The Negative Binomial-Lindley Generalized Linear Model: Characteristics and Application using Crash Data

Srinivas Reddy Geedipally[1]
Engineering Research Associate
Texas Transportation Institute
Texas A&M University
3135 TAMU
College Station, TX 77843-3135
Tel. (979) 862-1651
Fax. (979) 845-6006
Email: srinivas-g@ttimail.tamu.edu

Dominique Lord
Associate Professor
Zachry Department of Civil Engineering
Texas A&M University
3136 TAMU
College Station, TX 77843-3136
Tel. (979) 458-3949
Fax. (979) 845-6481
Email: d-lord@tamu.edu

Soma Sekhar Dhavala
Post Doctoral Researcher
Department of Statistics
Texas A&M University
3143 TAMU
College Station, TX 77843-3143
Tel. (979) 845-3141
Fax. (979) 845-3144
Email: soma@stat.tamu.edu

Forthcoming in Accident Analysis & Prevention

July 18, 2011

[1] Corresponding author

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.

Keywords: Poisson-gamma, negative binomial Lindley, generalized linear model, crash data.

1.0 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 (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[1]. Such models are particularly 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-

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

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.

2.0 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:

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

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;

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

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

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

The NB-L distribution[2] is defined as a mixture of NB and Lindley distributions such that:

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

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 Equation (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:

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

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):

[2] 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.

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

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 $\beta_0^{'} = \beta_0 + \log\left(\frac{\theta+2}{\theta(\theta+1)}\right)$, then the parameters in the equation above can be directly compared with the parameters described in Equation (3).

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

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

3.0 PARAMETER ESTIMATION

As discussed in the previous section, the likelihood function for the NB-L model is given by Equation (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 Equation (3).

A very important characteristic associated with this equation is related to the fact that the involved integral does not have a closed form. It can be solved elegantly using the hierarchical representation implicit both in the integrand and in the definition of the Lindley distribution itself. That is, the NB-L distribution conditional upon the unobserved site-specific frailty term ε, that explains additional heterogeneity, can be re-written as follows:

$$\begin{aligned} &P(Y = y, \mu, \phi \mid \varepsilon) = \text{NB}(y; \phi, \varepsilon\mu) \\ &\varepsilon \sim \text{Lindley}(\varepsilon; \theta) \end{aligned} \tag{12}$$

The above formulation can be thought of as an instance of the Generalized Linear Mixed model (GLMM) where the mixed effects (or the frailty terms) follow the Lindley distribution. However, considering that the Lindley is not a standard distribution, the hierarchical representation of the Lindley distribution can be further utilized.

Recall that the Lindley distribution is a two component mixture given by (Zamani and Ismail, 2010):

$$\varepsilon \sim \frac{1}{1+\theta}\text{Gamma}(2,\theta) + \frac{\theta}{1+\theta}\text{Gamma}(1,\theta) \tag{13}$$

Recognizing the special structure in the mixture components, the above equation can be re-written as follows:

$$\varepsilon \sim \sum \text{Gamma}(1+\text{z},\theta)\,\text{Bernoulli}\left(\text{z};\frac{1}{1+\theta}\right) \tag{14}$$

A hierarchical representation of the Lindley distribution can be represented as,

$$\begin{aligned} &\varepsilon \sim \text{Gamma}\,(1+z,\theta)\text{, and} \\ &z \sim \text{Bernoulli}\left(\frac{1}{1+\theta}\right), \end{aligned} \tag{15}$$

whose marginal distribution is the Lindley distribution. The complete multi-level hierarchical model can now be given as:

$$\begin{aligned} &P(Y=y,\mu,\phi\,|\,\varepsilon) = \text{NB}(y;\phi,\varepsilon\mu) \\ &\varepsilon \sim \text{Gamma}(\varepsilon;1+\text{z},\theta) \\ &\text{z} \sim \text{Bernoulli}\left(z;\frac{1}{1+\theta}\right) \end{aligned} \tag{16}$$

The above formulation has a nice Bayesian interpretation. That is, the observations follow the NB distribution and the site-specific frailty term follows a gamma distribution *a priori*. The shape parameter of the gamma distribution follows a Bernoulli distribution *a priori*. In this context, the above model can be seen as a hierarchical model involving a standard distribution at all stages. Consequentially, inference can be carried-out quite routinely using Markov chain Monte Carlo (MCMC) and an easy-to-use software tool, such as WinBUGS (Spiegelhalter et al., 2003). It should be pointed out, however, that the Bayesian formulation requires elicitation of priors on all the unknown parameters (in this case $\beta, \phi,$ and θ). In this study, normal priors for β, a beta prior for $1/1+\theta$ and a gamma prior for $1/\phi$ were used. The re-parameterization of the latter two parameters helps in improved convergence of the MCMC chains and hence the numerical accuracy of the estimates. Additional discussion about convergence issues is described in Section 6.0.

