Safety Science 70 (2014) 63–69

Contents lists available at ScienceDirect

Safety Science

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

Examining signalized intersection crash frequency using multivariate zero-inflated Poisson regression

Chunjiao Dong[a,*], Stephen H. Richards[a,1], David B. Clarke[a,2], Xuemei Zhou[b,3], Zhuanglin Ma[c]

[a] Center for Transportation Research, The University of Tennessee, 600 Henley Street, Knoxville, TN 37996, USA
[b] Key Laboratory of Road and Traffic Engineering of the Ministry of Education, Tongji University, 4800 Cao An Road 201804, Shanghai, PR China
[c] School of Automobile, Chang'an University, Nan Er Huang Zhong Duan, Xi'an 710064, Shaanxi, PR China

ARTICLE INFO

Article history:
Received 12 November 2013
Received in revised form 30 April 2014
Accepted 9 May 2014
Available online 2 June 2014

Keywords:
Crash frequency
Geometric design
MZIP model
Bayesian method

ABSTRACT

In crash frequency studies, correlated multivariate data are often obtained for each roadway entity longitudinally. The multivariate models would be a potential useful method for analysis, since they can account for the correlation among the specific crash types. However, one issue that arises with this correlated multivariate data is the number of zero counts increases as crash counts have many categories. This paper describes a multivariate zero-inflated Poisson (MZIP) regression model as an alternative methodology for modeling multivariate crash count data by severity. The Bayesian method is employed to estimate the model parameters. Using this Bayesian MZIP model, we can take into account correlations that exist among different severity levels. Our new method also can cope with excess zeros in the data, which is a common phenomenon found in practice. The proposed model is applied to the multivariate crash counts obtained from intersections in Tennessee for five years. The results reveal that, compared to the univariate ZIP models and multivariate Poisson-lognormal (MVPLN) models, the MZIP models provide the best statistic fit and have the smallest estimation bias. Apart from the improvement in goodness of fit, the results of the MZIP models show promise toward the goal of obtaining more accurate estimates by accounting for excess zeros in correlated count data.

1. Introduction

Intersections present a complicated and hazardous roadway environment to drivers. The presence of signals, guide signs for street names, indications of upcoming turn lanes, conflict traffic, exclusive left turn and right turn lanes, and other paraphernalia associated with intersections create a high degree of conflict that leads to higher crash numbers. In addition, intersections with closely spaced decision points, intensive land use, complex design features, and heavy traffic may cause information overload or driver confusion, resulting in an inadequate understanding of the driving situation and subsequently crashes. Based on the Fatality Analysis Reporting System (FARS) and National Automotive Sampling System-General Estimates System (NASS-GES) data, about 40% of the estimated 5,338,000 crashes during 2011 in the United States were intersection-related. Of those intersection crashes, about 36% occurred at signalized intersections. Furthermore, signalized intersections also tended to experience more severe crashes. Injury crashes accounted for 33.2% of reported signalized intersection crashes, compared to 25.2% for non-signalized intersection crashes. Many studies have shown that intersections are among the most dangerous locations of a roadway network. Therefore, there is a need to understand the factors that contribute to crashes at such locations.

However, some intersections might have only fatal crashes and others might have only incapacitating crashes, non-incapacitating crashes, possible injury crashes, or property damage only crashes. The differences might be caused by geometric features and traffic characteristics from intersection to intersection. Therefore, understanding which factor is significant to the specific crash types and what differences existing between crash types becomes a necessary subject needing to be studied. Since injury crashes cause very serious problems, the goal is to reduce the severe crash frequencies, such as fatal crashes and incapacitating crashes, when the total crash numbers are controlled at certain level.

This paper proposes a MZIP regression model to estimate the relationship between intersection geometric features, traffic

[*] Corresponding author. Tel.: +1 865 974 1826.
E-mail addresses: cdong5@utk.edu (C. Dong), stever@utk.edu (S.H. Richards), dbclarke@utk.edu (D.B. Clarke), zhouxm@tongji.edu.cn (X. Zhou), mazhuanglin@126.com (Z. Ma).
[1] Tel.: +1 865 974 0724.
[2] Tel.: +1 865 974 1813.
[3] Tel.: +86 021 6958 3001.

http://dx.doi.org/10.1016/j.ssci.2014.05.006
0925-7535/© 2014 Elsevier Ltd. All rights reserved.

factors, and crash counts across severity. The primary objective is to conduct a new alternative multivariate model to account for excess zeros in correlated count data. The secondary objective is to investigate which factors significantly contribute to the crash counts across severity. The last objective is to examine if there is any difference in the specific types of crashes for the same factors and how to control the factors to reduce severe crash frequencies under certain level crash frequencies.

