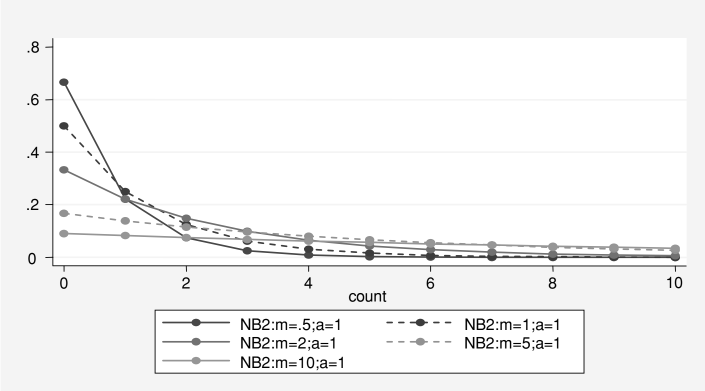

Figure 7.1. Negative binomial distributions: alpha = 1.

2.0 and −0.5 respectively. Hence the linear predictor is synthesized to be

$$\begin{aligned} \beta' x_i &= b_0 + b_1 x_1 + b_2 x_2 \\ & -1 + 2x_1 - 0.5x_2 \end{aligned}$$

Modeled using a maximum likelihood negative binomial algorithm produces the following output:

```
. nbreg y2 x1 x2, nolog

Negative binomial regression         Number of obs    =      50000
                                     LR chi2(2)       =   43212.16
Dispersion       =   mean            Prob > chi2      =     0.0000
Log likelihood   =   -57930.636      Pseudo R2        =     0.2716

----------------------------------------------------------------
        y2 |    Coef.   Std. Err.     z     P>|z|  [95% Conf. Interval]
-----------+----------------------------------------------------
        x1 |  1.992554 .0095348   208.98 0.000    1.973866  2.011242
        x2 | -.5023906 .0075823   -66.26 0.000   -.5172516 -.4875296
     _cons | -.9916718 .0107862   -91.94 0.000   -1.012812 -.9705312
-----------+----------------------------------------------------
  /lnalpha | -.0133656 .0147555                  -.0422859   .0155547
-----------+----------------------------------------------------
     Alpha |  .9867233 .0145596                   .9585957  1.015676
----------------------------------------------------------------
Likelihood-ratio test of alpha = 0: chibar2(01) = 9.9e+04 Prob>
= chibar2 = 0.000
AIC Statistic = 2.317
```

The parameter estimates are very close to what was specified in the simulation set-up. In addition, the fitted value of α is approximately 1.0, as would be restricted for a geometric model. Recall that unless a seed value is given to the random number simulator, the values of the parameter estimates will differ slightly from run to run. Also shown is a value for the separately calculated AIC statistic.

In the context of negative binomial modeling, an α of 0, or close to 0, indicates that the model is Poisson. The model will be considered as Poisson even if α is not exactly zero. There is no absolute criterion for determining when a model is to be classified as Poisson, since there are gradations of model fit. Given tests of negative binomial vs Poisson, such as the likelihood ratio test shown under the table of estimates above, a model can be classified as Poisson until it exceeds some specified p-value – usually 0.05 – for a given test. Models in which the above defined log-likelihood ratio test produce p-values equal to or greater than 0.05 are Poisson. Those under 0.05 are regarded as negative binomial, or rather, as a non-Poisson count model. The logically prior assumption is that the test is on a count response model belonging to the exponential family of distributions.

We next model the geometric data using Poisson regression. The aim is to see the extent of overdispersion in the Poisson model, as indicated by the dispersion statistic. Notice that the parameter estimates are close to those defined, but that the Pearson χ^2 dispersion is at 4.19. This is a particularly large value considering the number of observations in the data. If the data were estimated using a maximum likelihood Poisson algorithm that fails to provide information concerning fit, a user may well believe that model is in fact a good one. It is not.

```
. glm y2 x1 x2, nolog fam(poi)

Generalized linear                  No. of obs        =      50000
models
Optimization     :  ML              Residual df       =      49997
                                    Scale parameter   =          1
Deviance         =  161518.4435     (1/df) Deviance   =   3.230563
Pearson          =  209535.3003     (1/df) Pearson    =   4.190957
                                    AIC               =   4.288977
Log likelihood   =  -107221.4299    BIC               =    -379438

-----------------------------------------------------------------------------
      |             OIM
   y2 |    Coef.    Std. Err.      z      P>|z|   [95% Conf.  Interval]
------+----------------------------------------------------------------------
   x1 |  2.041422  .0024603    829.74  0.000    2.036599    2.046244
   x2 | -.5193739  .0024727   -210.04  0.000   -.5242204   -.5145274
_cons | -1.070714  .0057808   -185.22  0.000   -1.082045   -1.059384
-----------------------------------------------------------------------------
```

AIC and BIC statistics of the geometric model are calculated as 2.317 and −505568.5 respectively, indicating that the Poisson model is a comparatively poor fit.

7.1.3 The canonical geometric model

The geometric probability function can be expressed for $y \geq 0$, as (I shall forgo subscripts for the rest of this section for ease interpretation)

$$f(y; p) = p(1 - p)^y \tag{7.10}$$

In exponential family form we have

$$f(y;p) = \exp\{y * \ln(1-p) + \ln(p)\} \tag{7.11}$$

Following the same logic as in Chapter 5

$$\text{LINK} = \ln(1-p) \tag{7.12}$$

$$\text{CUMULANT} = \ln(p) \tag{7.13}$$

Differentiating the cumulant wrt ln(1−p)

$$\text{MEAN} = (1-p)/p = \mu \tag{7.14}$$

$$\text{VARIANCE} = (1-p)/p^2 = V(\mu) = \mu(1+\mu) \tag{7.15}$$

Parameterized in term of μ, the geometric probability distribution function is defined as

$$f(y;\mu) = 1/(1+\mu) * (\mu/(1-\mu))^y \tag{7.16}$$

The log-likelihood, expressed in exponential family form, appears as

$$\mathcal{L}(\mu;y) = \Sigma\exp\{y * \ln(\mu/(1+\mu)) - \ln(1+\mu)\} \tag{7.17}$$

with $\theta = \ln(\mu/(1+\mu))$. We define the link and inverse link in terms of μ as

$$\text{LINK} = \eta = \ln(\mu/(1+\mu)) = -\ln(1/\mu + 1) \tag{7.18}$$

$$\text{INVERSE LINK} = \mu = 1/(\exp(-\eta) - 1) \tag{7.19}$$

The canonical form of the geometric log-likelihood function in terms of βx, or xb, may be determined by substituting the value of the inverse link for every instance of μ.

CANONICAL LOG-LIKELIHOOD

$$\mathcal{L}(xb;y) = y * \ln(1/(\exp(-xb) - 1)) - (1+y)\ln(1 + (1/(\exp(-xb) - 1))) \tag{7.20}$$

Table 5.2 provides a sample GLM estimating algorithm for the canonical negative binomial. Using the same algorithm, but excluding the term α throughout the algorithm (or setting the value of α to 1.0), provides the appropriate canonical geometric parameter estimates.

In the last chapter, the fourth example concerned itself with attempting to determine the difference between pre- and post-reform visits to a physician by participants in the German health system. It was noted that the negative binomial value of α was close to 1.0. It might be of interest to compare the (log)geometric model, or traditional negative binomial with $\alpha = 1$, to a canonically linked geometric model. The parameter estimates are exponentiated, producing incidence rate ratios.

LOG-GEOMETRIC MODEL

```
. glm numvisit reform badh educ2 educ3 age2 age3 loginc,
nolog fam(nb 1) eform

Generalized linear models          No. of obs      =       2227
Optimization     :  ML             Residual df     =       2219
                                   Scale parameter =          1
Deviance         =  2410.140094    (1/df) Deviance =   1.086138
Pearson          =  2641.379859    (1/df) Pearson  =   1.190347
                                   AIC             =   4.103197
Log likelihood   = -4560.910337    BIC             =  -14694.82

------------------------------------------------------------------------------
         |                OIM
Numvisit |      IRR    Std. Err.      z    P>|z|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
  Reform | .8713684   .0445485    -2.69   0.007      .788287    .9632062
    Badh | 3.134881   .2333899    15.35   0.000     2.709253    3.627375
   educ2 | 1.085688   .0717399     1.24   0.213      .953805    1.235807
   educ3 | .9701217   .0686174    -0.43   0.668     .8445398    1.114377
    age2 | 1.049611   .0666594     0.76   0.446     .9267655    1.188741
    age3 | 1.206801   .0868511     2.61   0.009     1.048037    1.389617
  loginc | 1.134511   .0798964     1.79   0.073      .988244    1.302428
------------------------------------------------------------------------------
```

CANONICAL GEOMETRIC MODEL

```
. glm numvisit reform badh educ2 educ3 age2 age3 loginc,
nolog fam(nb 1) link(nb) eform

Generalized linear models           No. of obs      =       2227
Optimization     :  ML              Residual df     =       2219
                                    Scale parameter =          1
Deviance         =  2412.058271     (1/df) Deviance =   1.087002
Pearson          =  2654.882303     (1/df) Pearson  =   1.196432
                                    AIC             =   4.104059
Log likelihood   = -4561.869425     BIC             =   -14692.9

------------------------------------------------------------------------------
         |                OIM
Numvisit |   exp(b)    Std. Err.      z    P>|z|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
  Reform | .9691608   .0123938    -2.45   0.014     .9451713    .9937593
    Badh | 1.276637   .0168662    18.49   0.000     1.244004    1.310126
   educ2 | 1.017842   .0155696     1.16   0.248     .9877794     1.04882
   educ3 | .9802774   .0175087    -1.12   0.265     .9465547    1.015202
    age2 | 1.023043   .0156445     1.49   0.136     .9928357     1.05417
    age3 | 1.029421   .0167671     1.78   0.075     .9970771    1.062814
  loginc | 1.029692    .017289     1.74   0.081     .9963574    1.064141
------------------------------------------------------------------------------
```

Although the parameter estimates and standard errors are different, the fit statistics are all extremely close. The log-likelihood function, AIC and BIC statistics,

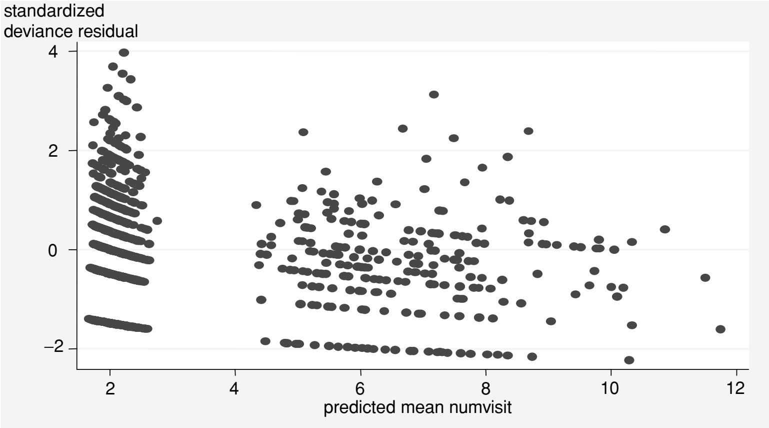

Figure 7.2. Canonical geometric model for predicted mean number of visits: standardized deviance residuals versus fitted values

and dispersion statistics are each nearly identical with one other. Only one predictor, *age3*, has ambivalent results with respect to significance. Scaling standard errors and applying a robust variance estimator did not affect the difference. As it is, because the coefficients of *age2* and *age3* are similar, one might attempt to combine levels 2–3 with age1. The three levels together then would serve as the referent. Education appears to be ruled out as contributory, and should likely be dropped. Moreover, assessing the possibility of interactions may lead to beneficial results. Modeling is far from complete.

A comparison of graphs of standardized deviance residuals by fit can apprise the reader if there are major distributional differences between the models, that somehow were not picked up by the fit statistics. Figure 6.6 displays the Poisson and negative binomial model residuals; Figure 7.1 shows the canonical geometric. However, since the negative binomial graph is based on an α of 1, it can be considered as (log) geometric.

7.2 NB-1: The linear constant model

7.2.1 NB-1 as QL-Poisson

Cameron and Trivedi (1986) were the first to make a distinction between the NB-1 and NB-2 models. The notion is based on the value of the exponent in

Table 7.2. *Creation of synthetic NB-1 data set <stata code>*

```
set obs 50000                      // Set data to 50,000 observations
gen x1 = invnorm(uniform())        // x1 random normal variate
gen x2 = invnorm(uniform())        // x2 random normal variate
gen xb1 =.5 + 1.25 *x1 - 1.5*x2    // define linear predictor
gen exb1 = exp(xb1)                // exponentiate linear predictor
gennbreg y1, mu(exb1) delta(0.5) dispersion(constant)
                                   // NB-1 random variates
```

the variance function. NB-2, the traditional parameterization of the negative binomial, has a variance function appearing as $\mu + \alpha\mu^2$, or equivalently, $\mu(1 + \alpha\mu)$. The first of these formulae is the most common representation of the variance. The square value of μ in the formula classifies the equation as quadratic. The NB-1 model, on the other hand, is called the linear parameterization, due to its form: $\mu + \alpha\mu$ or $\mu(1 + \alpha)$. The highest (implied) value of an exponent of μ in the formula is 1. A variable power negative binomial variance function has been given as $\mu + \alpha\mu^p$, with p taking the value of 2 if quadratic, and 1 if linear. In the next section we shall address the situation where p is considered as another ancillary parameter – NB-P.

Table 7.1 displayed the range of multiplicative extensions that have been applied to the Poisson variance function. The top five in the list included the following:

Poisson	:	$V = \mu$	NB-1	:	$V = \mu(1 + \alpha)$
QL Poisson	:	$V = \mu(\phi)$	NB-2	:	$V = \mu(1 + \alpha\mu)$
Geometric	:	$V = \mu(1 + \mu)$			

Expressed in the above manner, we see that the values within the parentheses are, taken together, multipliers of the basic Poisson variance. As such, there is little difference in the quasi-likelihood and NB-1 parameterizations. If α is entered into the estimating equation as a constant, then the two formulae are identical – both are quasi-likelihood models. On the other hand, if α is estimated (maximum likelihood), then the models are clearly different. An example may help clarify these relationships.

Parameters are defined as:

Constant =	0.50
X1 =	1.25
X2 =	-1.50

Modeling the data as NB-1, using maximum likelihood, we have:

```
. nbreg y1 x1 x2, nolog dispersion(cons)

Negative binomial regression       Number of obs    =      49770
                                   LR chi2(2)       =  108413.66
Dispersion        =   constant     Prob > chi2      =     0.0000
Log likelihood    =   -87238.249   Pseudo R2        =     0.3832

==========================================================================
        y1 |  Coef.    Std. Err.      z    P>|z|  [95% Conf.  Interval]
-----------+--------------------------------------------------------------
        x1 | 1.250363 .0016506   757.50 0.000   1.247128   1.253599
        x2 |-1.500081 .0016055  -934.37 0.000  -1.503228  -1.496934
     _cons | .5027491 .0035794   140.46 0.000   .4957337   .5097646
-----------+--------------------------------------------------------------
  /lndelta |-.6870266 .0222631                 -.7306615  -.6433917
-----------+--------------------------------------------------------------
     delta | .5030697 .0111999                  .4815903    .525507
--------------------------------------------------------------------------
Likelihood-ratio test of delta=0: chibar2(01) = 4039.17 Prob>
=chibar2 = 0.000
==========================================================================
```

We now model the data as a QL-Poisson with $\phi = (1 + \alpha) = 1.503+$. The estimating algorithm uses built-in quasi-likelihood capability, which is based on Fisher scoring. Fisher scoring in turn uses the expected rather than observed information matrix to calculate standard errors. This results in the comparative differences between the standard errors of the NB-1 and QL-Poisson models.

```
. glm y1 x1 x2, nolog fam(poi) disp(1.5030697) irls

Generalized linear models           No. of obs       =      49770
Optimization        : MQL Fisher    Residual df      =      49767
                      scoring
                      (IRLS EIM)    Scale parameter  =    1.50307
Deviance            = 43557.89747   (1/df) Deviance  =   .8752366
Pearson             = 50154.52375   (1/df) Pearson   =   1.007787
Quasi-likelihood    : 1.50307       BIC              =  -494680.6
model with
dispersion

=================================================================
             EIM
   y1 |  Coef.      Std. Err.      z      P>|z|  [95% Conf. Interval]
------+----------------------------------------------------------
   x1 |  1.250264  .0011009    1135.67  0.000    1.248106   1.252422
   x2 | -1.499982  .0010706   -1401.10  0.000    -1.50208  -1.497884
_cons |  .5030227  .0023855     210.87  0.000    .4983473   .5076982
-----------------------------------------------------------------
AIC = 3.586933
=================================================================
```

The foremost reason to use a maximum likelihood NB-1 model rather than a QL-Poisson model with data rests with the fact that NB-1 estimates ϕ as $(1 + \alpha)$, whereas ϕ, as a dispersion statistic, must be entered into the QL-Poisson model as a constant. On the down side, NB-1 is rarely supported in commercial software. As of this writing, only Stata and LIMDEP offer it as a capability.

7.2.2 Derivation of NB-1

The NB-1, or linear negative binomial, is derived as a Poisson–gamma mixture model; however, the manner of derivation differs from the traditional NB-2 model (see Chapter 5). As with the NB-2 model, the derivation begins with the usual count data, or Poisson, model (again, for the most part I shall forgo subscripts for ease of interpretation for the remainder of this section).

$$y_i \sim \text{Poisson}(\lambda_i) = f(y_\iota; \lambda_\iota) = e^{-\lambda}\mu^y/y! \tag{7.21}$$

However, in this case the mean of the Poisson is itself a random variable such that

$$\lambda_i \sim \text{gamma}(\delta, \mu_i) \tag{7.22}$$

through which covariates are introduced via $\mu_i = \exp(x_i\beta)$. If an offset is applied to the model, $\mu_i = \exp(x_i\beta) + \text{offset}$. However, unless specifically addressed, I shall forego including the offset with the linear predictor for simplification purposes.

From the definition of the Gamma distribution we know that the mean and variance are given by

$$E[\lambda_i] = \frac{\mu_i}{\delta} = \exp(x_i\beta)/\delta \tag{7.23}$$

$$V[\lambda_i] = \frac{\mu_i}{\delta^2} = \exp(x_i\beta)/\delta^2 \tag{7.24}$$

where δ is the gamma scale parameter.

The resulting mixture is described as

$$f(y|x) = \int_0^\infty \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}\frac{\delta^{\mu_i}}{\Gamma(\mu_i)}e^{-\lambda_i\delta}\,d\lambda_i \tag{7.25}$$

Solving to clear the integration, we have

$$f(y;\mu) = \int_0^\infty e^{-\mu}\mu^y/y!\ \ \delta^\mu/\Gamma(\mu)\quad \mu^{\mu-1}e^{-\mu\delta}d\mu \tag{7.26}$$

$$= \frac{\delta^\mu}{\Gamma(\mu+1)\Gamma(\mu)}\int_0^\infty \mu^{(\mathrm{y}+\mu)-1}e^{-\mu(\delta+1)}d\mu \tag{7.27}$$

$$= \frac{\delta^\mu}{\Gamma(\mu+1)\Gamma(\mu)}\frac{\Gamma(y+\mu)}{(\delta+1)^{y+\mu}} \times C \quad d\mu \tag{7.28}$$

where C reduces to the value of 1. It appears as

$$C = \int_0^\infty \frac{(\delta+1)^{y+\mu}}{\Gamma(y+\mu)}\lambda^{(y+\mu)-1}e^{-\lambda(\delta+1)}d\lambda \tag{7.29}$$

Continuing from Equation (7.28), less the value of C, we have

$$= \frac{\Gamma(y+\mu)}{\Gamma(y+1)\Gamma(\mu)}(\delta/(1+\delta))^{\mu}(1/1+\delta)^{y} \tag{7.30}$$

The mean and variance of Equation (7.30), the NB-1 distribution, are

$$\text{NB-1 mean} = E[y] = \exp(x\beta)/\delta \tag{7.31}$$

$$\text{NB-1 variance} = E[y] = \exp(x\beta)(1+\delta)/\delta^2 \tag{7.32}$$

The variance to mean ratio is $(1+\delta)/\delta^2$, which is constant for all observations. This feature of the distribution results in constant overdispersion within the model, unlike NB-2, in which δ is variable with a mean of 1. Defining $\alpha = 1/\delta$, the distribution may be re-expressed in more familiar terms as

$$\frac{\Gamma(y+\mu)}{\Gamma(y+1)\Gamma(\mu)}(1/(1+\alpha))^{\mu}(\alpha/1+\alpha)^{y} \tag{7.33}$$

As in the parameterization of NB-2, specified in Chapter 5, $\alpha = 0$ is Poisson.

Standardizing the coefficients, β, by the addition of $-\ln(\alpha)$ to the linear predictor

$$E(y) = \mu = \exp(x\beta - \ln(\alpha)) \tag{7.34}$$

the NB-1 distribution (Eq. (7.33)) may be expressed as the log-likelihood

$$\begin{aligned}\mathcal{L}(\mu; y) = {} & \Sigma[\ln\{\Gamma(\mu+y)\} - \ln\{\Gamma(y+1)\} \\ & - \ln\{\Gamma(\mu)\} + y\ln(\alpha) - (y+\mu)\ln(1+\alpha)]\end{aligned} \tag{7.35}$$

Additional discussion regarding the derivation of the NB-1, as well as of the logic of the standardization of the linear predictor, can be found in Hardin and Hilbe (2001). Also see Cameron and Trivedi (1998).

7.2.3 Modeling with NB-1

Using the same German health reform data as earlier in this chapter, number of visits is modeled as NB-1:

```
. nbreg numvisit reform badh educ2 educ3 age2 age3 loginc,
irr disp(constant)

Negative Binomial Type 1          Number of obs    =     2227
Regression
                                  Wald chi2(7)     =   318.49
Log likelihood   =   -4600.3458   Prob > chi2      =   0.0000
```

```
numvisit |   IRR    Std. Err.    z    P>|z|  [95% Conf. Interval]
---------+--------------------------------------------------------
  reform | .9019533  .0402234 -2.31 0.021    .8264641  .9843378
    badh | 2.607651   .147245 16.97 0.000    2.334453  2.912822
   educ2 | 1.050977  .0616253  0.85 0.396    .9368752  1.178974
   educ3 | 1.017186  .0629838  0.28 0.783    .9009369  1.148435
    age2 | .9936896  .0544585 -0.12 0.908    .8924856   1.10637
    age3 | 1.050923  .0655038  0.80 0.426    .9300703   1.18748
  loginc | 1.135741  .0686193  2.11 0.035    1.008907  1.278519
---------+--------------------------------------------------------
/lnalpha | .9881723    .05329 18.54 0.000    .8837258  1.092619
---------+--------------------------------------------------------
   alpha | 2.68632    .143154                2.419899  2.982073
---------+--------------------------------------------------------
AIC Statistic = 4.139
```

The following table lists the differences in AIC and alpha for the models we have thus far discussed. Note that the NB-2 and canonical linked model, NB-C, are nearly identical, even though NB-C does not use a log link in its algorithm. Both NB-2 and NB-1 employ the log link. On the other hand, parameter estimates for each of the models are similar, and indicate the same predictors as contributory to the model.

	AIC	alpha
NB-2	4.104	.998
NB-C	4.104	.998
NB-1	4.139	2.686

Of possible interest is a comparison of models based on the synthetic data created in the first section of this chapter. Using the same parameter specifications, synthetic data sets were created for both NB-2 and NB-1 models. Modeling combinations of NB-1 and 2 data with NB-1 and 2 models gives us:

```
PARAMETERS: X1= 2.0; X2= -.5; _CONS= -1.0; delta/alpha= 1.0

MODEL: NB-1
DATA: NB-1
-----------------------------------------------------------------
   y1 |  Coef.    Std. Err.     z     P>|z|  [95% Conf. Interval]
------+----------------------------------------------------------
   x1 |  2.001738 .0035015   571.68 0.000    1.994876  2.008601
   x2 | -.5045447 .0034868  -144.70 0.000   -.5113788 -.4977106
_cons | -1.002748 .0081231  -123.44 0.000   -1.018668 -.9868266
------+----------------------------------------------------------
Delta |  1.015196 .0201667                   .9764294  1.055501
------+----------------------------------------------------------
Likelihood-ratio test of delta=0:chibar2(01)= 9132.96 Prob>
=chibar2 = 0.000
AIC Statistic = 2.292
```

```
===============================================================
MODEL: NB-1
DATA: NB-2
---------------------------------------------------------------
   y2 |   Coef.    Std. Err.     z    P>|z|   [95% Conf.  Interval]
------+--------------------------------------------------------
   x1 |  1.541805   .0078804  195.65  0.000   1.526359    1.55725
   x2 | -.3880394   .0062148  -62.44  0.000  -.4002201  -.3758587
_cons | -.0982971    .014988   -6.56  0.000  -.1276729  -.0689212
------+--------------------------------------------------------
delta |  8.713806   .1342828                  8.454552   8.981011
---------------------------------------------------------------
Likelihood-ratio test of delta=0: chibar2(01) = 8.4e + 04 Prob>
=chibar2 = 0.000
AIC Statistic = 2.606
===============================================================
```

```
===============================================================
MODEL: NB-2
DATA: NB-1
---------------------------------------------------------------
   y1 |   Coef.    Std. Err.     z    P>|z|   [95% Conf.  Interval]
------+--------------------------------------------------------
   x1 |  1.989207    .006577  302.45  0.000   1.976316   2.002098
   x2 | -.4956426   .0051694  -95.88  0.000  -.5057745  -.4855108
_cons | -.9852668   .0088877 -110.86  0.000  -1.002686  -.9678472
------+--------------------------------------------------------
alpha |  .2370104   .0074614                  .2228283   .2520951
---------------------------------------------------------------
Likelihood-ratio test of alpha=0: chibar2(01) = 3761.32 Prob>
=chibar2 = 0.000
AIC Statistic = 2.408
===============================================================
```

A few observations regarding the above output: First, modeling NB-2 data with a NB-1 model substantially alters the specified parameter estimates, as well as the value of delta/alpha. Second, modeling NB-1 data with a NB-2 model does not substantially alter the specified parameter estimates, but the value of the ancillary parameter is changed. These relationships may be important in practical applications, but only if one knows a priori how the data are generated.

7.3 NB-H: Heterogeneous negative binomial regression

The heterogeneous negative binomial extends the negative binomial model by allowing observation-specific parameterization of the ancillary parameter, α. In other words, the value of α is partitioned by user-specified predictors. α takes the form $\exp(z_i \nu)$, which, like α, is positive.

There are two uses of the heterogeneous model. First, parameterization of α provides information regarding which predictors influence overdispersion. Second, it is possible to determine whether overdispersion varies over the significant predictors of α by observing the differential values of its standard errors. If the standard errors vary only little between parameters, then the overdispersion in the model can be regarded as constant.

```
. GENERALIZED NB-2:alpha==.5
-------------------------------------------------------------------------
            |            Std.                         [95% Conf.
            |  Coef.     Err.         z     P>|z|    Interval]
------------+------------------------------------------------------------
y1          |
         x1 |-1.503948 .0049501 -303.82 0.000     -1.51365 -1.494246
         x2 | .7491305  .004433  168.99 0.000      .740442   .757819
      _cons | .9939296  .004998  198.86 0.000     .9841336  1.003725
------------+------------------------------------------------------------
lnalpha     |
         x1 | .0041994 .0131544    0.32 0.750   -.0215828  .0299815
         x2 |-.0047799 .0110758   -0.43 0.666   -.0264881  .0169283
      _cons |-.6846148  .015551  -44.02 0.000   -.7150941 -.6541354
------------+------------------------------------------------------------
AIC Statistic = 4.584
```

The above NB-2 synthetic data set is created with parameters of $x_1 = -1.5$, $x_2 = 0.75$, and constant = 1.0. α is specified as 0.5. Parameterization of α produces estimates that have little variability; i.e., there is little difference in α parameter values as well as standard errors. Of course, in setting up the data, x_1 and x_2 were created using the same formula. Synthetic data sets having different αs produce similar results.

Parameterization of α for a NB-2 model of *numvisits* (*mdvisitsx*), as displayed in section 7.3 in this chapter, is given as:

```
---------------------------------------------------------------------
         |    Coef.   Std. Err.    z    P>|z|   [95% Conf. Interval]
---------+-----------------------------------------------------------
numvisit |
  reform | -.1331439  .0507271 -2.62 0.009   -.2325672 -.0337205
    badh |  1.127951  .0717968 15.71 0.000    .9872322   1.26867
   educ2 |  .0867164  .0686798  1.26 0.207   -.0478936  .2213264
   educ3 | -.0267123   .070227 -0.38 0.704   -.1643548  .1109302
    age2 |  .0429533  .0650913  0.66 0.509   -.0846232  .1705299
    age3 |  .1742675  .0751244  2.32 0.020    .0270264  .3215087
  loginc |  .1132818   .070934  1.60 0.110   -.0257463  .2523099
   _cons |  -.148793  .5421144 -0.27 0.784   -1.211318  .9137316
---------+-----------------------------------------------------------
lnalpha  |
  reform |  -.009731  .0971176 -0.10 0.920    -.200078   .180616
    badh | -.1890283  .1238452 -1.53 0.127   -.4317604  .0537037
   educ2 |  .0521736  .1223241  0.43 0.670   -.1875772  .2919245
   educ3 | -.3076519  .1383376 -2.22 0.026   -.5787886 -.0365151
    age2 |  .2675544  .1180651  2.27 0.023     .036151  .4989577
    age3 |  .3246583  .1284251  2.53 0.011    .0729499  .5763668
  loginc | -.0967873  .1338098 -0.72 0.469   -.3590498  .1654751
   _cons |  .7203326  1.025238  0.70 0.482   -1.289097  2.729762
---------+-----------------------------------------------------------
AIC Statistic =4.094
```

Having α parameterized tells us which predictors influence α. *educ3*, and *age* (*age2, age3*) influence the amount of overdispersion in the data. These two predictors also significantly contribute to the count aspect of a negative binomial-clog hurdle model on the same data, as shall be observed in Chapter 9. The AIC value of 4.094 is also the same value as that of the hurdle model.

