Accident Analysis and Prevention 117 (2018) 340–345

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Crash data modeling with a generalized estimator

Zhirui Ye[a,*], Yueru Xu[a], Dominique Lord[b]

[a] Jiangsu Key Laboratory of Urban ITS, Jiangsu Province Collaborative Innovation Center of Modern Urban Traffic Technologies, School of Transportation, Southeast University, 2 Sipailou, Nanjing, Jiangsu, 210096, China
[b] Zachry Department of Civil Engineering, Texas A&M University, 3136 TAMU, College Station, TX, 77843-3136, United States

ARTICLE INFO

Keywords:
Crash data analysis
Generalized event count model
Under-Dispersed data

ABSTRACT

The investigation of relationships between traffic crashes and relevant factors is important in traffic safety management. Various methods have been developed for modeling crash data. In real world scenarios, crash data often display the characteristics of over-dispersion. However, on occasions, some crash datasets have exhibited under-dispersion, especially in cases where the data are conditioned upon the mean. The commonly used models (such as the Poisson and the NB regression models) have associated limitations to cope with various degrees of dispersion. In light of this, a generalized event count (GEC) model, which can be generally used to handle over-, equi-, and under-dispersed data, is proposed in this study.

This model was first applied to case studies using data from Toronto, characterized by over-dispersion, and then to crash data from railway-highway crossings in Korea, characterized with under-dispersion. The results from the GEC model were compared with those from the Negative binomial and the hyper-Poisson models. The cases studies show that the proposed model provides good performance for crash data characterized with over- and under-dispersion. Moreover, the proposed model simplifies the modeling process and the prediction of crash data.

1. Introduction

More than one million people are killed in traffic crashes every year around the world (WHO, 2013). Traffic crashes result in enormous losses to society and the economy. Several researchers have been seeking methods for better understanding contributing factors that influence or are associated with crashes and develop effective strategies to improve road safety.

The relationships between traffic crashes and relative factors have been investigated for more than three decades (Lord and Mannering, 2010). Various kinds of methodologies have been proposed over the years to improve on predicting the likelihood of crashes and determine the variables or factors that significantly influence the number of crashes and their severities.

It has been shown that crash data usually exhibit over-dispersion. Initially, the negative binomial (NB) regression model was proposed to handle such datasets (Miaou, 1994; Poch and Mannering, 1996). The NB model is derived by rewriting the Poisson parameter as $\lambda_i = \mathrm{EXP}(\beta X_i + \varepsilon_i)$ in which $\mathrm{EXP}(\varepsilon_i)$ is a gamma-distributed error term with mean 1 and variance α. Given important limitations associated with the NB model, highway safety researchers have proposed new and innovative models, such as the random-effects (Hausman et al., 1984; Shankar et al., 1998) and its extension to random parameters models (Anastasopoulos and Mannering, 2009), bi-variate/multivariate models (Maher, 1990; Ma et al., 2006; Park and Lord, 2007; Barua et al., 2016), multiparameter models (Geedipally et al., 2012; Vangala et al., 2014), generalized additive models (Xie and Zhang, 2008), and semi-parametric models based on the Dirichlet process (Heydari et al., 2016b; Shirazi et al., 2016). These models can handle characteristics commonly found in crash data, such as excess zero responses and datasets with long tails among others. Readers are referred to Lord and Mannering (2010), Mannering and Bhat (2014), and Heydari et al. (2016a), who have provided a comprehensive review of existing methods with their advantages and disadvantages.

In addition to over-dispersion, some researchers have also encountered under-dispersion (Oh et al., 2006; Daniels et al., 2010; Lord et al., 2010). Although the models described above are able to capture or handle over-dispersion or unobserved heterogeneity, they cannot be used efficiently when the data are characterized by under-dispersion, either in the dataset itself or when the observations are conditioned upon the mean (Lord et al., 2010). To handle under-dispersion, Oh et al. (2006) first proposed the gamma model to analyze crash data exhibiting this unique characteristic. Although the gamma model can handle under-dispersion, the model suffers from an important

* Corresponding author.
E-mail address: yezhirui@seu.edu.cn (Z. Ye).

https://doi.org/10.1016/j.aap.2018.04.026
Received 5 December 2017; Received in revised form 24 April 2018; Accepted 28 April 2018

drawback since past observations are assumed to directly influence future observations (e.g., a crash that occurred in one year is directly correlated to a crash that will occur the following or a future year) (Lord et al., 2010). Subsequently, Lord et al., 2008, 2010, Lord and Mannering, 2010) have proposed the Conway-Maxwell-Poisson (COM-Poisson) generalized linear model for analyzing crash data. Recently, Huang (2017) proposed a re-parametrization of the COM-Poisson model, where the mean of the counts is modeled directly rather than using the mode as an approximation of the mean value. The COM-Poisson model can handle both under- and over-dispersion, similar to the gamma model, but without the key limitation of the latter, although it may provide erroneous estimates for very small sample size and low sample mean values (Lord et al., 2010). Along the same line, Zou et al. (2013) examined the applicability of double Poisson (DP) generalized linear model for analyzing crash data and compared its performance with the COM-Poisson model. Khazraee et al. (2015) applied the hyper-Poisson (hP) generalized linear model to analyze under-dispersed crash data. It should be pointed out that the COM-Poisson and hP models both allow the dispersion of the distribution to be observation-specific and dependent on model covariates and both the DP and hP models offered similar statistical performance than those associated with the COM-Poisson. The differences were seen with the complexity for estimating the coefficients of the models.

