Accident Analysis and Prevention 50 (2013) 73–82

Contents lists available at SciVerse ScienceDirect

Accident Analysis and Prevention

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

Modeling crash frequency and severity using multinomial-generalized Poisson model with error components

Yu-Chiun Chiou*, Chiang Fu

Institute of Traffic and Transportation, National Chiao Tung University, 4F, 118, Sec. 1, Chung-Hsiao W. Rd., Taipei 100, Taiwan

ARTICLE INFO

Article history:
Received 16 October 2011
Received in revised form 25 March 2012
Accepted 26 March 2012

Keywords:
Crash frequency
Crash severity
Multinomial-generalized Poisson
Error components

ABSTRACT

Since the factors contributing to crash frequency and severity usually differ, an integrated model under the multinomial generalized Poisson (MGP) architecture is proposed to analyze simultaneously crash frequency and severity—making estimation results increasingly efficient and useful. Considering the substitution pattern among severity levels and the shared error structure, four models are proposed and compared—the MGP model with or without error components (EMGP and MGP models, respectively) and two nested generalized Poisson models (NGP model). A case study based on accident data for Taiwan's No. 1 Freeway is conducted. The results show that the EMGP model has the best goodness-of-fit and prediction accuracy indices. Additionally, estimation results show that factors contributing to crash frequency and severity differ markedly. Safety improvement strategies are proposed accordingly.

1. Introduction

To improve traffic safety, numerous statistical models have been developed that identify factors contributing to crash frequency and severity. Most identify risk factors for either crash frequency or severity independently. When modeling crash frequency (the number of accidents on roadway segments or at intersections over a specified period), a considerable number of studies have used various methodological approaches. Due to the discrete and non-negative integer character of accident counts, count-data models such as the Poisson model (e.g., Jones et al., 1991; Miaou, 1994; Shankar et al., 1997), negative binomial model (e.g., Hadi et al., 1995; Shankar et al., 1995; Poch and Mannering, 1996; Milton and Mannering, 1998; Lord, 2006; Malyshkina and Mannering, 2010), Poisson lognormal model (e.g., Miaou et al., 2005; Lord and Miranda-Moreno, 2008), Gamma model (e.g., Oh et al., 2006), generalized Poisson model (e.g., Dissanayake et al., 2009; Famoye et al., 2004) as well as zero-inflated modeling and other flexible modeling techniques (e.g., Abdel-Aty and Radwan, 2000; Wang and Abdel-Aty, 2008; Park and Lord, 2009; Anastasopoulos and Mannering, 2009; see Lord and Mannering, 2010 for elaborate and complete reviews) have been applied to model crash counts.

Crash frequencies are commonly collected by severity on relatively homogenous roadway segments, supporting the development of crash count models. Thus, crash data are typically classified according to severity (e.g., property damage only, injury, and fatality) or collision type (e.g., rear-end, head-on, sideswipe, and right angle). With this data segmentation, separate severity–frequency models are developed for each accident severity level. In this way, a series of negative binomial accident frequency models were developed for each crash severity level to predict the number of accidents at each severity level on roadway segments. Unfortunately, such an approach can generate a statistical problem in that interdependence due to latent factors likely exists across crash rates at different severity levels for a specific roadway segment (Ma et al., 2008). For example, an increase in number of accidents that are classified as having a certain severity level is also associated with changes in the number of accidents that are classified with other severity levels, setting up a correlation among various injury-outcome crash frequency models (Lord and Mannering, 2010).

Considerable research effort has focused on modeling accident severity from an individual perspective using such methodological approaches as logistic regression (e.g., Lui et al., 1988; Yau, 2004), bivariate models (e.g., Saccomanno et al., 1996; Yamamoto and Shankar, 2004), the multinomial and nested logit structures to evaluate accident-injury severities (e.g., Shankar et al., 1996; Chang and Mannering, 1999; Carson and Mannering, 2001; Lee and Mannering, 2002; Ulfarsson and Mannering, 2004; Khorashadi et al., 2005), and the discrete ordered probit model (e.g., O'Donnell and Connor, 1996; Duncan et al., 1998; Renski et al., 1999; Kockelman and Kweon, 2002; Khattak et al., 2002; Kweon and Kockelman, 2003; Abdel-Aty, 2003). For more details on accident severity models may refer to Savolainen et al. (2011).

Although these models have been applied by a number of researchers with a considerable success, Milton et al. (2008) indicated that these studies relied heavily on detailed data in individual accident reports and they have been proved to be difficult to use

* Corresponding author. Tel.: +886 2 23494940; fax: +886 2 23494953.
E-mail address: ycchiou@mail.nctu.edu.tw (Y.-C. Chiou).

0001-4575/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.aap.2012.03.030

in safety programming because a large number of event-specific explanatory variables need to be estimated to produce useable severity forecasts. Moreover, significant contributory factors in the model are usually not closely related to traffic management strategies, roadway geometrics, and weather-related factors; therefore, the corresponding countermeasures are difficult to propose accordingly. Furthermore, as different data scales are used by frequency models and the severity model, integration is extremely difficult.

Obviously, crash frequency and severity are two key indices that measure risk for a roadway segment. Either one only generates partial insights for crash risk. Increased scope and in-depth insights cannot be obtained without considering both indices together. Thus, two possible integrated modeling approaches were attempted. The first approach uses a conventional frequency model to predict total number of crashes and a severity model, such as the multinomial logit model, nested logit model, ordered probit model, or mixed logit model, to predict aggregate severity probability (e.g., Yamamoto et al., 2008; Kim et al., 2008; Milton et al., 2008). However, the assumption that crash frequency and severity are mutually independent still exists. The second approach applies multivariate regression models to predict crash frequencies for different severity levels. Multivariate regression models simultaneously develop crash frequency models by severity (Bijleveld, 2005; Ma and Kockelman, 2006; Song et al., 2006; Park and Lord, 2007; Ma et al., 2008; Aguero-Valverde and Jovanis, 2009; El-Basyouny and Sayed, 2009; Ye et al., 2009) to overcome the correlation problem among crash frequencies at different severity levels. However, this approach requires a complex estimation procedure combined with a subjectively preset correlation matrix of severity levels, making field validation very difficult.

