Accident Analysis and Prevention 99 (2017) 6–19

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Multivariate poisson lognormal modeling of crashes by type and severity on rural two lane highways

Kai Wang[a,*], John N. Ivan[b], Nalini Ravishanker[c], Eric Jackson[d]

[a] Connecticut Transportation Safety Research Center, University of Connecticut, 270 Middle Turnpike, Unit 5202, Storrs, CT 06269-5202, USA
[b] Department of Civil and Environmental Engineering, University of Connecticut, 261 Glenbrook Road, Unit 3037, Storrs, CT 06269-3037, USA
[c] Department of Statistics, University of Connecticut, AUST 333, 215 Glenbrook Road, Storrs, CT 06269, USA
[d] Connecticut Transportation Safety Research Center, Department of Civil and Environmental Engineering, University of Connecticut, Longley Building Room 144, Storrs, CT 06269, USA

ARTICLE INFO

Article history:
Received 30 May 2016
Received in revised form 5 October 2016
Accepted 5 November 2016

Keywords:
Multivariate Poisson Lognormal (MVPLN) model
Integrated Nested Laplace (INLA) Bayesian approach
Rural two-lane highways
Crash type models
Crash severity models

ABSTRACT

In an effort to improve traffic safety, there has been considerable interest in estimating crash prediction models and identifying factors contributing to crashes. To account for crash frequency variations among crash types and severities, crash prediction models have been estimated by type and severity. The univariate crash count models have been used by researchers to estimate crashes by crash type or severity, in which the crash counts by type or severity are assumed to be independent of one another and modelled separately. When considering crash types and severities simultaneously, this may neglect the potential correlations between crash counts due to the presence of shared unobserved factors across crash types or severities for a specific roadway intersection or segment, and might lead to biased parameter estimation and reduce model accuracy. The focus on this study is to estimate crashes by both crash type and crash severity using the Integrated Nested Laplace Approximation (INLA) Multivariate Poisson Lognormal (MVPLN) model, and identify the different effects of contributing factors on different crash type and severity counts on rural two-lane highways. The INLA MVPLN model can simultaneously model crash counts by crash type and crash severity by accounting for the potential correlations among them and significantly decreases the computational time compared with a fully Bayesian fitting of the MVPLN model using Markov Chain Monte Carlo (MCMC) method. This paper describes estimation of MVPLN models for three-way stop controlled (3ST) intersections, four-way stop controlled (4ST) intersections, four-way signalized (4SG) intersections, and roadway segments on rural two-lane highways. Annual Average Daily traffic (AADT) and variables describing roadway conditions (including presence of lighting, presence of left-turn/right-turn lane, lane width and shoulder width) were used as predictors. A Univariate Poisson Lognormal (UPLN) was estimated by crash type and severity for each highway facility, and their prediction results are compared with the MVPLN model based on the Average Predicted Mean Absolute Error (APMAE) statistic. A UPLN model for total crashes was also estimated to compare the coefficients of contributing factors with the models that estimate crashes by crash type and severity. The model coefficient estimates show that the signs of coefficients for presence of left-turn lane, presence of right-turn lane, land width and speed limit are different across crash type or severity counts, which suggest that estimating crashes by crash type or severity might be more helpful in identifying crash contributing factors. The standard errors of covariates in the MVPLN model are slightly lower than the UPLN model when the covariates are statistically significant, and the crash counts by crash type and severity are significantly correlated. The model prediction comparisons illustrate that the MVPLN model outperforms the UPLN model in prediction accuracy. Therefore, when predicting crash counts by crash type and crash severity for rural two-lane highways, the MVPLN model should be considered to avoid estimation error and to account for the potential correlations among crash type counts and crash severity counts.

* Corresponding author.
E-mail addresses: kai.wang@uconn.edu (K. Wang), john.ivan@uconn.edu (J.N. Ivan), nalini.ravishanker@uconn.edu (N. Ravishanker), eric.d.jackson@uconn.edu (E. Jackson).

http://dx.doi.org/10.1016/j.aap.2016.11.006
0001-4575/

1. Introduction and motivation

In the United States, improving roadway safety is a high priority of the transportation agencies at the federal, state and local levels, and motor vehicle crashes bring one of the largest economic and societal losses (NHTSA, 2016). According the National Highway Traffic Safety Administration (NHTSA) (2016), there were 32,675 people killed in motor vehicle crashes in 2014, and the total economic losses are up to $836 billion. Given the importance of roadway safety and the substantial economic losses caused by motor vehicle crashes, there has been increasing interest in developing crash prediction models to estimate motor vehicle crash counts, identify crash contributing factors, and implement effective safety strategies and countermeasures to improve traffic safety.