Heterogeneous negative binomial regression is a valuable tool for assessing the source of overdispersion. It can be used to differentiate sources influencing the model parameter estimates versus sources influencing overdispersion. A reduced model, indicating such influences, is given as

```
--------------------------------------------------------------------------
         |     Coef.   Std. Err.      z    P>|z|  [95% Conf.   Interval]
---------+----------------------------------------------------------------
numvisit |
  reform | -.1359128  .0508994   -2.67  0.008   -.2356739   -.0361517
    badh |  1.156625  .0746261   15.50  0.000     1.01036    1.302889
    age3 |  .1775751  .0723686    2.45  0.014    .0357352     .319415
  loginc |  .1302475  .0689939    1.89  0.059   -.0049781    .2654731
   _cons | -.2445701  .5329164   -0.46  0.646   -1.289067    .7999268
---------+----------------------------------------------------------------
lnalpha  |
    age2 |  .2620468  .1147147    2.28  0.022    .0372102    .4868835
    age3 |  .3325014  .1244992    2.67  0.008    .0884876    .5765153
   _cons | -.1316017  .0657923   -2.00  0.045   -.2605521   -.0026512
--------------------------------------------------------------------------
```

7.4 The NB-P model

Building on the prior work of Winkelmann and Zimmermann (1995) and Cameron & Trivedi (1998), who discussed what was termed a Generalized Event Count [GEC(k)] model, Greene (2006) created the NB-P model to allow more flexibility in the NB-2 variance. Recall that the form of the NB-1 and NB-2 variance functions are, respectively

$$\begin{array}{llll} \text{NB-1} & \mu_i + \alpha\mu_i & \text{or} & \mu_i(1+\alpha) \\ \text{NB-2} & \mu_i + \alpha\mu_i^2 & \text{or} & \mu_i(1+\alpha\mu_i) \end{array}$$

Greene's generalization takes the form

$$\text{NB-P} \quad \mu_i + \alpha\mu_i^Q \quad \text{or} \quad \mu_i\left(1+\alpha\mu_i^{Q-1}\right) \tag{7.36}$$

where Q is a parameter to be estimated. This form of negative binomial is a three parameter model, with μ, α, and Q as parameters. It is not a simple reparameterization of the basic NB-2 model, but rather an entirely separate model.

The NB-2 model may be schematized as (see Greene, 2006 (E24.3.4))

$$\text{Prob}(Y = y_i|x_i) = \frac{\Gamma(y_i+\theta)}{\Gamma(\theta)\Gamma(y_i+1)} u_i^\theta (1-u_i)^{y_i} \tag{7.37}$$

with

$$u_i = \theta/(\theta + \mu_i) \tag{7.38}$$

and

$$\theta = 1/\alpha \tag{7.39}$$

NB-1 as

$$\text{Prob}(Y = y_i|x_i) = \frac{\Gamma(y_i + \mu_i\theta)}{\Gamma(\mu_i\theta)\Gamma(y_i + 1)} u_i^{\mu\theta}(1 - u_i)^{y_i} \tag{7.40}$$

with

$$u_i = \theta/(\theta + 1) \tag{7.41}$$

The NB-P distribution takes the same form as NB-2, as expressed in Equations (7.37)–(7.39). However, for each value of θ in Equations (7.37) and (**??**), we substitute the value $\theta\mu^{2\text{-}P}$. For ease of interpretation, however, Greene substitutes parameter 2-*P* as *Q*, with NB-2 having $Q = 0$ and NB-1 as $Q = 1$. Parameterized in this manner, the NB-P distribution replaces NB-2 values of θ with the value $\theta\mu^Q$.

The NB-P probability mass function is given by Greene (2006) as

$$\text{Prob}(Y = y_i|x_i) = \frac{\Gamma\left(\theta\lambda_i^Q + y_i\right)}{\Gamma\left(\theta\lambda_i^Q\right)\Gamma\left(y_i + 1\right)} \left(\frac{\theta\lambda_i^Q}{\theta\lambda_i^Q + \lambda_i}\right)^{\theta\lambda_i^Q} \left(\frac{\lambda_i}{\theta\lambda_i^Q + \lambda_i}\right)^{y_i} \tag{7.42}$$

An example will show the value of the NB-P model. We use the same German health data, called *rwm*, as used by Greene. It is taken from earlier years than the *mdvisitsx* data we have used in previous examples.

The response is *docvis*, number of visits to the doctor, with three predictors:

age : age from 25 through 64
hh : monthly net household income in German marks/1,000. Converted from *hhninc* by *hh*=*hhninc*/10
educ : years of schooling, ranging from 7 through 18(+).

First, we use a NB-2 model to estimate parameters and the value of alpha.

```
. gen hh = hhninc/10
. nbreg docvis age hh educ, nolog

Negative binomial regression      Number of obs    =     27326
                                  LR chi2(3)       =   1027.40
Dispersion        =   mean        Prob > chi2      =    0.0000
Log likelihood    =   -60322.021  Pseudo R2        =    0.0084
```

```
=========================================================================
  docvis |   Coef.    Std. Err.     z     P>|z|  [95% Conf.  Interval]
---------+---------------------------------------------------------------
     age |  .0204292  .0008006    25.52  0.000    .0188601   .0219984
      hh | -.4768144  .0522786    -9.12  0.000   -.5792785  -.3743503
    educ | -.0459575  .0042257   -10.88  0.000   -.0542398  -.0376752
   _cons |  .9132608  .0633758    14.41  0.000    .7890465   1.037475
---------+---------------------------------------------------------------
/lnalpha |  .6608039  .0115374                      .638191   .6834168
---------+---------------------------------------------------------------
   alpha |  1.936348  .0223404                     1.893053   1.980634
---------+---------------------------------------------------------------
Likelihood-ratio test of alpha=0: chibar2(01) = 8.9e + 04 Prob>
=chibar2 = 0.000
AIC Statistic =4.415
=========================================================================
```

We next model the data as NB-P, where both α as well as Q is estimated. LIMDEP 9.0 is used to model the data.

```
Negative Binomial (P) Model
Maximum Likelihood Estimates
Model estimated: Mar 31, 2006 at 09:13:20PM.
Dependent variable                   DOCVIS
Weighting variable                     None
Number of observations                27326
Iterations completed                     15
Log likelihood function           -60258.97
Number of parameters                      6
Info. Criterion: AIC =              4.41082
Finite Sample: AIC   =              4.41082
Info. Criterion: BIC =              4.41262
Info. Criterion:HQIC =              4.41140
Restricted log likelihood         -104814.1
McFadden Pseudo R-squared          .4250871
Chi squared                        89110.23
Degrees of freedom                        1
Prob[ChiSqd > value] =             .0000000
```

```
========================================================================
        |              Standard
Variable| Coefficient  Error       b/St.Er.  P[|Z|>z]     Mean of X
--------+---------------------------------------------------------------
Constant|   .77029290  .05940343    12.967   .0000
     AGE|   .02177762  .00074029    29.418   .0000      43.5256898
  HHNINC|  -.38749687  .05121714    -7.566   .0000       .35208362
    EDUC|  -.04127636  .00412037   -10.018   .0000      11.3206310
        | Dispersion parameter for count data model
   Alpha|  3.27757050  .14132403    23.192   .0000
        | Negative Binomial. General form, NegBin P
       P|  2.45563168  .03595933    68.289   .0000
========================================================================
```

Table 7.3 displays a comparison between NB-2 and NB-P estimates.

A likelihood ratio test between the NB-2 and NB-P models result in

$$-2\{\text{LL(NB-2)} - \text{LL(NB-P)}\}$$
$$-2\{(-60322.021) - (-60268.97)\}$$
$$106.10$$

Table 7.3. *Comparison: NB-2 and NB-P results*

	NB-2	NB-P
age	.0204	.0218
hh	-.4768	-.3875
educ	-.0460	-.0413
constant	.9133	.7703
alpha	1.936	3.275
power	2.000	2.456
AIC	4.415	4.411

A reverse cumulative upper tail χ^2 distribution with one degree of freedom gives us a probability value: chi2tail(1, 106.10) = 7.011e-25. The traditional cutoff point for a 0.05 significance level is chiprob(1, 3.84) = .05004352. Therefore, any χ^2 having a value exceeding 3.84 will be significant; i.e. the models significantly differ from one another. Greene presents a Wald t-test of significance for comparing the two values of power: (NB-P–NB-2)/SE$_{\text{NB-2}}$ or $(2.456 - 2)/.036 = 12.65$. Using a reverse cumulative upper tail Students' T distribution with one degree of freedom, we have: ttail(1,12.65) = .02511062. A test of power gives additional evidence that the two models are significantly different. Given the lower value of the NB-2 log-likelihood function, and a slightly lower value of the NB-P AIC statistic, it appears that the NB-P model may be slightly preferred over the NB-2.

NB-P models generalize the basic negative binomial model, providing more flexibility in fitting data. The usefulness of the model has yet to be explored, but I suspect that this fact will soon be remedied.

7.5 Generalized Poisson regression

A generalization to the basic Poisson model was developed by Consul and Jain (1973), which they aptly termed generalized Poisson regression. It has since undergone various modifications, with models created with names such as the restricted generalized Poisson, three parameterizations of a Hybrid Generalized Poisson, and so forth. Refer to Consul and Famoye (1992) for a good overview of the base generalized Poisson model and its derivation.

Generalized Poisson is similar to the negative binomial in that it incorporates an extra heterogeneity or dispersion parameter. However, unlike the negative binomial, the generalized Poisson, and its variations, allow modeling of both underdispersed as well as overdispersed data.

The probability distribution given below is the parameterization found in Famoye and Singh (2006), which corresponds to the manner in which we have expressed the mean and heterogeneity parameters throughout the text. It is given, without subscripts, as

$$f(y;\mu;\alpha) = (\mu/(1+\alpha\mu))^2(1+\alpha\mu)^{y-1}/y!\exp[-\mu(1+\alpha y)/(1+\alpha\mu)]$$

with α specifying the heterogeneity parameter.

The log-likelihood function can be given as

$$\begin{aligned}\mathcal{L}(\mu_i;\alpha,y_i) = \sum_{i=1}^{n} & y\ln(\mu_i/(1+\alpha\mu_i) + (y_i-1)\ln(1+\alpha y_i) \\ & - [\mu_i(1+\alpha y_i)/(1+\alpha\mu_i)] - 1\mathrm{n}\Gamma(y_i+1))\end{aligned}$$

or terms of $x\beta$ as

$$\begin{aligned}\mathcal{L}(\beta;\alpha,y_i) = \sum_{i=1}^{n} & y\ln(\exp(x_i\beta)/(1+\alpha\exp(x_i\beta))+(y_i-1)\ln(1+\alpha y_i) \\ & - [\exp((x_i\beta)(1+\alpha y_i)/(1+\alpha\exp(x_i\beta))] - \ln\Gamma(y_i+1)\end{aligned}$$

Like the negative binomial, as α approaches zero, the generalized Poisson reduces to the basic Poisson. That is, a generalized Poisson having an α zero is Poisson, and is equi-dispersed. The model is neither under-nor over-dispersed. In most real data situations this rarely occurs. However, since α of the generalized Poisson model can be both positive and negative, any value close to zero should be considered as Poisson. Tests modeled after NB-2 tests can assess if the generalized model is statistically different from a Poisson model.

We use a generalized Poisson regression procedure based on the above parameterization of the distribution to model the same German health care data as in the previous sections. LIMDEP is the only commercial software package to offer the generalized Poisson regression model to its users. However, the software implementation used here, authored by James Hardin, is written using Stata's higher programming language. It was first used in Hardin and Hilbe (2007) as well as in an unpublished manuscript authored by Yang, Hardin, and Addy (2007). Refer to Hardin and Hilbe (2007) for more extensive discussion of the model. In addition, an excellent overview of the various generalizations to the Poisson model, under the rubric of generalized Poisson regression, can be found in Drescher (2005).

We model the number of visits to the doctor with the aim of determining if there is a change in the count from before to after health care reform (*reform*).

The model is adjusted by health status, education level, age group, and the log of patient income. The result of employing a generalized Poisson model is:

```
. gpoisson numvisit reform badh educ2 educ3 age2 age3
loginc, nolog

Generalized Poisson regression      Number of obs   =     2227
                                    LR chi2(7)      =   212.17
                                    Prob > chi2     =   0.0000
Log likelihood   =   -4580.1188     Pseudo R2       =   0.0226

=============================================================
numvisit|   Coef.   Std. Err.    z    P>|z|  [95% Conf. Interval]
--------+----------------------------------------------------
  reform|-.0989636 .0444614  -2.23 0.026  -.1861064 -.0118208
    badh| .9377614 .0571122  16.42 0.000   .8258236  1.049699
   educ2| .0448823 .0586723   0.76 0.444  -.0701133   .159878
   educ3|  .026951 .0617273   0.44 0.662  -.0940324  .1479343
    age2|-.0209855 .0548281  -0.38 0.702  -.1284465  .0864755
    age3| .0440805 .0622676   0.71 0.479  -.0779618  .1661227
  loginc| .1224226 .0602443   2.03 0.042   .0043459  .2404993
   _cons|-.1382682 .4571795  -0.30 0.762  -1.034323  .7577872
--------+----------------------------------------------------
 /Zdelta| .0062482 .0243254                -.0414287  .0539252
--------+----------------------------------------------------
   delta| .5031241 .0121622                 .4792975  .5269365
--------+----------------------------------------------------
Likelihood-ratio test of delta=0:chibar2(1) = 2703.59 Prob>
=chibar2(1) = 0.000

. aic
AIC Statistic =4.121
=============================================================
```

The model indicates overdispersion ($\alpha = .50$). We compare with the NB-1 table of estimates.

```
=============================================================
numvisit|   IRR    Std. Err.   z    P>|z|  [95% Conf. Interval]
--------+----------------------------------------------------
  reform|.9019533   .0402234 -2.31 0.021   .8264641   .9843378
    badh|2.607651    .147245 16.97 0.000   2.334453   2.912822
   educ2|1.050977   .0616253  0.85 0.396   .9368752   1.178974
   educ3|1.017186   .0629838  0.28 0.783   .9009369   1.148435
    age2|.9936896   .0544585 -0.12 0.908   .8924856    1.10637
    age3|1.050923   .0655038  0.80 0.426   .9300703    1.18748
  loginc|1.135741   .0686193  2.11 0.035   1.008907   1.278519
--------+----------------------------------------------------
/lnalpha|.9881723     .05329 18.54 0.000   .8837258   1.092619
--------+----------------------------------------------------
   alpha| 2.68632    .143154                2.419899   2.982073
--------+----------------------------------------------------
AIC Statistic =4.139
=============================================================
```

and the NB-2 table of parameter estimates:

```
numvisit |   Coef.   Std. Err.    z    P>|z|   [95% Conf.  Interval]
---------+--------------------------------------------------------
  reform |-.1376897  .0510927  -2.69  0.007   -.2378295  -.0375499
    badh | 1.142588  .0743937  15.36  0.000    .9967791   1.288397
   educ2 | .0822126  .0660359   1.24  0.213   -.0472153   .2116406
   educ3 |-.0303508   .070688  -0.43  0.668   -.1688968   .1081951
    age2 | .0484328  .0634691   0.76  0.445   -.0759644     .17283
    age3 | .1879573  .0719228   2.61  0.009    .0469913   .3289233
  loginc | .1262033  .0703792   1.79  0.073   -.0117374   .2641439
   _cons |-.2478272  .5364981  -0.46  0.644   -1.299344   .8036898
---------+--------------------------------------------------------
/lnalpha | -.001789   .0476115                 -.0951058   .0915279
---------+--------------------------------------------------------
   alpha | .9982126   .0475264                  .9092767   1.095847
---------+--------------------------------------------------------
AIC Statistic =4.104
```

A reduced model may be developed by entering only those predictors that significantly contribute to the model. Care must be taken to check each combination of predictors in the full model. Significant interaction effects may still occur among predictors that appear non-contributory in the full model. In this case, however, no such effects were identified. The “best” model based on the full model predictors is:

```
. gpoisson numvisit reform badh loginc, nolog

Generalized Poisson regression     Number of obs    =      2227
                                   LR chi2(3)       =    210.80
                                   Prob > chi2      =    0.0000
Log likelihood    =   -4580.8043   Pseudo R2        =    0.0225

numvisit |   Coef.    Std. Err.    z    P<|z|  [95% Conf.  Interval]
---------+--------------------------------------------------------
  reform |-.0971893  .0442944  -2.19  0.028  -.1840048  -.0103739
    badh | .9392447  .0558072  16.83  0.000   .8298646   1.048625
  loginc | .1366256  .0576442   2.37  0.018   .0236451   .2496061
   _cons |-.2186771  .4473718  -0.49  0.625   -1.09551   .6581556
---------+--------------------------------------------------------
 /Zdelta | .0064582  .0243029                 -.0411745   .0540909
---------+--------------------------------------------------------
   delta | .5032291  .0121509                  .4794244   .5270191
---------+--------------------------------------------------------
Likelihood-ratio test of delta=0:chibar2(1) = 2748.22 Prob>
=chibar2(1) = 0.000

. aic
AIC Statistic = 4.118
```

The reduced model produces an AIC statistic that is less, but not significantly so, than the full generalized Poisson model. P-values for *reform* and *badh* are statistically identical between the full and reduced model. The p-value for *loginc* improves substantially in the reduced model, especially considering the relatively large number of observations. The heterogeneity parameter is statistically identical in the two models. It appears that there is no statistical reason to prefer one model over the other.

7.6 Summary

There are two major varieties of enhancements to the negative binomial model. One type relates to adjustments made to the basic NB-2 model in light of distributional abnormalities with the data. That is, the count data to be modeled do not always match the distributional assumptions of the negative binomial model. Likewise, count data do not always accord with the distributional assumnptions of the Poisson. Although NB-2 is used to adjust for Poisson overdispersion, it does so without a knowledge of the possible reasons for the overdispersion. However, certain distribuitional properties of the data violate both Poisson and NB-2 assumptions, and we can identify the source of overdispersion. For instance, both Poisson and NB-2 require that there at least be the possibility of zero counts in the data, or, if there are, that there is not an excessive number as compared with the assumptions of the respective distributions. When there are no possibilities of zero counts, or there are far more than what accords with the respective model assumptions, then one may construct zero-truncated or zero-inflated Poisson and negative binomial models. We address these models in the next chapter.

The second type of enhancement adjusts the Poisson variance function, thereby creating new statistical models. Alterations in variance functions are shown in Table 7.1. An overview of each of these models was presented in this chapter. The following is a very brief summary of how each of these models relates to the basic NB-2 model.

Geometric	:	NB-2 with α==1
NB-1	:	NB-2 with no exponent
NB-H	:	NB-2 with α parameterized
NB-P	:	NB-2 with the exponent parameterized

Generalized Poisson: α can be used for both over- and underdispersion. It is not a negative binomial model, but a Poisson with ancillary parameter not based on gamma distribution.

Unfortunately most commercial software does not offer these models. Only LIMDEP offers them all. Stata's `nbreg` command can be used to model both NB-2 and NB-1 by using the dispersion(mean) and dispersion(constant) options respectively. Dispersion(mean) is the default. Stata also has the NB-H model, but calls it *generalized negative binomial* instead. However, since there is a previous tradition in the literature regarding generalized negative binomial models that differs considerably from NB-H, Stata's usage is a misnomer. I do not discuss the generalized negative binomial in this text since the models that have been developed have thus far been proved to result in rather highly biased estimates. They are not found in any commercial software. On the other hand, the generalized Poisson is similar to the generalized negative binomial models, but it does not suffer from their difficulties. Stata offers it as a user authored program. NB-P, first developed in 2006, is unique to LIMDEP and can more properly be termed a generalized model. It will likely become a well-used model in the future.

Exercises

1 Construct an identity-geometric model; i.e. a geometric family with an identity link. Compare output with the identity-gamma model using the *cancer* data set with *studytime* as the response. Use *age* and levels of *drug* as explanatory predictors. Drug level 1 is the referent. Discuss the relationship between the two sets of parameter estimates.

2 Parameterize the NB-2 model with the variance function defined as $\mu + \mu^2/\nu$ rather than the standard parameterization. What will be the limiting value of ν for the model to be Poisson? Why?

3 Is the NB-2 model a member of the exponential family of distributions? What are the implications for estimating the model?

4 Develop a maximum likelihood NB-P program using a higher language such as SAS- IML, Stata, or R. How does the addition of a fractional power for the second term in the negative binomial variance function add to the robustness of the NB-2 binomial?

5 The NB-H model allows parameterization of the negative binomial ancillary parameter. Why may predictors differ between two sets of displayed parameter estimates? How do the results relate to model overdispersion?

6 Construct a maximum likelihood double-Poisson model for estimating underdispersed count data. Use it on the data of Exercise 7.1. How does it differ from the identity-Gaussian and identity-gamma models? The double Poisson PDF can be formulated as:

$$f(y_i \mu, \phi) = k(\mu, \phi)\phi^{0,5} \exp(-\mu, \phi) \exp(-y) y^y ((e\mu)/y)^{y\phi},$$

where

$$k(\mu, \phi)^{-1} \approx 1 + (1 - \phi)/(12\mu\phi)^*(1 + 1/(\mu\phi).$$

(See Cameron and Trivedi, 1998, p. 115.)

7 Show how a Poisson regression model can duplicate the output of a log-hazard parameterization right-censored exponential regression model. How does modeling negative binomial instead of a Poisson affect the results?

8
Problems with zero counts

I have indicated that extended negative binomial models are generally developed to solve either a distributional or variance problem arising in the base NB-2 model. Changes to the negative binomial variance function were considered in the last chapter. In this chapter, we address the difficulties that arise when there are either no possible zeros in the data, or when there are an excessive number.

8.1 Zero-truncated negative binomial

Often we are asked to model count data that structurally exclude zero counts. Hospital length of stay data are an excellent example of count data that cannot have a zero count. When a patient first enters the hospital, the count begins. Upon registration the length of stay is given as 1. There can be no 0 days – unless we are describing patients who do not enter the hospital, and this is a different model where there may be two generating processes. This type of model will be discussed later.

The Poisson and negative binomial distributions both include zeros. When data structurally exclude zero counts, then the underlying probability distribution must preclude this outcome to properly model the data. This is not to say that Poisson and negative binomial models are not commonly used to model such data, the point is that they should not. The Poisson and negative binomial probability functions, and their respective log-likelihood functions, need to be amended to exclude zeros, and at the same time provide for all probabilities in the distribution to sum to one.

With respect to the Poisson distribution, the probability of a zero count, based on the PDF given in Equation (3.1), is $\exp(-\mu)$. This value needs to be subtracted from 1 and then the remaining probabilities rescaled on this difference. The resulting log-likelihood function, with $\mu = \exp(x\beta)$, is

$$\mathcal{L}(\mu; y_i | y_i > 0) = \sum_{i=1}^{i=1} \{y_i(x_i\beta) - \exp(x_i\beta) \\ - \ln\Gamma(y_i + 1) - \ln[1 - \exp(-\exp(x_i\beta))]\} \tag{8.1}$$

The logic of the zero-truncated negative binomial is the same. The probability of a zero count is

$$(1 + \alpha\mu_i)^{-1/a} \tag{8.2}$$

Subtracting from 1 and together conditioning out of the negative binomial log-likelihood by rescaling, we have

$$\mathcal{L}(\mu; y_i | y_i > 0) = \sum_{i=1}^{n} \{\mathrm{LL}_{\mathrm{NB}} - \ln[1 - \{1 + \exp(x_i\beta)\}^{-1/\alpha}]\} \tag{8.3}$$

where $\mathrm{LL}_{\mathrm{NB}}$ is the negative binomial log-likelihood as given in Equation (5.41).

It is clear that an IRLS type estimating algorithm is not appropriate for zero-truncated models. IRLS or GLM-type models are based on a likelihood derived from a probability function that is a member of the exponential family of distributions. GLM methodology allows the canonical link function to be changed, but this does not affect the underlying probability function. In the case of the zero-truncated models, the likelihood itself has changed. The amendment to the PDF and log-likelihood is not a simple reparameterization, but rather an altogether new model. As with all models based on distributional violations of the base Poisson and negative binomial, estimation is by maximum likelihood, whether this be carried out by a type of Newton–Raphson method, quadrature, simulation, or similar set of estimating equations.

The effect of truncation on a negative binomial model can be substantial. Much depends on the shape of the observed distribution of counts. If the mean count is high, then the theoretical probability of obtaining a zero count is less than if the mean count is low. The graphs given in Section 5.3, provide clear evidence of the relationship of the mean count to the probability of a zero count. Given an observed distribution having a low mean count, the difference between using a negative binomial model and a zero-truncated model will be substantial. In such a case a truncated model should be used with the data.

We will again generate synthetic data to demonstrate the difference between models. We first create a 50 000 observation negative binomial data set with pre-established parameters of

$$x_1 = 0.75 \quad x_2 = -1.25 \quad \text{constant} = 0.5 \quad \text{alpha} = .75$$

Table 8.1. *Synthetic data distribution: counts 0−10*

y1	Freq.	Percent	Cum.
0	19,204	38.41	38.41
1	9,110	18.22	56.63
2	4,986	9.97	66.60
3	3,300	6.60	73.20
4	2,352	4.70	77.90
5	1,671	3.34	81.25
6	1,296	2.59	83.84
7	1,038	2.08	85.91
8	752	1.50	87.42
9	678	1.36	88.77
10	564	1.13	89.90

Next the data are modeled using a negative binomial algorithm. To find the estimated value of α, we first model the data using a maximum likelihood algorithm; then use a GLM model with α entered into the IRLS estimating algorithm as a constant. The results are given as:

```
. nbreg y1 x1 x2

Generalized linear models     No. of obs      =      50000
Optimization    : ML          Residual df     =      49997
                              Scale parameter =          1
Deviance        = 49382.34156 (1/df) Deviance =   .9877061
Pearson         = 50215.19709 (1/df) Pearson  =   1.004364
Alpha           = .7417374    AIC             =   3.809767
Log likelihood = −95241.17223 BIC             = −491574.1

----------------------------------------------------------------------------
      y1 |     Coef.   OIM Std. Err.      z    P>|z|   [95% Conf. Interval]
---------+------------------------------------------------------------------
      x1 |  .7458934   .0055059        135.47  0.000     .735102   .7566849
      x2 | -1.250949   .0058937       -212.25  0.000   -1.262501  -1.239398
   _cons |   .504105   .0060375         83.50  0.000    .4922717   .5159382
----------------------------------------------------------------------------
```

A tabulation of the first 11 counts, from 0 to 10, is given in Table 8.1.

The mean count of y_1 is 4.8, quite low considering the large range of counts. We would expect that there will be a marked difference in estimates, α, and remaining goodness-of-fit statistics when the data are modeled without zeros.

Specifying α as 0.75 resulted in the excess negative binomial zeros, although it is possible to check the observed versus theoretical negative binomial distributions as we previously did to determine the extent to which the two distributions vary. Next we drop all zero counts from the data, then model using a maximum likelihood algorithm to obtain a revised value of α, which is submitted to a GLM program to obtain new parameter estimates and goodness-of fit statistics.

```
. glm y1 x1 x2, nolog fam(nb.4065592)

Generalized linear models          No. of obs      =      30796
Optimization     : ML              Residual df     =      30793
                                   Scale parameter =          1
Deviance         = 27520.87201     (1/df) Deviance =   .8937379
Pearson          = 31951.97063     (1/df) Pearson  =   1.037637
Variance           V(u) =          [Neg. Binomial]
function        :  u+(.4065592)u^2
Link function :g(u) = ln(u)        [Log]
                                   AIC             =   5.064485

Log likelihood = -77979.93742      BIC             = -290729.1

------------------------------------------------------------------------------
    y1 |  Coef.    OIM Std. Err.     z     P>|z|   [95% Conf.  Interval]
------------------------------------------------------------------------------
    x1 |  .5836536 .0048454        120.45  0.000    .5741567   .5931504
    x2 | -.9780867 .0051875       -188.55  0.000    -.988254  -.9679193
 _cons |  .9911068 .0060602        163.54  0.000    .9792291   1.002985
------------------------------------------------------------------------------
```

Note the changes that have occurred by deleting zero counts from the full negative binomial distribution having those parameters.

Modeling the data *sans* zero counts with a zero-truncated negative binomial results in a model appearing very much like the model when zero counts were included. Recall, the zero-truncated model is determining parameter estimates based on a data without zero counts.

```
. ztnb y1 x1 x2

Zero Truncated Negative         Number of obs    =      30796
Binomial Regression
                                Wald chi2(2)     =   31709.22
Log likelihood = -70641.048     Prob > chi2      =     0.0000

------------------------------------------------------------------------------
    y1 | Coef.     Std. Err.     z      P>|z|   [95% Conf.  Interval]
------------------------------------------------------------------------------
    x1 | .7481833  .0067904    110.18  0.000    .7348743   .7614922
    x2 | -1.25509  .0077214   -162.55  0.000   -1.270224  -1.239956
 _cons | .4996702  .0096698     51.67  0.000    .4807178   .5186227
-----+------------------------------------------------------------------------
alpha |  .742314  .0127209                      .7177955     .76767
-----+------------------------------------------------------------------------
AIC Statistic = 4.588
------------------------------------------------------------------------------
```

We shall next use the *medpar* data set which we have used before to model length of stay (*los*) data. Noting that the mean value of *los* is 9.9, we should expect that zero counts would not have as great an impact on the model as with the synthetic data with its mean count value of 4.8.

NEGATIVE BINOMIAL

```
. nbreg los white died type2 type3
Log likelihood = -4781.7268
===========================================================================
   los |    Coef.    Std. Err.     z    P>|z|   [95% Conf.  Interval]
-------+-------------------------------------------------------------------
 white | -.1258405  .0677911   -1.86  0.063   -.2587085   .0070276
  died | -.2359093  .0404752   -5.83  0.000   -.3152393  -.1565792
 type2 |  .2453806  .0501704    4.89  0.000    .1470485   .3437128
 type3 |  .7388372  .0750077    9.85  0.000    .5918248   .8858496
 _cons |  2.365268  .0679452   34.81  0.000    2.232097   2.498438
-------+-------------------------------------------------------------------
 alpha |   .434539    .01947                    .398006   .4744253
---------------------------------------------------------------------------
AIC Statistic = 6.404               BIC Statistic = -9324.893
Pearson = 1690.433634               (1/df) Pearson = 1.134519
===========================================================================
```

ZERO-TRUNCATED NEGATIVE BINOMIAL

```
. ztnb los white died type2 type3
Log likelihood = -4736.7725
===========================================================================
   los |    Coef.    Std. Err.     z    P>|z|   [95% Conf.  Interval]
-------+-------------------------------------------------------------------
 white |  -.131835  .0746933   -1.77  0.078   -.2782312   .0145612
  died | -.2511928  .0446812   -5.62  0.000   -.3387663  -.1636193
 type2 |  .2601118  .0552939    4.70  0.000    .1517378   .3684858
 type3 |  .7691718  .0825861    9.31  0.000     .607306   .9310376
 _cons |  2.333412  .0749931   31.11  0.000    2.186428   2.480395
-------+-------------------------------------------------------------------
 alpha |  .5315121  .0292239                   .4772126     .59199
---------------------------------------------------------------------------
AIC Statistic = 6.344               BIC Statistic = -9545.967
===========================================================================
```

As expected, the two models are similar. Moreover, the AIC and BIC statistics tend to favor the zero-truncated model, as they should in such a situation.