A total of three Markov chains were used in the model estimation process with 30,000 iterations per chain. The first 15,000 iterations (burn-in samples) were discarded. Thus, the remaining

15,000 iterations were used for estimating the coefficients. The Gelman-Rubin (G-R) convergence statistic was used to verify that the simulation runs converged properly. In the analysis, the research team ensured that G-R statistic was less than 1.1. For comparison, Mitra and Washington (2007) suggested that convergence was achieved when the G-R statistic was less than 1.2.

4.0 DATA DESCRIPTION

This section describes the characteristics of the two datasets. The first part summarizes the characteristics of the Indiana data. The second part presents the summary statistics for the single-vehicle crash data that occurred on two-lane rural highways in Michigan. Both datasets contained several variables, which were used to minimize the omitted bias problem that can plague the development of crash prediction models (Lord and Mannering, 2010).

4.1 INDIANA DATA

The first dataset contained crash and traffic data collected for a five-year period (1995 to 1999) at 338 rural interstate road sections in the state of Indiana. The data have previously been used for estimating a model of accident rates using a tobit regression approach (Anastasopoulus et al., 2008; Washington et al., 2011). In this dataset, 120 out of the 338 highway segments did not have any reported crashes over the 5-year period (~36% are 0s). Table 1 presents the summary statistics of the variables used for developing the models in this study. For a complete and detailed list of variables, the interested reader is referred to Washington et al. (2011).

Table 1. Summary Statistics for the Indiana Data.

Variable	Min.	Max.	Average (std. dev)	Total
Number of Crashes (5 years)	0	329	16.97 (36.30)	5737
Average daily traffic over the 5 years (ADT)	9442	143,422	30237.6 (28776.4)	--
Minimum friction reading in the road segment over the 5-year period (FRICTION)	15.9	48.2	30.51 (6.67)	--
Pavement surface type (1 if asphalt, 0 if concrete) (PAVEMENT)	0	1	0.77 (0.42)	--
Median width (in feet) (MW)	16	194.7	66.98 (34.17)	--
Presence of median barrier (1 if present, 0 if absent) (BARRIER)	0	1	0.16 (0.37)	--
Interior rumble strips (RUMBLE)	0	1	0.72 (0.45)	--
Segment length (in miles) (L)	0.009	11.53	0.89 (1.48)	300.09

4.2 MICHIGAN DATA

The second dataset contained single-vehicle crashes that occurred on rural two-lane highways in Michigan for the year 2006. This database, which was originally collected for the Federal

Highway Administration's (FHWA) Highway Safety Information System (HSIS), was used by Qin et al. (2004) for developing zero-inflated regression models. The database included 33,970 segments. For this dataset, about 70% of the segments experienced no crash for the year 2006. The large number of zeros for this dataset can be explained by the sample that contains very short segments (about 87% are less than 0.3 mile). It would consequently be very difficult to change the spatial scale to reduce the number of zeros (see Lord and Geedipally, 2011, for further discussion on this topic). Table 2 presents the summary statistics for the Michigan data.

Table 2. Summary Statistics for the Michigan Data (1996).

	Min.	**Max.**	**Average (std. dev)**	**Total**
Number of Crashes (1 year)	0	61	0.68 (1.77)	23168
Annual average daily traffic (AADT)	160	20,994	4507.5 (3280.6)	--
Segment length (L) (miles)	0.001	54.54	0.18 (0.58)	6212
Shoulder width (in feet) (SW)	0	24	16.94 (5.26)	--
Lane width (in feet) (LW)	8	15	11.22 (0.78)	--
Speed limit (SPEED) (mph)	25	55	52.47 (6.39)	--

5.0 MODELING RESULTS

This section presents the modeling results for the NB-L GLMs as well as for the NB and zero-inflated models and is divided into three parts. The first part explains the modeling results for the simulated data. The second part provides details about the modeling results for the Indiana data. The last part documents the modeling results for the Michigan data.

5.1 SIMULATED DATA

This section presents the results of the simulation study intended to illustrate the general performance of the NB-L model. The simulation was performed for a large sample size in order to remove potential biases associated with the small sample size problem (Lord, 2006). Consequently, the sample used for estimating the model included 1,500 observations. The simulation design was carried out in several steps. In the first step, the independent variables were taken from an existing database (for this example, single-vehicle road departure crashes on rural two-lane horizontal curves in San Antonio, Texas were used. This is a subset of the data presented in Table 3 of Lord and Geedipally, 2011). A total of 1,909 observations were produced. From those, a sample of 1,500 observations was randomly selected. In the second step, the model coefficients were assumed in such a way that they are logical and comparable with existing literature for those variables. These coefficients are labeled as "true" parameters. During the third step, the crash mean was calculated from the independent variables and the "true" parameters using Equation (10). Crashes are then simulated using the NB-L distribution. Once the dataset was created, the parameters were re-estimated using the NB-L model and then compared with the "true" parameters.

