Analytic Methods in Accident Research 18 (2018) 33–44

Contents lists available at ScienceDirect

Analytic Methods in Accident Research

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

Developing a Random Parameters Negative Binomial-Lindley Model to analyze highly over-dispersed crash count data

Mohammad Razaur Rahman Shaon [a], Xiao Qin [a,*], Mohammadali Shirazi [b], Dominique Lord [b], Srinivas Reddy Geedipally [c]

[a] Department of Civil and Environmental Engineering, University of Wisconsin-Milwaukee, P.O. Box 784, Milwaukee, WI 53201-0784, USA
[b] Zachry Department of Civil Engineering, Texas A&M University, College Station, TX 77843, USA
[c] Texas A&M Transportation Institute, 110 N Davis Dr., Arlington, TX 76013, USA

ARTICLE INFO

Article history:
Received 18 August 2017
Received in revised form 9 April 2018
Accepted 9 April 2018

Keywords:
Excess zero observations
Over-dispersion
Unobserved heterogeneity
Mixed model
Random parameters model
Negative binomial-Lindley

ABSTRACT

The existence of preponderant zero crash sites and/or sites with large crash counts can present challenges during the statistical analysis of crash count data. Additionally, unobserved heterogeneity in crash data due to the absence of important variables could negatively impact the estimated model parameters. The traditional negative binomial (NB) model with fixed parameters might not adequately handle highly over-dispersed data or unobserved heterogeneity. Many research efforts that have involved the negative binomial–Lindley (NB-L) model or the random parameters negative binomial (RPNB) model, for example, have attempted to improve the inference of estimated coefficients by explicitly accounting for extra variation in crash data. The NB-L is a mixed modeling approach which provides flexibility to account for additional dispersion in data. The RP modeling approach accommodates the effect of unobserved variables by allowing the model parameters to vary from one observation to another. The following study proposes a combination of these models – the random parameters NB-L (RPNB-L) generalized linear model (GLM) – to account for underlying heterogeneity and address excess over-dispersion. The results show that the RPNB-L model not only provides a superior goodness-of-fit (GOF) with the sample data, but also offers a better understanding about the effects of potential contributing factors. The paper uses the Bayesian framework to provide a strategy for eliminating the potential for poor mixing in the Markov Chain Monte Carlo (MCMC) chains during the estimation of the RPNB-L model.

© 2018 Elsevier Ltd. All rights reserved.

1. Introduction

A roadway crash is a multifaceted event involving circumstances such as highway geometry, traffic exposure, contextual factors, driver characteristics, vehicle factors, as well as the interactions among them. Identifying key crash risk factors and understanding their effects is critical to finding cost-effective strategies for the prevention and reduction of traffic crashes and their severities. Typically, a quantitative safety analysis is performed through descriptive statistics to identify patterns, and regression models are used to identify factors associated with crashes. Once the association is properly established,

* Corresponding author.
E-mail addresses: mshaon@uwm.edu (M.R.R. Shaon), qinx@uwm.edu (X. Qin), alishirazi@tamu.edu (M. Shirazi), d-lord@tamu.edu (D. Lord), srinivas-g@tti.tamu.edu (S.R. Geedipally).

https://doi.org/10.1016/j.amar.2018.04.002
2213-6657/© 2018 Elsevier Ltd. All rights reserved.

additional insights about the crash can be revealed and evaluated. Lastly, the mean crash count can be estimated by mathematical formulation (Mitra and Washington, 2012).

Crash data are often characterized by the existence of a large sample variance compared with the sample mean[1] (Lord et al., 2005; Mitra and Washington, 2007). Extensive research has been devoted to modeling and analyzing this type of crash dataset (Lord and Mannering, 2010; Mannering and Bhat, 2014; Mannering et al., 2016). A notable accomplishment resulting from this research is the application of the negative binomial (NB) model in analyzing crash frequency data. The NB model can handle data over-dispersion by assuming a gamma distribution for the exponential function of the disturbance term in the Poisson mean. However, recent studies have pointed out that with a heavy-tailed crash dataset, the NB model can produce biased parameter estimates (Zou et al., 2015; Shirazi et al., 2016). A heavy-tailed distribution is a statistical phenomenon that occurs when sample observations have a few very high crash counts with preponderant zero observations; this shifts the overall sample mean to near zero (Shirazi et al., 2016). Failure to account for data over-dispersion could lead to biased and inconsistent parameter estimates, which in turn causes researchers to make erroneous inferences from models and also lead to inaccurate crash prediction values.

