Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

The negative binomial-Lindley generalized linear model: Characteristics and application using crash data

Srinivas Reddy Geedipally[a,*], Dominique Lord[b,1], Soma Sekhar Dhavala[c,2]

[a] Texas Transportation Institute, Texas A&M University, 3135 TAMU, College Station, TX 77843-3135, United States
[b] Zachry Department of Civil Engineering, Texas A&M University, 3136 TAMU, College Station, TX 77843-3136, United States
[c] Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843-3143, United States

ARTICLE INFO

Article history:
Received 28 June 2011
Received in revised form 18 July 2011
Accepted 18 July 2011

Keywords:
Poisson-gamma
Negative binomial Lindley
Generalized linear model
Crash data

ABSTRACT

There has been a considerable amount of work devoted by transportation safety analysts to the development and application of new and innovative models for analyzing crash data. One important characteristic about crash data that has been documented in the literature is related to datasets that contained a large amount of zeros and a long or heavy tail (which creates highly dispersed data). For such datasets, the number of sites where no crash is observed is so large that traditional distributions and regression models, such as the Poisson and Poisson-gamma or negative binomial (NB) models cannot be used efficiently. To overcome this problem, the NB-Lindley (NB-L) distribution has recently been introduced for analyzing count data that are characterized by excess zeros. The objective of this paper is to document the application of a NB generalized linear model with Lindley mixed effects (NB-L GLM) for analyzing traffic crash data. The study objective was accomplished using simulated and observed datasets. The simulated dataset was used to show the general performance of the model. The model was then applied to two datasets based on observed data. One of the dataset was characterized by a large amount of zeros. The NB-L GLM was compared with the NB and zero-inflated models. Overall, the research study shows that the NB-L GLM not only offers superior performance over the NB and zero-inflated models when datasets are characterized by a large number of zeros and a long tail, but also when the crash dataset is highly dispersed.

Published by Elsevier Ltd.

1. Introduction

Regression models play a significant role in highway safety. These models can be used for various purposes, such as establishing relationships between motor vehicle crashes and different covariates (i.e., understanding the system), predicting values or screening variables. As documented in Lord and Mannering (2010), there has been a considerable amount of work devoted by transportation safety analysts to the development and application of new and innovative models for analyzing count data. The development and application of new statistical methods are fostered by the unique characteristics associated with crash data. One important characteristic that has been documented in the literature is related to datasets that contained a large amount of zeros and a long or heavy tail (which creates highly dispersed data). For such datasets, the number of sites where no crash is observed is so large that traditional distributions and regression models, such as the Poisson and Poisson-gamma or negative binomial (NB) models, cannot be used efficiently.

In order to overcome this important problem, researchers have proposed the use of the zero-inflated model (both used for the Poisson and NB distributions) to analyze this kind of dataset (Miaou, 1994; Shankar et al., 1997; Kumara and Chin, 2003; Shankar et al., 2003). This type of model assumes that the zeros are generated using a two-state data generating process: zero or safe state and non-zero state. Although these models may offer a better statistical fit, a few researchers (Warton, 2005; Lord et al., 2005, 2007) have raised important methodological issues about the use of such models, including the fact that the safe state has a distribution with a long-term mean equal to zero. This latter characteristic is obviously theoretically impossible. So far, there has been no regression model that has been available for properly and fully analyzing crash data with an abundant number of zeros.[3] Such models are particularly

[*] Corresponding author. Tel.: +1 979 862 1651; fax: +1 979 845 6006.
E-mail addresses: srinivas-g@ttimail.tamu.edu (S.R. Geedipally), d-lord@tamu.edu (D. Lord), soma@stat.tamu.edu (S.S. Dhavala).
[1] Tel.: +1 979 458 3949; fax: +1 979 845 6481.
[2] Tel.: +1 979 845 3141; fax: +1 979 845 3144.

[3] Mayshkina and Mannering (2009) have proposed a zero-state Markov switching model, which overcomes some of the criticisms discussed above.

0001-4575/$ – see front matter. Published by Elsevier Ltd.
doi:10.1016/j.aap.2011.07.012

needed when changing the characteristics of the dataset cannot be done or is difficult to accomplish (see Lord and Geedipally, 2011). Under this scenario, the large number of zeros could still create many difficulties for adequately analyzing such dataset.

The objective of this paper is to document the application of a NB generalized linear model with Lindley mixed effects (NB-L GLM) for analyzing traffic crash data. This new model is based on the recently introduced NB-Lindley (NB-L) distribution for analyzing count data (Zamani and Ismail, 2010; Lord and Geedipally, 2011). The NB-L distribution is, as the name implies, a mixture of the NB and the Lindley distributions (Lindley, 1958; Ghitany et al., 2008). This two-parameter distribution has interesting and thorough theoretical properties in which the distribution is characterized by a single long-term mean that is never equal to zero and a single variance function, similar to the traditional NB distribution.

The study objective was accomplished using simulated and observed datasets. The simulated dataset was used to show the general performance of the model. The model was then applied to two datasets, one of which is characterized by a large amount of zeros. For both datasets, the observed dispersion was very large. The NB-L GLM was compared with the NB and zero-inflated models. The reader needs to bear in mind that regression models, such as the Poisson-gamma, Poisson-lognormal or the NB-L model are used as an approximation tool for analyzing the crash process (Lord et al., 2005). This process is known as the Poisson trials with unequal probability of events.