Zero-truncated models are subsets of the more general truncated count models we shall discuss in the next chapter. However, given the frequency of overdispersion in count data, as well as the frequency with which count models exclude the possibility of zero counts, zero-truncated negative binomial models are important to modeling counts and should therefore be part of the standard capabilities of commercial statistical packages.

8.2 Negative binomial with endogenous stratification

Negative binomial with endogenous stratification is a model that is perhaps most noted for its application in the area of recreation research (Shaw, 1988; Englin and Shonkwiler, 1995). The model simultaneously accommodates three features of on-site samples dealing with count data. The first accommodation is an overdispersion relative to the Poisson model; the second is truncation of zero counts; the third is endogenous stratification due to over-sampling.

Table 8.2. *Recreation model: predictors*

	income	Freq.	Percent	Cum.
income1	<=25000	53	13.91	13.91
income2	35000–55000	97	25.46	39.37
income3	65000–95000	87	22.83	62.20
income4	>95000	144	37.80	100.00
	Total	381	100.00	
	travel	Freq.	Percent	Cum.
travel1	<.25 hrs	95	25.89	25.89
travel2	.25-<4 hrs	142	38.69	64.58
travel3	>=4 hrs	130	35.42	100.00
	Total	367	100.00	
	gender	Freq.	Percent	Cum.
female	0	155	38.75	38.75
male	1	245	61.25	100.00
	Total	400	100.00	

Endogenous stratification occurs when the likelihood of sampling observations is dependent on the choice made by a subject of study which is in itself the response or dependent variable. For example, in the field of recreational demand analysis, one is more likely to interview subjects who visit a particular site more frequently than those who rarely visit it. This implies over-sampling of those who visit more frequently, and reports of more visits than are likely the case. This is termed endogeneity. Likewise, patients who visit a doctor more frequently are more likely to be sampled if the survey about number of visitations is conducted at the clinic.

If the data are in fact Poisson, and therefore equi-dispersed, but nevertheless are truncated and endogenously stratified, the model is equivalent to Poisson with the response subtracted by one, i.e. $y - 1$. The resulting re-scaled log-likelihood is expressed as

$$\begin{aligned}\mathcal{L}(x\beta; y_i) = {} & y_i \ln(\alpha) + (y_i - 1)\ln(\exp(x_i\beta)) \\ & - (y_i + 1/\alpha)\ln(1 + \alpha(\exp(x_i\beta))) \\ & - \ln\Gamma(y_i + 1) - \ln\Gamma(1/\alpha) + \ln\Gamma(y_i + 1/\alpha) + \ln(y_i);\ y_i > 0\end{aligned} \tag{8.4}$$

The following data *loomis* are taken from Loomis (2003). The study relates to a survey taken on reported frequency of visits to national parks during the year. The survey was taken at park sites, thus incurring possible effects of endogenous stratification. I shall model a subset of the data, with *anvisits*, annual count of reported visits to parks, as the response. Predictors include gender, distance traveled to the closest park, and annual income. Travel and income are factored as shown in Table 8.2.

Modeling the data with a negative binomial algorithm with endogenous stratification (Martinez-Espiñeira, Amoako-Tuffour, and Hilbe, 2006) produces the following output:

```
. nbstrat anvisits gender travel2 travel3 income2 income3
income4

Negative Binomial with Endogenous  Number of obs =    342
Stratification
                                   Wald chi2(6)  = 519.24
Log likelihood = -1232.0184        Prob > chi2   = 0.0000

==============================================================
anvisits |    Coef.   Std. Err.     z    P>|z|   [95% Conf. Interval]
--------------------------------------------------------------
  gender | -.6011335 .1272006    -4.73 0.000  -.8504421   -.3518249
 travel2 | -.5569893 .1453207    -3.83 0.000  -.8418127   -.2721658
 travel3 | -3.080732 .1580607   -19.49 0.000  -3.390525   -2.770939
 income2 |  .4045486 .1919525     2.11 0.035   .0283287    .7807686
 income3 | -.7505286 .1953772    -3.84 0.000  -1.133461   -.3675962
 income4 |  -.599445 .1827182    -3.28 0.001  -.9575661    -.241324
   _cons | -12.10614 124.4169    -0.10 0.922  -255.9588    231.7465
---------+----------------------------------------------------
/lnalpha |  16.60685 124.4169     0.13 0.894  -227.2457    260.4594
---------+----------------------------------------------------
   alpha |1.63e + 07 2.03e + 09                2.03e-99 1.3e + 113
---------+----------------------------------------------------
AIC Statistic = 7.252
==============================================================
```

Modeling the same data using a zero truncated negative binomial gives us:

```
. ztnb anvisits gender travel2 travel3 income2 income3
income4
Zero Truncated Negative Binomial  Number of obs  =    342
Regression

                                  Wald chi2(6)   =  264.39
Log likelihood = -1188.9244       Prob > chi2    =  0.0000

==============================================================
anvisits |   Coef.    Std. Err.    z     P>|z|   [95% Conf. Interval]
--------------------------------------------------------------
  gender |-.7376444   .208367    -3.54  0.000   -1.146036  -.3292526
 travel2 |-.5795156  .2498082    -2.32  0.020   -1.069131  -.0899005
 travel3 |-3.646618  .2709248   -13.46  0.000   -4.177621  -3.115615
 income2 | .7553192  .3404561     2.22  0.027    .0880375   1.422601
 income3 |-.8828776  .3153164    -2.80  0.005   -1.500886  -.2648688
 income4 | -.510249  .3077044    -1.66  0.097   -1.113339   .0928406
   _cons | 4.233715  .3241964    13.06  0.000    3.598301   4.869128
---------+----------------------------------------------------
/lnalpha | 1.339343  .2412675     5.55  0.000    .8664672   1.812218
---------+----------------------------------------------------
   alpha | 3.816534  .9208055                    2.378493   6.124017
--------------------------------------------------------------
AIC Statistic = 6.994             BIC Statistic = -1749.631
==============================================================
```

The AIC statistic of two models indicate that the zero-truncated model without adjustment for endogenous stratification is preferred. Moreover, the value of α for the latter model is far too high. This indicates that the model does not fit well with the data. The first 11 counts are:

```
anvisits                   Freq.              Percent              Cum.
 1                         133                32.44                32.44
 2                          30                 7.32                39.76
 3                          15                 3.66                43.41
 4                           8                 1.95                45.37
 5                           5                 1.22                46.59
 6                          15                 3.66                50.24
 7                           3                 0.73                50.98
 8                           3                 0.73                51.71
 9                           2                 0.49                52.20
10                          10                 2.44                54.63
```

The extreme number of 1s in the data is the likely cause of the model not fitting.

The negative binomial with endogenous stratification model has seen relatively little use and is apparently fragile in the presence of ill-structured data. However, if the data are appropriate for the model, it performs better than its strictly zero-truncated counterpart.

8.3 Hurdle models

We have emphasized the necessity of creating models that remedy variance and distributional problems with the data that can plague Poisson and negative binomial NB-2 models. Zero-truncated models attempt to accommodate the data when zeros are structurally excluded from the model. But what happens when there are far more zero counts in the data than are allowed on the basis of negative binomial distributional assumptions? This topic, with graphs, was discussed earlier. But we did not attempt to address ways to deal with the situation.

Hurdle and zero-inflated count models are the two foremost methods used to deal with count data having excessive zero counts. Neither of these models is well supported in current commercial software; however LIMDEP supports both and Stata incorporates zero-inflated models. User designed hurdle models have been created in Stata by the author and have been published at the Boston School of Economics statistical software repository (see: http://ideas.repec.org/s/boc/bocode.html). I discuss hurdle models first because of their greater intuitional sense, and ease of interpretation. Note that hurdle models are sometimes referred to as zero altered models (Heilbron, 1989). Zero altered Poisson and negative binomial models are thus referred to, respectively, as ZAP and ZANB.

The essential idea of a hurdle model is to partition the model into two parts – first, a binary process generating positive counts (1) versus zero counts (0); second, a process generating positive only counts. The binary process is modeled using a binary model, the positive count process is modeled using a zero-truncated count model. There have been nine commonly used hurdle models: the binary part modeled by logit, probit, or complememtary loglog, and the count part modeled using Poisson, geometric, or negative binomial. The first hurdle models were designed by Mullahy (1986) and were later popularized by Cameron and Trivedi (1986, 1998). Mullahy used logit and cloglog binary models with Poisson and geometric count models.

Again, the notion of hurdle comes from considering the data as being generated by a process that commences generating positive counts only after crossing a zero barrier, or hurdle. Until the hurdle is crossed, the process generates a binary response (1/0). The nature of the hurdle is left unspecified, but may simply be considered as the data having a positive count. In this sense, the hurdle is crossed if a count is greater than zero. In any case, the two processes are conjoined using the following log-likelihood

$$\mathrm{LL} = \ln(f(0)) + \{\ln[1 - f(0)] + \ln P(t)\}$$

where $f(0)$ represents the probability of the binary part of the model and $P(j)$ represents the probability of a positive count. For ease of interpretation I shall forgo the use of subscripts for the remainder of this section.

In the case of a logit model, the probability of zero is

$$f(0) = P(y = 0; x) = 1/(1 + \exp(x\beta_b))$$

and $1-f(0)$ is

$$\exp(x\beta_b)/(1 + \exp(x\beta_b))$$

The zero-truncated negative binomial loglikelihood is

$$\begin{aligned} P(y|x > 0) = {} & y^{*}\ln(\exp(x\beta)/(1 + \exp(x\beta))) - \ln(1 + \exp(x\beta))/\alpha \\ & + \ln\Gamma(y + 1/\alpha) - \ln\Gamma(y + 1) - \ln\Gamma(1/\alpha) \\ & - \ln(1 - (1 + \exp(\mathrm{x}\beta))^{\wedge}(-1/\alpha)) \end{aligned}$$

Putting the above together, we have the two-part NB-logit hurdle model likelihood.

$$\text{If } y == 0 = 1/(1+\exp(x\beta_b))$$

$$\begin{aligned}\text{If } y > 0 = {} & \exp(x\beta_b)/(1+\exp(x\beta_b)) + y^{*}\ln(\exp(x\beta)/(1+\exp(x\beta))) \\ & - \ln(1+\exp(xb))/\alpha + \ln\Gamma(y+1/\alpha) - \ln\Gamma(y+1) \\ & - \ln\Gamma(1/\alpha) - \ln(1-(1+\exp(xb))\hat{}(-1/\alpha))\end{aligned}$$

For the negative binomial – complementary loglog hurdle model,

$$\text{If } y == 0 = -\exp(x\beta_b)$$

$$\begin{aligned}\text{If y} > 0 = {} & \ln(1-\exp(-\exp(x\beta_b))) + y^{*}\ln(\exp(x\beta)/(1+\exp(x\beta))) \\ & - \ln(1+\exp(x\beta))/\alpha + \ln\Gamma(y+1/\alpha) - \ln\Gamma(y+1) \\ & - \ln\Gamma(1/\alpha) - \ln(1-(1+\exp(x\beta))\hat{}(-1/\alpha))\}\end{aligned}$$

Using the German health reform data, *mdvisitsx*, we model *numvisit* using a negative binomial – complementary loglog model.

NEGATIVE BINOMIAL – COMPLEMEMTARY LOGLOG
HURDLE MODEL

```
. hnbclg numvisit reform badh educ2 age3

Log Likelihood = -1331.9847
```

```
===================================================================
           |   Coef.    Std. Err.    z    P>|z|  [95% Conf. Interval]
-------------------------------------------------------------------
cloglog    |
    reform | -.1073391  .0543833   -1.97  0.048   -.2139284  -.0007497
      badh |  .5771608  .0853904    6.76  0.000    .4097988   .7445228
     educ3 |  .1123926  .0573353    1.96  0.050    .0000175   .2247678
      age3 |  .0180807  .0742734    0.24  0.808   -.1274926    .163654
     _cons |  .1445252  .0454318    3.18  0.001    .0554804   .2335699
-------------------------------------------------------------------
negbinomial|
    reform | -.1182993  .0639946   -1.85  0.065   -.2437265   .0071278
      badh |  1.159176  .0862206   13.44  0.000    .9901863   1.328165
     educ3 | -.1960328  .0682512   -2.87  0.004   -.3298028  -.0622629
      age3 |  .2372101   .084559    2.81  0.005    .0714776   .4029426
     _cons |  .7395257  .0671674   11.01  0.000    .6078801   .8711714
     alpha | 1.1753772
-------------------------------------------------------------------
AIC Statistic = 4.096
===================================================================
```

The model output provides parameter estimates for both the binary complementary loglog model and the negative binomial. The joint model can be separated into partitioned models by first creating a second response variable, which we shall call *visit*, by the following logic

visit = 1 if numvisit >0, i.e. is a positive count
visit = 0 if numvisit==0.

```
. gen int visit = numvisit>0
. tab visit
```

```
visit |        Freq.          Percent          Cum.
------+--------------------------------------------
    0 |        665            29.86            29.86
    1 |        1,562          70.14            100.00
------+--------------------------------------------
Total |        2,227          100.00
```

The binary part, a complementary loglog regression, is modeled as:

COMPLEMENTARY LOGLOG MODEL

```
. cloglog numvisit reform badh educ2 age3

Complementary log-log             Number of obs      =     2227
regression                        Zero outcomes      =      665
                                  Nonzero outcomes   =     1562
                                  LR chi2(4)         =    51.55
Log likelihood = -1331.9847       Prob > chi2        =   0.0000
------------------------------------------------------------------
 visit|   Coef.    Std. Err.    z    P>|z|    [95% Conf. Interval]
------+-----------------------------------------------------------
reform| -.1073391 .0543833   -1.97  0.048   -.2139284   -.0007497
  badh|  .5771608 .0853904    6.76  0.000    .4097988    .7445228
 educ3|  .1123926 .0573353    1.96  0.050    .0000175    .2247678
  age3|  .0180807 .0742734    0.24  0.808   -.1274926     .163654
 _cons|  .1445252 .0454318    3.18  0.001    .0554805    .2335699
------------------------------------------------------------------
AIC Statistic = 1.202
```

The parameter estimates are identical to that of the hurdle model.

We next model the data using a zero-truncated negative binomial, making certain to exclude zero counts from the modeling process.

ZERO TRUNCATED NEGATIVE BINOMIAL MODEL

```
. ztnb numvisit reform badh educ3 age3 if numvisit>0,
Zero-truncated negative binomial  Number of obs  =    1562
regression
                                  LR chi2(4)     =  233.36
Dispersion = mean                 Prob > chi2    =  0.0000
Log likelihood = -3223.6195       Pseudo R2      =  0.0349
------------------------------------------------------------------
numvisit|   Coef.    Std. Err.    z    P>|z|    [95% Conf. Interval]
--------+---------------------------------------------------------
  reform| -.1182993 .0639946  -1.85  0.065    -.2437265    .0071278
    badh|  1.159176 .0862206  13.44  0.000     .9901863    1.328165
   educ3| -.1960328 .0682512  -2.87  0.004    -.3298028   -.0622629
    age3|  .2372101  .084559   2.81  0.005     .0714776    .4029426
   _cons|  .7395257 .0671674  11.01  0.000     .6078801    .8711714
--------+---------------------------------------------------------
/lnalpha|  .1615891 .1106842         -.055348   .3785262
--------+---------------------------------------------------------
   alpha|  1.175377 .1300957                    .9461558    1.460131
------------------------------------------------------------------
Likelihood-ratio test of alpha=0: chibar2(01) = 1677.79 Prob>
=chibar2 = 0.000
------------------------------------------------------------------
AIC Statistic = 4.134             BIC Statistic = -10615.008
```

Again, the parameter estimates are identical to the hurdle model. Notice, however, that the AIC statistic is lower for the conjoined hurdle model than for the zero-truncated model.

The same relationship maintains for the NB-logit model. The logit model appears as:

LOGISTIC MODEL

```
. logit visit reform badh educ3 age3, nolog
Logistic regression               Number of obs    =      2227
                                  LR chi2(4)       =     51.96
                                  Prob > chi2      =    0.0000
Log likelihood = -1331.7768       Pseudo R2        =    0.0191
==========================================================
 visit|   Coef.    Std. Err.    z    P>|z|    [95% Conf. Interval]
----------------------------------------------------------
reform| -.1879245 .0939389  -2.00 0.045   -.3720413  -.0038076
  badh|  1.144087 .1940181   5.90 0.000    .7638189   1.524356
 educ3|  .2018225 .1003517   2.01 0.044    .0051367   .3985082
  age3|  .0238393 .1301832   0.18 0.855   -.2313152   .2789937
 _cons|  .7795456 .078196    9.97 0.000    .6262844   .9328069
----------------------------------------------------------
AIC Statistic = 1.201
==========================================================
```

The zero-truncated negative binomial is the same as before. Submitting the data of a negative binomial-logit model produces the following output.

NEGATIVE BINOMIAL – LOGIT

```
. hnblogit numvisit reform
badh educ3 age3, nolog

Negative Binomial-Logit Hurdle   Number of obs    =     2227
Regression
                                 Wald chi2(4)     =    42.65
Log likelihood = -4555.3963      Prob > chi2      =   0.0000
===========================================================
          |    Coef.   Std. Err.    z    P>|z|    [95% Conf. Interval]
-----------------------------------------------------------
logit     |
    reform|-.1879245 .0939389  -2.00 0.045   -.3720413  -.0038076
      badh| 1.144088 .1940181   5.90 0.000    .7638189   1.524356
     educ3| .2018225 .1003517   2.01 0.044    .0051367   .3985082
      age3| .0238393 .1301832   0.18 0.855   -.2313152   .2789937
     _cons| .7795456  .078196   9.97 0.000    .6262844   .9328069
-----------------------------------------------------------
negbinomial
    reform|-.1182993 .0639946  -1.85 0.065   -.2437264   .0071278
      badh| 1.159175 .0862205  13.44 0.000    .9901862   1.328164
     educ3|-.1960328 .0682512  -2.87 0.004   -.3298027  -.0622629
      age3|    23721 .0845589   2.81 0.005    .0714776   .4029424
     _cons| .7395271 .0671672  11.01 0.000    .6078818   .8711725
-----------------------------------------------------------
  /lnalpha| .1615866  .110684   1.46 0.144   -.0553499   .3785232
-----------------------------------------------------------
     alpha| 1.175374 .1300951                   .946154   1.460127
-----------------------------------------------------------
AIC Statistic = 4.096
===========================================================
```

The logit models are the same. All hurdle models can be partitioned or broken apart in this fashion.

It is interesting to note that the AIC statistic for the hurdle model can be calculated from a knowledge of the AIC statistics of both constituent models and the percentage of response values greater than zero. Although the hurdle model algorithm does not use this approach to calculate an AIC statistic, it can be calculated by hand as

$$\text{AIC\_hurdle} = ((\text{AIC\_zero.trunc.count}^{*} N(>0)/N)) + \text{AIC\_binary}$$

Calculating the AIC statistic for the hurdle model based on the above formula provides:

```
. di 4.13* (1−665/2227) + 1.20
4.096749
```

which is nearly the same as the AIC observed for the negative binomial–logit model. Rounding errors will at times produce minor discrepancies between the hand-calculated value and the model-calculated value. Regardless, we see that the AIC statistic for the hurdle model is proportioned between both constituent model AIC statistics, with the count model adjusted by the percentage of non-zero counts in the response.

Interpretation is now considered. Each predictor is evaluated in terms of the contribution it makes to each respective model. For example, with respect to the NB-logit model, a positive significant coefficient in the negative binomial frame indicates that the predictor increases the rate of physician visits, in the same manner as any negative binomial model is interpreted. A positive coefficient in the logit frame is interpreted in such a manner that a one-unit change in a coefficient decreases the odds of no visits to the doctor by exp(β). If a logistic coefficient is 0.2018, then the odds of no visits is decreased by exp(.2018) = 1.2236033, or about 22%.

Parameterizing the estimates in exponential form so that the counts can be interpreted as incidence rate ratios, and the logistic model as odds ratios, we have:

```
. hnblogit numvisit reform badh educ3 age3, nolog eform

Negative Binomial-Logit Hurdle  Number of obs    =     2227
Regression
                                Wald chi2(4)     =    42.65
Log likelihood = -4555.3963     Prob > chi2      =   0.0000
```

```
-----------------------------------------------------------------------
            |  exp(b)  Std. Err.    z    P>|z|  [95% Conf.  Interval]
------------+----------------------------------------------------------
logit       |
     reform | .8286773   .077845 -2.00 0.045    .6893258    .9961997
       badh | 3.139575  .6091346  5.90 0.000    2.146458    4.592186
      educ3 | 1.223631  .1227934  2.01 0.044     1.00515    1.489601
       age3 | 1.024126   .133324  0.18 0.855    .7934893    1.321799
------------+----------------------------------------------------------
negbinomial |
     reform | .8884301  .0568547 -1.85 0.065     .783702    1.007153
       badh | 3.187304   .274811 13.44 0.000    2.691735    3.774109
      educ3 | .8219853  .0561015 -2.87 0.004    .7190656    .9396358
       age3 | 1.267707  .1071959  2.81 0.005    1.074094    1.496221
------------+----------------------------------------------------------
   /lnalpha | .1615866   .110684  1.46 0.144   -.0553499    .3785232
------------+----------------------------------------------------------
      alpha | 1.175374  .1300951                  .946154    1.460127
------------+----------------------------------------------------------
AIC Statistic = 4.096
-----------------------------------------------------------------------
```

Predicted values, μ, may also be calculated for the count model.

```
. hnblogit_p mu, eq(#2) irr

. l mu numvisit reform badh educ3 age3 in 1/5

      mu          numvisit    reform    badh    educ3    age3
1.    2.359472    30          1         0       0        1
2.    2.094944    25          0         0       0        0
3.    2.183009    25          0         0       1        1
4.    2.094944    25          0         0       0        0
5.    2.359472    20          1         0       0        1
```

To check the first fitted value

```
. di exp(−.1182993 +.23721 +.7395271)
2.3594718
```

which is consistent with the value shown on line one of the above table. Note that the hurdle models used in this book are not part of an official commercial package, but rather were written by the author using Stata's higher programming language.

Checking output displayed by software is recommended. Reviewing software for many years has made me a bit hesitant about simply accepting all output, especially with procedures that are seldom used. The more a procedure is used by statisticians, the more likely mistakes are identified. No commercial software is immune from error.

8.4 Zero-inflated count models

Zero-inflated count models were first introduced by Lambert (1992) to provide another method of accounting for excessive zero counts. Like hurdle models, they are two-part models, consisting of both binary and count model sections.

Unlike hurdle models, though, zero-inflated models provide for the modeling of zero counts using both binary and count processes. The hurdle model separates the modeling of zeros from the modeling of counts, entailing that only one process generates zeros. This mixture of modeling zeros is reflected in the log-likelihood function.

Commercial software implementations typically allow the zero-inflated binary process to be either probit or logit. Counts, including zeros, are estimated using either a Poisson or negative binomial regression. The log-likelihood functions of the NB-logit and NB-probit models are listed below, without subscripts.

ZERO-INFLATED NEGATIVE BINOMIAL-LOGIT

$$\begin{aligned}
&\text{If } y == 0: \Sigma\{\ln(1/(1+\exp(-x\beta_1)) + 1/(1+\exp(x\beta_1)))^* \\
&\qquad \times (1/(1+\alpha^*\exp(x\beta)))^\wedge(1/\alpha)\} \\
&\text{If } y > 0: \Sigma\{\ln(1/(1+\exp(x\beta_1))) + \ln\Gamma(1/\alpha + y) - \ln\Gamma(y+1) \\
&\qquad - \ln\Gamma(1/\alpha) + (1/\alpha)^*\ln(1/(1+\alpha^*\exp(x\beta))) \\
&\qquad + y^*\ln(1 - (1/(1+\alpha^*\exp(x\beta))))\}
\end{aligned}$$

ZERO-INFLATED NEGATIVE BINOMIAL-PROBIT

$$\begin{aligned}
&\text{If } y == 0: \Sigma\{\ln(\Phi x\beta_1) + (1 - \Phi(x\beta_1))^*(1/(1+\alpha^*\exp(x\beta)))^\wedge(1/\alpha)\} \\
&\text{If } y > 0: \Sigma\{\ln(1 - \Phi(x\beta_1)) + \ln\Gamma(1/\alpha + y) - \ln\Gamma(y+1) \\
&\qquad - \ln\Gamma(1/\alpha) + (1/\alpha)^*\ln(1/(1+\alpha^*\exp(x\beta))) \\
&\qquad + y^*\ln(1 - (1/(1+\alpha^*\exp(x\beta))))\}
\end{aligned}$$

where $\exp(x\beta_1)$ is the fit, or μ, from the binary process, and $\exp(x\beta)$ is the same with respect to the count process. Φ represents the normal or Gaussian cumulative distribution function.

Inflation refers to the binary process. Unlike hurdle models, the binary process may include different predictors than in the count process. The important point is for the statistician to use the model to determine which variables or items in the data have a direct bearing on zero counts. This is why the zero-inflated model has its count process, unlike hurdle models, predict zeros. Note in the first equation of the zero-inflated NB-logit model, the three terms predicting zero counts are: (1) logistic inverse link, i.e. μ, the prediction that y==0, (2) $1-\mu$, and (3) the negative binomial prediction of a zero count. If the latter formula is unfamiliar, recall that the formula has been expressed in a variety of ways, e.g. $\{\alpha^{-1}/(\alpha^{-1}+\mu)\}\alpha^{-1}$.

I shall again use the German health reform data used in the previous section. Modeled as a NB-logit, we have:

```
. zinb numvisit reform badh age3, nolog inflate(badh age3
loginc)

Zero-inflated negative binomial  Number of obs     =    2227
regression
                                 Nonzero obs       =    1562
                                 Zero obs          =     665
Inflation model = logit          LR chi2(3)        =  297.48
Log likelihood = -4562.087       Prob > chi2       =  0.0000
---------------------------------------------------------------
         |    Coef.    Std. Err.    z    P>|z|  [95% Conf.   Interval]
---------+-----------------------------------------------------
numvisit |
  reform |-.1397194  .0509428  -2.74  0.006   -.2395655   -.0398733
    badh | 1.127971  .0733397  15.38  0.000    .9842281    1.271715
    age3 | .2243277  .0725813   3.09  0.002    .0820709    .3665845
   _cons | .7742719  .0397866  19.46  0.000    .6962916    .8522523
---------+-----------------------------------------------------
inflate  |
    badh |-11.54223  775.4551  -0.01  0.988   -1531.406    1508.322
    age3 | 2.443905  1.372909   1.78  0.075   -.2469473    5.134757
  loginc |-3.029019  1.255131  -2.41  0.016   -5.489031   -.5690067
   _cons | 17.87723  8.419252   2.12  0.034    1.375797    34.37866
---------+-----------------------------------------------------
/lnalpha |-.0287578   .052818  -0.54  0.586   -.1322792    .0747637
---------+-----------------------------------------------------
   alpha | .9716518  .0513207                  .8760964    1.077629
---------+-----------------------------------------------------
AIC Statistic = 4.1034
```

Post-reform, bad health, and age from 50–60 are all predictors of positive counts; e.g. the number of visits to the physician. Therefore, patients made [exp(−.1397194) = .86960221] about 13% fewer visits to the doctor following reform, which was a marked goal of reform legislation. The (log) income of patients had an inverse relationship with not visiting their doctor. That is, the greater the patient income, the more likely they were to visit a physician.

```
zinb (N=2227): Factor Change in Expected Count
Observed SD: 4.0161991

Count Equation: Factor Change in Expected Count for Those Not Always 0
----------------------------------------------------------------------
numvisit |      b         z      P>|z|     e^b     e^bStdX    SDofX
----------------------------------------------------------------------
  reform |  -0.13972   -2.743   0.006   0.8696   0.9325    0.5001
    badh |   1.12797   15.380   0.000   3.0894   1.4305    0.3174
    age3 |   0.22433    3.091   0.002   1.2515   1.0861    0.3682
---------+------------------------------------------------------------
ln alpha |  -0.02876
---------+------------------------------------------------------------
   alpha |   0.97165   SE(alpha) = 0.05132
----------------------------------------------------------------------
b = raw coefficient
z = z-score for test of b=0
P>|z| = p-value for z-test
e^b = exp(b) = factor change in expected count for unit increase in X
e^bStdX = exp(b*SD of X) = change in expected count for SD increase in X
SDofX = standard deviation of X
Binary Equation: Factor Change in Odds of Always 0
```

```
=========================================================================
Always0 |       b         z      P>|z|       e^b    e^bStdX     SDofX
--------+----------------------------------------------------------------
   badh | -11.54223   -0.015    0.988     0.0000    0.0256     0.3174
   age3 |   2.44390    1.780    0.075    11.5179    2.4593     0.3682
 loginc |  -3.02902   -2.413    0.016     0.0484    0.3178     0.3785
-------------------------------------------------------------------------
b = raw coefficient
z = z-score for test of b=0
P>|z| = p-value for z-test
e^b = exp(b) = factor change in odds for unit increase in X
e^bStdX = exp(b*SD of X) = change in odds for SD increase in X
SDofX = standard deviation of X
=========================================================================
```

During the post-reform period (1998), there is a decrease in the expected rate of visits to the doctor by a factor of .87, holding all other predictors constant. Patients having the opportunity to visit their doctor and who are in bad health increased visits to their doctor some three fold. The binary equation frame describes the change in odds for always having zero visitations versus not always having zero visitations. As such, we can interpret the output as showing that a higher (log)income decreases by some 6% the odds of not having the opportunity of visiting a doctor. An excellent discussion of the logic of this interpretation and related probabilities can be found in Long and Freese (2006).

Those using zero-inflated models on their data must take special care to correctly interpret the model. It is somewhat trickier than interpreting the hurdle model, which is fairly straightforward. A source of mistakes relates to the fact that the model predicts zero counts in two quite different ways. First, zero counts are predicted as usual on the basis of a standard Poisson or negative binomial model. Secondly, within the framework of the binary process, a prediction of success is a prediction that the response has a zero count. A response of zero indicates a positive count – or rather, a non-zero-count. Unfolding the relationships can be a source of confusion.

Zero inflated models may be tested to determine if they are statistically different from their base model; that is, we may evaluate the model displayed above using a Vuong test to determine if the data are negative binomial, or if the excess zeros come from a different generating process. The test is valid for both zero-inflated Poisson and zero-inflated negative binomial. The test has been applied to generalized Poisson models as well.

The Vuong test, developed by Greene (1994), is generally formulated as $V = (\text{sqrt}(N)^{*}\text{mean}(m))/Sm.\, m$ symbolizes the result of $\ln(\mu_1/\mu_2)$ where μ_1 is the predicted probability of y for the zero-inflated model and μ_2 is the predicted probability of y for the base model, e.g. negative binomial. Sm is the standard deviation of m. N is the number of observations in each model. It is important that both models handle the same observations.

