CAMBRIDGE

Cambridge University Press
978-0-521-19815-8 - Negative Binomial Regression: Second Edition
Joseph M. Hilbe
Excerpt
More information

1

Introduction

1.1 What is a negative binomial model?

The negative binomial regression model is a truly unusual statistical model. Typically, those in the statistical community refer to the negative binomial as a single model, as we would in referring to Poisson regression, logistic regression, or probit regression. However, there are in fact several distinct negative binomial models, each of which are referred to as being a negative binomial model. Boswell and Patil (1970) identified 13 separate types of derivations for the negative binomial distribution. Other statisticians have argued that there are even more derivations. Generally, those who are using the distribution as the basis for a statistical model of count data have no idea that the parameterization of the negative binomial they are employing may differ from the parameterization being used by another. Most of the time it makes little difference how the distribution is derived, but, as we shall discover, there are times when it does. Perhaps no other model has such a varied pedigree.

I will provide an outline here of the intertwining nature of the negative binomial. Unless you previously have a solid background in this area of statistics, my overview is not likely to be completely clear. But, as we progress through the book, its logic will become evident.

The negative binomial model is, as are most regression models, based on an underlying probability distribution function (PDF). The Poisson model is derived from the Poisson PDF, the logistic regression model is derived from the binomial PDF, and the normal linear regression model (i.e. ordinary least squares), is derived from the Gaussian, or normal, PDF. However, the traditional negative binomial, which is now commonly symbolized as NB2 (Cameron and Trivedi, 1986), is derived from a Poisson–gamma mixture distribution. But such a mixture of distibutions is only one of the ways in which the negative binomial PDF can be defined. Unless otherwise specified, when I

Cambridge University Press
978-0-521-19815-8 - Negative Binomial Regression: Second Edition
Joseph M. Hilbe
Excerpt
More information

refer to a negative binomial model, it is the NB2 parameterization to which I refer.

The nice feature of this parameterization is that it allows us to model Poisson heterogeneity. As we shall discover, the mean and variance of the Poisson PDF are equal. The greater the mean value, the greater is the variability in the data as measured by the variance statistic. This characteristic of the data is termed equidispersion and is a distributional assumption of Poisson data. Inherent in this assumption is the requirement that counts are independent of one another. When they are not, the distributional properties of the Poisson PDF are violated, resulting in extra-dispersion. The mean and variance can no longer be identical. The form of extra-dispersion is nearly always one of overdispersion. That is, the variance is greater in value to that of the mean.

The negative binomial model, as a Poisson–gamma mixture model, is appropriate to use when the overdispersion in an otherwise Poisson model is thought to take the form of a gamma shape or distribution. The same shape value is assumed to hold across all conditioned counts in the model. If different cells of counts have different gamma shapes, then the negative binomial may itself be overdispersed; i.e the data may be both Poisson and negative binomial overdispersed. Random-effects and mixed-effects Poisson and negative binomial models are then reasonable alternatives.

What if the shape of the extra correlation, or overdispersion, is not gamma, but rather another identifiable shape such as inverse Gaussian? It is possible to construct a Poisson-inverse Gaussian distribution, and model. This distribution is formally known as a Holla distribution, but is better known to most statisticians as a PIG function. Unfortunately there is no closed form solution for the PIG model; estimation is therefore typically based on quadrature or simulation. It is not, however, a negative binomial, but can be used when the data takes its form.

What if we find that the shape of overdispersion is neither gamma nor inverse Gaussian? Poisson-lognormal models have been designed as well, but they too have no closed form. If overdispersion in the data takes no identifiable shape, most statisticians employ a negative binomial. There are other alternatives though – for instance, quantile count models. We spend some time later in the text evaluating models that address these data situations.

It may appear that we have gotten off track in discussing non-negative binomial methods of handling Poisson overdispersion. However, these methods were derived specifically because they could better model the data than available negative binomial alternatives. Each one we mentioned is based on the mixture approach to the modeling of counts. Knowledge of the negative binomial regression model therefore entails at least a rudimentary acquaintance with its alternatives.

