Accident Analysis and Prevention 45 (2012) 258–265

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