The test statistic V is asymptotically normal. Referring to the zero-inflated model as the first model, which produced the predicted probabilities, μ_1, and the base model as the second, if $V > 1.96$, the zero-inflated model is preferred. If $V < -1.96$, the base model is preferred. Values of V between –1.96 and 1.96 indicate that neither model is preferred; that is, the excessive zeros for the model response are not sufficient to warrant adjustment by a zero-inflated model. The data are likely to be either Poisson or negative binomial, depending on models involved in the test. In the case of our example zero-inflated model, V is 1.07 with a p-value of 0.1425. There is a moderate, but insignificant, preference for the zero-inflated model over the base NB-2 negative binomial (see Long and Freese, 2006 for a additional discussion of the Vuong statistic).

8.5 Summary

We discussed two data situations which we know give rise to overdispersion in both Poisson and negative binomial (NB-2) models. The first relates to when there is no possibility of a zero count in the data. Hospital length of stay is a good example of this type of data. Zero-truncated models adjust the probability functions of the Poisson and NB-2 models so that respective zero counts are excluded, but the sum of probabilities is still one.

The data situation of having an excess of zero counts is probably more frequent than data having no possibility of zeros. Two models have been developed to adjust for excessive zeros, both based on different reasoning. Each differs in how it accounts for the origin or generation of the extra zeros. This accounting is then reflected in the estimating algorithm.

We next turn to a discussion of models involving censoring and truncation.

Exercises

1 Zero-inflated and hurdle models typically assume a non-distributional value for the number of zero counts in the response. Does having the number of zeros in the response being under-represented or over-represented make a difference when selecting a zero-inflated rather than a hurdle model?

2 Using the following data *edsurvey* from an educational survey (amended), model *passed* on *suburbs* and *minority*. Check if the number of zero counts differs substantially from the Poisson assumption. If so, then adjust using a zero-inflated or hurdle model. Attempt to construct a well-fitted model.

suburbs	minority	passed	suburbs	minority	passed
0	0	4	0	0	5
0	1	9	1	0	0
1	1	3	0	0	9
0	0	9	1	0	0
1	0	0	0	1	5
1	0	4	0	0	5
1	0	10	0	0	0
1	0	3	0	1	10
0	0	5	1	0	11
0	0	6	0	0	10
0	0	4	0	0	4
0	1	9	0	0	0
0	0	11	1	0	3
0	0	29	0	1	4
0	0	5	0	0	15
0	0	11	0	0	6
1	0	0	0	1	9
0	0	13	0	1	9
0	1	4	0	1	3
0	0	3	0	1	5
1	0	0	0	0	3
0	0	10	0	0	16
0	0	21	0	0	4
0	0	5	1	0	0
0	0	10	1	0	15

3 List various example situations in the fields of economics, education, health analysis, and physical sciences where the negative binomial with endogenous stratification model is appropriate.

4 Using the *azprocedure* data found on the text web site, model *los* on *procedure*, *admit*, *sex*, and *age 75*. There are 3589 observations in the model. Compare results of a NB-2 model with a zero-truncated model. Then take a random sample of 100 cases and compare the two models again. Repeat this procedure five times. Are there significant differences between the standard NB-2 model and the truncated model when using the entire data compared with the 100 observation random sample? Discuss the reasons for the results you discover.

5 Derive the zero-truncated NB-1 log-likelihood function. Do the same for the zero-inflated NB-1 log-likelihood function.

9

Negative binomial with censoring, truncation, and sample selection

There are many times when certain data elements are lost, discarded, ignored, or are otherwise excluded from analysis. Truncated and censored models have been developed to deal with these types of data. Both models take two forms, truncation or censoring from below, and truncation or censoring from above. Count model forms take their basic logic from truncated and censored continuous response data, in particular from Tobit (Amemiya, 1984) and censored normal regression (Goldberger, 1983) respectively.

Count sample selection models also deal with data situations in which the distribution is confounded by an external condition. We shall address sample selection models at the end of the chapter.

The traditional parameterization used for truncated and censored count data can be called the econometric parameterization. This is the form of model discussed in standard econometric texts and is the form found in current econometric software implementations. I distinguish this from what I term a survival parameterization, the form of which is derived from standard survival models. This parameterization only relates to censored Poisson and censored negative binomial models. I shall first address the more traditional econometric parameterization. In addition, I shall not use subscripts for this chapter; they are understood as presented in the earlier chapters.

9.1 Censored and truncated models – econometric parameterization

Censored and truncated count models are related, with only a relatively minor algorithmic difference between the two. The essential difference relates to how response values beyond a user-defined cut point are handled. Truncated models eliminate the values altogether; censored models revalue them to the value of the cut point. In both cases the probability function and log-likelihood functions must be adjusted to account for the change in distribution of the response. We begin by considering truncation.

9.1.1 Truncation

In order to understand the logic of truncation, we begin with the basic Poisson probability distribution function, defined as

$$\text{Prob}(Y = y) = e^{-\mu}\mu^{y}/y! \quad y = 0, 1, \ldots \tag{9.1}$$

Recall that when we discussed zero-truncated Poisson in the last chapter, we adjusted the basic Poisson PDF to account for the structural absence of zeros. Given the probability of a zero count as $e^{-\mu}$, or $\exp(-\mu)$, it is subtracted from one to obtain the probability of a non-zero positive count. The Poisson probability distribution function is then rescaled by the resultant value, $1-\exp(-\mu)$, to obtain the zero-truncated Poisson PDF. The same logic maintains for zero-truncated negative binomial regression. The probability of a negative binomial count of zero is $(1 - \alpha\mu)^{-1/\alpha}$. Subtracting this value from 1 gives the negative binomial formula of a non-zero positive count. The negative binomial PDF is then rescaled by $1 - (1 - \alpha\mu)^{-1/\alpha}$ to obtain the zero-truncated negative binomial.

In the more general case, zero-truncated count models can be considered as left- or lower-truncated count models. The lower cut point, C, is at 1. If we wish to extend C to any lower value in the observed distribution, the value to be divided from the basic PDF must reflect the total probability of counts up to the cut. The smallest response value in the observed distribution is $C + 1$. For example, if C is specified as 1, then both the probability of zero counts and counts of 1 need to be calculated, summed, subtracted from 1, and used to rescale the resulting basic count PDF. In the case of Poisson

$$\text{Prob}(Y = (y = 0)) = e^{-\mu}, \text{ and} \tag{9.2}$$

$$\text{Prob}(Y = (y = 1)) = \mu e^{-\mu} \tag{9.3}$$

These values are then summed and subtracted from 1.

$$\text{Prob}(Y = (y = 0, 1)) = 1 - (e^{-\mu} + \mu e^{-\mu}), \tag{9.4}$$

This value is then divided from the Poisson PDF to obtain a one-truncated Poisson, or more accurately, a left-truncated at 1 Poisson PDF. Remaining values in the distribution have a minimum at $C + 1$, or, in this case, $1 + 1 = 2$.

The same logic applies with respect to repeatedly greater left values, depending on the value of the defined cut point. A left-truncated cut at $C = 3$ specifies that the response values in the model have values starting at $C + 1$, or 4. This does not mean that the distribution must have a value of 4, only that no values have non-zero probability for the distribution less than 4. The left-truncated negative binomial follows the same reasoning.

We can formalize the left-truncated Poisson PDF as

$$\text{Prob}(Y = y|Y > C) = \frac{\exp(-\mu)\mu^{y}/y!}{\text{Prob}(y > C)} = \frac{\exp(-\mu)\mu^{y}/y!}{1 - \sum_{j=0}^{C} \exp(-\mu)\mu^{j}/j!},$$
$$\text{for } y = C + 1, C + 2, \ldots \tag{9.5}$$

where C is the user defined cut point and j is the running index in the summations. See Greene (2006) for details of derivation as well as formulae for gradients and marginal effects.

A little algebraic manipulation allows formulation of the left-truncated Poisson log-likelihood function as

$$\mathcal{L}(\beta; y) = \sum \Big[y \ln(\mu) - \mu - \ln\Gamma(y + 1)$$
$$- \Big(1 - \sum_{j=0}^{C} j^{*} \ln(-\mu) - \mu - \ln\Gamma(j + 1)\Big)\Big] \tag{9.6}$$

An example may help show the differences in parameter estimates and associated model statistics for data in which the left-side numbers have been dropped up to a specified point and a left-truncated model with a cut defined at the same point. The first set of models will come from the *mdvisitsx* data. The left truncation is defined with a cut of 3, meaning that the sample response, *numvisit*, starts with the count value of 4. A tabulation of counts is provided in Table 9.1.

POISSON: ALL DATA

```
Variable Coefficient Standard Error b/St.Er. P[}Z}>z] Mean of X
Constant   .83196836 .02194310       37.915 .0000
REFORM    −.12911602 .02649991       −4.872 .0000    .50606197
BADH      1.15700714 .02904310       39.838 .0000    .11360575
EDUC3     −.12997795 .02892196       −4.494 .0000    .33767400
```

POISSON: DROPPED VALUES 1–3

```
Variable Coefficient Standard Error b/St.Er. P[}Z}>z] Mean of X
Constant  2.00581075 .02771067       72.384 .0000
REFORM    −.05894525 .03277276       −1.799 .0721    .45154639
BADH       .39973779 .03310038       12.077 .0000    .31134021
EDUC3     −.21994661 .03684811       −5.969 .0000    .31340206
```

Table 9.1. *MDVISITSX data, truncated at 3*

numvisit	Freq.
0	665
1	447
2	374
3	256
CUT	
4	117
5	101
6	76
7	21
8	27
9	9
10	61
11	1
12	20
13	5
14	3
15	19
16	2
20	10
24	1
25	3
30	4
36	1
40	2
50	1
60	1
Total	485

LEFT-TRUNCATED POISSON, CUT = 3

Variable	Coefficient	Standard Error	b/St.Er.	P[|Z|>z]	Mean of X
Constant	1.93809303	.03189251	60.770	.0000	
REFORM	−.07655409	.03724341	−2.056	.0398	.45154639
BADH	.48318788	.03670938	13.163	.0000	.31134021
EDUC3	−.29882651	.04462514	−6.696	.0000	.31340206

Notice that the model for which counts 0–3 were simply dropped results in a p-value for *reform* that is not significant at the 0.05 level, whereas it is significant in the left-truncated model. The reason for the difference is due to the fact that the truncated model adjusts the log-likelihood. The models were estimated using LIMDEP.

Right-truncated models have a cut on the upper or right side of the distribution. The right-truncated Poisson PDF may be specified as

$$\text{Prob}(Y = y|Y < C) = \frac{\exp(-\mu)\mu^y/y!}{\text{Prob}(y < C)} = \frac{\exp(-\mu)\mu^y/y!}{\sum_{j=0}^{C-1}\exp(-\mu)\mu^j/j!}, \quad \text{for } y = 0, 1, 2, \ldots, C-1 \tag{9.7}$$

A right cut at 10 provides that values up to and including 9 will have non-zero probabilities in the truncated model, i.e. $C - 1$.

The left-truncated negative binomial PDF may be expressed as

$$H = \frac{\Gamma(y + \alpha^{-1})}{\Gamma(\alpha^{-1})\Gamma(y + 1)}(\alpha\mu)^y(1 + \alpha\mu)^{-(y+1/\alpha)} \tag{9.8}$$

$$I_j = \frac{\Gamma(j+\alpha^{-1})}{\Gamma(\alpha^{-1})\Gamma(j + 1)}(\alpha\mu)^j(1 + \alpha\mu)^{-(j+1/a)} \tag{9.9}$$

$$\text{Prob}(Y = y|Y > C) = \frac{H}{1 - \sum_{j=0}^{C} I_j} \tag{9.10}$$

The right-truncated negative binomial PDF is formulated in the same manner as is the right-truncated Poisson

$$\text{Prob}(Y = y|Y < C) = \frac{H}{\sum_{j=0}^{C-1} I_j} \tag{9.11}$$

We shall use the same *mdvisitsx* data for an example of a right-truncated negative binomial model. Using a cut of 15, only values of *numvisit* up to 14 will be included in the model.

RIGHT-TRUNCATED NEGATIVE BINOMIAL: CUT = 15

```
Variable  Coefficient  Standard Error  b/St.Er.  P[|Z|>z]  Mean of X

Constant    .73384757  .04065421        18.051  .0000
REFORM    -.13716032  .05221156        -2.627  .0086    .50664224
BADH      1.18445765  .12792755         9.259  .0000    .10352726
EDUC3     -.01416990  .05690943         -.249  .8034    .34127348
            Dispersion parameter for count data model
Alpha       .87479181  .04757499        18.388  .0000
```

9.1.2 Censored models

Censored models have a similar form to the truncated, but there are important differences. For left-censored models, a cut of *C* indicates that *C* is the smallest recordable response value of the censored model. As such, all values of the original response that are actually less than *C* are measured and recorded as the value of *C*; thus this value in the data actually means “less than or equal to C." If $C = 3$, the lowest measurable response is a 3, and all values of the original response less than 3 now have a value recorded as 3. On the other hand, a $C = 3$ value for truncated models specify that 4 is the lowest possible value of the truncated response and that there are no response values under 4; if there are, then they are dropped from a truncated analysis.

There is a like difference of interpretation for right censoring. A right censored cut at C indicates that the largest recordable value of the response is C and that all values greater than C are recorded as C; thus this value in the data actually means "greater than or equal to C". If $C = 15$, then all response values greater than 15 are re-valued to 15. No values are dropped from the model analysis.

To summarize this point, censored cut points differ from truncated cuts in two ways.

Truncated Left: If $C = 3$, only values >3 are supported by the underlying distribution; lower values, if they exist, are dropped.

Right: If $C = 15$, only values <15 are supported by the underlying distribution; higher values, if they exist, are dropped.

Censored Left: If $C = 3$, 3 is smallest observable value in the model; this value is inexact and means only that the observation is less than or equal to 3. Any response in the data that is less than 3 is also considered to be less than or equal to 3.

Right: If $C = 15$, 15 is highest observable value in model; this value is inexact and means only that the observation is greater than or equal to 15. Any response in the data that is greater than 15 is also considered to be greater than or equal to 15.

The disparity of meanings for what a cut value indicates may give rise to considerable confusion. We must keep the difference clearly in mind when engaging truncated and censored count models.

We shall use the *mdvisitsx* data used for truncated models to compare with censored value output. We begin with respective examples of left-truncated and left-censored negative binomial regression with a cut of 3.

LEFT-TRUNCATED NEGATIVE BINOMIAL: CUT = 3

```
==========================================================================
Variable  Coefficient  Standard Error  b/St.Er.  P[|Z|>z]  Mean of X
--------------------------------------------------------------------------
Constant  −.34794507  1.76257038       −.197  .8435
REFORM    −.11831607   .13485136       −.877  .3803    .45154639
BADH        .90189762   .16928103       5.328  .0000    .31134021
EDUC3     −.48182675   .14780265      −3.260  .0011    .31340206
                Dispersion parameter for count data model
Alpha      7.44973611 14.2858346         .521  .6020
==========================================================================
```

LEFT-CENSORED NEGATIVE BINOMIAL: CUT = 3

```
Variable Coefficient  Standard Error  b/St.Er. P[|Z|>z] Mean of X

Constant    .73734541  .05285178         13.951  .0000
REFORM    −.14399988  .05908096         −2.437  .0148    .50606197
BADH      1.23620067  .09635204         12.830  .0000    .11360575
EDUC3     −.15383921  .06588073         −2.335  .0195    .33767400
              Dispersion parameter  for count  data model
Alpha     1.47560302  .11474405         12.860  .0000
```

The parameter estimates have the same signs, but quite different values. Moreover, reform is not contributory to the truncated model, whereas it is for the censored model. The values of α are substantially different, i.e. 7.45 to 1.48. It is likely that the deletion of values less than 4, which consist of 1742 out of the original 2227 patients, or 78% of the cases, result in considerable overdispersion in the remaining data. This situation does not exist for censored models. All observations are kept; censored values are just revalued.

Modeling a right-truncated negative binomial with a cut of 15 may be compared with the truncated model at the end of the previous section.

RIGHT-CENSORED NEGATIVE BINOMIAL: CUT = 15

```
Variable Coefficient  Standard Error  b/St.Er. P[|Z|>z] Mean of Xπ

Constant    .79319680  .03867885         20.507  .0000
REFORM    −.10697298  .04788241         −2.234  .0255    .50606197
BADH      1.13470532  .07883350         14.394  .0000    .11360575
EDUC3     −.07048948  .05299836         −1.330  .1835    .33767400
              Dispersion parameter  for count  data model
Alpha       .97148925  .04434086         21.910  .0000
```

Unlike the left-truncated model, the right deletes only 44 out of 2227 observations, or only 2%. The differences between the right-censored and right-truncated models, as expected, do not substantially differ. In fact, the models display quite similar output.

For completeness, mention should be made about the probability and log-likelihood functions. The logic is quite simple. All four of the relevant censored models

right-censored Poisson   right-censored negative Binomial
left-censored Poisson   left-censored negative Binomial

take the same form as the truncated models we discussed in the last section. The difference is, however, in how the cuts are managed, as well as how values beyond the cut points are handled. The values at the cut, C, are included in the model, unlike the case with truncation, and values beyond cut points are revalued to the value of *C*. With truncation, values of *C* and beyond are dropped. The only alteration this causes in the PDF and log-likelihood functions, is at the cut.

Formula for the above functions, together with score functions and the observed information matrix for the left-truncated Poisson and negative binomial, are discussed in Cameron and Trivedi (1998, Chapter 4.5). Overdispersion tests based on score functions for both left- and right-truncated Poisson and negative binomial models are discussed by Gurmu and Trivedi (1992). Greene (2006) provides an excellent discussion of both truncated and censored Poisson and negative binomial models.

9.2 Censored Poisson and NB-2 models – survival parameterization

A prime motivating feature of survival models is the capability to censor certain observations. Censoring in this sense generally relates to the time when information about the observation is part of the model. For example, suppose we are following two groups of cancer patients for a 10-year period. Patients are registered into the study upon diagnosis. One group is given a new type of treatment; the other, called control, is given the standard treatment. Patients are followed until they either die or the study closes. What happens, though, if a study patient withdraws from the study following eight years of participation? The patient has not died, but rather has moved away from the study area. They have contributed a substantial amount of information to the study, and we know that the patient has survived through eight years. Patients who withdraw in such a fashion are said to be right censored. On the other hand, if another patient who has been taking the treatment protocol, or who has been on the standard treatment, enters the study well after diagnosis, they are said to be left censored. Potential contributing information is lost to the study results. In a single study, patients can be both right and left censored, as well as not censored at all. Moreover, they can be censored at a variety of times. This situation is vastly different from the econometric sense of censoring in which a single cut point defines censoring for all affected cases.

The majority of survival models have a continuous time response, as would be the case in the example above. However, there is a separate type of model called discrete response survival model. Rather than having the response defined in terms of time, the response can be construed as counts. These types of models have traditionally been modeled as piecewise models, but can in fact be modeled using a count model that has the capability of accounting for censored observations. Censored Poisson and negative binomial models can be devised for such count response data. The models may be used for any type of count response, regardless of whether it is being used in a survival context. The only difference is that censoring – in the survival sense – is allowed as a capability.

The essential difference between the two approaches to censoring is that the survival parameterization considers censor points as observation defined, whereas the econometric parameterization considers them as dataset defined. The econometric parameterization uses cut points at specified values in the data set, with all values above or below the cuts censored. The survival parameterization is more general in that cut points may be defined by observations above or below an assigned or specified value, but may also be defined for individual observations within the cuts. The econometric parameterization defines ranges of truncated and censored values, whereas the survival parameterization defines only individual observations. Values are censored by virtue of their place in the data set, or values are censored by virtue of external reasons, e.g. lost to study, late entry, and so forth.

CENSORED POISSON LOG-LIKELIHOOD FUNCTION

$$\begin{aligned}\mathcal{L}(x\beta; y) = {} & \delta\{-\exp(x\beta) + y(x\beta) - \ln\Gamma(y+1)\} + \zeta\{\ln\Gamma_{\mathrm{I}}(y, \exp(x\beta))\}\\ & + \tau\{\ln(1 - \ln\Gamma_{\mathrm{I}}(y+1, \exp(x\beta)))\}\end{aligned} \tag{9.13}$$

where

δ : 1 if observation not censored; 0 otherwise
ζ : 1 if observation is left censored; 0 otherwise
τ : 1 if observation is right censored; 0 otherwise

and $\ln\Gamma_{\mathrm{I}}$ is the 2 parameter incomplete gamma function.

CENSORED NEGATIVE BINOMIAL LOG-LIKELIHOOD FUNCTION

$$\begin{aligned}\mathcal{L}(x\beta; y, \alpha) = {} & \delta\{y\ln(\mu/(1+\mu)) - \ln(1+\mu)/\alpha + \ln\Gamma(y + 1/\alpha)\\ & - \ln\Gamma(y+1) - \ln\Gamma(1/\alpha)\}\\ & + \zeta\{\ln(\Gamma_I(y, \exp(x\beta))), \ln(\beta_I(y, n-y+1, \exp(x\beta)))\}\\ & + \tau\{\ln(\beta_I(y+1, n-y, \exp(x\beta)))\}\end{aligned} \tag{9.14}$$

with $\alpha = \exp(\alpha)$ and $\mu = \alpha * \exp(x\beta)$.

Other terms are n = number of observations in the data and β_{I} = incomplete beta function. The three parameter β_{I} function returns the cumulative beta distribution, or incomplete beta function, for censored responses.

Table 9.2. *MEDPAR: censor variable*

die	Freq.	Percent	Cum.
−1	58	3.88	3.88
0	924	61.81	65.69
1	513	34.31	100.00
Total	1,495	100.00	

Using the *medpar* data used earlier in the text, we model length of hospital stay (*los*) on *white*, being a member of an HMO (*hmo*), and whether the patient is over 80 years of age (*age80*). A right-censor indicator, *die*, was created from the variable, *died*, such that 1 specifies that the patient has died, 0 = patient is alive and −1 that the patient is lost from the study. Patients lost from the study after participating are, therefore, right censored.

Values of −1 were randomly assigned from among patients who were alive.

A censored Poisson model on the data gives the following output:

```
. cpoisson los white hmo age90, cen(died)

Censored Poisson Regression                  Number of obs =    1495
                                             Wald chi2(3)  =  193.88
Log likelihood = −4623.6239                  Prob > chi2   =  0.0000

-----------------------------------------------------------------------
   los |    Coef.    Std. Err.     z     P>|z|    [95% Conf. Interval]
-----------------------------------------------------------------------
 white | −.2879192  .0326395    −8.82   0.000    −.3518915   −.2239469
   hmo | −.1753515   .027617    −6.35   0.000    −.2294799    −.121223
 age80 | −.1865871  .0228685    −8.16   0.000    −.2314086   −.1417656
 _cons |  2.901129  .0312814    92.74   0.000     2.839819    2.962439
-----------------------------------------------------------------------
AIC Statistic = 6.191
-----------------------------------------------------------------------
```

Comparing the model to a standard Poisson model on the same data, we note that the parameter estimates and standard errors are somewhat similar. However, the Pearson Chi2 dispersion is extremely high at 7.71.

```
. glm los white hmo ago80, fam(poi)

Generalized linear models              No. of obs          =      1495
Optimization    : ML                   Residual df         =      1491
                                       Scale parameter     =         1
Deviance        = 8800.483496          (1/df) Deviance     =  5.902403
Pearson         = 11490.80115          (1/df) Pearson      =  7.706775
                                       AIC                 =  9.714805
Log likelihood = −7257.816534          BIC                 = −2098.55

-----------------------------------------------------------------------
   los |    Coef.   OIM Std. Err.     z    P>|z|    [95% Conf. Interval]
-----------------------------------------------------------------------
 white | −.1858699  .0273143        −6.80  0.000    −.2394049  −.1323349
   hmo | −.1448544   .023748        −6.10  0.000    −.1913997  −.0983091
 age80 | −.0712421  .0203222        −3.51  0.000    −.1110729  −.0314113
 _cons |  2.493478  .0260726        95.64  0.000     2.442377   2.544579
-----------------------------------------------------------------------
```

Generally speaking, negative binomial models tend to dampen any overdispersion that may reside in a Poisson model. However, in so doing, the significance of various model predictors may be affected. Typically the negative binomial inflates the standard errors of overdispersed Poisson parameter estimates. This results in one or more predictors showing a non-contributory relationship to the model, whereas they appeared significant in the Poisson model. The same is the case with censored models.

```
. censornb los white hmo age80, cen(died)

Censored Negative Binomial Regression   Number of obs   =   1495
                                        Wald chi2(3)    = 128.83
Log likelihood = −1981.6047             Prob > chi2     = 0.0000

============================================================================
      los |    Coef.    Std. Err.     z    P>|z|     [95% Conf. Interval]
----------+-----------------------------------------------------------------
xb        |
    white |   .0026736  .1153047    0.02  0.982    −.2233194    .2286667
      hmo |  −.4839241  .1019323   −4.75  0.000    −.6837077   −.2841404
    age80 |  −.7338938  .0814828   −9.01  0.000    −.8935971   −.5741904
    _cons |   4.027752  .1104592   36.46  0.000     3.811256    4.244248
----------+-----------------------------------------------------------------
lnalpha   |
    _cons |   .7238855   .058532   12.37  0.000     .6091649    .8386061
----------+-----------------------------------------------------------------
    alpha |   2.062431  .1207182                     1.838895     2.31314
----------+-----------------------------------------------------------------
AIC Statistic = 2.656
============================================================================
```

Note also the markedly reduced AIC statistic. The censored negative binomial appears to be the preferred model of the two.

It may be of interest to compare the censored negative binomial, parameterized as an econometric model, with the survival parameterization. Since there is a substantial drop in values after LOS = 24, we shall set a cut at 24. For the survival model we shall create a right censor variable that has a code of −1 for *los* values of 24 or greater.

CENSORED NB-2: ECONOMETRIC

```
RIGHT Censored Data: Threshold = 24.
NegBin form 2; Psi(i) = theta

=========================================================================
Variable Coefficient  Standard Error b/St.Er. P[|Z|>z] Mean of X
-------------------------------------------------------------------------
Constant  2.40569591  .06564927       36.645  .0000
WHITE     −.13858051  .06807046       −2.036  .0418    .91505017
HMO       −.11228740  .05392468       −2.082  .0373    .15986622
AGE80     −.05182604  .04747138       −1.092  .2750    .22073579
               Dispersion parameter for count data model
Alpha       .42171203  .02139959       19.707  .0000
=========================================================================
```

CENSORED NB-2: SURVIVAL

```
gen rgtc= −1 if los>=24
replace rgtc = 1 if rgtc==.
tab rgtc
```

```
rgtc   |        Freq.         Percent            Cum.
-------+----------------------------------------------
    -1 |           74            4.95            4.95
     1 |        1,421           95.05          100.00
-------+----------------------------------------------
Total  |        1,495          100.00
```

```
. censornb los white hmo age80, cen(rgtc)

Censored Negative Binomial Regression    Number of obs   =    1495
                                         Wald chi2(3)    =    5.13
Log likelihood = -4299.2669              Prob > chi2     =  0.1627

       los |    Coef.   Std. Err.     z    P>|z|    [95% Conf. Interval]
-----------+-----------------------------------------------------------
xb         |
     white | -.1162715  .0635172   -1.83  0.067   -.2407629   .0082199
       hmo | -.0340281  .0480665   -0.71  0.479   -.1282368   .0601805
     age80 |  .0459967  .0421791    1.09  0.275   -.0366727   .1286661
     _cons |  2.238545  .0612939   36.52  0.000    2.118411   2.358679
-----------+-----------------------------------------------------------
lnalpha    |
     _cons | -1.116643  .0524391  -21.29  0.000   -1.219422  -1.013865
-----------+-----------------------------------------------------------
     alpha |  .3273768  .0171673                    .2954008   .3628141
-----------+-----------------------------------------------------------
AIC Statistic = 5.757
```

The parameter estimates, standard errors, and alpha are similar between the two parameterizations. The difference, of course, is that the survival parameterization has a substantially greater scope of censoring capabilities. Censoring can be anywhere in the data, not only at the tails.

The difference between econometric and survival parameterizations of the right-censored Poisson show the same close list of parameter estimates and standard errors.

CENSORED POISSON: ECONOMETRIC

```
RIGHT Censored Data: Threshold = 24.
Overdispersion tests: g=mu(i): 6.439
Overdispersion tests: g = mu(i)^2 : 6.537

Variable  Coefficient  Standard Error  b/St.Er.  P[|Z|>z]  Mean of X
--------------------------------------------------------------------
Constant   2.36486935   .02775028       85.220   .0000
WHITE     -.12829066   .02897987       -4.427   .0000     .91505017
HMO       -.09474473   .02406297       -3.937   .0001     .15986622
AGE80     -.02868762   .02071530       -1.385   .1661     .22073579
```

CENSORED POISSON: SURVIVAL

```
. cpoisson alos white hmo age80, cen(rgtc)

Censored Poisson Regression         Number of obs          =    1495
                                    Wald chi2(3)           =   20.65
Log likelihood = -5188.4209         Prob > chi2            =  0.0001
```

```
----------------------------------------------------------------------
   alos |     Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------+-------------------------------------------------------------
  white | -.1173776   .0315764    -3.72   0.000    -.1792662   -.0554889
    hmo | -.0346934   .0249728    -1.39   0.165    -.0836391    .0142524
  age80 |  .0477127   .0215141     2.22   0.027     .0055458    .0898795
  _cons |  2.239263   .0302942    73.92   0.000     2.179888    2.298639
-----------------------------------------------------------------------
AIC Statistic = 6.946
======================================================================
```

9.3 Sample selection models

A variety of sample selection models can be found in statistical literature. The most common usage of sample selection has been within the domain of continuous-response models. Heckman selection models, bivariate probit, and normal models with censoring have been most commonly used in research. However, as discussed in Cameron and Trivedi (1998) these models are not appropriate for count response models.

Greene (1994) and Terza (1998) have recently developed maximum likelihood and two-step algorithms for count response sample selection models. We follow Green's example using credit card reports to show how the model works and how it is to be interpreted.

The data contain records of major derogatory reports about credit card holders, with the goal of predicting the probability of a default on a credit card loan. Selection bias is inherent in the data since the reports are only gathered on those who already have credit cards. There is no information on individuals who have not yet been issued cards, but who would default if they had them. Since these individuals are excluded from the sample data, the sample is not completely random and exhibits selection bias. In order to remedy the bias it is necessary to model both the manner in which credit cards are issued as well as the actual counts of derogatory reports. The process of issuing cards may be modeled using a binary response model with 1 indicating that a card has been issued and 0 that it has not. All potential applicants are thereby made part of the model, although individuals classified by the binary process as 0 are latent members of the data.