I should mention here that the form of the mixture of variances that constitute the core of the Poisson–gamma mixture is $\mu + \mu^2/\nu$ where μ is the Poisson variance and μ^2/ν is the gamma variance; ν is the gamma shape parameter, and corresponds to extra dispersion in the mixture model meaning of negative binomial, as described above. Conceived of in this manner, there is an indirect relationship between ν and the degree of overdispersion in the data. A negative binomial model based on this joint definition of the variance becomes Poisson when ν approaches infinity. However, it is perfectly permissible, and more intuitive, if ν is inverted so that there is a direct relationship between the parameter and extra correlation. The standard symbol for the heterogeneity or overdispersion parameter given this parameterization of the negative binomial variance is α. Sometimes you may find r or k symbolizing ν, and confusingly some have used k to represent α. We shall use the symbols r for ν for the indirect relationship and α for the directly related heterogeneity parameter. All current commercial software applications of which I am aware employ the α parameterization. R's **glm** and **glm.nb** functions employ θ, or $1/\alpha$.

The origin of the negative binomial distribution is not as a Poisson–gamma mixture, which is a rather new parameterization. The earliest definitions of the negative binomial are based on the binomial PDF. Specifically, the negative binomial distribution is characterized as the number of failures before the rth success in a series of independent Bernoulli trials. The Bernoulli distribution is, as you may recall, a binomial distribution with the binomial denominator set at one (1). Given r as an integer, this form of the distribution is also known as a *Pascal* distribution, after mathematician Blaise Pascal (1623–1662). However, for negative binomial models, r is taken as a real number greater than 0, although it is rarely above four.

It is important to understand that the traditional negative binomial model can be estimated using a standard maximum likelihood function, or it can be estimated as a member of the family of generalized linear models (GLM). A negative binomial model is a GLM only if its heterogeneity parameter is entered into the generalized linear models algorithm as a constant. We shall observe the consequences of this requirement later in the text.

In generalized linear model theory the link function is the term that linearizes the relationship of the linear predictor, $x'\beta$, and the fitted value, μ or $\hat{y}$. In turn, μ is defined in terms of the inverse link. Given that generalized linear models are themselves members of the single parameter exponential family of distributions, the exponential family log-likelihood for count models can be expressed as

$$\sum y\theta + b(\theta) + c(y) \tag{1.1}$$

with θ as the link, $b(\theta)$ as the cumulant from which the mean and variance functions are derived, and $c(y)$ as the normalization term guaranteeing that the probability sums to 1. For the GLM negative binomial, the link, $\theta = -\ln((1/(\alpha\mu)) + 1)$, and the inverse link, which defines the fitted value, is $b'(\theta) = \mu$, or $1/(\alpha(\exp(x\beta) - 1))$. The GLM inverse link is also symbolized as η.

The traditional NB2 negative binomial amends the canonical link and inverse link values to take a log link, $\ln(\mu)$, and exponential inverse link, $\exp(\mathrm{x}'\beta)$. These are the same values as the canonical Poisson model. When the negative binomial is parameterized in this form, it is directly related to the Poisson model. As a GLM, the traditional NB2 model is a log-linked negative binomial, and is distinguished from the canonical form, symbolized as NB-C.

We shall discover when we display the derivations of the negative binomial as a Poisson–gamma mixture, and then from the canonical form defined as the number of failures before the *r*th success in a series of independent Bernoulli trials, that both result in an identical probability function when the mean is given as μ. When the negative binomial PDF is parameterized in terms of $x'\beta$, the two differ.

There are very good reasons to prefer the NB2 parameterization of the negative binomial, primarily because it is suitable as an adjustment for Poisson overdispersion. The NB-C form is not interpretable as a Poisson type model, even though it is the canonical form derived directly from the PDF. We shall discuss its interpretation later in the text. We shall also show how the negative binomial variance function has been employed to generalize the function.