Table 3 presents the modeling results for the simulated dataset. This table shows that the NB-L model was able to reproduce the "true" parameter values. All coefficients were statistically significant at the 5% level.

Table 3. Modeling Results for the Simulated Data.

Parameters	True value	Estimated Values
β_0	-6.0	-6.019 (0.415)
β_1	0.6	0.6082 (0.059)
β_2	0.2	0.2045 (0.025)
β_3	-0.05	-0.038 (0.027)
$\alpha = 1/\phi$	0.6	0.6616 (0.194)

5.2 INDIANA DATA

Table 4 summarizes the results for the Indiana data. The segment length variable is considered as an offset which means that the crashes increase linearly with the increase in segment length. To compare the NB-L GLM with zero-inflated models, zero-inflated Poisson (ZIP) and zero-inflated NB (ZINB) models were also estimated (although the dataset did not contain as many zeros as for the other dataset). However, the ZINB model provided a better fit for this dataset. This table shows that the coefficients for the flow parameters are below one for the NB-L and NB, which indicates that the crash risk increases at a decreasing rate as traffic flow increases. The ZINB model shows that crashes increase almost linearly with the increase in flow. It should be pointed out that the 95% marginal posterior credible intervals for each of the coefficients did not include the origin. Except for the Pearson's chi-square, all other GOF statistics in Table 4 shows that NB-L provides superior fit to the NB and ZINB models. The estimated coefficients between the models all have the same sign, but their values are not always very close. The standard errors for the estimated coefficients are slightly larger for the NB-L when compared to that of NB. It should be noted that the NB-L is a multi-level hierarchical model and the effective number of parameters could be larger than that in a simple parametric model alternative, such as the NB model. As a result, the effective degrees of freedom could be smaller, leading to increased standard errors. However, due to the frailty terms that explain additional heterogeneity, it compensates for the increased model complexity by improving the predictive modeling ability, which is reflected in the MSPE that considers both bias and variance.

Table 4. Modeling Results for the Indiana Data.

Variable	NB		NB-L		ZINB†	
	Value	Std. dev	Value	Std. dev	Value	Std. dev
INTERCEPT (β_0)	-4.779	0.979	-3.739	1.115	-8.3381	1.126
Ln(ADT) (β_1)	0.7219	0.091	0.630	0.106	1.0845	0.105
FRICTION (β_2)	-0.02774	0.008	-0.02746	0.0111	-0.0205	0.008
PAVEMENT (β_3)	0.4613	0.135	0.4327	0.217	0.2306	0.151
MW (β_4)	-0.00497	0.001	-0.00616	0.002	-0.0023	0.002
BARRIER (β_5)	-3.195	0.234	-3.238	0.326	-1.5095	0.389
RUMBLE (β_6)	-0.4047	0.131	-0.3976	0.213	-0.511	0.151
$\alpha = 1/\phi$	0.934	0.118	0.238	0.083	0.375	0.056
DIC[1]	1900		1701		1850‡	
MAD[2]	6.91		6.89		8.04	
MSPE[3]	206.76		195.54		268.01	
Pearson χ^2	1174		978		851	
MCPD[4]	454		261		778	

[1] Deviance Information Criterion; [2] Mean Absolute Deviance (Oh et al, 2003); [3] Mean Squared Predictive Error (Oh et al, 2003); [4] Maximum Cumulative Residual Plot Deviation (Geedipally et al, 2010).
†Estimated using the MLE and the inflated parameters are not presented here; ‡AIC

A cumulative residual (CURE) plot presents how the model fits the data with respect to each covariate by plotting the cumulative residuals in the increasing order for each key covariate (Hauer and Bamfo, 1997). A better fit occurs when the cumulative residuals oscillate more closely around the value of zero for a given covariate. The figure can also be used to identify potential biases within the range of the variable investigated (i.e., when the predicted values almost always over- or under-estimate the observed values for the entire or a large portion of the range for the variable investigated).

Figure 1 shows the CURE plot for the ADT variable. The plots were adjusted for the final cumulative value to be equal to zero. The figure clearly shows that the NB-L fits the data better even when the curves are adjusted. Although not shown here, it should be pointed out that the final values for the unadjusted NB, ZINB and NB-L curves are equal to 398, 292, and 155 respectively. Finally, as shown in last row in Table 4, the maximum deviation, calculated from the unadjusted curves, for the NB-L model is much smaller than that one calculated for the NB and ZINB models.