2. Literature review

To deal with the data and methodological issues associated with crash frequency data, a wide variety of count data models (Zhang et al., 2012; Chang and Chen, 2005) have been applied over the years. Models to estimate crash frequencies on roadway segments or at intersections fall into two broad categories. One category includes conventional univariate regression models, such as the Poisson models, Poisson-gamma (negative binomial) models, Poisson-lognormal models, zero-inflated models, Conway–Maxwell–Poisson models, gamma models, and generalized estimating equation models. The second category includes potentially more realistic specifications such as generalized additive models, random-effects models, negative multinomial models, random-parameters models, bivariate/multivariate models, finite mixture/ Markov switching models, duration models, and hierarchical/multilevel models (for a complete review of this literature see Lord and Mannering, 2010).

In crash frequency analysis, modeling the frequencies of specific types of crashes cannot be done with independent count models, since the frequencies of specific crash types are not independent. When one wishes to model specific types of crash counts (for example, the number of crashes resulting in fatalities, injuries, etc.), multivariate models become necessary because they explicitly consider the correlation among the severity levels for each intersection (Miaou and Song, 2005; Bijleveld, 2005; Song et al., 2006). Ma and Kockelman (2006) used a multivariate Poisson (MVP) regression model to estimate the injury count by severity level. Positive correlation in unobserved factors affecting count outcomes was found across severity levels, resulting in a statistically significant assistive latent term. Several researchers (Park and Lord, 2007; Ma et al., 2008; El-Basyouny and Sayed, 2009; Dong et al., 2014) employed a multivariate Poisson-lognormal (MVPLN) approach as an improved model to describe the relationship between factors and crash counts. Anastasopoulos et al. (2012) proposed a multivariate tobit regression model to handle left-censored at zero issues. These efforts demonstrate that multivariate models become a potential useful method for analysis when modeling the counts of specific types of crashes. However, as finer crash categorization became available, a significant amount of zeros appeared which are challenging the ability of the conventional multivariate regression model (i.e. MVP, multivariate negative binomial (MVNB), and MVPLN). As the crash data with excess zeros become an issue, there is a need of model development that can describe data characterized with a preponderance of zeros.

Zero-inflated models have been applied to cope with data characterized by a significant amount of zero observations or more zero observations than the one would expect in a traditional model (Poisson or negative binomial model). Zero-inflated models operate on the principle that the excess zeros is accounted for by a splitting regime that models a virtually safe state versus a crash-prone propensity of a roadway entity. The probability of an intersection being in zero or non-zero states can be determined by a binary logit or probit model (Lambert, 1992; Lee and Mannering, 2002; Kumara and Chin, 2003). Despite zero-inflated model has been used broadly under the situations where the observed data are characterized by large zero densities, Lord et al. (2005, 2007) have criticized the application of this model in highway safety. They argued that zero-inflated model cannot properly reflect the crash-data generating process, since the zero or safe state has a long-term mean equal to zero. Recently, the problems associated with a dual-state data generating process have been discussed by other researchers. As an alternative, Malyshkina and Mannering (2010) have proposed a zero-state Markov switching count-data model for circumventing such problem. Markov switching models offer considerable potential for providing important new insights into the analysis of crash data. However, these models are quite complex to estimate. Furthermore, there is no evidence that the Markov switching model can be extended to bivariate/multivariate formulation.

Though the zero-inflated model had limitations, we chose it as the baseline model based on the following considerations. First, literature indicates that zero-inflated models provide improved statistical fit compared to traditional Poisson and NB models (Lee and Mannering, 2002; Kumara and Chin, 2003; Shankar et al., 2003). Second, the crash data in our dataset were characterized by a preponderance of zeros, which is caused by one or more of the following conditions: (1) the subject of the study is the intersections, which represent small spatial scales; (2) analysis of intersections are characterized by a combination of low exposure, high heterogeneity, and sites categorized as high risk; (3) the number of zero observations increases as crash counts have many categories; and (4) it is possible that the sample data contain a relatively high percentage of non-reported crashes. To deal with the problems associated with using crash data characterized by a large number of zeros, we employed zero-inflated models as the best modeling approach. Nevertheless, univariate zero-inflated models cannot handle the correlation problem among specific types of crashes. So there is a need to develop a multivariate zero-inflated regression model to handle the situation which involves more than one type of crash and crash data were characterized by a significant amount of zeros.

In the current paper, a multivariate approach is introduced for jointly modeling data on crash counts by severity on the basis of MZIP distributions. Using this MZIP specification, as well as Bayesian estimation techniques, our study models correlated traffic crash counts simultaneously at different levels of severity by using crash data for signalized intersections in Tennessee. In addition, the paper investigates the performances of MVPLN, ZIP, and MZIP regression models in establishing the relationship between crashes, traffic factors, and geometric design features of roadway intersections.

3. Model structure and estimation

The following section presents the general forms of MZIP regression models and provides brief descriptions of its estimation procedures. Bivariate ZIP (BZIP) distributions are presented first, since we adopted the ideas of constructing the BZIP to construct the MZIP distribution.

3.1. BZIP distributions

There are at least three methods to construct a BZIP model (Li et al., 1999; Walhin, 2001; Wang et al., 2003). In this paper, the BZIP distribution is constructed as a mixture of a bivariate Poisson, two univariate Poisson, and a point mass at (0, 0), which has the property that the marginal distributions are univariate ZIP's. The probability mass function (pmf) of the BZIP is given by