In the current Highway Safety Manual (HSM, 2010), crash counts are estimated in total, even though the crash patterns may vary by crash type and by crash severity. To account for crash frequency variations among crash types, some studies estimated crash counts separately and independently by crash type (Hauer et al., 1988; Shankar et al., 1995; Poch and Mannering 1996; Hauer 2000; Ivan et al., 2000; Qin 2002; Abdel-Aty et al., 2005; Geedipally and Lord 2010). Similarly, in order to accommodate crash frequency variations among crash severities, crash prediction models have been considered by severity level (Abdel-Aty and Radwan 2000; Lord and Persaud 2000; Ulfarsson and Shankar 2002; Lyon et al., 2003; Lord et al., 2008; Tarko et al., 2008; Geedipally et al., 2010). The Poisson regression model has been used in crash prediction and exploration of effects of roadway geometric factors on crash counts. The limitation of the Poisson model is that the variance of the data is constrained to be equal to the mean. This constraint might be questionable as the variance of crash data is usually greater than the mean, which is known as over-dispersion (Washington et al., 2011). To address the over-dispersion issue, the Univariate Poisson Lognormal (UPLN) regression and Negative Binomial (NB) regression models are two commonly used approaches to predict the total crash counts or crash counts by crash type or severity (Lord and Mannering, 2010; Mannering and Bhat, 2014). However, all of these models assume crash counts by crash type or severity to be independent. When crash type or severity counts are considered simultaneously, modeling them independently might be questionable, because the crash counts among different crash types or severities may be correlated, due to the presence of shared unobserved factors across crash types or severities for a specific roadway intersection or segment. Neglecting their correlations might lead to biased parameter estimation, and reduce model accuracy (Ma et al., 2008; Mannering and Bhat 2014; Serhiyenko et al., 2016).

Researchers have applied a large number of methodologies to jointly estimate crashes by crash type or severity, such as simultaneous equations model (Ye et al., 2009 and Ye et al., 2013), Bayesian joint-probability model (Pei et al., 2011), two-stage mixed multivariate model (Wang et al., 2011), Multinomial generalized Poisson with spatiotemporal dependence Model (Chiou and Fu 2013, 2015; Chiou et al., 2014) and Artificial Neural Network (Zeng et al., 2016) In recent years, the Multivariate regression model has been increasingly used to estimate crash counts simultaneously by severity level, and explore the effects of explanatory variables on crash severity counts, as the Multivariate regression model is able to account for potential correlation between crashes among different crash severities (Mannering and Bhat, 2014). Ma and Kockelman (2006) applied a Multivariate Poisson (MVP) model to estimate crash counts by severity level. They found the crash counts are significantly correlated at different levels of injury severity. However, the MVP model cannot account for the overdispersion issue that is usually observed in crash data. Then the Multivariate Poisson Lognormal (MVPLN) model (Chib and Winkelmann, 2001) has been applied over the MVP model to accommodate the overdispersion issue in crash severity estimation (Park and Lord 2007; Ma et al., 2008). Similarly, Anastasopoulos et al. (2012) used both Multivariate Tobit (MVT) regression model and Multivariate Negative Binomial (MVNB) regression model to estimate the injury severity rates on multilane divided highways in Washington State. They found the prediction accuracy between MVT and MVNB model are very close, and the Multivariate model outperforms the Univariate model based on the goodness-of-fit.

The Bayesian framework using Markov Chain Monte Carlo (MCMC) simulation method for Multivariate models is one of the most popular alternatives to simultaneously estimate crash counts by severity level (Ma et al., 2008; Aguero-Valverde and Jovanis, 2009; El-Basyouny and Sayed, 2009). All of these studies indicated that there is a significant correlation across different crash severity counts. Furthermore, research found that unobserved heterogeneity in the crash data can result in biased and inconsistent parameter estimates (Mannering et al., 2016), the random-parameter Multivariate models have been increasingly used to address this issue. Dong et al. (2014) applied the Multivariate random-parameter zero-inflated negative binomial model to simultaneously estimate car-only crashes, car-truck crashes and truck-only crashes at urban intersections. They found the Multivariate random-parameter zero-inflated negative binomial can not only address the issue of excess zero crash counts, but can also account for the unobserved heterogeneity by allowing the coefficients of covariates to be random parameters. Some studies used the spatial Multivariate models to estimate severity counts by accounting for the spatial correlation, especially for zonal level crash prediction models. Barua et al. (2014) used the Multivariate Poisson Lognormal spatial model to estimate crash severity counts at urban segment in city of Richmond and Vancouver. The spatial correlation among crash severity counts were found to be statistically significant, and the Multivariate model outperforms the univariate model based on the goodness-of-fit. Wang and Kockelman (2013) used a Poisson Lognormal conditional-autoregressive model to estimate pedestrian crashes by severity at Census level in Austin, Texas. They found the spatial correlation of pedestrian crashes is positive across neighborhoods. To account for both the unobserved heterogeneity and spatial correlation of crash data, Barua et al. (2016) developed a Multivariate random parameter spatial Poisson Lognormal model to estimate crash counts by crash severity in the city of Vancouver. They found the unobserved heterogeneity issue to be present and can be addressed by this model, and the spatial correlation between severe and no-injury crashes is very high.

Despite all these Multivariate regression models in estimating crashes by crash severity, only a few studies that implement the MVPLN model have been conducted in estimating crash counts by crash type, and examine the different effects of variables on different crash types. Lee et al. (2015) used a MVPLN model to simultaneously estimate crash counts for motor vehicle crashes, bicycle crashes and pedestrian crashes by traffic analysis zone (TAZ) level in central Florida. The study illustrated that the MVPLN model outperforms the univariate model, and there is a significant correlation across the three crash type counts. Serhiyenko et al. (2016) used a MVPLN model to estimate freeway crashes by crash type in Connecticut. They verified that the crash counts are correlated among different crash types.

Although these MVPLN models can account for both the overdispersion of crash data and correlation among crash types or crash severities, all of these approaches using MCMC simulation method are computationally challenging and time consuming, especially for large data sets with a dependent variable containing many categories (Mannering and Bhat, 2014). In order to improve the computational time, instead of the MCMC simulation method, Serhiyenko et al. (2016) developed an Integrated Nested Laplace