The mixed model is a well-known methodology used to incorporate heterogeneity into statistical analysis. Safety literature shows that mixed distribution NB models expanded the linear mixed model for continuous responses to discrete responses (e.g., crash count) by incorporating correlated non-normally distributed outcomes. Several mixed NB models have been proposed, including the NB-Lindley (NB-L), NB-Generalized Exponential (NB-GE), and NB-Dirichlet process (NB-DP) generalized linear models (GLMs) (Geedipally et al., 2012; Vangala et al., 2015; Rahman Shaon and Qin, 2016; Shirazi et al., 2016). The advantage of using a mixed model is that it adds a mixed distribution to account for extra variance in the crash data which is caused by preponderant zero crash responses and/or a heavy-tail of crash counts (Shirazi et al., 2016). The underlying hypothesis is that the crash datasets are comprised of distinct subpopulations which have different probabilistic distributions. Accessing all data items associated with the likelihood of crash occurrence and/or injury severity is nearly impossible, but omitting important variables causes data heterogeneity which adds extra variation in the effects of explanatory variables. Random parameters (RP) models can account for unobserved heterogeneity by allowing the parameter of variables to vary from one observation to the next and by estimating the unbiased mean effect of explanatory variables (Mannering et al., 2016). Therefore, incorporating both random parameters and mixed probabilistic distributions within a single model can be a viable alternative for handling crash data with high over-dispersion and unobserved heterogeneity.

The objective of this study was to develop and document an RPNB model with Lindley mixed effect for heterogeneous count data that features an excess number of zero responses and/or a heavy-tail. The proposed RPNB-L model was developed in a Bayesian hierarchical framework that is expanded from fixed-coefficients NB-L GLM (Geedipally et al., 2012; Rahman Shaon and Qin, 2016). The study utilized two crash datasets, one from Indiana and one from South Dakota, to calibrate the parameters in RPNB-L GLM. The datasets were characterized by over-dispersion with a very high percentage of zero responses and a heavy-tail. The model fitting and the modeling results were compared with the traditional NB, RPNB and NB-L models.

2. Literature review

The existence of preponderant zero crash sites with a heavy tail can create highly over-dispersed data. The NB distribution has been used to model crash frequencies for decades because it can handle data over-dispersion, a unique attribute of crash frequency data. However, some studies have noted that the NB distribution cannot adequately handle over-dispersion caused by a heavy tail in the crash data (Guo and Trivedi, 2002; Park et al., 2010; Zou et al., 2015; Shirazi et al., 2016). Guo and Trivedi (2002) noted that a negligible probability is usually assigned to higher crash counts in the NB model during the modeling of highly over-dispersed data with a heavy tail. Lord et al. (2005) pointed out that over-dispersion arises from the actual nature of the crash process. One limitation of the NB distribution is that it assumes that only one underlying process affects the likelihood of crash frequency (Shankar et al., 1997).

A mixture model is a very popular statistical modeling technique that is often used to account for data over-dispersion because it is flexible and extensible (Shankar et al., 1997; Aguero-Valverde and Jovanis, 2008; Lord et al., 2008; Lord and Geedipally ,2011; Geedipally et al., 2012; Cheng et al., 2013; Mannering and Bhat, 2014; Rahman Shaon and Qin, 2016; Shirazi et al., 2016). The mixture model is comprised of a convex combination of a finite number of different distributions. The NB-L GLM is a mixture of the NB and Lindley distribution in which the Lindley distribution itself is a mixture of two gamma distributions (Lindley, 1958). The NB-L GLM was recently introduced to model crash frequency data (Geedipally et al., 2012; Rahman Shaon and Qin, 2016). The count data mixture model works well when the dataset contains a large number of zero responses, is skewed, or is highly dispersed. Zamani and Ismail showed that the NB-L distribution provides a better fit compared to the Poisson and NB models when there is a large probability of crash frequency at zero (Zamani and Ismail, 2010). Lord and Geedipally (2011) applied the NB-L distribution to estimate the predicted probability and frequency of crashes using both simulated and observed crash data. The authors concluded that the NB-L distribution can handle crash

[1] In a statistical term, the sample data is over-dispersed when the variance is greater than the mean. Data over-dispersion is often caused by unobserved data heterogeneity due to unobserved, unavailable, or unmeasurable variables that are important to explain model responses.

Download English Version:

https://daneshyari.com/en/article/7534257

Download Persian Version:

https://daneshyari.com/article/7534257

Daneshyari.com