The next section describes the characteristics of the NB-L GLM.

2. Characteristics of the NB-L GLM

This section describes the characteristics of the NB-L distribution and the GLM for modeling crash data.

The NB-L distribution is a mixture of negative binomial and Lindley distributions. This mixed distribution has a thick tail and works well when the data contains large number of zeros or is highly dispersed. In other situations (e.g., less dispersed data, etc.), it works similar to that of the NB distribution.

Before tackling the NB-L, it is important to first define the NB distribution. The NB distribution can be parameterized in two different manners, either as a mixture of the Poisson and gamma distributions or based on a sequence of independent Bernoulli trials. Using the latter parameterization, the probability mass function (pmf) of the NB distribution can be given as:

$$P(Y=y;\ \ \phi,p)=\frac{\Gamma(\phi+y)}{\Gamma(\phi)\times y!}(p)^{\phi}(1-p)^{y};\quad \phi>0,\ 0<p<1 \tag{1}$$

The parameter 'p' is defined as the probability of failure in each trial and is given as:

$$p=\frac{\phi}{\mu+\phi} \tag{2}$$

where μ = mean response of the observation and φ = inverse of the dispersion parameter α (i.e. $\phi = 1/\alpha$).

In the context of the NB GLM, the mean response for the number of crashes is assumed to have a log-linear relationship with the covariates and is structured as:

$$\ln(\mu)=\beta_0+\sum_{i=1}^{q}\beta_i X \tag{3}$$

where X = traffic and geometric variables of a particular site, β_s = regression coefficients to be estimated and q = total number of covariates in the model.

Then, it can be shown that the variance is equal to (Casella and Berger, 1990):

$$var(Y)=\phi\frac{p}{(1-p)^2}=\frac{1}{\phi}\mu^2+\mu \tag{4}$$

Using Eqs. (1) and (2), the pmf of the NB distribution and its GLM can be re-parameterized this way (as a Poisson-gamma model):

$$\begin{aligned}P(Y=y,\mu,\phi) &= NB(y;\ \ \phi,\mu)\\ &=\frac{\Gamma(\phi+y)}{\Gamma(\phi)\Gamma(y+1)}\left(\frac{\phi}{\mu+\phi}\right)^{\phi}\left(\frac{\mu}{\mu+\phi}\right)^{y}\end{aligned} \tag{5}$$

The pmf in Eq. (5) is the one normally used for analyzing crash count data.

The NB-L distribution[4] is defined as a mixture of NB and Lindley distributions such that:

$$P(Y=y,\mu,\phi,\theta)=\int \mathrm{NB}(y;\phi,\varepsilon\mu)\,\mathrm{Lindley}(\varepsilon;\theta)\ \ d\varepsilon \tag{6}$$

here $f(u;a,b)$ means that f is the distribution of the variable μ, with parameters a and b. The parameter μ is similar to the one described in Eq. (3) and ε follows the Lindley distribution.

The Lindley distribution is a mixture of exponential and gamma distributions (Lindley, 1958; Ghitany et al., 2008; Zamani and Ismail, 2010; Lord and Geedipally, 2011). The pmf of the Lindley distribution can be defined as follows:

$$f(X=x;\theta)=\frac{\theta^2}{\theta+1}(1+x)e^{-\theta x};\quad \theta>0,\quad x>0 \tag{7}$$

The first moment (i.e., the mean) of the Lindley distribution is given as (Ghitany et al., 2008):

$$E(\varepsilon)=\frac{\theta+2}{\theta(\theta+1)} \tag{8}$$

The second moment of the Lindley distribution is given as (Ghitany et al., 2008):

$$E(\varepsilon^2)=\frac{2(\theta+3)}{\theta^2(\theta+1)} \tag{9}$$

Thus, if the number of crashes Y is assumed to follow a NB-L (ϕ, p) distribution, then the mean function can be given as:

$$E(Y)=\mu\times E(\varepsilon)=e^{\beta_0+\sum_{i=1}^{p}\beta_i X}\,\frac{\theta+2}{\theta(\theta+1)} \tag{10}$$

If $\beta'_0=\beta_0+\log((\theta+2)/(\theta(\theta+1)))$, then the parameters in the equation above can be directly compared with the parameters described in Eq. (3).

The crash variance is given by the Eq. (11) below:

$$\begin{aligned}Var(Y)&=\mu\times\frac{\theta+2}{\theta(\theta+1)}+\mu^2\times\frac{2(\theta+3)}{\theta^2(\theta+1)}\times\frac{(1+\phi)}{\phi}\\ &\quad-\left(\mu\times\frac{\theta+2}{\theta(\theta+1)}\right)^2\end{aligned} \tag{11}$$

3. Parameter estimation

As discussed in the previous section, the likelihood function for the NB-L model is given by Eq. (6), where the mean response for the number of crashes 'μ' is assumed to have a log-linear relationship with the covariates as given in Eq. (3).

A very important characteristic associated with this equation is related to the fact that the involved integral does not have a

[4] The NB-L distribution in this work has slightly been re-parameterized from the original paper by Lord and Geedipally (2011) in order to fully develop the GLM.