The sample selection model is therefore a two-part process, somewhat in the tradition of hurdle and zero-inflated models. Unfortunately, maximization of the two-part likelihood is much more complex than it is with any of the models we have thus far discussed. It must be maximized using either numerical integration or by using by simulation. LIMDEP, authored by Greene, uses simulation.

The model is structured so that the binary part, here a probit model, provides estimates of the probability of being issued a credit card. The count part is then estimated, but as adjusted by the probability values from the probit model. The

Table 9.3. *Sample selection models*

PROBIT		
Response		
cardhldr	=	1: card been issued; 0: card not been issued
Predictors		
agec	=	age in years and twelfths of a year when applied
income	=	income in $10,000s
ownrent	=	own or rent home (1/0)
curr_add	=	months residing at same address when applied for card
POISSON/NEGATIVE BINOMIAL		
Response		
majordrg	=	number of derogatory reports
Predictors		
avgexp	=	Average monthly expenditure
inc_per	=	income per dependent, in $10,000 units
major	=	1/0 if applicant had another credit card at application

count model is said to be a selection-corrected Poisson or negative binomial. Predictors of each model are rarely identical since they are predicting different processes.

For our example, the models are specified as shown in Table 9.3.

The probit model is entered into LIMDEP as:

```
probit; lhs=cardhldr; rhs=one, agec, income, ownrent,
cur_add; hold
```

followed by the selection-corrected Poisson:

```
pois; lhs=majordrg; rhs=one, avgexp, inc_per, major;
sel; mle$
```

The relevant models are displayed as:

SAMPLE CORRECTED POISSON

```
Poisson Model with Sample Selection.
Mean of LHS Variable = .12903
Restr. Log-L is Poisson+Probit (indep).
Log L for initial probit = −682.33401
Log L for initial Poisson = −430.22927
Means for Psn/Neg.Bin. use selected data.
```

```
Variable  Coefficient  Standard Error  b/St.Er.  P[|Z|>z]   Mean of X
          Parameters of Poisson/Neg. Binomial Probability
Constant −3.65979959      .59816464    −6.118  .0000
AVGEXP     .00078020      .00031097     2.509  .0121   238.602421
INC_PER    .16237072      .07091514     2.290  .0220   2.21873662
MAJOR      .22733512      .30623298      .742  .4579    .83968719
          Parameters of Probit Selection Model
Constant   .74148859      .15225187     4.870  .0000
AGEC     −.01027791       .00485007    −2.119  .0341   33.3853297
INCOME     .06174082      .02416541     2.555  .0106   3.36537604
OWNRENT    .45569930      .08791258     5.184  .0000    .44048522
CUR_ADD  −.00046311       .00063456    −.730   .4655   55.2676270
          Standard Deviation of Heterogeneity
Sigma     1.16180092      .22171151     5.240  .0000
          Correlation of Heterogeneity & Selection
Rho        .39658662     1.06023875      .374  .7084
```

Major does not contribute to the probit model and *cur_add* is not contributory to the selected corrected Poisson.

The sample corrected negative binomial model is displayed following the command:

```
  negb; lhs=majordrg; rhs=one, avgexp, inc_per, major;
  sel; mle$
```

```
Neg.Bin.Model with Sample Selection.
Maximum Likelihood Estimates
Model estimated: May 29, 2006 at 08:18:15AM.
Dependent variable                                 MAJORDRG
Weighting variable                                     None
Number of observations                                 1319
Iterations completed                                    101
Log likelihood function                           −1081.161
Number of parameters                                     12
Info. Criterion: AIC =                              1.65756
Finite Sample: AIC   =                              1.65774
Info. Criterion: BIC =                              1.70473
Info. Criterion:HQIC =                              1.67524
Restricted log likelihood                         −1112.600
McFadden Pseudo R-squared                          .0282573
Chi squared                                        62.87818
Degrees of freedom                                        2
Prob[ChiSqd > value] =                             .0000000
```

```
Neg.Bin.Model with Sample Selection.
Mean of LHS Variable =                            .12903
Restr. Log-L is Poisson+Probit (indep).
Log L for initial probit =                        −682.37037
Log L for initial Poisson =                       −430.22927
Means for Psn/Neg.Bin. use selected data.
```

```
Variable  Coefficient  Standard Error  b/St.Er.  P[|Z|>z]   Mean of X
          Parameters of Poisson/Neg. Binomial Probability
Constant  −3.07204954       .81142913   −3.786  .0002
AVGEXP      .00082084       .00035875    2.288  .0221   238.602421
INC_PER     .15419880       .07648030    2.016  .0438   2.21873662
MAJOR       .23419460       .30789939     .761  .4469    .83968719
          Parameters of Probit Selection Model
Constant    .73676537     1.88621080      .391  .6961
AGEC      −.01022903       .02646802    −.386  .6992   33.3853297
INCOME      .06194595       .15953633     .388  .6978   3.36537604
OWNRENT     .45529765     1.16379189      .391  .6956    .44048522
CUR_ADD   −.00041768       .00125190    −.334  .7387   55.2676270
          Overdispersion Parameter for Negative Binomial
Theta      1.60803441     1.31003489     1.227  .2196
          Standard Deviation of Heterogeneity
Sigma       .49395748     1.15588222      .427  .6691
          Correlation of Heterogeneity & Selection
Rho         .61978139     2.31339106      .268  .7888
```

The parameters of the probit and the Poisson/negative binomial are fit at the same time.

The Sigma statistic that appears in both model output is the standard deviation of ν in $\lambda = \exp(\beta x + \nu)$. Rho is the correlation between ν and u in Prob$[d = 1] =$ Prob$(d'z + u > 0)$ in the probit model. The variance of u is 1. According to the model results, Rho is not significant in the sample selection Poisson model, which is interpreted as meaning that selection is not an issue in these data. The interpretation of the negative binomial model is the same.

The negative binomial selection model here is apparently only weakly identified, which is no surprise given the Poisson results (rho approx = 0). What is going on is something like collinearity, but with the derivatives. It is likely that this model is overspecified. The reason is that the negative binomial model as initially constructed is the Poisson model with an additional term for heterogeneity. The selection model adds yet another source of latent heterogeneity to what is already intrinsic to the negative binomial. That is, the negative binomial is used to accommodate overdispersed Poisson models. But then selection adds an additional layer of accommodation to overdispersion. In this case I suspect it is too much, i.e. it is overspecified. The model may in fact not be adequate to pick up all this latent activity. Other modeling situations may require the extra accommodation. But in this case, the probit selection model predictors are not significant, even though they were so when modeled alone. Thus the probit selection criteria with the Poisson model provides no support to the selection process.

We next discuss negative binomial panel models, including fixed and random effects models, GEE or population averaged models, and random intercept and random parameter models. Several of these types of models have only recently been developed, and represent some of the more interesting, as well as useful, applications of the negative binomial to study data.

9.4 Summary

Truncation and censoring primarily deal with how data are gathered. When there is missing data due to late entry into a study, or because data elements at a given point have been excluded from a study, we have truncation and censoring.

Truncation occurs when we do not have knowledge of counts at either a point near the beginning or at the end of the counting process. In either case, the truncated data are actually excluded from the model. The probability function of either the Poisson or NB-2 model is adjusted to account for the missing counts.

Censoring may be parameterized in either an econometric or a survival sense. In the former, censoring is similar to truncation, but the censored data to the outside of the cut point(s) are re-valued to that nearest value included in the model. For instance, if the cut point is at 3, counts of 0, 1, and 2 are not dropped as in truncation, but are set to 3. Censored data sets can be identified with numerous values at either cut point.

The survival parameterization of censoring allows censoring to take place anywhere in the data, not at cut points. This type of censoring is based on survival models such as Cox regression and the various parametric survival models. The survival parameterization of censored Poisson and negative binomial regression is available only with Stata; refer to the user authored procedures `cpoisson` and `censornb` respectively.

Sample selection models are common in econometric literature. The model is structured as a two-part process, with one process required to be existent, or to have occurred, before the main process occurs. The selection criterion is typically defined as a probit model, with the main counting process consisting of either a Poisson or negative binomial. There are many variations of sample selection when the main process is continuous – the foremost model being the Heckman two-step approach. However, Heckman and similar approaches are not appropriate for use with count response models. Sample selection count models have only recently come into fruition with the work of Greene and Terza.

We next discuss the nature and evaluation of count models for panel data. These models are appropriate for handling both clustered and longitudinal data.

Exercises

1 Model *time* on *died* and levels of *drug*, with drug level 1 as the referent. Censoring is indicated by the variables *cen1* and *cenx*. Observations having a censor value of 1 are not censored. Left-censored observations are designated by a censor value of 0, right-censored observations by a value of −1. Many

algorithms require a censor value regardless of the censoring status of the observations; thus the inclusion of *cen1* in the data. For the modeling task, censoring is indicated using the variable *cenx*.

Compare parameter estimates and AIC values for both the survival and econometric parameterizations of censored Poisson and censored negative binomial regression. Determine the optimal model. This data set is named *cancercen*.

time	died	drug	age	cen1	cenx	probit
1	1	1	65	1	0	0
1	1	1	61	1	1	0
2	1	1	59	1	1	1
3	1	1	52	1	0	1
4	1	1	67	1	0	0
4	1	1	56	1	0	1
5	1	1	58	1	1	0
5	1	1	63	1	1	1
6	0	2	65	1	1	0
6	1	3	55	1	1	0
6	1	2	67	1	1	1
7	1	2	58	1	0	1
8	0	1	58	1	1	1
8	1	1	56	1	1	1
8	1	1	49	1	1	1
8	1	1	52	1	0	1
9	0	2	56	1	0	1
10	0	2	49	1	1	0
10	1	3	54	1	1	1
11	0	2	61	1	1	0
11	1	1	55	1	−1	0
11	1	1	50	1	1	0
12	1	1	49	1	1	0
12	1	1	62	1	1	1
13	1	2	62	1	1	1
15	0	2	50	1	1	1
15	1	1	51	1	−1	1
16	1	2	67	1	−1	1
17	0	3	60	1	1	1
17	1	1	49	1	1	1
19	0	3	49	1	1	1
19	0	2	50	1	1	0
20	0	2	55	1	1	1
22	1	2	58	1	1	1
22	1	1	57	1	1	0
23	1	2	47	1	1	1
23	1	1	52	1	1	1
24	1	3	58	1	0	1
25	0	3	50	1	1	1

time	died	drug	age	cen1	cenx	probit
25	1	3	55	1	1	1
28	0	3	48	1	1	1
28	1	3	57	1	1	1
32	0	3	56	1	1	1
32	0	2	52	1	1	1
33	1	3	60	1	1	0
34	0	3	62	1	1	0
35	0	3	48	1	1	1
39	0	3	52	1	1	0

2 Use the data in Exercise 1 to develop a sample selection model with *probit* as the indicator of being selected into the main model. Model both a Poisson and negative binomial selection model of time on *died*, *age*, and levels of *drug*. You may find that leveling *age* into three of four groups will assist in model fit, as well as interpretation.

3 Compare the negative binomial selection model developed for Exercise 2 with a survival parameterized censored negative binomial model where probit is the censor variable. Employ probit as both a left censor, with a value of 0 indicating left censoring, and then as a right-censored variable. One must convert all probit values of 0 to −1 prior to modeling with probit as a right censor. Discuss differences in output.

4 Using the data from Example 1, create an indicator variable called *rgtrun*. Assign values of 1 for all values of *rgtrun* where time is less than 30. For values of time greater than 40, assign *rgtrun* a value of −1. Compare the modeling results of a negative binomial truncated model with that of a survival parameterized censored negative binomial model. Then model the same data using an econometric parameterized censored negative binomial. Which model results are more likely to be similar? Discuss why.

5 Discuss the type of data that are appropriate for survival parameterized censoring in difference to econometric parameterized censoring.

6 Model the following data using a truncated Poisson and truncated negative binomial. What reasons are indicated to prefer the negative binomial over the Poisson model? (data *contacts* from Sikkel and Jelierse, 1988, Table 1).

Contacts in months	Count
0	4244
1	1719
2	3337
3	461
4	190
>4	267

10

Negative binomial panel models

A basic assumption in the construction of models from likelihood theory is that observations in the model are independent. This is a reasonable assumption for perhaps the majority of studies. However, for longitudinal studies this assumption is not feasible; nor does it hold when data are clustered. For example, observations from a study on student drop-out can be clustered by the type of schools sampled. If the study is related to intervention strategies, schools in affluent suburban, middle-class suburban, middle-class urban, and below poverty level schools have more highly correlated strategies within the school type than between types or groups. Likewise, if we have study data taken on a group of individual patients over time (e.g., treatment results obtained once per month for a year), the data related to individuals in the various time periods are likely to be more highly correlated than are treatment results between patients. Any time the data can be grouped into clusters, or panels, of correlated groups, we must adjust the likelihood-based model (based on independent observations) to account for the extra-correlation.

We have previously employed robust variance estimators and bootstrapped standard errors when faced with overdispersed count data. Overdispersed Poisson models were adjusted by using different types of negative binomial models, or by extending the basic Poisson model by adjusting the variance or by designing a new log-likelihood function to account for the specific cause of the overdispersion. Examples we have previously discussed include zero-inflated models, zero-truncated models, hurdle models, and censored or truncated models.

In this chapter we shall describe a group of models that:

1 add at least one extra parameter to the linear predictor, specifying how observations within panels are to be construed, and
2 derive new log-likelihoods based on panels of correlated observations.

The type of parameters that are added to the linear predictor, and the manner in which panels are treated will determine the type of panel model described. We shall first discuss fixed effects count models, differentiating between unconditional and conditional varieties. Following an examination of fixed-effects count models, we address random-effects models, followed by generalized estimating equations, or population-averaged models. Each of these types of panel models has software support in several commercial packages. Our final group of panel models has only recently found commercial software support, with only one package providing support for negative binomial models (LIMDEP). These regression models are commonly referred to as multilevel models. The two foremost members of this variety of panel model are random intercept and random coefficient or parameter models. More complex multilevel models are, for the most part, built on their basis. With respect to multilevel negative binomial models, current research has only begun in earnest within the last couple of years. They are still in the developmental stage.

10.1 Unconditional fixed-effects negative binomial model

Fixed-effects count models may be estimated in two ways – unconditionally and conditionally. We begin with a consideration of the unconditional fixed-effects Poisson model since it is the basis on which we can understand the negative binomial parameterizations.

Unconditional estimation of the fixed effects Poisson model can be obtained using standard GLM software as well as the traditional maximum likelihood Poisson procedure. The model is specified by including a separate fixed effect for each defined panel in the data. The fixed effects are specified by indicator variables, just as is done when estimating factor or categorical predictors. We represent this relationship as

$$\ln(\mu_{ik}) = \exp(\beta x_{ik} + \delta_i) \tag{10.1}$$

where δ is the fixed effect associated with individual, i, and subscript k indexes the observations associated with individual i. When a panel relates observations collected over a time period, it is customary to use the subscript t instead of k. We shall use k throughout our discussion, but with the knowledge that t is commonly used for longitudinal models.

The log-likelihood for the unconditional fixed effects Poisson takes the form of

$$\mathcal{L}(x\beta_i; y_i) == \sum_{i=1}^{n} \sum^{N}{}_{k=1} [y_{ik}(X_{ik}\beta + \delta_i) - \exp(X_{ik}\beta + \delta_i) - \ln\Gamma(y_{ik} + 1)] \tag{10.2}$$

Note the similarity to that of the standard Poisson log-likelihood function defined in Chapter 3 as

$$\mathcal{L}(x\beta_i; y_i) = \sum_{i=1}^{n} \{y_i(x_i\beta) - \exp(x_i\beta) - \ln\Gamma(y_i + 1)\}$$

I shall use the well-known *ships* data set that was used in McCullagh and Nelder (1989), Hardin and Hilbe (2003), and other sources. The dataset contains values on the number of reported accidents for ships belonging to a company over a given time period. The variables are defined as:

```
accident    : number of accidents (reponse)
ship        : ship identification (1–8)
op          : ship operated between the years 1975 and 1979 (1/0)
co65_69     : ship was in construction between 1965 and 1969 (1/0)
co70_74     : ship was in construction between 1970 and 1974 (1/0)
co75_79     : ship was in construction between 1975 and 1979 (1/0)
service     : months in service
```

With the natural log of the months of service specified as the offset, a basic Poisson model of the data is given as:

```
. glm accident op co_65_69 – co_75_79, nolog fam(poi)
lnoffset(service)

Generalized linear models            No. of obs        =         34
Optimization     :   ML              Residual df       =         29
                                     Scale parameter   =          1
Deviance         =   62.36534078     (1/df) Deviance   =   2.150529
Pearson          =   82.73714004     (1/df) Pearson    =   2.853005
                                     AIC               =   5.006819
Log likelihood   =   -80.11591605    BIC               =  -39.89911

---------------------------------------------------------------------
accident |    Coef.   OIM Std. Err.    z     P>|z|  [95% Conf. Interval]
---------------------------------------------------------------------
      op |  .3874638     .118107      3.28 0.001   .1559783   .6189494
co_65_69 |  .7542017    .1487697      5.07 0.000   .4626185   1.045785
co_70_74 |   1.05087      .15757      6.67 0.000   .7420385   1.359701
co_75_79 |  .7040507    .2203103      3.20 0.001   .2722504   1.135851
   _cons |  -6.94765    .1269363    -54.73 0.000  -7.196441   -6.69886
 service |(exposure)
---------------------------------------------------------------------
```

Predictors appear to be significant; however, the model is clearly overdispersed. We have purposefully ignored the correlation of values within each panel of ships in the above model. A negative binomial model can be used to generically account for the overdispersion, appearing as:

```
. glm accident op co_65_69 – co_75_79, fam(nb.1303451)
lnoffset(service)

Generalized linear models              No. of obs        =         34
Optimization    :  ML                  Residual df       =         29
                                       Scale parameter   =          1
Deviance        =  36.84717336         (1/df) Deviance   =   1.270592
Pearson         =  42.24099154         (1/df) Pearson    =   1.456586
Variance        :  V(u) =              [Neg. Binomial]
function           u+(.1303451)u^2
                                       AIC               =   4.874952
Log likelihood  =  -77.87418504        BIC               =  -65.41728
```

```
accident |     Coef.   OIM Std. Err.    z    P>|z|  [95% Conf. Interval]
---------+--------------------------------------------------------------
      op |  .3536459    .2347302      1.51 0.132   -.1064169  .8137087
co_65_69 |  1.012518     .329175      3.08 0.002    .3673472   1.65769
co_70_74 |  1.255125    .3086897      4.07 0.000    .6501045  1.860146
co_75_79 |  .7595303    .3854008      1.97 0.049    .0041585  1.514902
   _cons | -6.933539    .2849396    -24.33 0.000   -7.492011 -6.375068
 service |(exposure)
```

Much of the overdispersion has been accommodated by the negative binomial model, but there is still evidence of extra correlation in the data. We also know from the data what may be causing overdispersion – the panel-specific effect of the individual ships.

We assign a specific indicator to each panel. Each ship will have a separate slope. This type of model is called an unconditional fixed-effects model. As a Poisson model we have:

```
. glm accident op co_65_69 – co_75_79 ship2-ship5,fam(poi)
lnoffset(service)

Generalized linear models            No. of obs         =        34
Optimization      :   ML             Residual df        =        25
                                     Scale parameter    =         1
Deviance          =   38.69505154    (1/df) Deviance    =  1.547802
Pearson           =   42.27525312    (1/df) Pearson     =   1.69101
                                     AIC                =  4.545928
Log likelihood    =   -68.28077143   BIC                = -49.46396
```

```
accident |     Coef.   OIM Std. Err.    z    P>|z|  [95% Conf. Interval]
---------+--------------------------------------------------------------
      op |   .384467    .1182722      3.25 0.001    .1526578  .6162761
co_65_69 |  .6971404    .1496414      4.66 0.000    .4038487  .9904322
co_70_74 |  .8184266    .1697736      4.82 0.000    .4856763  1.151177
co_75_79 |  .4534266    .2331705      1.94 0.052   -.0035791  .9104324
   ship2 | -.5433443    .1775899     -3.06 0.002   -.8914141 -.1952745
   ship3 | -.6874016    .3290472     -2.09 0.037   -1.332322  -.042481
   ship4 | -.0759614    .2905787     -0.26 0.794   -.6454851  .4935623
   ship5 |  .3255795    .2358794      1.38 0.168   -.1367357  .7878946
   _cons | -6.405902    .2174441    -29.46 0.000   -6.832084 -5.979719
 service |(exposure)
```

A substantial amount of the overdispersion present in the original Poisson model has been accounted for. However, the negative binomial handles the overdispersion better than the unconditional fixed effects Poisson. We next attempt to model an unconditional fixed-effects negative binomial.

```
. glm accident op co_65_69-co_75_79 ship2-ship5,
fam(nb 0.0000000253) lnoffset(service)

Generalized linear models                No. of obs       =        34
Optimization    : ML                     Residual df      =        25
                                         Scale parameter  =         1
Deviance        = 38.69504594            (1/df) Deviance  =  1.547802
Pearson         = 42.27517882            (1/df) Pearson   =  1.691007
Variance        : V(u) =                 [Neg. Binomial]
function          u+(0.0000000253) u^2
                                         AIC              =  4.545928
Log likelihood = -68.28077281            BIC              = -49.46397
```

```
accident |     Coef.   OIM Std. Err.     z    P>|z|  [95% Conf.  Interval]
---------+----------------------------------------------------------------
      op |   .384467    .1182722       3.25  0.001    .1526577   .6162762
co_65_69 |  .6971404    .1496415       4.66  0.000    .4038485   .9904323
co_70_74 |  .8184266    .1697737       4.82  0.000    .4856763   1.151177
co_75_79 |  .4534266    .2331705       1.94  0.052   -.0035792   .9104324
   ship2 | -.5433445    .1775899      -3.06  0.002   -.8914144  -.1952747
   ship3 | -.6873984    .3290468      -2.09  0.037   -1.332318  -.0424786
   ship4 | -.0759617    .2905787      -0.26  0.794   -.6454854    .493562
   ship5 |  .3255795    .2358794       1.38  0.168   -.1367356   .7878946
   _cons | -6.405901    .2174441     -29.46  0.000   -6.832084  -5.979719
 service | (exposure)
```

From a previous maximum likelihood estimation of the model it is discovered that the negative binomial value of α is approximately 0.0. This indicates that the model is in fact Poisson, and that the extra overdispersion is likely to come from a source other than the fixed panel effect of the individual ships. Nevertheless, the fit, as indicated by the AIC and BIC statistics, appears to favor the unconditional fixed-effects Poisson model over a model not accounting for the panel effect of individual ships.

A caveat on using this form of fixed-effects regression: use it only if there are a relatively few number of panels in the data. If there are more than 20 panels, it is preferred to use the conditional fixed-effects model. The greater the number of panels, the greater the possible bias in parameter estimates for the levels or panels of the effect variable. This is called the 'incidental parameters problem', first defined by Neyman and Scott (1948). It is interesting that a number of econometricians have thought that the incidental parameters problem, which we shall refer to as the IP problem, affects the unconditional fixed-effects Poisson model. Woutersen (2002) attempted to ameliorate the IP problem with Poisson models by employing an integrated moment estimator. Other attempts include Lancaster (2002) and Vadeby (2002). Most of these "solutions" are based on separating the main model parameters from the array of fixed-effects parameters. However, it has been demonstrated by Greene (2006) and others that the IP problem is not real when applied to the Poisson model. This conclusion is based on the observation that the Poisson conditional fixed-effects estimator is numerically equal to the unconditional estimator, which means that there is no IP problem. On the other hand, the IP problem does affect the unconditional fixed-effects negative binomial. But the fixed-effects negative binomial model has a different problem. It is intrinsically different from the Poisson. Recall that the Poisson fixed-effects has a mean, μ_{ik}, value of $\exp(\beta x_{ik} + \delta_i)$. This means that the fixed effect is built into the Poisson mean parameter. The negative binomial fixed-effects model, though, builds the fixed-effects into the distribution of the gamma heterogeneity, α, not the mean. This makes it rather difficult to interpret the IP problem with the negative binomial. One result is that the estimator is inconsistent in the presence of a large number of fixed effects. But exactly how it is inconsistent is still a matter of debate.

There is good evidence that in the presence of a large number of fixed effects, the unconditional negative binomial will underestimate standard errors, resulting in insufficient coverage of the confidence intervals. That is, negative binomial predictors appear to enter the model as significant when in fact they do not. Simulation studies (Greene, 2006) have demonstrated that scaling the unconditional fixed-effects negative binomial model standard errors by the deviance-based dispersion statistic produces standard errors that are closer to the nominal values. This is not the case when using Pearson χ^2-based dispersion as the basis for scaling standard errors, as is the norm for non-panel models. On the other hand, using the deviance-based dispersion statistic for scaling unconditional fixed effects Poisson models does not improve coverage, and the Pearson χ^2 dispersion should be used. These facts need to be kept in mind when modeling unconditional fixed-effects count models.

10.2 Conditional fixed-effects negative binomial model

Panel data models are constructed in order to control for all of the stable predictors in the model and to account for the correlation resulting from observations being associated within groups or panels. The value of conditional fixed-effects models is that a near infinite number of panels may be adjusted, while at the same time being conditioned out of the actual model itself. We do not have to deal with a host of dummy slopes.

A conditional fixed-effects model is derived by conditioning out the fixed effects from the model estimation. Like unconditional fixed-effects models, there is a separate fixed effect, δ, specified in the linear predictor. Hence, $\eta = x\beta + \delta$. However, unlike the unconditional version, a revised log-likelihood function is derived to affect the conditioning out of the panel effects through a sufficient statistic.

The conditional log-likelihood function is conditioned on the sum of the responses within each panel:

$$\sum y_{ik} \tag{10.3}$$

Prior to defining the Poisson log-likelihood, we shall first present the Poisson probability function.

CONDITIONAL FIXED EFFECTS POISSON PROBABILITY FUNCTION

$$f(y_{it}; x_{it}\beta)\left(\sum_{t=1}^{n_i} y_{it}\right)! \prod_{t=1}^{n_i} \frac{\exp(x_{it}\beta)^{y_{it}}}{y_{it}!\left\{\sum_k \exp(x_{it}\beta)\right\}^{y_{it}}}$$

CONDITIONAL FIXED EFFECTS POISSON LOG-LIKELIHOOD FUNCTION

$$\mathcal{L}(\beta; y_{it}) = \sum_{i=1}^{n_i} \left[\ln\Gamma\left(\sum_{t=1}^{n_i} y_{it} + 1\right) - \sum_{t=1}^{n_i}(\ln\Gamma(y_{it}+1)) \right.$$
$$\left. + \sum_{t=1}^{n_i}\left\{ y_{it}(x_{it}\beta) - y_{it}\ln\left(\sum_{l=1}^{n_i}(\exp(x_{il}\beta))\right)\right\}\right] \qquad (10.4)$$

or

$$\mathcal{L}(\mu; y_{it}) = \sum_{i=1}^{n_i} \left[\ln\Gamma\left(\sum_{t=1}^{n_i} y_{it} + 1\right) - \sum_{t=1}^{n_i}(\ln\Gamma(y_{it}+1)) \right.$$
$$\left. + \sum_{t=1}^{n_i}\left\{ y_{it}x_{it}\beta - y_{it}\ln\left(\sum_{l=1}^{n_i}\mu_{il}\right)\right\}\right]$$

Following the derivation of the model as proposed by Hausman, Hall and Griliches (1984), the conditional fixed-effects negative binomial probability and log-likelihood functions are shown as:

CONDITIONAL FIXED EFFECTS NEGATIVE BINOMIAL PROBABILITY FUNCTION

$$f(y_{it}; x_{it}\beta) = \prod_{t=1}^{n_i}\left(\frac{\Gamma(\mu_{it}+y_{it})}{\Gamma(\mu_{it})y_{it}!}\right)\left(\frac{\Gamma\left(\sum_{t=1}^{n_i}\mu_{it}\right)\Gamma\left(\sum_{t=1}^{n_i}y_{it}+1\right)}{\Gamma\left(\sum_{t=1}^{n_i}\mu_{it}+\sum_{t=1}^{n_i}y_{it}\right)}\right)$$

CONDITIONAL FIXED EFFECTS NEGATIVE BINOMIAL LOG-LIKELIHOOD

$$\mathcal{L}(\mu_{it}; y_{it}) = \sum_{t=1}^{n_i}\ln\Gamma\left(\sum_{t=1}^{n_i}\mu_{it}\right) + \ln\Gamma\left(\sum_{t=1}^{n_i}y_{it}+1\right) - \sum_{t=1}^{n_i}(\ln\Gamma(y_{it}+1))$$
$$- \ln\Gamma\left(\sum_{t=1}^{n_i}y_{it} + \sum_{t=1}^{n_i}(\mu_{it}) + \sum_{t=1}^{n_i}(\ln\Gamma(\mu_{it}+y_{it})) - \sum_{t=1}^{n_i}(\ln\Gamma(\mu_{it}))\right) \qquad (10.5)$$

or

$$\begin{aligned}\mathcal{L}(\beta; y_{it}) = \sum_{t-1}^{n_i} \ln\Gamma\left[\sum_{t-1}^{n_i}(\exp(x_{it}\beta))\right] + \ln\Gamma\left(\sum y_{it} + 1\right)\\ - \sum_{t-1}^{n_i}(\ln\Gamma(y_{it} + 1)) - \ln\Gamma\left(\sum_{t-1}^{n_i} y_{it} + \sum_{t-1}^{n_i}(\exp(x_{it}\beta))\right)\\ + \sum_{t-1}^{n_i}(\ln\Gamma(\exp(x_{it}\beta) + y_{it})) - \sum_{t-1}^{n_i}(\ln\Gamma(\exp(x_{it}\beta)))\end{aligned} \qquad (10.6)$$

Note that the heterogeneity parameter, δ or α, does not appear in the log likelihood. It does not, as a result, appear in the model output.