The characteristic form of the canonical and NB2 variance functions is $\mu + \alpha\mu^2$. This value can be determined as the second derivative of the cumulant, or $b''(\theta)$. A linear negative binomial has been constructed, termed NB1, that parameterizes the variance as $\mu + \alpha\mu$. The NB1 model can also be derived as a form of Poisson–gamma mixture, but with different properties resulting in a linear variance. In addition, a generalized negative binomial has been formulated as $\mu + \alpha\mu^p$, where p is a third parameter to be estimated. For NB1, $p = 1$; for NB2, $p = 2$. The generalized negative binomial provides for any reasonable value of p. Another form of negative binomial, called heterogeneous negative binomial, NB-H, is typically a NB2 model, but with α parameterized. A second table of estimates is presented that displays coefficients for the influence of predictors on the amount of overdispersion in the data.

We have seen that the negative binomial can be understood in a variety of ways. All of the models we have discussed here are negative binomial; both the NB2 and NB1 are commonly used when extending the negative binomial to form models such as mixtures of negative binomial models, or when employed in panel models. Knowing which underlying parameterization of negative binomial is being used in the construction of an extended negative binomial model is essential when we are evaluating and interpreting it, which is our subject matter.

Now that we have provided an overview of the landscape of the basic negative binomial models, we take a diversion and provide a brief history of the negative binomial. Such an historical overview may help provide a sense of how the above varieties came into existence, and inform us as to when and why they are most effectively used.

1.2 A brief history of the negative binomial

If we are to believe Isaac Todhunter's report in his *History of the Mathematical Theory of Probability from the Time of Pascal to that of Laplace* (1865), Pierre de Montmort in 1713 mentioned the negative binomial distribution in the context of its feature as the number of failures, y, before the kth success in a series of binary trials. As a leading mathematician of his day, Montmort was in constant communication with many familiar figures in the history of statistics, including Nicholas and Jacob Bernoulli, Blaise Pascal, Brook Taylor (Taylor series) and Gottfried Leibniz (credited, along with Newton, with the discovery of the Calculus). He is said to have alluded to the distribution in the second edition of his foremost work, *Essay on the Analysis of Games of Chance* (1708), but did not fully develop it for another five years.

The Poisson distribution upon which Poisson regression is based, originates from the work of Siméon Poisson (1781–1840). He first introduced the distribution as a limiting case of the binomial in his, *Research on the Probability of Judgments in Criminal and Civil Matters* (1838). Later in the text we derive the Poisson from the binomial, demonstrating how the two distributions relate. Poisson regression developed as the foremost method of understanding the distribution of count data, and later became the standard method used to model counts. However, as previously mentioned, 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*.

Little was done with either the earliest definition of negative binomial as derived by Montmort, or with Poisson's distribution for describing count data, until the early twentieth century. Building on the work originating with Gauss (1823), who developed the normal, or Gaussian, distribution, upon which ordinary least squares (OLS) regression is based, the negative binomial was again derived by William Gosset, under his pen name, Student, in 1907 while working under Karl Pearson at his Biometric Laboratory in London (Student, 1907). In the first paper he wrote while at the laboratory, he derived the negative binomial while investigating the sampling error involved in the counting of yeast cells with a haemocytometer. The paper was published in *Biometrika*, and appeared a year earlier than his well regarded papers on the sampling error of the mean and correlation coefficient (Jain, 1959). However, G. Udny Yule is generally, but arguably, credited with formulating the first negative binomial distribution based on a 1910 article dealing with the distribution of the number of deaths that would occur as a results of being exposed to a disease (i.e. how many deaths occur given a certain number of exposures). This formulation stems from what is called inverse binomial sampling. Later Greenwood and Yule (1920) derived the negative binomial distribution as the probability of observing y failures before the rth success in a series of Bernoulli trials, replicating in a more sophisticated manner the work of Montmort. Three years later the contagion or mixture concept of the negative binomial originated with Eggenberger and Polya (1923). They conceived of the negative binomial as a compound Poisson distribution by holding the Poisson parameter, λ, as a random variable having a gamma distribution. This was the first derivation of the negative binomial as a Poisson–gamma mixture distribution. The article also is the first to demonstrate that the Poisson parameter varies proportionally to the *chi*2 distribution with 2 degrees of freedom.