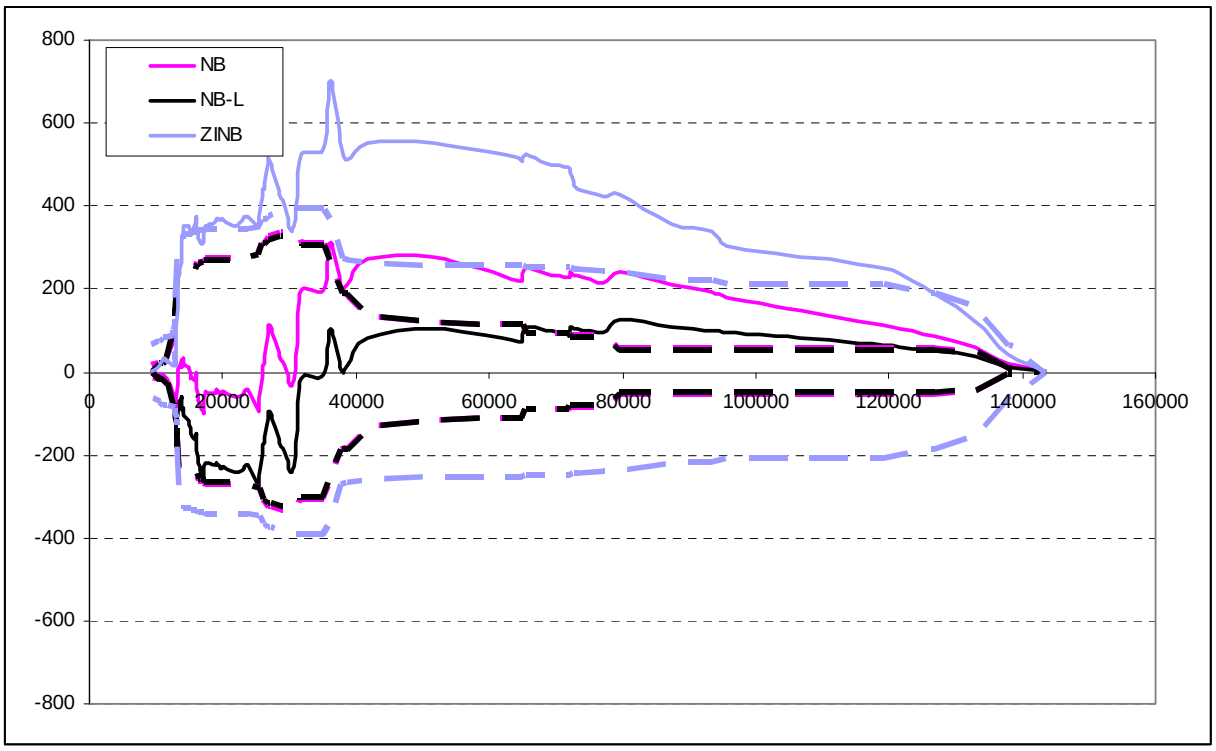

Figure 1. Cumulative Residual Plot for Indiana data (ADT Variable).
Note: Dotted lines represent ± 2 Std. Dev.

At this point, it is important to discuss the issue related to the use of GOF statistics for comparing different models. As discussed in Miaou and Lord (2003) and Lord et al. (2005; 2007), the primary goal for analyzing regression models should not be solely based on finding the absolute best statistical fit. It is also very important to look at the data generating process, the relationship between the variables and whether the distribution or model is logically or theoretically sound. Miaou and Lord (2003) referred to the latter characteristic as "goodness-of-logic" (GOL). Consequently, the transportation safety analyst needs to consider both the GOF and GOL when assessing competitive models.

5.3 MICHIGAN DATA

Table 5 summarizes the results for the Michigan data. To be consistent with the functional form used by Qin et al. (2004) for this dataset, the segment length was used as a covariate rather than as an offset. However, the parameter for segment length is almost equal to one which suggests that the crashes increase linearly with the increase in segment length. Similar to the first dataset, the 95% marginal posterior credible intervals for each of the coefficients did not include the origin. For this dataset, Qin et al. (2004) only estimated a ZIP model. Since it was found that ZINB model fitted the data much better than the ZIP model, only the former model is presented in Table 5.

Table 5. Modeling Results for the Michigan Data.