Another model that we should mention, but that has had very little application and currently has no commercial software support, is the NB-1 conditional fixed effects model. Its probability and log-likelihood functions can be derived as:

CONDITIONAL FIXED EFFECTS NB-1 PROBABILITY FUNCTION

$$f(y_{it}; \beta) = \prod_{t=1}^{n_I}\left(\frac{\Gamma(\mu_{it} + y_{it})}{\Gamma(y_{it} + 1)}\right)\left(\frac{\Gamma\left(\sum_{t=1}^{n_i}\mu_{it}\right)\Gamma\left(\sum_{t=1}^{n_i} y_{it} + 1\right)}{\Gamma\left(\sum_{t=1}^{n_i}\mu_{it} + \sum_{t=1}^{n_i} y_{it}\right)}\right) \qquad (10.7)$$

CONDITIONAL FIXED EFFECTS NB-1 LOG-LIKELIHOOD

$$\begin{aligned}\mathcal{L}(\mu_{it}; y_{it}) = \sum_{t=1}^{n_t}\{\ln\Gamma(\mu_{it} + y_{it}) - \ln\Gamma(\mu_{it}) - \ln\Gamma(y_{it} + 1)\}\\ + \ln\Gamma\left[\sum_{t=1}^{n_t}(x_{it}\beta) + \ln\Gamma\left(\sum_{t=1}^{n_i}(y_{it}) + 1\right)\right.\\ \left. - \ln\Gamma\left(\sum_{t=1}^{n_i}(x_{it}\beta) + \sum_{t=1}^{n_t}(y_{it}\beta)\right)\right]\end{aligned} \qquad (10.8)$$

or

$$\begin{aligned}\mathcal{L}(\beta; y_{it}) = \sum_{t=1}^{n_t}\{\ln\Gamma(\exp(x_{it}\beta) + y_{it}) - \ln\Gamma(\exp(x_{it}\beta)) - \ln\Gamma(y_{it} + 1)\}\\ + \ln\Gamma\left[\sum_{t=1}^{n_t}(x_{it}\beta) + \ln\Gamma\left(\sum_{i=1}^{n_i}(y_{it} + 1)\right) - \ln\Gamma\left(\sum_{i=1}^{n_i}(x_{it}\beta) + \sum_{t=1}^{n_t}(y_{it})\right)\right]\end{aligned}$$

Complete derivations of both the Poisson and negative binomial log likelihood functions can be found in Cameron and Trevedi (1998), and Hardin and Hilbe (2003).

We next model the same data using conditional fixed-effects as we did with unconditional fixed-effects.

```
. xtpoisson accident op co_65_69 — co_75_79, nolog i(ship)
exposure(service) fe

Conditional fixed-effects            Number of obs       =      34
Poisson regression
Group variable (i)   :  ship         Number of groups    =       5
                                     Obs per group: min  =       6
                                                    avg  =     6.8
                                                    max  =       7
                                     Wald chi2(4)        =   48.44
Log likelihood      =  -54.641859    Prob > chi2         =  0.0000

------------------------------------------------------------------------
accident |     Coef.   Std. Err.    z     P>|z|   [95% Conf.  Interval]
---------+--------------------------------------------------------------
      op |   .384467   .1182722   3.25   0.001    .1526578    .6162761
co_65_69 |  .6971405   .1496414   4.66   0.000    .4038487    .9904322
co_70_74 |  .8184266   .1697737   4.82   0.000    .4856764    1.151177
co_75_79 |  .4534267   .2331705   1.94   0.052   -.0035791    .9104324
 service | (exposure)
------------------------------------------------------------------------
```

The associated AIC statistic is 3.567, which is a full one unit lower in value than the unconditional model. Compare the above list of parameter estimates and standard errors output with that of the unconditional results:

```
. glm accident op co_65_69 — co_75_79 ship2-ship5,fam(poi)
lnoffset(service)
-------------------------------------------------------------------------
accident |    Coef.   OIM Std. Err.     z    P>|z|  [95% Conf.  Interval]
---------+---------------------------------------------------------------
      op |   .384467    .1182722     3.25  0.001    .1526577   .6162762
co_65_69 |  .6971404    .1496415     4.66  0.000    .4038485   .9904323
co_70_74 |  .8184266    .1697737     4.82  0.000    .4856763   1.151177
co_75_79 |  .4534266    .2331705     1.94  0.052   -.0035792   .9104324
   ship2 | -.5433445    .1775899    -3.06  0.002   -.8914144  -.1952747
   ship3 | -.6873984    .3290468    -2.09  0.037   -1.332318  -.0424786
   ship4 | -.0759617    .2905787    -0.26  0.794   -.6454854    .493562
   ship5 |  .3255795    .2358794     1.38  0.168   -.1367356   .7878946
   _cons | -6.405901    .2174441   -29.46  0.000   -6.832084  -5.979719
 service |(exposure)
-------------------------------------------------------------------------
```

The first thing that can be noticed is that the parameter estimates and standard errors for the unconditional and conditional fixed-effects Poisson models are identical, though this equality is not an equivalence of the two approaches. The conditional fixed-effects Poisson model does not include a constant, whereas the unconditional does. Of interest to note as well is the fact that the respective log-likelihoods differ (−54.64 to −68.28). We previously pointed out that the AIC statistics differ as well (3.57 to 4.55). We may conclude from this that, although the estimates and standard errors are the same, the two models intrinsically differ, with the preferred fit being that of the conditional fixed-effects Poisson model.

We now turn to modeling the same data using the conditional fixed-effects negative binomial model. It can be displayed as:

```
. xtnbreg accident op co_65_69 – co_75_79, nolog i(ship)
exposure(service) fe

Conditional FE negative              Number of obs       =        34
binomial regression
Group variable (i)   :   ship        Number of groups    =         5
                                     Obs per group: min  =         6
                                                    avg  =       6.8
                                                    max  =         7
                                     Wald chi2(4)        =     34.81
Log likelihood       =   -53.08425   Prob > chi2         =    0.0000
==================================================================
accident     Coef.    Std. Err.    z     P>|z|  [95% Conf.   Interval]
------------------------------------------------------------------
      op |  .3835077  .1423269    2.69  0.007     .104552    .6624634
co_65_69 |  .6772064  .1769717    3.83  0.000    .3303482    1.024065
co_70_74 |  .8186325  .2009683    4.07  0.000     .424742    1.212523
co_75_79 |  .4774847  .2773654    1.72  0.085   -.0661414    1.021111
   _cons | -5.995012  .8205518   -7.31  0.000   -7.603264    -4.38676
 service | (exposure)
---------------------------------------------------------------------
AIC Statistic =3.476
==================================================================
```

Compare the above with the unconditional model:

```
. glm accident op co_65_69-co_75_79 ship2-ship5, fam(nb
0.0000000253) lnoffset(service)

Log likelihood = -68.28077281                AIC = 4.545928
=====================================================================
accident |    Coef.   OIM Std. Err.     z    P>|z|  [95% Conf. Interval]
---------------------------------------------------------------------
      op |   .384467    .1182722       3.25 0.001    .1526577   .6162762
co_65_69 |  .6971404    .1496415       4.66 0.000    .4038485   .9904323
co_70_74 |  .8184266    .1697737       4.82 0.000    .4856763   1.151177
co_75_79 |  .4534266    .2331705       1.94 0.052   -.0035792   .9104324
   ship2 | -.5433445    .1775899      -3.06 0.002   -.8914144  -.1952747
   ship3 | -.6873984    .3290468      -2.09 0.037   -1.332318  -.0424786
   ship4 | -.0759617    .2905787      -0.26 0.794   -.6454854    .493562
   ship5 |  .3255795    .2358794       1.38 0.168   -.1367356   .7878946
   _cons | -6.405901    .2174441     -29.46 0.000   -6.832084  -5.979719
 service | (exposure)
=====================================================================
```

The conditional and unconditional fixed-effects negative binomial models do not normally have the same parameter estimates. However, in this case the data are very close to Poisson, so the results will not be too dissimilar. Note though that, unlike the Poisson, both the conditional and unconditional negative binomial models have a constant in the model. And, as previously indicated, the conditional version does not have a value for α. For a discussion on unconditional fixed effects constants see Greene (2003).

Unfortunately it has been discovered that the conditional fixed-effects negative binomial model is not a true fixed-effects model since it fails to control for all of its predictors. In addition, the α parameter that is conditioned out of the log-likelihood does not correspond to the different intercepts in the decomposition of μ. Allison and Waterman (2002) provide a full discussion, together with alternative models. The negative multinomial model has been suggested as an alternative for the conditional negative binomial. However, the negative multinomial produces the same estimators as a conditional Poisson, so does not provide any additional capability for handling overdispersion over what is available with Poisson options. The other foremost alternative is to revert to the unconditional negative binomial model. In fact, they recommend that the unconditional negative binomial be used rather then the conditional. But, as previously discussed, it should also be accompanied by scaling the standard errors by the Pearson Chi2 dispersion. If this strategy is unsatisfactory, then one should consider using other panel models; e.g. random-effects models or GEE models.

10.3 Random-effects negative binomial

Random-effects models begin with the same notation as fixed-effects models in that a heterogeneity parameter is added to the linear predictor. Moreover, the fixed-effects parameter, δ, is now considered to be an iid random parameter rather than a fixed parameter. It is derived from a known probability distribution. In the case of Poisson, the random parameter can follow the usual Gaussian distribution, the gamma distribution, or the inverse Gaussian distribution. Gamma is the preferred random distribution to use since it is the conjugate prior to Poisson. The gamma distribution also allows an analytic solution of the integral in the likelihood. Other random distributions do not have these favorable features.

We shall use the term ν rather than δ for depicting the random parameter for random-effects count models. In so doing we shall be consistent with common terminology. We shall also use the standard GLM term μ rather than λ for the Poisson and negative binomial fitted value. λ is commonly found in the literature on count response models. But as with our choice of using ν, we shall use the term μ to maintain consistency for all count models that in some respect emanate from a GLM background. The framework for the random-effects Poisson is

$$\ln(\mu_{ik}) = \beta x_{ik} + \nu_i \tag{10.9}$$

with $\nu_i = \nu + \varepsilon_i$

Following the derivation of the random gamma effects Poisson model by Hausman, Hall, and Griliches (1984), we assume a random multiplicative effect on μ specified as

$$\begin{aligned}\Pr(y_{ik}; \nu_i, x) &= \{\Pi(\mu_{ik}\nu_i)^y / y_{ik}!\} \exp(-\Sigma\mu_{ik}\nu_i) \\ &= \left(\Pi\left(\mu_{ik}^y\right)/y_{ik}!\right) \exp(-\nu_i \Sigma\mu_{ik})\nu^{\Sigma y}\end{aligned} \tag{10.10}$$

Summing of subject-specific observations are over panels with a mean given for each panel of

$$\mu_{ik} = \exp(x_{ik}\beta) \tag{10.11}$$

where each panel has separately defined means given as

$$\nu_i \mu_{ik} = \exp(x_{ik}\beta + \eta_{ik}) \tag{10.12}$$

With ν following a gamma distribution with a mean of one and a variance of θ, we have the mixture

$$\Pr(\nu_i, \mu_{ik}) = \theta^\theta / \Gamma(\theta)\nu_i^{\theta-1} \exp(-\theta\nu_i) \quad \Pi \exp(-\nu_i \mu_{ik})(\nu_i \mu_{ik})^y / y_{ik}!) \tag{10.13}$$

where the terms prior to the product sign specify the gamma distributed random component and the terms from the product sign to the right provide the Poisson probability function. This mixture is of the same structural form as we derived for the NB-1 probability function in Chapter 5.

Each panel is independent of one another, with their joint density combined as the product of the individual panels. The log-likelihood for the gamma distributed Poisson random effects model can be calculated by integrating over ν_i. The result is:

RANDOM EFFECTS POISSON WITH GAMMA EFFECT

$$\begin{aligned}\mathcal{L}(\beta; y_{it}) = \sum_{i=1}^{n_i} \Bigg\{ & \ln\Gamma\left(\theta + \sum_{k=1}^{n_k} y_{ik}\right) - \ln\Gamma(\theta) \\ & - \sum_{k=1}^{n_k}(\ln\Gamma(y_{ik}+1) + \theta\ln(u_i) + \left(\sum_{k=1}^{n_k} y_{ik}\right)\ln(1-u_i) \\ & - \left(\sum_{k=1}^{n_k} y_{ik}\right)\ln\left(\sum_{k=1}^{n_k}(\exp(x_{ik}\beta))\right) + \sum_{k=1}^{n_k}(y^*(x_{ik}\beta))\Bigg\}\end{aligned} \tag{10.14}$$

where $\theta = 1/\nu$
and $u_i = \theta/(\theta + \Sigma(\exp(x_{ik}\beta))$

We shall use the same data that were used for examining fixed-effects models for the examples of random-effects Poisson and negative binomial. Random effects Poisson, with a gamma effect, is shown below as:

```
. xtpoisson accident op co_65_69-co_75_79, nolog exposure(service)
i(ship) re

Random-effects Poisson regression   Number of obs       =      34
Group variable (i)   :   ship       Number of groups    =       5
Random effects u_i ~ Gamma          Obs per group: min  =       6
                                                   avg  =     6.8
                                                   max  =       7
                                    Wald chi2(4)        =   50.90
Log likelihood      =  -74.811217   Prob > chi2         =  0.0000

=================================================================
accident |    Coef.    Std. Err.     z     P>|z|  [95% Conf.  Interval]
---------+-------------------------------------------------------
      op |  .3827453  .1182568    3.24  0.001    .1509662    .6145244
co_65_69 |  .7092879  .1496072    4.74  0.000    .4160633    1.002513
co_70_74 |  .8573273  .1696864    5.05  0.000    .5247481    1.189906
co_75_79 |  .4958618  .2321316    2.14  0.033    .0408922    .9508313
   _cons | -6.591175  .2179892  -30.24  0.000   -7.018426   -6.163924
 service | (exposure)
---------+-------------------------------------------------------
/lnalpha | -2.368406  .8474597                    -4.029397   -.7074155
---------+-------------------------------------------------------
   alpha |  .0936298  .0793475                     .0177851    .4929165
=================================================================

Likelihood-ratio test of alpha=0: chibar2(01) = 10.61
Prob>=chibar2 = 0.001

AIC Statistic = 4.754
```

The likelihood ratio tests whether the data are better modeled using a panel structure or whether a pooled structure is preferred. Here we find that the random effects (panel) parameterization is preferred over the pooled, or standard, Poisson model.

It is interesting to compare this output with that of a NB-1 model on the same data. Recall that mixing the gamma random parameter with the Poisson probability function resulted in a NB-1 PDF. Of course the NB-1 PDF does not account for the panel structure of the data as does the gamma distributed random effects Poisson. However, because of the base similarity of the two models, we should expect that the outputs of the respective models will be similar, but not identical. This suspicion is indeed confirmed. Note also that the output above specifies “alpha” as the heterogeneity parameter. It is the same as what we have referred to as ν. Interpret δ in the same manner.

CONSTANT NEGATIVE BINOMIAL (NB-1)

```
. nbreg accident op co_65_69-co_75_79, exposure(service)
cluster(ship) disp(constant)

Negative binomial regression            Number of obs    =    34
Dispersion              = constant      Wald chi2(2)     =     .
Log pseudolikelihood    = -74.801716    Prob > chi2      =     .
                      (Std. Err. adjusted for 5 clusters in ship)
```

```
------------------------------------------------------------------------------
             |               Robust
    accident |      Coef.   Std. Err.      z    P>|z|  [95% Conf.  Interval]
-------------+----------------------------------------------------------------
          op |   .3824838   .0903809     4.23  0.000    .2053404    .5596272
    co_65_69 |   .7174666   .0996523     7.20  0.000    .5221517    .9127814
    co_70_74 |   1.025627   .2156908     4.76  0.000     .602881    1.448373
    co_75_79 |   .7266669   .1996568     3.64  0.000    .3353468    1.117987
       _cons |  -6.924931   .0522819  -132.45  0.000   -7.027402   -6.822461
     service | (exposure)
-------------+----------------------------------------------------------------
    /lndelta |  -.1042511   .4995717                   -1.083394    .8748916
-------------+----------------------------------------------------------------
       delta |   .9009991   .4501137                     .338445    2.398615
-------------+----------------------------------------------------------------
AIC Statistic = 4.635
------------------------------------------------------------------------------
```

We next compare the above with the standard NB-2.

NEGATIVE BINOMIAL (NB-2)

```
. nbreg accident op co_65_69-co_75_79, exposure(service)
cluster(ship)

Negative binomial regression                  Number of obs    =     34
Dispersion                =   mean            Wald chi2(2)     =      .
Log pseudolikelihood      =   -77.874185      Prob > chi2      =      .
                           (Std. Err. adjusted for 5 clusters in ship)
------------------------------------------------------------------------------
             |               Robust
    accident |      Coef.   Std. Err.      z    P>|z|  [95% Conf.   Interval]
-------------+----------------------------------------------------------------
          op |   .3536459   .2161704     1.64  0.102   -.0700402     .777332
    co_65_69 |   1.012518   .6365455     1.59  0.112   -.2350879    2.260125
    co_70_74 |   1.255125   .3774548     3.33  0.001    .5153274    1.994923
    co_75_79 |   .7595303   .2988691     2.54  0.011    .1737576    1.345303
       _cons |  -6.933539   .0955349   -72.58  0.000   -7.120784   -6.746294
     service | (exposure)
-------------+----------------------------------------------------------------
    /lnalpha |  -2.037569   1.517455                   -5.011727    .9365884
-------------+----------------------------------------------------------------
       alpha |   .1303451   .1977929                    .0066594    2.551263
-------------+----------------------------------------------------------------
AIC Statistic = 4.816
------------------------------------------------------------------------------
```

When deriving the random effects negative binomial, we begin with the same Poisson–gamma mixture as Equation (10.13). By rearranging terms and not integrating out ν as we did for the Poisson, we have

$$f(y_{it};\mu_{it}\nu_i) = \prod_{i=1}^{n_i} \frac{\Gamma(\mu_{ik}+y_{ik})}{\Gamma(\mu_{ik})\Gamma(y_{ik}+1)}\left(\frac{1}{1+\nu_i}\right)^{\mu_{ik}}\left(\frac{\nu_i}{1+\nu_i}\right)^{y_{ik}} \tag{10.15a}$$

which is the panel structure form of the NB-1 model. For the random effect we select the beta distribution, which is the conjugate prior of the negative binomial, as gamma was the conjugate prior of the Poisson. With the dispersion defined as the variance divided by the mean, or $1 + \nu$, it is stipulated that the inverse dispersion is distributed following a Beta distribution. We have, therefore

$$\nu_i/(1+\nu_i) \quad \sim \text{Beta}(a, b)$$

which layers the random panel effect onto the negative binomial model. Deriving the probability and the log-likelihood function results in the following forms of the function

RANDOM EFFECTS NEGATIVE BIOMIAL WITH BETA EFFECT PDF

$$\begin{aligned} f(y_{it}; x_{it}\beta, a, b) = {} & \frac{\Gamma(a+b) + \Gamma\left(a + \sum_t \exp(x_{it}\beta)\right)\Gamma\left(b + \sum_t y_{it}\right)}{\Gamma(a)\Gamma(b)\Gamma\left(a+b+\sum_t \exp(x_{it}\beta) + \sum_t y_{it}\right)} \\ & \prod_{t=1}^{n_i} \frac{\Gamma(\exp(x_{it}\beta)+y_{it})}{\Gamma(\exp(x_{it}\beta))\Gamma(y_{it}+1)} \end{aligned} \tag{10.15b}$$

and

RANDOM EFFECTS NEGATIVE BINOMIAL WITH BETA EFFECT LOG-LIKELIHOOD FUNCTION

$$\begin{aligned} \mathcal{L}(\beta; y_{it}, a, b) = {} & \sum_{i=1}^{n_i} \ln\Gamma(a+b) + \ln\Gamma\left(a + \sum_{k=1}^{n_k}(\exp(x_{ik}\beta))\right) \\ & + \ln\Gamma\left(b + \sum_{k=1}^{n_k} y_{ik}\right) - \ln\Gamma(a) - \ln\Gamma(b) \\ & - \ln\Gamma\left(a + b + \sum_{k=1}^{n_k}(\exp(x_{ik}\beta)) + \sum_{k=1}^{n_k} y_{ik}\right) \\ & + \sum_{t=1}^{n_t}(\ln\Gamma((\exp(x_{it}\beta)) + y_{it}) \\ & - \ln\Gamma(y_{it}+1) - \ln\Gamma(\exp(x_{it}\beta))) \end{aligned} \tag{10.16}$$

Derivatives of the conditional fixed-effects and random-effects Poisson and negative binomial models are given in Greene (2006).

Output of the beta distributed random effect negative binomial is shown below for the data we have used in this chapter.

```
. xtnbreg accident op co_65_69-co_75_79, exposure(service)
i(ship) re

Random-effects Negative binomial    Number of obs        =      34
regression
Group variable (i)   :  ship        Number of groups     =       5
Random effects u_i ~ Beta           Obs per group: min   =       6
                                                   avg   =     6.8
                                                   max   =       7
                                    Wald chi2(4)         =   37.15
Log likelihood      =  -73.222498   Prob > chi2          =  0.0000
```

```
------------------------------------------------------------------------------
accident     Coef.     Std. Err.    z     P>|z|  [95% Conf.   Interval]
------------------------------------------------------------------------------
      op |  .3815599  .1426935    2.67  0.007    .1018857     .661234
co_65_69 |  .6935116  .1777982    3.90  0.000    .3450335     1.04199
co_70_74 |  .8766236  .2024878    4.33  0.000    .4797548    1.273492
co_75_79 |  .5452717   .277925    1.96  0.050    .0005488    1.089995
   _cons | -6.039451  .8228179   -7.34  0.000   -7.652145   -4.426758
 service | (exposure)
---------+--------------------------------------------------------------------
   /ln_r |  3.641897  1.097047                    1.491725    5.792069
   /ln_s |  3.029242  1.108291                    .8570312    5.201453
---------+--------------------------------------------------------------------
       r |  38.16417  41.86788                    4.444755    327.6904
       s |  20.68155  22.92118                    2.356155    181.5358
------------------------------------------------------------------------------

Likelihood-ratio test vs. pooled: chibar2(01) = 3.16
Prob>=chibar2 = 0.038
AIC Statistic = 4.660
```

A likelihood ratio test accompanies the output, testing the random-effects panel estimator with the pooled NB-1, or constant dispersion, estimator. Here the random-effects model is preferred. *r* and *s* refer to the beta distribution values for the *a* and *b* parameters respectively. Note the extremely wide confidence intervals.

We shall use another example of a random effects model. It is not only a good random effects example in its own right, but it will prove useful when discussing generalized estimating equations.

The data come from Thall and Vail (1990) and are used in Hardin and Hilbe (2003). Called the *progabide* data set, the data are from a panel study of seizures in patients with epilepsy. Four successive two-week counts of seizures were taken for each patient. The response is *seizure*, with explanatory predictors consisting of the *progabide* treatment (1/0), a follow-up indicator called *time* (1/0), and an interaction of the two, called *timeXprog*. An offset, called *Period*, is given for weeks in the study, which are either two or eight. Since *Period* is converted by a natural log, the two values of *lnPeriod* are 2.079442 and 0.6931472. There are 295 observations on 59 epileptic patients (panels), with five observations, *t*, each.

Results of modeling the data are:

GAMMA DISTRIBUTED RANDOM EFFECTS POISSON

```
. xtpoisson seizures time progabide timeXprog, nolog
offset(lnPeriod) re i(t)

Random-effects Poisson regression   Number of obs      =      295
Group variable (i)   :   t          Number of groups   =        5
Random effects u_i ~ Gamma          Obs per group: min =       59
                                                   avg =     59.0
                                                   max =       59
                                    Wald chi2(3)       =     4.42
Log likelihood      =  -2318.1938   Prob > chi2        =   0.2192
```

```
-----------------------------------------------------------------------------
  seizures |   Coef.    Std. Err.    z     P>|z|  [95% Conf.   Interval]
-----------+-----------------------------------------------------------------
      time |   .111836  .0634911    1.76  0.078   -.0126042    .2362763
progabide  |  .0275345  .0466847    0.59  0.555   -.0639658    .1190348
timeXprog  | -.1047258  .0650304   -1.61  0.107    -.232183    .0227314
     _cons |  1.347609  .0512546   26.29  0.000    1.247152    1.448066
 lnPeriod  |  (offset)
-----------+-----------------------------------------------------------------
 /lnalpha  | -6.524573   1.629814                   -9.71895   -3.330195
-----------+-----------------------------------------------------------------
     alpha |  .0014669   .0023908                   .0000601    .0357861
-----------------------------------------------------------------------------
Likelihood-ratio test of alpha=0: chibar2(01) = 0.62
Prob>=chibar2 = 0.216
=============================================================================
```

The model does not appear to favor the panel specification of the data. Note as well that alpha is very close to 0, indicating equi-dispersed Poisson data.

BETA DISTRIBUTED RANDOM EFFECTS NEGATIVE BINOMIAL

```
. xtnbreg seizures time progabide timeXprog, nolog
offset(lnPeriod) re i(t)

Random-effects Negative binomial    Number of obs        =     295
regression
Group variable (i)   :  t           Number of groups     =       5
Random effects u_i ~ Beta           Obs per group: min   =      59
                                                   avg   =    59.0
                                                   max   =      59
                                    Wald chi2(3)         =    6.19
Log likelihood       =  -1005.9032  Prob > chi2          =  0.1026

=============================================================================
  seizures |   Coef.    Std. Err.    z     P>|z|  [95% Conf.   Interval]
-----------+-----------------------------------------------------------------
      time |  .3111093  .1411709    2.20  0.028    .0344195    .5877992
progabide  |   .060761  .1524915    0.40  0.690   -.2381168    .3596388
timeXprog  | -.2456651  .1931765   -1.27  0.203   -.6242841    .1329539
     _cons | -1.111632  .1432822   -7.76  0.000    -1.39246   -.8308041
 lnPeriod  |  (offset)
-----------+-----------------------------------------------------------------
     /ln_r |  16.95649  573.0725                   -1106.245    1140.158
     /ln_s |  19.32986  573.0725                   -1103.872    1142.531
-----------+-----------------------------------------------------------------
         r |  2.31e+07  1.33e+10                           0           .
         s |  2.48e+08  1.42e+11                           0           .
-----------------------------------------------------------------------------
Likelihood-ratio test vs. pooled: chibar2(01) = 0.00
Prob>=chibar2 = 1.000
=============================================================================
```

Again, the panel structure of the data is questionable, and not supported by the model. However, this conclusion holds only with respect to the subject-specific parameterization. The same is the case for the random-effects Poisson model above. A thorough discussion of the derivation of the random-effects Poisson and negative binomial models can be found in Frees (2004) and Greene (2006).

A disadvantage of a random-effects model is that it assumes that the subject-specific effects are uncorrelated with other predictors. The Hausman test is commonly used to evaluate whether data should be modeled using a fixed- or a random-effects model. The test is based on work done by Mundlak (1978) who argued that the fixed-effects model is more robust than the random-effects model to important predictors left out of the model. That the subject-specific effects are not correlated highly with model predictors is specified as the null hypothesis. See Frees (2004), p 247, for additional discussion.

Random effects estimators are more efficient than fixed-effects estimators when the data come from within a larger population of observations, as well as when there are more panels in the data. Data coming from a smaller complete data set, with relatively few panels, prefer the fixed-effects estimator.

Random-effects models are subject-specific models in that the log-likelihood models individual observations rather than the average of panels, or marginal distribution. GEEs are population averaging models; care must be taken when interpreting GEE against random effects model output, as we shall turn to next.

10.4 Generalized estimating equation

10.4.1 The GEE algorithm

Generalized estimating equation (GEE) refers to a population averaging panel method first proposed by Liang and Zeger in 1986. It is developed as an extension to the standard generalized linear models algorithm. Unlike the random-effects model, which is subject-specific, GEE is a population-averaged approach in which the marginal effects of the model are averaged across individuals. Essentially, GEE models the average response of individuals sharing the same predictors across all of the panels.

The GEE algorithm is structured such that observations are grouped in panels, in a similar manner to fixed-effects and random-effects models. At the heart of the model specification, the variance function is factored to include an identity matrix operating as a within-panel correlation structure. This panel form of the variance function appears as

$$V(\mu_i) = [D(V(\mu_{ik}))^{1/2} I_{(nxn)} D(V(\mu_{ik}))^{1/2}]_{nxn} \tag{10.17}$$

where $V(\mu_{ik})$ is the GLM variance function defined from the family being modeled. For example, the variance function of the Poisson family is μ and the variance of the NB-2 model is $\mu + \alpha\mu^2$. This structure (represented by the identity matrix) is called the independent correlation structure.

The benefit of the GEE approach is that the identity matrix, which is sandwiched between the factored GLM variance functions, can be replaced by a parameterized matrix containing values other than one and zero. The structure of values that are substituted into this alternative matrix define the various GEE correlation structures. These include:

FOREMOST GEE CORRELATION STRUCTURES

```
1 Independent           2 Exchangeable       3 Unstructured
4 Autoregressive        5 Stationary         6 Non-stationary
```

The most commonly used correlation structure is the exchangeable, which we shall later describe in more detail. All of the structures define constraints on the values to be estimated. Those values are estimated from Pearson residuals obtained using the regression parameters. Pearson residuals are defined, in panel format, as

$$r_{ik} = \Sigma(y_{ik} - \mu_{ik})^2 / V(\mu_{ik}). \tag{10.18}$$

The Poisson Pearson residual is defined as $\Sigma(y_{ik} - \mu_{ik})^2/\mu_{ik}$. The exchangeable correlation is then defined as, without subscripts (see Hardin and Hilbe, 2007),

$$\alpha = \frac{1}{\phi}\Sigma\left\{\frac{\Sigma\Sigma rr - \Sigma r^2}{n(n-1)}\right\} \tag{10.19}$$

The exchangeable correlation structure has also been referred to as the compound symmetry matrix and the equal correlation structure. All off-diagonal values have a single scalar constant.

Other correlation matrices are defined in different manners, depending on the purport of the structure. However, all are inserted into the variance function as

$$V(\mu_i) = [D(V(\mu_{ik}))^{1/2} R(a)_{(nxn)} D(V(\mu_{ik}))^{1/2}]_{nxn} \tag{10.20}$$

The GEE algorithm begins by estimating a model from the member families, e.g. Poisson. After the initial iteration, the Pearson residuals are calculated (Equation (10.18)) and put into the formula for calculating R(a) (Equation (10.19)). $R(a)$ is then inserted into Equation (10.20) in place of the identity matrix. The updated variance function is then used as such in the second iteration. Again, another updated variance function is calculated, and so on until the algorithm converges as does any GLM model.

Since the resulting GEE model is not based on a pure probability function, the method is called a quasi-likelihood model. Recall that we used a similar appellation for instance when an otherwise GLM variance function is multiplied by either a constant, or by another non-constant variable. In either case the working likelihood function is not based on a probability function.

An example GEE Poisson model using the *Progabide* data can be shown as:

GEE POISSON (EXCHANGEABLE)

```
. xtpoisson seizures time progabide timeXprog, nolog
offset(lnPeriod) pa i(t)

GEE population-averaged model     Number of obs        =      295
Group variable:                t  Number of groups     =        5
Link:                        log  Obs per group: min   =       59
Family:                  Poisson                 avg   =     59.0
Correlation:       exchangeable                  max   =       59
                                  Wald chi2(3)         =    34.21
Scale parameter:               1  Prob > chi2          =   0.0000
==================================================================
 seizures |   Coef.   Std. Err.    z    P>|z|  [95% Conf. Interval]
----------+-------------------------------------------------------
     time |  .111836  .0357489   3.13  0.002    .0417694  .1819026
progabide | .0275345   .047044   0.59  0.558   -.0646701  .1197391
timeXprog |-.1047258  .0655263  -1.60  0.110    -.233155  .0237034
    _cons | 1.347609  .0259747  51.88  0.000      1.2967  1.347609
 lnPeriod | (offset)
==================================================================
```

If you recall from the last section, the random-effects Poisson model on the same data resulted in identical parameter estimates to the above GEE model. The table of estimates and related statistics are displayed below.