Much of the early work on the negative binomial during this period related to the *chi*2 distribution, which seems somewhat foreign to the way in which we now understand the distribution. During the 1940s, most of the original work on count models came from George Beall (1942), F. J. Anscombe (1949), and Maurice Bartlett (1947). All three developed measures of transformation to normalize non-normal data, with Bartlett (1947) proposing an 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 an intercept-only non-linear regression. Anscombe (1950), later derived the negative binomial as a series of logarithmic distributions, and discussed alternative derivations as well, for example: (1) inverse binomial sampling; (2) heterogeneous Poisson sampling where λ is considered as proportional to a *chi*2; (3) the negative binomial as a population growth model; and (4) the negative binomial derived from a geometric series. Evans (1953) developed what we now refer to as the NB1, or linear negative binomial, parameterization of the negative binomial.

Leroy Simon (1961), following his seminal work differentiating the Poisson and negative binomial models (1960), was the first to publish a maximum likelihood algorithm for fitting the negative binomial. He was one of the many actuarial scientists at the time who were engaged in fitting the Poisson and negative binomial distributions to insurance data. His work stands out as being the most sophisticated, and he was perhaps cited more often for his efforts in the area than anyone else in the 1960s. Birch (1963) is noted as well for being the first to develop a single predictor maximum likelihood Poisson regression model which he used to analyze tables of counts. It was not until 1981 that Plackett first developed a single predictor maximum likelihood negative binomial while working with categorical data which he could not fit using the Poisson approach.

Until the mid-1970s, parameterizing a non-linear distribution such as logit, Poisson, or negative binomial, so that the distributional response variable was conditioned on the basis of one of more explanatory predictors, was not generally conceived to be as important as understanding the nature of the underlying distribution itself, i.e. determining the relationships that obtain between the various distributions. When considering the negative binomial distribution, for example, the major concern was to determine how it related to other distributions – the *chi*2, geometric, binomial and Bernoulli, Poisson, gamma, beta, incomplete beta, and so forth. Regression model development was primarily thought to be a function of the normal model, and the transformations that could be made to both the least squares response and predictors.

It was not until 1981 that the first IBM personal computer became available to the general public, an event that changed forever the manner in which statistical modeling could be performed. Before that event, most complex statistical analyses were done using mainframe computers, which were usually at a remote site. Interactive analyses simply did not occur. Computer time was both time-consuming and expensive.

The emphasis on distributional properties and relationships between distributions began to change following the development of generalized linear models (GLM) by John Nelder and R. W. M. Wedderburn (1972). The new emphasis was on the construction of non-linear models that incorporated explanatory predictors. In 1974 Nelder headed a team of statisticians, including Wedderburn and members of the statistical computing working party of the Royal Statistical Society, to develop GLIM (Generalized Linear Interactive Modeling), a software application aimed to implement GLM theory. GLIM software allowed users to estimate GLM models for a limited set of exponential family members, including, among others, the binomial, Poisson, and, for a constant value of its heterogeneity parameter, the negative binomial. Although GLIM did not have a specific option for negative binomial models, one could use the *open* option to craft such a model.

Incorporated into GLIM software was the ability to parameterize Poisson models as rates. Nelder had developed the notion of offsets as a side exercise, only to discover that they could be used to model counts as incident rate ratios, which as we shall discover was a considerable advancement in statistical modeling. The traditional negative binomial model can also be parameterized in this manner.

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, albeit briefly. The second edition of the text appeared in 1989 (McCullagh and Nelder, 1989), and is still regarded as the premiere text on the subject. GLM-based negative binomial regression software was only available as a user-defined macro in GLIM until 1992 when Nelder developed what he called the *kk* system for estimating the negative binomial as a GenStat macro. The first implementation of the negative binomial as part of a GLM software algorithm did not occur until 1993 (Hilbe), with the software developed for both Stata and Xplore. The algorithm included links for the estimation of the traditional log-linked negative binomial, NB2, the canonical model, NB-C, and a negative binomial with an identity link. We shall discuss the use of an identity linked negative binomial later in the text.