Variable	NB		NB-L		ZINB†	
	Value	Std. dev	Value	Std. dev	Value	Std. dev
INTERCEPT (β_0)	-3.412	0.239	-3.2607	0.193	-3.1503	0.225
Ln(AADT) (β_1)	0.4267	0.014	0.4243	0.015	0.4205	0.013
L (β_2)	0.9571	0.009	0.9615	0.009	0.9579	0.008
SW (β_3)	-0.00009	0.002	-0.0003	0.002	0.0002	0.002
LW (β_4)	0.0589	0.013	0.0508	0.011	0.0516	0.012
SPEED (β_5)	0.0098	0.002	0.0091	0.002	0.0077	0.002
$\alpha = 1/\phi$	0.5727	0.019	0.1024	0.002	0.5588	0.024
DIC	59354		56046		59341‡	
MAD	0.651		0.648		0.651	
MSPE	2.831		2.884		2.863	
Pearson χ^2	49911		44774		60614	
MCPD	701		657		702	

†Estimated using the MLE and the inflated parameters are not presented here; ‡AIC

The results show that the NB-L model also performed better than the NB and ZINB models. Table 5 shows that the NB and ZINB are very close, but it can be demonstrated that the ZINB actually provides a slightly better fit (although the MLE result for NB is not presented here, the difference in AIC values between ZINB and NB is greater than 10). The magnitude of the difference between these two, the NB and ZINB models, is actually very similar to what has been documented in previous work on this topic (see, e.g., Kumara and Chin, 2003; Yau et al., 2003). With the exception of one highly insignificant variable for the ZINB model, all the variables have the same sign.

Figure 2 shows the (adjusted) CURE plots for the AADT variable. This figure seems to show that the NB and ZINB offers a better fit, but when the unadjusted NB, ZINB and NB-L curves are examined, the final values for the cumulative curves are equal to -322.1, -281.1 and -3.9, respectively. In other words, the NB-L curve almost arrives at zero naturally. Furthermore, based on the unadjusted curves, the NB and ZINB seem to show a biased estimate since the difference between the predicted and observed values is almost always negative over the entire range of the flow variable. Similar to the previous dataset, the maximum deviation for the NB-L model is smaller than that of the NB and ZINB models.

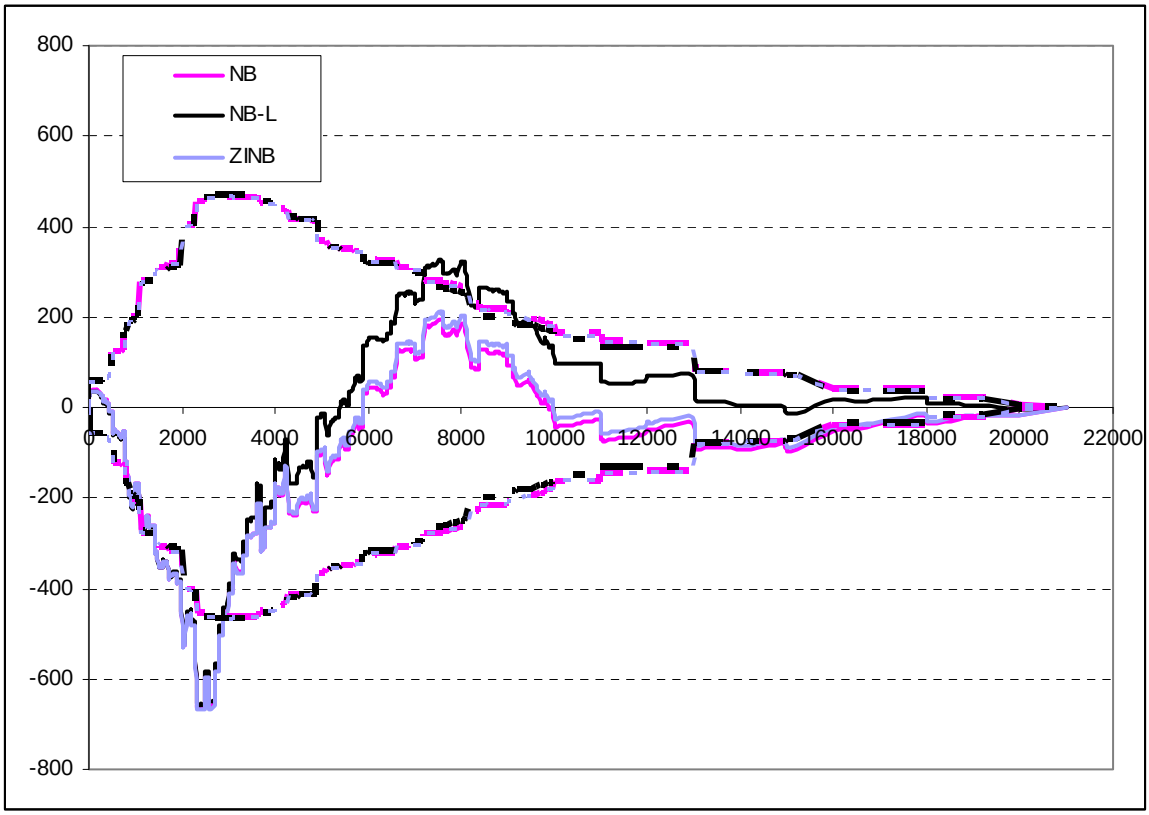

Figure 2. Cumulative Residual Plot for Michigan data (AADT Variable).
Note: Dotted lines represent ± 2 Std. Dev.