GAMMA DISTRIBUTED RANDOM-EFFECTS POISSON

```
. xtpoisson seizures time progabide timeXprog, nolog
offset(lnPeriod) re i(t)

Random-effects Poisson regression   Number of obs        =      295
Group variable (i)   :  t           Number of groups     =        5
Random effects u_i ~ Gamma          Obs per group: min   =       59
                                                   avg   =     59.0
                                                   max   =       59
                                    Wald chi2(3)         =     4.42
Log likelihood      = −2318.1938    Prob > chi2          =   0.2192
====================================================================
 seizures |   Coef.   Std. Err.    z    P>|z|  [95% Conf.  Interval]
----------+---------------------------------------------------------
     time |  .111836  .0634911   1.76  0.078   −.0126042   .2362763
progabide | .0275345  .0466847   0.59  0.555   −.0639658   .1190348
timeXprog |-.1047258  .0650304  -1.61  0.107    -.232183   .0227314
    _cons | 1.347609  .0512546  26.29  0.000    1.247152   1.448066
 lnPeriod | (offset)
----------+---------------------------------------------------------
 /lnalpha |-6.524573  1.629814                  -9.71895  -3.330195
----------+---------------------------------------------------------
    alpha | .0014669  .0023908                  .0000601   .0357861
----------+---------------------------------------------------------
Likelihood-ratio test of alpha=0: chibar2(01) = 0.62
Prob>=chibar2 = 0.216
AIC Statistic = 15.750
====================================================================
```

Identical parameter estimate values between a gamma distributed Poisson random effects model and the GEE Poisson with exchangeable correlation structure do not normally occur. Note also that the standard errors of *progabide* and *timeXprog* are similar, but not identical between the two models. This occurrence happens only if both the variance of the random effect is zero and the exchangeable correlation parameter is zero. The near zero values of alpha and its standard error above indicates that this situation occurs here. The values of the parameter estimates diverge if we add another predictor.

We next model the data using a negative binomial model with an exchangeable correlation structure.

GEE NEGATIVE BINOMIAL (EXCHANGEABLE)

```
. xtnbreg seizures time progabide timeXprog, nolog
offset(lnPeriod) pa i(t)

GEE population-averaged model       Number of obs        =     295
Group variable:                t    Number of groups     =       5
Link:                        log    Obs per group: min   =      59
Family:                 negative                   avg   =    59.0
                   binomial(k=1)
Correlation:        exchangeable                   max   =      59
                                    Wald chi2(3)         =    1.52
Scale parameter:               1    Prob > chi2          =  0.6775

------------------------------------------------------------------
 seizures |   Coef.    Std. Err.    z    P>|z|  [95% Conf. Interval]
----------+-------------------------------------------------------
     time |  .111836  .1663624    0.67  0.501   -.2142283   .4379004
progabide | .0275345   .266861    0.10  0.918   -.4955036   .5505725
timeXprog |-.1047258  .3009084   -0.35  0.728   -.6944954   .4850439
    _cons | 1.347609  .1476073    9.13  0.000    1.058304   1.636914
 lnPeriod | (offset)
------------------------------------------------------------------
```

Again we find that the heterogeneity parameter is not estimated as a separate parameter. It is apportioned across panels as was the conditional fixed-effects negative binomial.

Note that an AIC statistic is not associated with either the Poisson or negative binomial GEE models. Since the model is a quasi-likelihood model, the software uses the deviance statistic as the basis of convergence. The likelihood function is not directly calculated.

This is done for theoretical purity; a quasi-likelihood value can be calculated, and a (quasi)-AIC statistic calculated. The majority of GEE implementations provide the log-likelihood function as output.

We shall now summarize the various major correlation structures and provide some insight as to when each should best be used.

10.4.2 Correlation structures

Although GEE models are robust to the use of incorrect correlation structures, it is nevertheless preferable to select the structure most appropriate to the data or to the goal of the study. One may check the observed correlation matrix if there is no known reason to select a specific matrix based on previous clinical studies. This might not provide a definitive solution as to which is the best correlation structure for the data, but it can nevertheless inform you about which type of structure is not appropriate.

The QIC statistic can be used to quantitatively decide on the preferred correlation structure. The statistic, created by Pan (2001), is called the *quasi-likelihood under the independence model information criterion*. It is similar to the AIC statistic, but tests correlation structures within the scope of generalized estimating equations. The QICu statistic, also developed by Pan (2001a), helps the user to decide on the best subset of model predictors for a particular correlation structure.

If two or more correlation structures result in nearly the same QIC statistic, and no other factor can be used to help decide which structure to use, the preferred choice is to employ the simplest structure – one with the least parameters – which fits that data. Hardin and Hilbe (2003) provide complete details regarding the use of both statistics.

There are a few summary guidelines that may be helpful to deciding which correlation structure to use. If the data relates to first-level clustered data, then it is likely that either the exchangeable or unstructured correlation structure should be used. When panel data relate to measurements over time periods, then the autoregressive, non-stationary, and stationary or m-dependent structures are generally the most appropriate. The autoregressive structure is usually associated with longitudinal time-series data.

A listing of the major correlation structures follows. 5×5 matrix schematics of the respective correlation structures are displayed, together with representative negative binomial GEE models. Only the lower half of the symmetric matrix is completed. I also provide additional guidelines on when each structure should be used.

INDEPENDENT CORRELATION STRUCTURE

SCHEMATIC

```
1
0             1
0             0             1
0             0             0             1
0             0             0             0             1
```

```
. xtgee seizures time progabide timeXprog, offset(lnPeriod) i(t)
fam(nb) corr(indep)

Pearson chi2(295):     520.34   Deviance    = 321.92
Dispersion (Pearson): 1.763876  Dispersion = 1.091264
```

```
------------------------------------------------------------------------
  seizures |   Coef.    Std. Err.     z     P>|z|  [95% Conf.  Interval]
-----------+------------------------------------------------------------
      time |   .111836  .2164262    0.52  0.605   -.3123515   .5360236
 progabide |  .0275345  .2648619    0.10  0.917   -.4915852   .5466542
 timeXprog | -.1047258  .2986543   -0.35  0.726   -.6900775   .4806259
     _cons |  1.347609   .192027    7.02  0.000    .9712432   1.723975
  lnPeriod |  (offset)
------------------------------------------------------------------------
```

```
. xtcorr
         c1         c2         c3         c4         c5         c6
r1     1.0000
r2     0.0000     1.0000
r3     0.0000     0.0000     1.0000
r4     0.0000     0.0000     0.0000     1.0000
r5     0.0000     0.0000     0.0000     0.0000     1.0000
r6     0.0000     0.0000     0.0000     0.0000     0.0000     1.0000
```

The independent correlation structure imposes on the GEE model the same structure as the standard variance–covariance matrix of a generalized linear model. The observations are considered to be independent of one another. The use of this model is to set a base for evaluation of other GEE correlation structures. The structure assumes a zero correlation between subsequent measures of a subject within panels.

Use this correlation structure if the size of the panels are small and if there is evidently no panel effect in the data. Adjust standard errors by a robust or sandwich variance estimator.

EXCHANGEABLE CORRELATION STRUCTURE

SCHEMATIC

```
1
a              1
a              A              1
a              A              a              1
a              A              a              A              1
```

EXAMPLE

```
1
.29            1
.29            .29            1
.29            .29            .29            1
.29            .29            .29            .29            1
```

```
. xtgee seizures time progabide timeXprog, offset(lnPeriod) i(t)
fam(nb) corr(exch)

GEE population-averaged model          Number of obs         =     295
Group variable:                   t    Number of groups      =       5
Link:                           log    Obs per group: min    =      59
Family:                    negative                   avg    =    59.0
                     binomial(k=1)
Correlation:           exchangeable                   max    =      59
                                       Wald chi2(3)          =    1.52
Scale parameter:                  1    Prob > chi2           =  0.6775
```

```
---------------------------------------------------------------------------
 seizures |   Coef.    Std. Err.     z    P>|z|  [95% Conf.  Interval]
----------+----------------------------------------------------------------
     time |  .111836  .1663624    0.67  0.501   -.2142283   .4379004
progabide | .0275345   .266861    0.10  0.918   -.4955036   .5505725
timeXprog |-.1047258  .3009084   -0.35  0.728   -.6944954   .4850439
    _cons | 1.347609  .1476073    9.13  0.000    1.058304   1.636914
 lnPeriod | (offset)
---------------------------------------------------------------------------

. xtcorr

Error structure: exchangeable
Estimated within-t correlation: -0.0152

[correlation structure not shown. The value, -0.0152, is
displayed in each cell of the correlation matrix]
```

The exchangeable correlation structure is the most commonly used structure. It is the default for several of the major commercial software implementations and it is generally the appropriate model to use with clustered or first-level nested data. Moreover, it is assumed that the correlations between subsequent measurements within a panel are the same, irrespective of the time interval. The value of *a*, displayed in the schematic matrix above, is a scalar. It does not vary between panels.

Use this correlation structure when the observations are clustered and not collected over time.

UNSTRUCTURED CORRELATION STRUCTURE

SCHEMATIC

```
1
C1             1
C2             C5             1
C3             C6             C8             1
C4             C7             C9             C10             1
```

EXAMPLE

```
1
.34            1
.29            .28            1
.33            .14            .24            1
.21            .07            .11            .23             1
```

```
. xtgee seizures time progabide timeXprog, offset(lnPeriod) i(t)
fam(nb) corr(unstr) t(id)

GEE population-averaged model        Number of obs      =     295
Group and time vars:           t id  Number of groups   =       5
Link:                           log  Obs per group: min =      59
Family:                    negative                 avg =    59.0
                      binomial(k=1)
Correlation:           unstructured                 max =      59
                                     Wald chi2(3)       =    0.24
Scale parameter:                  1  Prob > chi2        =  0.9710
```

```
  seizures |     Coef.   Std. Err.      z    P>|z|   [95% Conf.  Interval]
-----------+----------------------------------------------------------------
      time | -.0037579   .1964401    -0.02  0.985    -.3887734    .3812576
 progabide | -.0565913   .2412008    -0.23  0.815    -.5293361    .4161535
 timeXprog |  .0025841   .2720929     0.01  0.992    -.5307082    .5358763
     _cons |  1.422683   .1741719     8.17  0.000     1.081312    1.764054
  lnPeriod |  (offset)
```

```
. xtcorr    c1         c2         c3         c4         c5         c6
-----------------------------------------------------------------------------
r1           1.0000
r2           0.1752     1.0000
r3           0.2055     0.2112     1.0000
r4           0.1919     0.1919     0.2351     1.0000
r5          -0.2380    -0.2067    -0.2014    -0.2158     1.0000
r6           0.0868     0.0783     0.0923     0.0833    -0.1237     1.0000
```

In the unstructured correlation structure all correlations are assumed to be different; correlations are freely estimated from the data. This can result in the calculation of a great many correlation coefficients for large matrices. Because it (can have) has a different coefficient for each cell, the correlation structure optimally fits the data. However, it loses efficiency, and hence interpretability, when models have more than about three predictors. The number of coefficients to be estimated is based on the size of the largest panel of observations

$$\#\text{coefficients} = p(p-1)/2$$

where p is the number of observations in the largest panel

Use this correlation structure when the size of the panels is small, there are relatively few predictors, and there are no missing values

AUTOREGRESSIVE CORRELATION STRUCTURE

SCHEMATIC

```
1
C^1            1
C^2            C^1            1
C^3            C^2            C^1            1
C^4            C^3            C^2            C^1            1
```

EXAMPLE

```
1
.48            1
.23            .48            1
.11            .23            .48            1
.05            .11            .23            .48            1
```

```
. xtgee seizures time progabide timeXprog, offset(lnPeriod)
i(t) fam(nb) corr(ar2) t(id)

GEE population-averaged model        Number of obs        =      295
Group and time vars:          t id   Number of groups     =        5
Link:                          log   Obs per group: min   =       59
Family:                   negative                  avg   =     59.0
                     binomial(k=1)
Correlation:                 AR(2)                  max   =       59
                                     Wald chi2(3)         =     0.69
Scale parameter:                 1   Prob > chi2          =   0.8749

------------------------------------------------------------------------
 seizures |   Coef.     Std. Err.    z     P>|z|  [95% Conf.  Interval]
----------+-------------------------------------------------------------
     time |  .1043002   .1750602   0.60   0.551   -.2388115   .4474118
progabide |  .0135439   .2149714   0.06   0.950   -.4077922   .4348801
timeXprog | -.0895488   .2423872  -0.37   0.712   -.5646189   .3855213
    _cons |  1.360433   .1553269   8.76   0.000    1.055998   1.664868
 lnPeriod |  (offset)
------------------------------------------------------------------------

. xtcorr
Error structure             :            AR(2)
Estimated within-t correlations
lag 1                       :            -0.1367
lag 2                       :            -0.0624
lag>2                       :            0
```

Estimated within-t correlation matrix *R*:

```
        c1          c2          c3          c4          c5          c6
r1    1.0000
r2   -0.1367      1.0000
r3   -0.0624     -0.1367      1.0000
r4    0.0205     -0.0624     -0.1367      1.0000
r5    0.0021      0.0205     -0.0624     -0.1367      1.0000
r6   -0.0020      0.0021      0.0205     -0.0624     -0.1367      1.0000
```

The autoregressive correlation structure assumes that there is a marked decrease in correlation coefficient values with the corresponding increase in measurements within panel time intervals. Each off-diagonal from the main diagonal decreases by the square of the previous diagonal. One might consider the decrease in values to be increasing powers of the first off diagonal. Large matrices produce very small coefficient values. The depiction here is true for AR(1), and it is for this example, but not for all AR levels.

Use this correlation structure when the panels are collections of data over time for the same person.

STATIONARY OR m-DEPENDENT CORRELATION STRUCTURE

SCHEMATIC

```
1
C1              1
C2              C1              1
0               C2              C1              1
0               0               C2              C1              1
```

EXAMPLE

```
1
.27            1
.18            .27            1
0              .18            .27            1
0              0              .18            .27            1
```

```
. xtgee seizures time progabide timeXprog, offset(lnPeriod)
i(t) fam(nb) corr(sta2) t(id)

GEE population-averaged model            Number of obs       =     295
Group and time vars:              t id   Number of groups    =       5
Link:                              log   Obs per group: min  =      59
Family:                       negative                  avg  =    59.0
                          binomial(k=1)
Correlation:             stationary(2)                  max  =      59
                                         Wald chi2(3)        =    0.72
Scale parameter:                     1   Prob > chi2         =  0.8694

------------------------------------------------------------------
 seizures |   Coef.    Std. Err.    z    P>|z|  [95% Conf. Interval]
----------+-------------------------------------------------------
     time | .1019404   .1698023   0.60  0.548   -.230866   .4347469
progabide |  .010137   .2087913   0.05  0.961  -.3990864   .4193604
timeXprog |-.0845498   .2354149  -0.36  0.719  -.5459545    .376855
    _cons | 1.364083   .1506621   9.05  0.000    1.06879   1.659375
 lnPeriod | (offset)
------------------------------------------------------------------


. xtcorr
Error structure      :                       stationary(2)
Estimated within-t correlations
lag 1                :                       -0.1368
lag 2                :                       -0.0625
lag>2                :                       0

------------------------------------------------------------------
Estimated within-t correlation matrix R:
        c1           c2           c3           c4           c5
------------------------------------------------------------------
r1       1.0000
r2      -0.1368       1.0000
r3      -0.0625      -0.1368       1.0000
r4       0.0000      -0.0625      -0.1368       1.0000
r5       0.0000       0.0000      -0.0625      -0.1368       1.0000
------------------------------------------------------------------
```

The stationary correlation structure specifies a constant correlation for each off-diagonal. The diagonals are then interpreted as lags or measurements. Correlations *c* lags apart are equal in value to one another, *c* + 1 lags apart are also equal to one another, and so forth until a defined stop, *m*, is reached. Correlations greater than *m* are defined as zero, hence the meaning of *m*-dependent. In larger matrices the correlation structure appears as a band.

Use this correlation structure when the off-diagonals, or lags, are thought of as time intervals.

NON-STATIONARY CORRELATION STRUCTURE

SCHEMATIC

```
1
C1              1
C5              C2              1
0               C6              C3              1
0               0               C7              C4              1
```

EXAMPLE

```
1
.99             1
.71             .84             1
0               .78             .73             1
0               0               .56             .70             1
```

```
. xtgee seizures time progabide timeXprog, offset(lnPeriod)
i(t) fam(nb) corr(non) t(id)

GEE population-averaged model        Number of obs       =     295
Group and time vars:           t id  Number of groups    =       5
Link:                           log  Obs per group: min  =      59
Family:                    negative                 avg  =    59.0
                     binomial(k=1)
Correlation:                  nonst                 max  =      59
                                     Wald chi2(3)        =    6.24
Scale parameter:                  1  Prob > chi2         =  0.1007
-------------------------------------------------------------------
 seizures |    Coef.  Std. Err.      z  P>|z|  [95% Conf. Interval]
----------+--------------------------------------------------------
     time |  3740533  .2033533   1.84  0.066   -.0245119   .7726185
progabide | .0234849  .2562165   0.09  0.927   -.4786902     .52566
timeXprog |-.2675043  .2882112  -0.93  0.353   -.8323879   .2973793
    _cons | 1.446338  .1809807   7.99  0.000    1.091622   1.801053
 lnPeriod | (offset)
-------------------------------------------------------------------

. xtcorr
Estimated within-t correlation matrix R:
        c1           c2           c3           c4            c5
-------------------------------------------------------------------
r1      1.0000
r2      0.3924       1.0000
r3      0.0000       0.4392       1.0000
r4      0.0000       0.0000       0.4667        1.0000
r5      0.0000       0.0000       0.0000       -0.1549       1.0000
-------------------------------------------------------------------
```

The non-stationary correlation structure is the same as the stationary except that the values of each lag or off-diagonal are not constant. Of course, correlation values beyond *m* are all 0.

Some statisticians use the non-stationary correlation structure when they have ruled out the others, but still have a limit to the range of measurement error or lags in the data.

The advantage of GEE over random-effects models relates to the ability of GEE to allow specific correlation structures to be assumed within panels. Parameter estimates are calculated without having to specify the joint distribution of the repeated observations.

In random-effects models a between-subject effect represents the difference between subjects conditional on them having the same random effect. Such models are thus termed conditional or subject-specific models. GEE parameter estimates represent the average difference between subjects, and thus are known as marginal, or population-averaged models. Which to use depends on the context, i.e. on the goals of the study.

Our final section will discuss an emerging area of study – multilevel models. Two of the basic multilevel models are random intercept and random coefficient models. Since other more complex multilevel models are built on their bases, we shall restrict our discussion to these two models.

10.5 Multilevel negative binomial models

Multilevel models are also called hierarchical models, particularly in educational and social science research. The idea behind multilevel models is to model the dependence that exists between nested levels in the data. For instance, we may model visits to the doctor within groups of different hospitals. Unlike GEE models, multilevel models are not based on the framework of generalized linear models. Rather, they are an extension to the random-effects models we discussed in the previous section.

Until recently, nearly all discussion, and application, of multilevel models have been of continuous response models. Binary response models, especially logistic models, were introduced about ten years ago. Only in the last few years have Poisson models been discussed within the domain of multilevel regression. Negative binomial models have been largely ignored. As of this writing, only LIMDEP provides the capability of modeling negative binomial random coefficient models. We shall use it in this section to examine both random intercept and random coefficient models. Random intercept models are considered to be the most elementary, and fundamental, of multilevel models.

10.5.1 Random intercept negative binomial models

Suppose that we are studying student performance on statewide exit examinations. Schools are funded, and school administrators retained, on the basis of how their students perform on the examination. When studying such data it is evident that student performance within schools is more highly correlated than between schools. Likewise, average school performance within various school districts is likely correlated. Here we have student performance nested in schools, which itself is nested within school districts. Correlation effects exist for students within schools, and correlation effects exist for schools within school districts. We may add other levels, such as types of school programs within schools, but three levels are sufficient to see the problem. The levels of dependency must be adjusted by the models if the resulting levels of overdispersion are to be accommodated.

Multilevel models handle nested levels of dependency by allowing the regression coefficients to vary within levels. Because the multilevel algorithm permits the coefficients to vary, statisticians have come to use the more specific term, *random coefficient model*, for this type of multilevel model.

The most basic random coefficient model is one in which only the regression intercept is allowed to vary. Such a model is called a *random intercept model*. It is a subset of random coefficient models.

The random intercept model may be expressed as an equation

$$y_{ik} = \beta_{0i} + \beta_1 X + \varepsilon_{ik} \tag{10.21}$$

y_{it} is the response for individual i in group k, or at time k (which we would change to t). β_{0i} refers to the regression intercept, varying over the individual, i. β_1 X is the coefficient for predictor x, and ε_{ik} is the error term, varying over both individuals and groups.

The following example comes from the German Health data set. It is from 1996, prior to the later reform data that we have used in previous discussions. Model variables include the following:

RESPONSE

```
docvis The number of visits to the doctor by a patient
       recorded over seven time periods.
```

PREDICTORS INCLUDE

```
age      Age (25−64)
female   1=Female; 0=Male
educ     Years of schooling (7−18)
married  1=Married; 0 = Not married
hhninc   Net monthly house income in Marks/10000 (0−30.67)
hsat     Health satisfaction evaluation (0−10)
_groupti periods in which data were recorded (1−7)
```

The majority of LIMDEP procedures are available in a point-and-select format. Random coefficient models, however, require the use of the command line. The code for the model is, followed by output,

```
--> Negb; lhs=docvis; rhs=one, age, female, educ, married,
hhninc, hsat; pds=_groupti; rpm; fcn=one(n); pts=20; halton $
Random Coefficients NegBnReg Model
Maximum Likelihood Estimates
Model estimated: May 29, 2006                    at 05:35:36PM.
Dependent variable                                       DOCVIS
Weighting variable                                         None
Number of observations                                     6209
Iterations completed                                         24
Log likelihood function                               -12723.32
Number of parameters                                          9
Info. Criterion: AIC =                                  4.10125
Finite Sample: AIC =                                    4.10125
Info. Criterion: BIC =                                  4.11101
Info. Criterion:HQIC =                                  4.10463
Restricted log likelihood                             -46669.83
Chi squared                                            67893.02
Degrees of freedom                                            1
Prob[ChiSqd > value] =                                 .0000000
Unbalanced panel has                           887 individuals.
Negative binomial regression model
Simulation based on 20 Halton draws
```

```
                          Standard
Variable   Coefficient     Error     b/St.Er.  P[|Z|>z]   Mean of X

Nonrandom parameters
AGE          .01798893  .00163076    11.031    .0000    44.3351586
FEMALE       .39665379  .03117120    12.725    .0000     .42277339
EDUC        -.04380941  .00829631    -5.281    .0000    10.9408707
MARRIED      .09475447  .04169113     2.273    .0230     .84538573
HHNINC      -.00497977  .08748656     -.057    .9546     .34929630
HSAT        -.21528106  .00596570   -36.087    .0000    6.69640844
                      Means for random parameters
Constant    1.52593656  .13369248     11.414   .0000
         Scale parameters for dists. of random parameters
Constant     .80989785  .01676004     48.323   .0000
           Dispersion parameter for NegBin distribution
ScalParm    1.18747048  .02671618     44.448   .0000

Implied standard deviations of random parameters
Matrix S.D_Beta has 1 rows and 1 columns.
                                 1
1              .80990
```

ScalParm is the negative binomial heterogeneity parameter. At 1.19, overdispersion still appears to remain in the data.

It appears that all predictors but *hhninc* significantly contribute to the model.

10.5.2 Random coefficient negative binomial models

We previously provided the formula for a random intercept model, where the intercept varies over periods. The random coefficient model expands this analysis to allow the regression coefficients to vary. The equation for a model in which the coefficient, β_{1i}, varies, but not the intercept, can be expressed as

$$y_{ik} = \beta_0 + \beta_{1i}x + \varepsilon_{ik} \tag{10.22}$$

We may allow both the intercept and coefficient to vary, giving us

$$y_{ik} = \beta_{0i} + \beta_{1i}x + \varepsilon_{ik} \tag{10.23}$$

Both of the above models are random coefficient models. They are also referred to as random parameter and random slope models. More complex models can exist depending on the number of nested levels in the data.

We now use a random coefficient model on the same data, allowing the coefficient on health satisfaction, *hsat*, to vary. *hsat* has 11 levels.

```
--> Negb; lhs=docvis; rhs=one, age, female, educ, married,
hhninc, hsat; pds=_groupti; rpm; fcn=one(n), hsat(n); cor;
pts=20; halton $
Random Coefficients NegBnReg Model
Maximum Likelihood Estimates
Model estimated: May 29, 2006                      at 05:38:24PM.
Dependent variable                                         DOCVIS
Weighting variable                                           None
Number of observations                                       6209
Iterations completed                                           28
Log likelihood function                                 -12694.26
Number of parameters                                           11
Info. Criterion: AIC =                                    4.09253
Finite Sample: AIC =                                      4.09254
Info. Criterion: BIC =                                    4.10446
Info. Criterion:HQIC =                                    4.09666
Restricted log likelihood                               -46669.83
Chi squared                                              67951.15
Degrees of freedom                                              3
Prob[ChiSqd > value] =                                   .0000000
Unbalanced panel has                             887 individuals.
Negative binomial regression model
Simulation based on 20 Halton draws
```

```
                          Standard
Variable  Coefficient      Error      b/St.Er.  P[|Z|>z]   Mean of X
---------------------------------------------------------------------
                        Nonrandom parameters
AGE        .01680197    .00159059     10.563    .0000    44.3351586
FEMALE     .37039825    .02990835     12.384    .0000     .42277339
EDUC      -.03147513    .00813808     -3.868    .0001    10.9408707
MARRIED    .09847528    .04141046      2.378    .0174     .84538573
HHNINC     .01194585    .08915694       .134    .8934     .34929630
                     Means for random parameters
Constant  1.65242649    .12955696     12.754    .0000
HSAT      -.24630006    .00653340    -37.699    .0000    6.69640844
                Diagonal elements of Cholesky matrix
Constant   .78345690    .04466106     17.542    .0000
HSAT       .11609985    .00249648     46.505    .0000
```

```
                          Standard
Variable  Coefficient    Error      b/St.Er.  P[|Z|>z]  Mean of X
             Below diagonal elements of Cholesky matrix
lHSA_ONE   -.08790569   .00670183    -13.117    .0000
            Dispersion parameter for NegBin distribution
ScalParm  1.24095398   .02905699     42.708    .0000
           Implied covariance matrix of random parameters
              Matrix Var_Beta has 2 rows and 2 columns.
                    1           2
1             .61380     -.06887
2            -.06887      .02121
          Implied standard deviations of random parameters
              Matrix S.D_Beta has 2 rows and 1 columns.
1             .78346
2             .14562
Implied correlation matrix of random parameters
Matrix Cor_Beta has 2 rows and 2 columns.
                    1           2
1           1.00000     -.60364
2           -.60364     1.00000
```

The negative binomial by construction already picks up some heterogeneity that manifests itself in the overdispersion. The random coefficients formulation is an extension that gathers together other time invariant heterogeneity across individuals. Random coefficient models allow the randomness of the coefficients to explain the heterogeneity across individuals as well as the heterogeneity across groups. This heterogeneity, in sum, results in the differences found in the responses, y_{ik}, due to changes in the predictors. The fact that levels of nesting, or of additional unexplained heterogeneity, can be explained by these models, make them attractive to those who wish to model count data having such a structure.

The only caveat to keep in mind when using negative binomial random coefficient models is to be careful of over-specification; i.e., multiple adjustments are being given to the otherwise Poisson counts. Care must be taken to assure that our model does not make too much adjustment.

10.6 Summary

Panel data, consisting of data in clustered and longitudinal format, violate the basic maximum likelihood assumption of the independence of observations. Whether the data are clustered by groups or recorded by observations over periods of time, the same methods are used to estimate the respective panel models.

In this chapter I presented overviews of the foremost panel models: unconditional and conditional fixed effects models, random effects models, and generalized estimating equations. Multilevel mixed models are a comparatively new area of research, with multilevel count models being the most recent. LIMDEP is the only commercial software supporting negative binomial linear mixed models, with its initial application in 2006. The software limits use to random intercept and the more detailed random coefficient negative binomial models. Examples of both models are provided in this chapter.

Hierarchical GLMs, called HGLMs, and double HGLMs have recently been developed, primarily by John Nelder and Yougjo Lee. However, they have not employed HGLM theory to the negative binomial, and we do not discuss them here. HGLMs are supported only by GENSTAT software.

Exercises

1 Why does a conditional fixed-effects negative binomial allow a model constant, unlike fixed-effects Poisson and most other fixed-effects models?

2 It has been demonstrated that an unconditional fixed effects negative binomial is statistically preferably to the conditional fixed effects negative binomial, which produces biased estimates. Discuss why is this the case.

3 GEE models are population averaged models. How does this differ from random-effects models?

4 Why is the conjugate prior the most appropriate random-effect mixing distribution for random-effects models?

5 Why do most GEE software packages set the robust variance estimator as the default method of calculating standard errors?

6 Using the *ships* data set that is provided on the text web site, model *accident* on the levels of *construction* and *operation* with the natural log of *service* as the offset. Model as both a Poisson and negative binomial GEE using the exchangeable correlation matrix. Which model is the preferred model? Compare the models after the creation of interactions between predictors. Does this make a difference?

7 Use a stationary correlation matrix to model the ship data as in Question 6. Compare it with a random-beta effects negative binomial model. How do the values in the correlation structure help interpret the model?

8 Using the *absenteeism* data set on the text web site, model *days* on binary predictors *aborig* (is aborigine), *girl* (is girl), and *slow* (is slow learner), as well as the four-level *schoolyr* (8th grade, freshman, sophomore, senior). Attempt to determine the source of overdispersion. Model with *id* as a random intercept. Compare AIC statistics to determine if there is a substantial improvement in fit.

9 Derive the conditional fixed-effects NB-1 log-likelihood function. Express it in a manner similar to the formulae listed in Appendix A.

10 Use the *mdvisitsx* data set, employing a random coefficient negative binomial model to the data. Let *id* be the random intercept and *educ* the random coefficient. Compare the results to a random Gaussian-effects negative binomial model. Note the differences in AIC statistic and significance of the predictors.