In 1994 Hilbe (1994) developed a SAS macro for the NB2 model using SAS's GENMOD procedure, SAS's GLM modeling tool. The macro estimated the negative binomial heterogeneity parameter using a damping method adapted from a method first advanced by Breslow (1984) of reducing the Pearson dispersion to a value approximating 1. In late 1994, Venables posted a GLM-based NB2 model to StatLib using S-Plus, and SAS (Johnston) incorporated the negative binomial into its GENMOD procedure in 1998, with the same links offered in Stata and Xplore. SPSS did not offer a GLM procedure until 2006 with the release of version 15. A negative binomial option with all three links was included.

R has offered its users GLM-based negative binomial models through the **glm.nb** and **negative.binomial** functions, which are functions in the **MASS** package that is normally included when installing R from the web. Packages such as **gamlss** and **pscl** also provide negative binomial options, but they are based on full maximum likelihood estimation.

Currently nearly all GLM software includes a negative binomial family, and several of the major statistical applications, like Stata and LIMDEP, offer independent maximum likelihood negative binomial commands.

Full maximum likelihood models were also being developed for extended negative binomial models during this time. Geometric hurdle models were developed by Mullahy (1986), with a later enhancement to negative binomial hurdle models. Prem Consul and Felix Famoye have developed various forms of

Cambridge University Press
978-0-521-19815-8 - Negative Binomial Regression: Second Edition
Joseph M. Hilbe
Excerpt
More information

generalized negative binomial models using generalized maximum likelihood, as well as other mixture models. They have worked singly as well as jointly for some 30 years investigating the properties of such models – but the models have never gained widespread popularity.

William Greene's LIMDEP was the first commercial package to offer maximum likelihood negative binomial regression models to its users (2006a [1987]). 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 NB1 and NB2 parameterizations. In 1998, Stata offered a generalized negative binomial, **gnbreg**, in which the heterogeneity 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, NB-H, since the model extends NB2 to permit observed sources of heterogeneity in the overdispersion parameter.

Gauss and MATLAB also provide their users with the ability to estimate maximum likelihood negative binomial models. In Matlab one can use the maximum likelihood functions to rather easily estimate NB2 and NB1 models. Gauss provides modules in their *Count* library for handling NB2, as well as truncated and censored negative binomial models. Only LIMDEP and R provide both truncated and censored negative binomial modeling capability. In the meantime, LIMDEP has continuously added to its initial negative binomial offerings. It currently estimates many of the negative binomial-related models that shall be discussed in this monograph. In 2006 Greene developed a new parameterization of the negative binomial, NB-P, which estimates both the traditional negative binomial ancillary parameter, as well as the exponent of the second term of the variance function.

I should perhaps reiterate 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; this is the case when the Poisson variance defined as μ and the gamma as μ^2/ν, is used to create the negative binomial variance characterized as $\mu + \mu^2/\nu$. This parameterization is the same as that originally derived by Greenwood and Yule (1920). An inverse relationship between μ and ν 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 and Lee (1992) developed his *kk* system, 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 (1994) continued to prefer the direct relationship in his subsequent writings. Still, referencing 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 first suggested by Bliss and Owen (1958) and was favored by Breslow (1984) and Lawless (1987) in their highly influential seminal articles on the negative binomial. In 1990s, the direct relationship was used in the major software implementations of the negative binomial: Hilbe (1993b, 1994b) for XploRe and Stata, Greene (2006a) for LIMDEP, and Johnston (1997) for SAS. The direct parameterization was also specified in Hilbe (1994b), 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. Since a negative binomial algorithm cannot estimate $\alpha = 0$, owing to division by zero in the estimating algorithm, values for data seen in practice typically range from 0.01 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 strictly speaking a GLM.

Finally, it should be reiterated that Stata's **glm** command, R's **glm.nb** function, and SAS's **GENMOD** procedure are IRLS (iteratively reweighted least squares)-based applications in which the negative binomial heterogeneity parameter, α, is estimated using an external maximum likelihood mechanism, which then inserts the resulting value into the GLM algorithm as a constant. This procedure allows the GLM application to produce maximum likelihood