6.0 DISCUSSION

Since this is a brand new GLM that has never been applied before, there are several interesting findings that need to be discussed. First, the modeling results showed that the NB-L model performs much better than the traditionally used NB model for modeling traffic crashes, at least for these two datasets. As shown by various the GOF measures and CURE plots, the NB-L model always provided a superior fit compared to the NB model, while still maintaining sound theoretical properties (Miaou and Lord, 2003). The modeling results show that not only did the NB-L offers superior performance when datasets are characterized by a large number of zeros and a long tail, but also was better when the crash dataset is highly dispersed even for high sample mean values. When the dispersion becomes smaller, the NB-L model actually resorts back to a NB model (Lord and Geedipally, 2011). Thus, at worse, the NB-L model performs as well as the NB, which means that one could theoretically only need to use the NB-L model when data are over-dispersed. Obviously, more work needs to be done in this regards.

Second, the original purpose for developing and using the NB-L distribution was to handle datasets that contain a large number of zeros (Zamani and Ismail, 2010, Lord and Geedipally, 2011). The distribution did indeed work much better than the NB distribution for datasets characterized as such. The next step consisted in examining whether the model performed better

than previous models that have been proposed in the past (e.g. zero-inflated models). To accomplish this objective, the authors developed zero-inflated models using both datasets, even though the Indiana data are not characterized by excess zeros. The results clearly showed that the NB-L GLM was much better than the ZINB estimated in this work. Although not shown here, the authors compared the NB-L model with the ZIP and ZINB models with two other flow-only datasets and the results were consistent with those shown above. Although the NB-L model provided a better fit and is a more theoretically sound, further work needs to be done on this topic. For instance, when the data are truly generated by a multi-state process, zero-inflated and finite mixture models may still be the model of choice even if the NB-L fits the data better (see Kadane et al., 2006; Park and Lord, 2009).

Third, as discussed in Section 2.0, the parameterization of the NB-L used in this work is slightly different than the parameterization described in Lord and Geedipally (2011) and Zamani and Ismail (2010). When a GLM is considered with the original formulation of the NB-L likelihood, the mean response is a non-linear, non-invertible function of the covariates and the parameters. This can be understood by considering the following GLM modeling framework:

$$
\begin{aligned}
P(Y = y;\phi, p) &= \frac{\Gamma(\phi + y)}{\Gamma(\phi)\times y!}(p)^{\phi}(1-p)^{y} \\
\lambda &= \text{-ln(p)} \sim \text{Lindley}(\theta) \\
\ln(\theta) &= \beta_0 + \sum_{i=1}^{q} \beta_i X
\end{aligned}
\qquad (17)
$$

The advantage with this characterization is that the likelihood is available in closed form after integrating λ in the above equation. However, the mean response, as given by Equation (7) in Lord and Geedipally (2011), is nonlinear in θ, which makes it difficult to characterize the predicted response. On the contrary, the parameterization based on Equation (5) is easily interpretable. For example, as given in Equation (10), the predicted mean response is scaled by a certain factor. This is even more evident by looking at the hierarchical representation shown in Equation (16): each site-specific mean response is multiplied by its own frailty term. This is equivalent to adding site-specific offset terms in the log-transformed domain of the mean response, similar to an additional random covariate at each site.

Despite the nice interpretability offered by this characterization, MCMC chains still suffer from poor mixing. This often results from two scenarios: some parameters are not identifiable or they are strongly correlated. This problem can be mitigated by some kind of regularization or re-parameterization of the parameters or a combination of both. One way to solve the problem is to model the dispersion parameter (α) instead of the inverse dispersion parameter (ϕ). However, some strong correlation still exists between β_0 and θ, as can be seen by absorbing all constant

terms into the intercept: $\beta_0^{'} = \beta_0 + \log\left(\frac{\theta+2}{\theta(\theta+1)}\right)$. To solve this problem, the underlying hierarchical representation of the Lindley distribution can be resorted to offer some insight. The scale parameter of the Gamma distributions in the mixture is θ , whereas the mixing probability is $\frac{1}{1+\theta}$. Clearly, the latter parameter is restricted to lie in the unit interval, since it can be interpreted as the probability of choosing one of the Gamma distributions for each site. A Beta distribution can be elicited and the hyper-priors are chosen such that, *a priori*, $E(Y) = \mu \times E(\varepsilon) = \mu$. This re-parameterization also helps in regularization, since eliciting informative priors is now possible. That is, suppose a Beta distribution is elicited for $\frac{1}{1+\theta}$, then its conditional posterior distribution is also Beta. To further illustrate this point, suppose a Beta(1/3,1/2) is chosen, it still ensures that, *a priori*, $E(\varepsilon) = 1$. However, in the presence of the likelihood, this becomes completely irrelevant. Thus, a reasonable choice for the prior distribution is Beta(*n*/3, *n*/2), where *n* is the total number of observations.