Appendix A

Negative binomial log-likelihood functions

NEGATIVE BINOMIAL

$$y^{*}\ln(\alpha\exp(xb)/(1+\alpha\exp(xb)) - \ln(1+\alpha\exp(xb))/\alpha + \ln\Gamma(y+1/\alpha)$$
$$- \ln\,\Gamma(y+1) - \ln\Gamma(1/\alpha)$$

NEGATIVE BINOMIAL TYPE 1 (NB-1)

$$y^{*}\ln(\alpha) - (y - \alpha\exp(xb))^{*}\ln(1+\alpha)$$
$$+ \ln\Gamma(y+\alpha\exp(xb)) - \ln\Gamma(y+1) - \ln\Gamma(\alpha\exp(xb))$$

NEGATIVE BINOMIAL CANONICAL

$$y^{*}\ln(\{\alpha(1/(\alpha\exp(-xb)-1))\}/(1-\{\alpha(1/(\alpha\exp(-xb)-1))\}))$$
$$- \ln(1+\{\alpha(1/(\alpha\exp(-xb)-1))\})/a + \ln\Gamma(y+1/\alpha) - \ln\Gamma(y+1)$$
$$- \ln\Gamma(1/\alpha)$$

ZERO-TRUNCATED NEGATIVE BINOMIAL

$$y^{*}\ln(\alpha\exp(xb)/(1+\alpha\exp(xb)) - \ln(1+\alpha\exp(xb))/\alpha + \ln\Gamma(y+1/\alpha)$$
$$- \ln\Gamma(y+1) - \ln\Gamma(1/\alpha) - \ln(1-(1+\alpha\exp(xb))\hat{}(-1/\alpha))$$

NEGATIVE BINOMIAL WITH ENDOGENOUS STRATIFICATION

$$y^{*}\ln(\alpha) + (y-1)^{*}xb - (y+1/\alpha)\ln(1+\alpha\exp(xb)) + \ln\Gamma(y+1/\alpha)$$
$$- \ln\Gamma(y+1) - \ln\Gamma(1/\alpha) + \ln\Gamma(y)$$

ZERO-INFLATED POISSON – logit

$$y{=}{=}0:\ln(1/(1+\exp(-x\beta_b)) + 1/(1+\exp(x\beta_b))^{*}\exp(-\exp(x\beta)))$$
$$y>0:\ln(1/(1+\exp(-x\beta_b))) - \exp(x\beta) + y(x\beta) - \ln\Gamma(y+1)$$

ZERO-INFLATED POISSON – probit

$$\begin{aligned}&\text{If } y{=}{=}0 : \ln(\Phi x\beta_b) + (1-\Phi(x\beta_b)) * \exp(-\exp(x\beta)))\\ &y > 0 : \ln(1-\Phi(x\beta_b)) - \exp(x\beta) + y(x\beta) - \ln\Gamma(y+1))\end{aligned}$$

ZERO-INFLATED NEGATIVE BINOMIAL – logit

$$\begin{aligned}&\text{cond}\{y{=}{=}0, \ln(1/(1+\exp(-xb1)) + 1/(1+\exp(xb1))\\ &{}^*(1/(1+\alpha^*\exp(xb)))^\wedge(1/\alpha),\\ &\ln(1/(1+\exp(xb1)) + \ln\Gamma(1/\alpha+y) - \ln\Gamma(y+1) - \ln\Gamma(1/\alpha)\\ &+(1/\alpha)^*\ln(1/(1+\alpha^*\exp(xb))) + y^*\ln(1-(1/(1+\alpha^*\exp(xb))))\}\end{aligned}$$

ZERO-INFLATED NEGATIVE BINOMIAL – probit

$$\begin{aligned}&\text{cond}\{y == 0, \ln(\Phi(xb1) + (1-\Phi(xb1))\\ &{}^*(1/(1+\alpha*\exp(xb)))\ (1/\alpha),\\ &\ln(1-\Phi(xb1)) + \ln\Gamma(1/\alpha+y) - \ln\Gamma(y+1) - \ln\Gamma(1/\alpha) + (1/\alpha)\\ &{}^*\ln(1/(1+\alpha*\exp(xb))) + y*\ln(1-(1/(1+\alpha*\exp(xb))))\}\\ &< \Phi = \text{normal CDF} >\end{aligned}$$

GENERALIZED POISSON

$$\begin{aligned}&y\ln(\exp(x\beta)/(1+\alpha\exp(x\beta)) + (y-1)\ln(1+\alpha y)\\ &\quad - [\exp(x\beta)(1+\alpha y)/(1+\alpha\exp(x\beta))] - \ln\Gamma(y+1)\end{aligned}$$

NEGATIVE BINOMIAL-LOGIT HURDLE

$$\begin{aligned}&\text{cond}\{y == 0, \ln(1/(1-\exp(xb1)),\\ &\ln(\exp(xb1)/(1+\exp(xb1))) + y*\ln(\exp(xb)/(1+\exp(xb)))\\ &-\ln(1+\exp(xb))/\alpha + \ln\Gamma(y+1/\alpha) - \ln\Gamma(y+1) - \ln\Gamma(1/\alpha)\\ &-\ln(1-(1+\exp(xb))\ (-1/\alpha))\}\end{aligned}$$

NEGATIVE BINOMIAL-CLOGLOG HURDLE

$$\begin{aligned}&\text{cond}\{y{=}{=}0, -\exp(xb1), \ln(1-\exp(-\exp(xb1)))\\ &+y^*\ln(\exp(xb)/(1+\exp(xb))) - \ln(1+\exp(xb))/\alpha\\ &+\ln\Gamma(y+1/\alpha) - \ln\Gamma(y+1) - \ln\Gamma(1/\alpha) - \ln(1-(1+\exp(xb))^\wedge(-1/\alpha))\}\end{aligned}$$

CONDITIONAL FIXED EFFECTS POISSON

$$\text{sum}((\ln\Gamma(\text{sum}(y)+1) - \text{sum}(\ln\Gamma(y+1) + \text{sum}(y^*xb - y^*\ln(\text{sum}(\exp(xb))))$$

CONDITIONAL FIXED EFFECTS NEGATIVE BINOMIAL

$$\ln\Gamma(\text{sum}(\exp(xb))) + \ln\Gamma(\text{sum}(y)+1) - \text{sum}(\ln\Gamma(y+1))$$
$$-\ln\Gamma(\text{sum}(y) + \text{sum}(\exp(xb))) + \text{sum}(\ln\Gamma(\exp(xb)+y)) - \text{sum}(\ln\Gamma(\exp(xb)))$$

RANDOM EFFECTS POISSON WITH GAMMA EFFECT

$$\ln\Gamma(1/\alpha - \Sigma y) - \ln\Gamma(1/\alpha) - (1/\alpha + \Sigma y)^{*}\ln(1+\alpha^{*}\Sigma(\exp(xb)))$$
$$+\ln(\alpha)^{*}\Sigma y - \ln\Gamma(y+1) + y^{*}xb$$

RANDOM EFFECTS NEGATIVE BINOMIAL WITH BETA EFFECT

$$\ln\Gamma(a+b) + \ln\Gamma(a+\Sigma(\exp(xb))) + \ln\Gamma(b+\Sigma y) - \ln\Gamma(a) - \ln\Gamma(b)$$
$$-\ln\Gamma(a+b+\Sigma(\exp(xb)) + \Sigma y) + \Sigma(\ln\Gamma((\exp(xb)) + y)$$
$$-\ln\Gamma(y+1) - \ln\Gamma(\exp(xb)))$$

GEOMETRIC

$$y^{*}xb - (1+\exp(xb)) - \ln(1+\exp(xb)) \quad \text{or } y^{*}xb - (1+y)\ln(1+\exp(xb))$$

ZERO-TRUNCATED GEOMETRIC

$$y^{*}xb - (1+y)\ln(1+\exp(xb)) - \ln(1+\ln(1+\exp(xb)))$$

CANONICAL GEOMETRIC

$$y^{*}\ln(1/(\exp(-xb)-1)) - (1+y)\ln(1+(1/(\exp(-xb)-1)))$$

GEOMETRIC-LOGIT HURDLE

$$\text{cond}\{y{=}{=}0, \ln(1/(1+\exp(xb1))),$$
$$\ln(\exp(xb1)/(1+\exp(xb1))) + y^{*}xb - (1+y)^{*}\ln(1+\exp(xb))$$
$$-\ln(1+\ln(1+\exp(xb)))\}$$

GEOMETRIC-CLOGLOG HURDLE

$$\text{cond}\{y{=}{=}0, -\exp(xb1),$$
$$\ln(1-\exp(-\exp(xb1))) + y^{*}xb - (1+y)^{*}\ln(1+\exp(xb))$$
$$-\ln(1+\ln(1+\exp(xb)))\}$$

Appendix B

Deviance functions

Stata code for four count model deviance functions: Poisson, Geometric, NB-2, NB-1

POISSON

```
tempvar dev sdev
egen ‘dev’ = sum(‘y’ *ln(‘y’/‘mu’) – (‘y’ − ‘mu’))
local deviance = 2*‘dev’
```

GEOMETRIC

```
tempvar y lp mu dev sdev
predict ‘lp’, xb
gen double ‘mu’ = exp(‘lp’)
gen ‘y’ = ‘lhs’
egen ‘dev’ = sum(‘y’*ln(‘y’/‘mu’) – (1 + ‘y’)*ln((1 + ‘y’)/(1 + ‘mu’)))
local deviance = 2*‘dev’
```

NEGATIVE BINOMIAL 2

```
tempvar y lp mu dev sdev alpha
predict ‘lp’, xb
gen double ‘mu’ = exp(‘lp’)
gen ‘y’ = ‘lhs’
local alpha r(est)
egen ‘dev’= sum((‘y’ *ln(‘y’/‘mu’)) – (((1 + ‘alpha’*‘y’)/‘alpha’) *ln((1 + ‘alpha’*‘y’)/
(1 + ‘alpha’*‘mu’))))
local deviance = 2*‘dev’
```

NEGATIVE BINOMIAL 1

```
tempvar y lp mu dev sdev alpha
predict ‘lp’, xb
gen double ‘mu’ = exp(‘lp’)
gen ‘y’ = ‘lhs’
local alpha r(est)
egen ‘dev’ = sum((‘mu’ − ‘y’) *ln(1 + ‘alpha’))
local deviance = 2*‘dev’
```

Appendix C

Stata negative binominal – ML algorithm

```
*! Version 1.0.3
* NEGATIVE BINOMIAL REGRESSION: Joseph Hilbe: 8Sep2005
program jhnbin, eclass properties(svyb svyj svyr)
    version 9.1
    syntax [varlist] [if] [in] [fweight pweight aweight iweight] [, ///
        Level(cilevel) IRr Robust noLOG  ///
        OFFset(passthru) EXposure(passthru)  ///
        CLuster(passthru) FROM(string asis) *]
    gettoken lhs rhs: varlist
    mlopts mlopts, 'options'
    if (“'weight'"!= "") local weight “['weight' 'exp']"
    if ('“'from'"'!= '""') local initopt '“init('from')"'

    ml model lf jhnb_ll (xb: 'lhs' = 'rhs', 'offset' 'exposure')  ///
            /lnalpha    ///
        'if' 'in' 'weight',    ///
        'mlopts' 'robust' 'cluster'    ///
        title(“Negative Binomial Regression")  ///
        maximize 'log' 'initopt'    ///
        diparm(lnalpha, exp label(alpha))

    ereturn scalar k_aux = 1
    ml display, level('level') 'irr'

qui {
* AIC
    tempvar aic
    local nobs e(N)
    local npred e(df_m)
    local df = e(N) -- e(df_m) −1
    local llike e(ll)
    gen 'aic' = ((−2*'llike') + 2*('npred' + 1))/'nobs'
}

* DISPLAY
di in gr _col(1) “AIC Statistic =" in ye%11.3f 'aic'

end
```

LOG-LIKELIHOOD FUNCTION FOR NEGATIVE BINOMIAL

```
*! version 1.0.1 7Sep2005
* Negative binomial: log likelihood function: Joseph Hilbe
program define jhnb_ll
version 9.1
args lnf xb alpha

tempvar a mu
qui gen double 'a' = exp('alpha')
qui gen double 'mu' = exp('xb') * 'a'
qui replace 'lnf' = $ML_y1 * ln('mu'/(1 + 'mu')) -- ///
ln(1 + 'mu')/'a' + lngamma($ML_y1 + 1/'a') -- ///
lngamma($ML_y1 + 1) -- lngamma(1/'a')

end
```

Appendix D

Negative binomial variance functions

Given that the negative binomial with $\alpha = 0$ is Poisson, and negative binomial with $\alpha = 1$ is geometric

`Poisson:`	$\mathtt{V} = \mu$
`QL Poisson:`	$\mathtt{V} = \mu\phi$
`Geometric:`	$\mathtt{V} = \mu(1 + \mu)$
`NB-1:`	$\mathtt{V} = \mu(1 + \alpha)$
`NB-2:`	$\mathtt{V} = \mu(1 + \alpha\mu)$
`NB-H:`	$\mathtt{V} = \mu(1 + (\alpha\nu)\mu)$
`NB-P:`	$\mathtt{V} = \mu + \alpha\mu^{\rho}$

Log and canonical links are available for above models beginning with the geometric.

Appendix E

Data sets

All data sets, in Stata, Excel, and ASCII format, are posted at the Cambridge University Press web site for this text: http://www.cambridge.org/9780521857727

Below are listed the data sets used in the text, except for those that are created as simulated data. However, I shall post the constructed simulated data sets to the text web site for your use. Note that many of the smaller data sets have the data presented in the text. Whether this is the case for a particular data set is indicated by a *Yes* in the `DATA IN BOOK` column. The format of the data is also provided. *Case* specifies that the data are presented in observation form; *group* indicates that the data are either grouped or are frequency weighted.

DATA SETS USED IN TEXT

	DATA IN			
NAME	BOOK	FORMAT	OBS	
affairs	No	case	601	(Dict. in Table 6.5)
azprocedure	No	case	3589	(Sum stats in Table 6.9)
fastrak	No	case	5388	(Ch 3.2)
lbw	No	case	189	(Ch 6.2; Dict. in Table 6.3)
lbwcases	No	group	23	(Ch 6.2)
loomis	No	case	410	(Ch 8.2; Table 8.2)
medpar	No	case	1495	(Ch 3.1)
mdvisits	No	case	2227	(Dict. in Table 6.12)
mdvisitsx	No	case	2227	(Ch 7.3)
progabide	No	case	295	(Ch 10.3)
rwm	No	case	27326	(Ch 7.4)
ships	No	group	40	(Ch. 10.1)
titanic	Yes	group	12	(Data in Table 6.11)

SIMULATED DATA SETS USED IN TEXT (MAJOR)

NAME	OBS	CHAPTER CREATED
odtest	10000	Ch 4.2
nboverex	10000	Ch 4.3.5
syn_nb	10000	Ch 6.1
geo_simul	50000	Ch 7.1
syn_nb1	50000	Ch 7.2.1

Data sets used in end-chapter examples are listed below. Some data sets are used more than once. The chapter and question number where the data are first used is given in the second column. Ten of the 17 exercise data sets are shown in full, either as part of the question, or referenced to another part of the text where the data can be found.

DATA SETS USED IN EXERCISES

NAME	1^{ST} USED	DATA IN BOOK	FORMAT	OBS	COMMENTS
absenteeism	(Ch 10, # 8)	No	case	146	
azprocedure	(Ch 8, # 4)	No	case	3589	
cancer	(Ch 7, # 1)	No	case	48	
cancercen	(Ch 9, # 1)	Yes	case	48	(cancer modified)
contacts	(Ch 9, # 6)	Yes	group	6	
doll	(Ch 6, # 7)	Yes	group	10	
drg112az	(Ch 4, # 5)	No	case	1798	
edsurvey	(Ch 8, # 2)	Yes	case	50	
ex2_4	(Ch 2, # 4)	Yes	case	10	
gss2002_educ	(Ch 6, # 3)	No	case	2500	
hiv	(Ch 3, # 5)	Yes	group	8	
horsekick	(Ch 3, # 7)	Yes	group	6	
kyp	(Ch 6, # 1)	Yes	case	24	(modified)
mdvisitsx	(Ch 10, #10)	No	case	2227	
pyears	(Ch 5, # 3)	Yes	group	10	
ships	(Ch 10, # 6)	No	group	40	
ticks	(Ch 4, # 8)	Yes	group	26	

References

Allison, P. D. and R. Waterman (2002). Fixed-effects negative binomial regression models, unpublished manuscript.

Amemiya, T. (1984). Tobit models: a survey, *Journal of Econometrics* **24**: 3–61.

Anscombe, F. J. (1948). The transformations of Poisson, binomial, and negative binomial data, *Biometrika* **35**: 246–254.

Anscombe, F. J. (1949). The statistical analysis of insect counts based on the negative binomial distribution, *Biometrics* **5**: 165–173.

Anscombe, F. J. (1972). Contribution to the discussion of H. Hotelling's paper, *Journal of the Royal Statistical Society – Series B* **15**(1): 229–230.

Bartlett, M. S. (1947). The use of transformations, *Biometrics* **3**: 39–52.

Beall, G. (1942). The transformation of data from entomological field experiments so that that analysis of variance becomes applicable, *Biometrika* **29**: 243–262.

Blom, G. (1954). Transformations of the binomial, negative binomial, Poisson, and χ^2 distributions, *Biometrika* **41**: 302–316.

Breslow, N. E. (1984). Extra-Poisson variation in log-linear models, *Applied Statistics* **33**(1): 38–44.

Breslow, N. (1985). Cohort analysis in epidemiology, in *Celebration of Statistics*, ed. A. C. Atkinson and S. E. Fienberg, New York: Springer-Verlag.

Cameron, A. C. and P. K. Trivedi (1986). Econometric models based on count data: comparisons and applications of some estimators, *Journal of Applied Econometrics* **1**: 29–53.

Cameron, A. C. and P. K. Trivedi (1990). Regression-based tests for overdispersion in the Poisson model, *Journal of Econometrics* **46**: 347–364.

Cameron, A. C. and P. K. Trivedi (1998). *Regression Analysis of Count Data*, New York: Cambridge University Press.

Collett, D. (1989). *Modelling Binary Data*, London: Chapman & Hall.

Consul, P. and G. Jain (1973). A generalization of the Poisson distribution, *Technometrics* **15**: 791–799.

Consul, P. C. and R. C. Gupta (1980). The generalized binomial distribution and its characterization by zero regression, *SIAM Journal of Applied Mathematics* **39**(2): 231–237.

Consul, P. and F. Famoye (1992). Generalized Poisson regression model, *Communications in Statistics – Theory and Method* **21**: 89–109.

Drescher, D. (2005). Alternative distributions for observation driven count series models, Economics Working Paper No. 2005–11, Christian-Albrechts-Universitat, Kiel, Germany.

Edwards, A. W. F. (1972). *Likelihood*, Baltimore, MD: John Hopkins University Press.

Eggenberger F. and G. Polya (1923). Über die Statistik Verketteter Vorgänge, *Journal of Applied Mathematics and Mechanics* **1**: 279–289.

Englin, J. and J. Shonkwiler (1995). Estimating social welfare using count data models: an application under conditions of endogenous stratification and truncation, *Review of Economics and Statistics* **77**: 104–112.

Fair, R. (1978). A theory of extramarital affairs, *Journal of Political Economy* **86**: 45–61.

Famoye, F. and K. Singh (2006). Zero-truncated generalized Poisson regression model with an application to domestic violence, *Journal of Data Science* **4**: 117–130.

Famoye, F (1995). Generalized binomial regression model, *Biometrical Journal* **37**(5): 581–594.

Faraway, J. (2006). *Extending the Linear Model with R*, Boca Raton, FL: Chapman & Hall/CRC Press.

Frees, E. (2004). *Longitudinal and Panel Data*, Cambridge: Cambridge University Press.

Goldberger, A. S. (1983). Abnormal selection bias, in *Studies in Econometrics, Time Series, and Multivariate Statistics*, ed. S. Karlin, T. Amemiya, and L. A. Goodman, New York: Academic Press, pp. 67–85.

Gould, W., J. Pitblado, and W. Sribney (2006). *Maximum Likelihood Estimation with Stata*, Third Edition, College Station: Stata Press.

Greene, W. H. (1992). Statistical models for credit scoring, Working Paper, Department of Economics, Stern School of Business, New York University.

Greene, W. H. (1994). Accounting for excess zeros and sample selection in Poisson and negative binomial regression models, EC-94–10, Department of Economics, Stern School of Business, New York University.

Greene, W. H. (2003). *Econometric Analysis*, Fifth Edition, New York: Macmillan.

Greene, W. H. (2006). *LIMDEP Econometric Modeling Guide*, Version 9, Plainview, NY: Econometric Software Inc.

Greene, W. H. (2006a). A general approach to incorporating 'selectivity' in a model, Working Paper, Department of Economics, Stern School of Business, New York University.

Greenwood, M. and G. U. Yule (1920) An inquiry into the nature of frequency distributions of multiple happenings, with particular reference to the occurrence of multiple attacks of disease or repeated accidents, *Journal of the Royal Statistical Society* A, **83**: 255–279.

Gurmu, S. and P. K. Trivedi (1992). Overdispersion tests for truncated Poisson regression models, *Journal of Econometrics* **54**: 347–370.

Hardin, J. W. and J. M. Hilbe (2001). *Generalized Linear Models and Extensions*, College Station, TX: Stata Press

Hardin, J. W. and J. M. Hilbe (2002). *Generalized Estimating Equations*, Boca Raton, FL: Chapman & Hall/CRC Press.

Hardin, J. W. and J. M. Hilbe (2007). *Generalized Linear Models and Extensions*, Second Edition, College Station, TX: Stata Press

Hardin, J. W. (2003). The sandwich estimate of variance, in *Advances in Econometrics: Maximum Likelihood of Misspecified Models: Twenty Years Later*, ed. T. Fomby and C. Hill, Elsevier, **17**: 45–73.

Hausman, J., B. Hall, and Z. Griliches (1984). Econometric models for count data with an application to the patents – R&D relationship, *Econometrica* **52**: 909–938.

Heckman, J. (1979). Sample selection bias as a specification error, *Econometrica*, **47**: 153–161.

Heilbron, D. (1989). Generalized linear models for altered zero probabilities and overdispersion in count data, Technical Report, Department of Epidemiology and Biostatistics, University of California, San Francisco.

Hilbe, J. M. (1993a). Log negative binomial regression as a generalized linear Model, technical Report COS 93/945–26, Department of Sociology, Arizona State University.

Hilbe, J. (1993b). Generalized linear models, *Stata Technical Bulletin*, STB-11, sg16.

Hilbe, J. (1993c). Generalized linear models using power links, *Stata Technical Bulletin*, STB-12, sg16.1

Hilbe, J. (1994a). Negative binomial regression, *Stata Technical Bulletin*, STB-18, sg16.5

Hilbe, J. (1994b). Generalized linear models, *The American Statistician*, **48**(3): 255–265.

Hilbe, J. (2000). Two-parameter log-gamma and log-inverse Gaussian models, in *Stata Technical Bulletin Reprints*, College Station, TX: Stata Press, pp. 118–121.

Hilbe, J. (2005a). CPOISSON: Stata module to estimate censored Poisson regression, Boston College of Economics, Statistical Software Components, http://ideas.repec.org/c/boc/bocode/s456411.html

Hilbe, J. (2005b), CENSORNB: Stata module to estimate censored negative binomial regression as survival model, Boston College of Economics, Statistical Software Components, http://ideas.repec.org/c/boc/bocode/s456508.html

Hilbe, J. and B. Turlach (1995). Generalized linear models, in *XploRe: An Interactive Statistical Computing Environment*, ed. W. Hardle, S. Klinke, and B. Tulach, New York: Springer-Verlag, pp. 195–222.

Hilbe, J. (1998). Right, left, and uncensored Poisson regression, *Statistical Bulletin*, **46**: 18–20.

Hilbe, J. and W. Greene (2007). Count response regression models, in *Epidemiology and Medical Statistics*, ed. C. R. Rao, J. P Miller, and D. C. Rao, Elsevier Handbook of Statistics Series, London, UK: Elsevier.

Hoffmann, J. (2004). *Generalized Linear Models*, Boston, MA: Allyn & Bacon.

Hosmer, D. and S. Lemeshow (2003). *Applied Logistic Regression*, second edition, New York: Wiley

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions, in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, pp. 221–233.

Ihaka, R. and R. Gentleman (1996). "R: A language for data analysis and graphics," *Journal of Computational and Graphical Statistics* **5**: 299–314.

Jain, G. C. and P. C. Consul (1971). A generalized negative binomial distibution, *SIAM Journal of Applied Mathematics* **21**(4): 501–513.

Johnston, G. (1998). SAS/STAT/GENMOD procedure, SAS Institute

Karim, M. R. and S. Zeger (1989). A SAS macro for longitudinal data analysis, Department of Biostatistics, Technical Report **674**, The John Hopkins University.

Katz, E. (2001). Bias in conditional and unconditional fixed effects logit estimation, *Political Analysis* **9**(4): 379–384.

King, G. (1988). Statistical models for political science event counts: bias in conventional procedures and evidence for the exponential Poisson regression model, *American Journal of Political Science* **32**: 838–863.

King, G. (1989). Event count models for international relations: generalizations and applications, *International Studies Quarterly* **33**: 123–147.

Lambert, D. (1992). Zero-inflated Poisson regression with an application to defects in manufacturing, *Technometrics* **34**: 1–14.

Lancaster, T. (2002). Orthogonal parameter and panel data, *Review of Economic Studies* **69**: 647–666.

Lawless, J. F. (1987). Negative binomial and mixed Poisson regression, *Canadian Journal of Statistics* **15**, (3): 209–225.

Lee, Y., J. Nelder, and Y. Pawitan (2006). *Generalized Linear Models with Random Effects*, Boca Raton, FL: Chapman & Hall/CRC Press.

Liang, K.-Y. and S. Zeger (1986). Longitudinal data analysis using generalized linear models, *Biometrika*, **73**: 13–22.

Long, J. S. (1997). *Regression Models for Categorical and Limited Dependent Variables*, Thousand Oaks, CA: Sage.

Long, J. S. and J. Freese (2003, 2006). *Regression Models for Categorical Dependent Variables using Stata*, Second Edition, College Station, TX: Stata Press.

Loomis, J. B. (2003). Travel cost demand model based river recreation benefit estimates with on-site and household surveys: comparative results and a correction procedure, *Water Resources Research* **39**(4): 1105.

Marquardt, D. W. (1963). An algorithm for least-squares estimation of nonlinear parameters, *Journal of the Society for Industrial and Applied Mathematics* **11**: 431–441.

Martinez-Espiñeira, R., J. Amoako-Tuffour, and J. M. Hilbe (2006), Travel cost demand model based river recreation benefit estimates with on site and household surveys: comparative results and a correction procedure – revaluation, *Water Resource Research*, **42**.

McCullagh, P. (1983). Quasi-likelihood functions, *Annals of Statistics* **11**: 59–67.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models*, Second Edition, New York: Chapman & Hall.

Mullahy, J. (1986). Specification and testing of some modified count data models, *Journal of Econometrics* **33**: 341–365.

Mundlak, Y. (1978). On the pooling of time series and cross section data, *Econometrica* **46**: 69–85.

Mwalili, S., E. Lesaffre, and D. DeClerk (2005). The zero-inflated negative binomial regresson model with correction for misclassification: an example in Caries Research, Technical Report 0462, LAP Statistics Network Interuniversity Attraction Pole, Catholic University of Louvain la Neuve, Belgium, www.stat.ucl.ac.be/IAP.

Nelder, J. A. (1994). Generalized linear models with negative binomial or beta-binomial errors, unpublished manuscript.

Nelder, J. A. and R. W. M. Wedderburn (1972). Generalized linear models, *Journal of the Royal Statistical Society*, Series A, **135**(3): 370–384.

Nelder, J. and D. Pregibon (1987). An extended quasi-likelihood function, *Biometrika* **74**: 221–232.

Nelder, J. A. and Y. Lee (1992). Likelihood, quasi-likelihood, and pseudo-likelihood: some comparisons, *Journal of the Royal Statistical Society*, Series B, **54**: 273–284.

Nelson, D. L. (1975). Some remarks on generalized of negative binomial and Poisson distributions, *Technometrics* **17**: 135–136.

Neyman, O, and E. L. Scott (1948). Consistent estimation from partially consistent observations, *Econometrica* **16**(1): 1–32.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations, *Biometrics* **57**: 120–125.

Pan, W. (2001a). On the robust variance estimator in generalized estimating equations, *Biometrika* **88**(3): 901–906.

Pierce, D. A. and D. W. Schafer (1986). "Residuals in Generalized Linear Models," *Journal of the American Statistical Association*, **81**: 977–986.

Plackett, R. L. (1981). *The Analysis of Categorical Data* (2nd ed.), London: Chapman and Hall.

Rabe-Hesketh, S. and A. Skrondal (2004). *Generalized Latent Variable Modeling*, Boca Raton, FL: Chapman & Hall/CRC Press.

Rabe-Hesketh, S. and A. Skrondal (2005). *Multilevel and Longitudinal Modeling Using Stata*, College Station, TX: Stata Press.

Sikkel, D. and G. Jelierse (1998). Renewal theory and retrospective questions, *Applied Statistics* **37**: 312–420.

Shaw, D. (1988). On-site samples' regression, *Journal of Econometrics* **37**: 211–223.

Terza, J. V. (1998). A Tobit-type estimator for the censored Poisson regression model, *Econometric Letters* **18**: 361–365.

Thall, P. and S. Vail (1990). Some covariance models for longitudinal count data and overdispersion, *Biometrika* **46**: 657–671.

Twisk, J. W. R. (2003). *Applied Longitudinal Data Analysis for Epidemiology*, Cambridge: Cambridge University Press.

Vadeby, A. (2002). Estimation in a model with incidental parameters, LiTH-MAT-R-2002-02 working paper.

Venables, W. and B. Ripley (2002). *Modern Applied Statistics with S*, Fourth Edition, New York: Springer-Verlag.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses, *Econometrica* **57**: 307–333.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models and the Gauss–Newton method. *Biometrika* **61**: 439–447.

White H. (1980). A heteroskestasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity, *Econometrica* **48**(4): 817–838.

Winkelmann, R (1995). Duration dependence and dispersion in count-data models, *Journal of Business and Economic Statistics* **13**: 467–474.

Winkelmann, R. (2003). *Econometric Analysis of Count Data*, Fourth Edition, Heidelberg, Germany: Springer-Verlag.

Winkelmann, R., and K. F. Zimmermann (1995). Recent developments in count data modelling: theory and application, *Journal of Economic Surveys*, **9**: 1–24.

Woutersen, T. (2002). Robustness against incidental parameters, Department of Economics Working paper 2008, University of Western Ontario.

Yang, Z., J. Hardin, and C. Abby (2006). A score test for overdispersion based on generalized Poisson model, unpublished manuscript.

Zeger, S. L., K.-Y. Liang, and P. S. Albert (1988). Models for longitudinal data: A generalized estimating equation approach, *Biometrics* **44**: 1049–1060.

Author index

Subject index