The research documented in this paper therefore continues the work performed on the development of tools that would allow the analysis of both over- and under- dispersion. More specifically, the main goal is to apply the generalized event count (GEC) model developed by King (1989) for crash analysis and prediction. Similar to the COM-Poisson, DP and hP models introduced above, this model also handles over-, under- and equi-dispersion and has been shown to provide good statistical performances in other fields, such as the evaluation of congressional challenges of presidential votes and superpower conflicts (King, 1989). So far, this model has not been applied for analyzing crash data. Overall, the GEC model is easy to implement since the coefficients can be estimated using maximum likelihood estimation (MLE) and can handle over-, under- and equi-dispersion with good performance. The next section presents the GEC model for crash data analysis. Subsequently, case studies are presented; they were used to evaluate the performance of the proposed method by comparing the model with existing models, such as NB regression model or HP model. Finally, the findings and conclusions are summarized.

2. Methodology

This section first briefly introduces the Poisson model as background. It is followed by a more detailed description of the GEC model.

2.1. The Poisson regression model

The Poisson regression model is the basic model for analyzing count data. It aims at modeling a count (or crash) variable Y, which is assumed to follow a Poisson distribution with a parameter (or mean) λ (Lord and Mannering, 2010; Myers et al., 2012). The Poisson distribution usually implies that the probability of an event occurring at any instant is constant and independent of all previous events during the observation period (King, 1988). In highway safety, the probability that the number of crashes takes the value y_i on the *i*th entity can be expressed as Eq. (1).

$$P(Y_i = y_i) = f(y_i|\lambda_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!},\ i = 1,2,\ .....n. \tag{1}$$

In the Poisson regression model, the mean can be written as $\lambda_i = \mathrm{EXP}(\beta X_i)$, where X_i is a vector of k explanatory variables and β is a 1 × k parameter vector that indicates the effect of the explanatory variables on the dependent variable. To estimate β, the method of maximum likelihood can be used. The likelihood function is presented in Eq. (2).

$$L(\boldsymbol{\beta}y) = \prod f(y_i|\lambda_i) = \prod \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} \tag{2}$$

For a Poisson regression model, the variance of Y_i is equal to its expected value: $V(Y_i) = E(Y_i) = \lambda_i$. In practice, this model is not used frequently in safety research since the main assumption between the mean and variance is violated. This model is presented here since its characteristics are expanded in the next section.

2.2. The generalized event count model

In most cases, road crash data display the characteristic of over-dispersion and, on rare occasions, could exhibit under-dispersion. Considering all possible situations, the relationship between mean and variance is defined by $V(Y_i) = \lambda_i\sigma^2$ for $\lambda_i > 0$ and $\sigma^2 > 0$. σ^2 is called the dispersion parameter. If the crash variable follows a Poisson distribution, then $\sigma^2 = 1$ and $V(Y_i) = E(Y_i) = \lambda_i$; if $\sigma^2 > 1$, the data are over-dispersed; and if $0 < \sigma^2 < 1$, the data are regarded as under-dispersed. With the introduction of the parameter σ^2, the GEC model is developed and is able to model event counts with unknown degrees of dispersion. To construct this model, a GEC probability distribution with parameters λ_i and σ^2 is established. In this model, σ^2 can take on any value greater than zero. Special cases occur when the dispersion parameter falls into different ranges. When $0 < \sigma^2 < 1$, the GEC distribution can handle under-dispersed data. When $\sigma^2 = 1$, the model has the same probability function as the Poisson regression model; and when $\sigma^2 > 1$, it's probability function is similar to the NB regression model. This GEC's probability distribution offers smooth transitions between these scenarios. To derive the GEC's probability distribution, a concept taken from theoretical statistics called "bilinear recurrence relationship" was introduced (Katz, 1965). The relationship is shown in Eq. (3).

$$\frac{f_k(y_i + 1|\theta_i, y_i)}{f_k(y_i|\theta_i, y_i)} = \frac{\theta_i + \gamma_i y_i}{y_i + 1} \text{for } y_i = 0,1,2,\ \ldots \text{and } \theta_i + \gamma_i y_i \geq 0 \tag{3}$$

where θ_i and γ_i are ancillary parameters. In this case, Eq. (3) should be re-parameterized in order to make the relationship suitable for the previous definitions.

Statistical analysis reveals that the expected value $E(Y_i)$ and variance $V(Y_i)$ of a random variable Y_i that adheres to the relationship in Eq. (3) are as follows (Lee, 1986):

$$E(Y_i) = \lambda_i = \frac{\theta_i}{1-\gamma_i} \tag{4}$$

$$V(Y_i) = \lambda_i\sigma^2 = \frac{\theta_i}{(1-\gamma_i)^2} \tag{5}$$

Solving the above two equations, we then get:

$$\gamma_i = 1-\frac{1}{\sigma^2},\ \theta_i = \frac{\lambda_i}{\sigma^2} \tag{6}$$

Then Eq. (3) becomes:

$$f_{gec}(y_i|\lambda_i, \sigma^2) = \left(\frac{\lambda_i + (\sigma^2-1)(y_i-1)}{\sigma^2 y_i}\right) f_{gec}(y_i-1|\lambda_i, \sigma^2) \tag{7}$$

In Eq. (7), f_{gec} represents the GEC distribution. The expected value and variance of the distribution is consistent with the previous definitions: $E(Y_i) = \lambda_i$ and $V(Y_i) = \lambda_i\sigma^2$.

At this point, Eq. (7) is still in the form of a recurrence relationship, so it should be transformed into more traditional probability distributions. To achieve this, the term $f_{gec}(y_i-1|\lambda_i, \sigma^2)$ must be recursively substituted by λ_i and σ^2. To better explain this process, Eq. (8) shows a simple example where given $y_i = 2$ (King, 1989)

$$\Pr(Y_i = 2|\lambda_i, \sigma^2) = f_{gec}(2\lambda_i, \sigma^2) \tag{8}$$