Fourth, since the NB-L model involved additional parameters when compared to NB model, the computational time for MCMC runs was increased. However, the difference in computational times between the two models was not very large. Hence, there is nothing that prevents a transportation safety analyst to use the NB-L GLM for analyzing crash data.

There are several avenues for further work. First, given the fact that crash data are often subjected to low sample mean values and small sample size, the stability of the NB-L GLM should be investigated. Second, since the EB method is now used frequently in highway safety analyses, an EB modeling framework should be developed for the NB-L model. Third, the difference for identifying hazardous sites between the NB-L and NB models should be investigated. Fourth, although the analysis carried out in this research was conducted by assuming a fixed dispersion parameter α (independent of the covariates), further research should be done to examine the effects of a covariate-dependent dispersion parameter on NB-L GLMs. Fifth, the parameterization of the NB-L model used in this study is slightly different than the one documented in Lord and Geedipally (2011) and Zamani and Ismail (2010). Thus, a well-defined likelihood function and the related moments for the NB-L model should be built. This way, the maximum likelihood estimation (MLE) method could be used for estimating NB-L GLMs. Sixth, a time-dependent NB-L model should be developed and compared with the zero-state Markov switching models proposed by Mayshkina and Mannering (2009).

Finally, the recently introduced random-parameters count model (Anastasopoulos and Mannering, 2009) has been shown to significantly improve the statistical fit compared to that of the NB model. Because of this property, this model has become quite popular since its introduction (e.g., El-Basyouny and Sayed, 2009; Dinu and Veeraragavan, 2011). Interestingly, the proposed NB-L GLM is also a random-parameters model, albeit with a notable difference. In the model proposed by these authors, the coefficients are random and are allowed to vary from site to site. In addition, the coefficients for each site could follow any distribution and are independent of each other. In the model proposed in this study, the intercept and the coefficients

are also random (because of the Bayesian framework), but only the intercept varies from site to site. It should be noted that it would be possible to allow the coefficients to vary from site to site for the NB-L. Given the characteristics of both models, it would therefore be of a great interest to compare their performance, especially when datasets are characterized by a large amount of zeros.

7.0 SUMMARY AND CONCLUSIONS

This paper has described the application of the NB-L GLM for analyzing crash data. The model was evaluated using simulated and observed crash datasets. For the two crash datasets, both were characterized by very high dispersion and one was also characterized by a large number of zeros. Traditional statistical methods (i.e., zero inflated models) that have been proposed for analyzing the datasets characterized by a large number of zeros have been found to suffer from important numerical and methodological problems. The newly introduced NB-L distribution offers the advantage of being able to handle datasets with a large number of zeros and/or high dispersion, while still maintaining similar characteristics as the traditional NB distribution. That is, the NB-L distribution is a two-parameter distribution and the long-term mean is never equal to zero. The results for the simulated data showed that the NB-L model was able to reproduce the "true" parameter values. The results have also shown that the NB-L always provided a better statistical fit relative to the NB and ZINB models for the two observed crash datasets. In conclusion, it is believed that the NB-L distribution and its GLM may offer a viable alternative to the traditionally used NB model for analyzing over-dispersed datasets.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Fred Mannering from Purdue University and Dr. Xiao Qin from South Dakota State University for providing us with the data.

REFERENCES

Anastasopoulos, P.C., Mannering, F.L., 2009. A note on modeling vehicle accident frequencies with random-parameters count models. Accident Analysis and Prevention 41(1), 153-159.

Anastasopoulos, P., Tarko, A., Mannering, F., 2008. Tobit analysis of vehicle accident rates on interstate highways. Accident Analysis and Prevention 40(2), 768-775.

Casella, G., Berger, R.L., 1990. Statistical Inference. Wadsworth Brooks/Cole, Pacific Grove, CA.

Dinu, R.R., Veeraragavan, A., 2011. Random parameter models for accident prediction on two-lane undivided highways in India Journal of Safety Research 42 (1), 39-42.