Another drawback of the multivariate modeling approach is its inability to grasp associated changes related to severity and frequency variables only. If one fails to observe separately the effects of factors contributing to crash frequency and severity, the multivariate model may be partly limited for practical program evaluation. An appealing idea is to view risk factors according to their accident descriptive components (i.e., severity and frequency) individually under an integrated framework. However, expected difficulties arise when analyzing subjects and procedures. Consequently, using a conceptual model combining both crash frequency and severity is worthwhile.

Thus, this paper aims to develop a novel multinomial generalized Poisson (MGP) model to simultaneously model crash frequency (count data) and severity (ratio data). Furthermore, the proposed model considers the substitution pattern among severity levels and constructs a shared error structure as a correlation matrix through error components specified under an integrated model framework. A case study of Taiwan freeway crash data is utilized to assess the applicability of the proposed model. The remainder of this paper is organized as follows. Section 2 presents the proposed MGP model. Section 3 addresses data collection and descriptive statistics of the accident dataset for Taiwan's No. 1 Freeway. Section 4 presents model estimation results and comparisons. Section 5 discusses safety implications based on estimation results. Section 6 gives concluding remarks and suggestions to further research.

2. The proposed models

The MGP model is an extension of the multinomial-Poisson (MP) regression model (Terza and Wilson, 1990). In the context of crash frequency and severity modeling, we assume that accidents can be classified into a finite number of clusters according to severity levels and that the frequency of each severity level follows a conditional multinomial distribution, which is expressed as follows:

$$f\left(Y\,\middle|\,\sum_{j=1}^{J}y_j=N\right)=\frac{N!\prod_{j=1}^{J}\pi_j^{y_j}}{\prod_{j=1}^{J}y_j!} \tag{1}$$

where $f(\cdot)$ is the conditional probability of Y; $Y=[y_1, y_2,\ldots y_j,\ldots,y_J]$ and $\sum_{j=1}^{J}y_j=N$; $y_j=0,1,2,\ldots,\infty$, for $j=1,2,\ldots,J$, is a random vector representing the observed crash counts of segment t within a given period (e.g., 1 year) at severity level j; J is the total number of severity levels determined in advance; π_1, $\pi_2,\ldots$, π_J are multinomial probabilities of severity levels $1,2,\ldots,J$, respectively; $\pi_j=y_j/N$ and $\pi_1+\pi_2+\ldots+\pi_J=1$; and N is the total number of accidents across different severity levels of segment m within a given period. Thus, the conditional multinomial distribution can be used to determine crash frequencies at various severity levels, i.e., $y_1, y_2,\ldots, y_J$, given total number of accidents, N. Furthermore, the joint probability of these crash frequencies $h(y_1, y_2,\ldots, y_J)$ can be expressed as the product of conditional probability and marginal probability:

$$h(y_1,y_2,\ldots,y_J)=f\left(Y\,\middle|\,\sum_{j=1}^{J}y_j=N\right)\cdot g\left(\sum_{j=1}^{J}y_j=N\right) \tag{2}$$

where $g(\cdot)=g\left(\sum_{j=1}^{J}y_j=N\right)$ is the marginal probability of crash counts. Terza and Wilson (1990) assumed that the marginal (unconditional) probability has the following Poisson distribution:

$$g(\cdot)=\frac{\lambda^N\exp(-\lambda)}{N!} \tag{3}$$

where $g(\cdot)$ is the probability that N accidents occurred, and λ is the expected number of accidents. For estimation purposes, λ is usually specified as

$$\lambda=\exp(\beta' X) \tag{4}$$

where X and β' are vectors of explanatory variables and estimated parameters, respectively. The formulation of the multinomial Poisson (MP) model is then derived by substituting Eqs. (1) and (3) into Eq. (2).

The Poisson model assumes that variance equals mean. If observed data exhibit over-dispersion (under-dispersion), this assumption does not hold. This leads to estimation inefficiency because inference was invalidated by unreliable estimated standard errors. We can relax this assumption using the generalized Poisson (GP) model (Famoye et al., 2004; Dissanayake et al., 2009). The probability function of total accidents at any segment, N, can be written as Eq. (5):

$$g(\cdot)=\left(\frac{\lambda}{1+\eta\lambda}\right)^N\frac{(1+\eta N)^{N-1}}{N!}\exp\left(\frac{-\lambda(1+\eta N)}{1+\eta\lambda}\right) \tag{5}$$

where η is the dispersion parameter. If $\eta>0$, the GP model indicates the over-dispersion feature in the empirical data. If $\eta=0$, the probability function degenerated to the Poisson model. In contrast, if $\eta<0$, the GP model denotes the under-dispersion feature in the empirical data. All other involved arguments associated with Eq. (5) are as defined previously. The mean and variance of N are represented by Eqs. (6) and (7), respectively:

$$E(N|X)=\lambda \tag{6}$$

$$V(N|X)=\lambda(1+\eta\lambda)^2 \tag{7}$$

According to Eq. (6), the probability function in Eq. (5) degenerates into the original Poisson model as $\eta=0$. Hence, the GP model is a generalized Poisson model. Interested readers can refer to Famoye (1993) for detailed proofs. In accordance with the derivation by