Analytic Methods in Accident Research 7 (2015) 29–36

Contents lists available at ScienceDirect

Analytic Methods in Accident Research

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

Exploring the application of the Negative Binomial–Generalized Exponential model for analyzing traffic crash data with excess zeros

Prathyusha Vangala [a], Dominique Lord [a,*], Srinivas Reddy Geedipally [b,1]

[a] Zachry Department of Civil Engineering, Texas A&M University, United States
[b] Texas A&M Transportation Institute, Texas A&M University System, United States

ARTICLE INFO

Article history:
Received 25 August 2014
Received in revised form
12 June 2015
Accepted 12 June 2015
Available online 28 July 2015

Keywords:
Regression model
Generalized Exponential
Negative Binomial
Excess zeros

ABSTRACT

In order to analyze crash data, many new analysis tools are being developed by transportation safety analysts. The Negative Binomial–Generalized Exponential distribution (NB–GE) is such a tool that was recently introduced to handle datasets characterized by a large number of zero counts and is over-dispersed. As the name suggests, this three-parameter distribution is a combination of both Negative binomial and Generalized Exponential distributions. So far, nobody has used this distribution in the context of a regression model for analyzing datasets with excess zeros. This paper therefore describes the application of the NB–GE generalized linear model (GLM). The distribution and GLM were applied to four datasets known to have large dispersion and/or a large number of zeros. The NB–GE was compared to the Poisson, NB as well as the Negative Binomial–Lindley (NB–L) model, another three-parameter recently introduced in the safety literature. The study results show that for datasets characterized by a sizable over-dispersion and contain a large number of zeros, the NB–GE performs similarly as the NB–L, but significantly outclass the traditional NB model. Furthermore, the NB-GE model has a simpler modeling framework than the NB–L, which makes its application relatively straight forward.

1. Introduction

Crash data are usually characterized by over-dispersion, which means that the sample variance is larger than the sample mean (Lord and Mannering, 2010). Depending on the characteristics of the data, the level of dispersion can be so large that traditional statistical models, such as the Poisson-gamma or the Poisson-lognormal, cannot effectively be used for analyzing such datasets (Lord and Geedipally, 2011). The high level of dispersion can be influenced by the proportion of observations that have a value equal to zero as well as data that are characterized by very long tails (observations with a large number of crashes).

To overcome this problem, researchers have proposed different tools for analyzing datasets with a large number of zeros and long tails. They include the zero-inflated (ZI) models (Shankar et al., 1997, 2003), the Negative Binomial–Lindley (NB–L) model (Geedipally et al., 2012; Hallmark et al., 2013; Xu and Sun, 2015), the Poisson-weighted exponential model (Zamani

* Corresponding author. Tel.: +1 979 458 3949.
E-mail addresses: prathyusha89@neo.tamu.edu (P. Vangala), d-lord@tamu.edu (D. Lord), srinivas-g@ttimail.tamu.edu (S.R. Geedipally).
[1] Tel.: +1 817 462 0519.

http://dx.doi.org/10.1016/j.amar.2015.06.001
2213-6657/

et al., 2014), the Poisson Inverse Gaussian (PIG) (Zha et al., In press), the Negative Binomial–Crack (NB–CR) distribution (Saengthong and Bodhisuwan, 2013), and the Sichel (SI) model (Zou et al., 2013; 2015). Lord et al. (2005), (2007) and Lord and Geedipally (2011), (2014) provide discussions about the advantages and limitations of these distributions and models.

To continue the research on the application of three-parameter models for analyzing datasets with excess zeros, this paper documents the development and application of the Negative Binomial Generalized Exponential (NB–GE) generalized linear model (GLM). This model, which has never been developed before, is based on the recently introduced NB–GE distribution (Aryuyuen and Bodhisuwan, 2013). The distribution and model were evaluated using four different datasets. The NB–GE model was compared to the Poisson, NB and NB–L models. Finally, this paper provides information about the application of NB–GE GLM under a Bayesian framework.

1.1. Background

This section is divided into two parts and describes the characteristics of the NB–GE distribution and the GLM.

1.1.1. NB–GE distribution

As the name implies, the NB–GE distribution is a combination of NB and GE distributions. The NB–GE distribution was first introduced by Aryuyuen and Bodhisuwan (2013), where they presented the basic properties of the new distribution such as the mean, variance, skewness and kurtosis. They also presented an approach to find the parameters of the mixed distribution via the maximum likelihood estimate (MLE), but not for the GLM, similar to how Zamani and Ismail (2010) showed how to estimate the parameters of the NB–L distribution. Estimating the coefficients of the NB–L GLM turned out to be more complex than solving the parameters of the distribution because of the non-linear relationship between the covariates and the estimated parameters, as described in Geedipally et al. (2012).

Similar to the NB–L distribution (Lord and Geedipally, 2011), this mixed distribution also has a thick tail and works well when the data contains a large number of zeros or is highly dispersed. The traditional NB and NB-exponential distributions are the special cases of the NB–GE distribution.

The probability mass function (pmf) of Negative Binomial distribution is given by

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

where, $\mu=$ mean response of the observation; and, $\phi=$ inverse of the dispersion parameter ϖ (i.e., $\varpi=1/\phi$).

The Generalized Exponential distribution has the probability density function (pdf) as follows (Aryuyuen and Bodhisuwan, 2013):

$$f(Z=z;\alpha,\lambda)=\alpha\lambda(1-e^{\lambda z})^{\alpha-1}e^{-\lambda z};\ \alpha,\lambda>0, z>0 \tag{2}$$

where, $\alpha=$ shape parameter; and, $\lambda=$ scale parameter.

The exponential distribution is the special case of the GE distribution (i.e., when $\alpha=1$).

The moment generating function of the GE distribution is given as (Aryuyuen and Bodhisuwan, 2013):

$$M_z(t)=\frac{\Gamma(\alpha+1)\Gamma\left(1-\frac{t}{\lambda}\right)}{\Gamma\left(\alpha-\frac{t}{\lambda}+1\right)} \tag{3}$$

The mean and variance of the GE distribution are given as (Gupta and Kundu, 1999):

$$E(Z)=\frac{1}{\lambda}(\psi(\alpha+1)-\psi(1)) \tag{4}$$

$$\mathrm{Var}(Z)=\frac{1}{\lambda^2}(\psi'(\alpha+1)-\psi'(1)) \tag{5}$$

where, $\psi(.)=$digamma function; and $\psi'(.)=$ derivative of the digamma function $\psi(.)$.

The NB–GE distribution arises by combining the NB and GE distributions. The pmf of NB–GE distribution, is given as (Aryuyuen and Bodhisuwan, 2013):

$$f(X=x;r,\alpha,\lambda)=\binom{r+x-1}{x}\sum_{j=0}^{x}\binom{x}{j}(-1)^j\left(\frac{\Gamma(\alpha+1)\Gamma\left(1+\frac{r+j}{\lambda}\right)}{\Gamma\left(\alpha+\frac{r+j}{\lambda}+1\right)}\right);\ r,\ \alpha,\ \lambda>0 \tag{6}$$

The shape parameter '*r*' dictates the mean and variance of the NB–GE distribution. For the moments, skewness, and other details, the reader is referred to Aryuyuen and Bodhisuwan (2013).

1.1.2. NB–GE generalized linear model

The NB–GE distribution is defined as a mixture of NB and GE distributions such that

$$P(X=x;\mu,\phi,\alpha)=\int \mathrm{NB}(x;\phi,z\mu)\mathrm{GE}(z;\alpha,\lambda)dz \tag{7}$$