El-Basyouny, K., Sayed, T., 2009. Accident prediction models with random corridor parameters. Accidents Analysis and Prevention 41 (5), 1118–1123.

Geedipally, S.R., Patil, S., Lord, D. 2010. Examination of Methods for Estimating Crash Counts According to their Collision Type. Transportation Research Record 2165, pp. 12-20.

Ghitany, M.E., Atieh, B., Nadarajah, S., 2008. Lindley distribution and Its application. Mathematics and Computers in Simulation (78), 39-49.

Hauer, E., and J. Bamfo, 1997.Two Tools for Finding What Function Links the Dependent Variable to the Explanatory Variables. Proceedings of the ICTCT 1997 Conference, Lund, Sweden.

Kadane, J.B., Shmueli, G., Minka, T.P., Borle, S., Boatwright, P., 2006. Conjugate analysis of the Conway-Maxwell-Poisson distribution. *Bayesian Analysis*, Vol. 1, pp. 363-374.

Kumara, S.S.P., Chin, H.C., 2003. Modeling accident occurrence at signalized tee intersections with special emphasis on excess zeros. Traffic Injury Prevention 3(4), 53-57.

Lindley, D.V., 1958. Fiducial distributions and Bayes' theorem. J. R. Stat. Soc. (20), 102-107. http://www.jstor.org/stable/2983909

Lord, D., 2006. Modeling motor vehicle crashes using Poisson-gamma models: Examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter. Accident Analysis & Prevention 38 (4), 751-766.

Lord, D., Geedipally, S.R., 2011. The negative binomial–Lindley distribution as a tool for analyzing crash data characterized by a large amount of zeros. Accident Analysis & Prevention 43 (5), 1738-1742.

Lord, D., Mannering, F.L., 2010. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. Transportation Research - Part A 44(5), pp. 291-305.

Lord, D., Washington, S.P., Ivan, J.N., 2005. Poisson, Poisson-gamma and zero inflated regression models of motor vehicle crashes: Balancing statistical fit and theory. Accident Analysis & Prevention 37 (1), 35-46.

Lord, D., Washington, S.P., Ivan, J.N., 2007. Further notes on the application of zero inflated models in highway safety. Accident Analysis & Prevention 39 (1), 53-57.

Mayshkina, N.V., Mannering, F.L., 2009. Zero-state Markov switching count-data models: An empirical assessment. Accident Analysis & Prevention Vol. 42, No. 1, pp. 122-130.

Miaou, S.-P., 1994. The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. Accident Analysis & Prevention 26 (4), pp. 471-482.

Miaou, S.P., Lord, D., 2003. Modeling traffic crash-flow relationships for intersections: dispersion parameter, functional form, and Bayes versus Empirical Bayes. Transportation Research Record 1840, 31-40.

Mitra, S., Washington, S., 2007. On the nature of over-dispersion in motor vehicle crash prediction models. Accident Analysis & Prevention 39 (3), 459-468.

Oh, J., Lyon, C., Washington, S.P., Persaud, B.N., Bared, J., 2003. Validation of the FHWA Crash Models for Rural Intersections: Lessons Learned. *Transportation Research Record 1840*, pp. 41-49.

Park, B.-J., Lord, D., 2009. Application of Finite Mixture Models for Vehicle Crash Data Analysis. Accident Analysis & Prevention 41(4), 683-691.

Qin, X., Ivan, J.N., Ravishanker, N., 2004. Selecting exposure measures in crash rate prediction for two-lane highway segments. Accident Analysis & Prevention 36 (2), 183–191.

Shankar, V., Milton, J., Mannering, F.L., 1997. Modeling accident frequency as zero-altered probability processes: An empirical inquiry. Accident Analysis & Prevention 29 (6), 829-837.

Spiegelhalter, D.J., Thomas, A., Best, N.G., Lun, D., 2003. WinBUGS Version 1.4.1 User Manual. MRC Biostatistics Unit, Cambridge. Available from: <http://www.mrcbsu. cam.ac.uk /bugs/welcome.shtml>.

Warton, D.I., 2005. Many zeros does not mean zero inflation: Comparing the Goodness-of-Fit of parametric models to multivariate abundance data. Environmetrics 16(2), 275–289.

Washington, S., Karlaftis, M., Mannering, F., 2011. Statistical and econometric methods for transportation data analysis. Chapman and Hall/CRC, Boca Raton, FL, Second edition.

Yau, K.K.W., Wang, K., Lee, A.H., 2003. Zero-Inflated Negative Binomial Regression Modeling of Over-Dispersed Count Data with Extra Zeros. Biometrical Journal 45 (4), 437-452.

Zamani, H., Ismail, N., 2010. Negative binomial-Lindley distribution and Its application. Journal of Mathematics and Statistics 6 (1), 4-9.