Accident Analysis and Prevention 57 (2013) 140–149

Contents lists available at SciVerse ScienceDirect

Accident Analysis and Prevention

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

A simultaneous equations model of crash frequency by severity level for freeway sections

Xin Ye[a,*], Ram M. Pendyala[b,1], Venky Shankar[c,2], Karthik C. Konduri[d,3]

[a] *Civil Engineering Department, California State Polytechnic University, Room 17-2679, 3801 West Temple Avenue, Pomona, CA 91768, United States*
[b] *Department of Civil and Environmental Engineering, Arizona State University, Room ECG252, Tempe, AZ 85287-5306, United States*
[c] *Department of Civil and Environmental Engineering, The Pennsylvania State University, 212 Sackett Building, University Park, PA 16802, United States*
[d] *Department of Civil and Environmental Engineering, University of Connecticut, 261 Glenbrook Road, Unit 3037, Storrs, CT 06269-3037, United States*

ARTICLE INFO

Article history:
Received 19 October 2012
Received in revised form 25 February 2013
Accepted 21 March 2013

Keywords:
Crash severity
Crash frequency
Simultaneous equations model
Freeway safety
Maximum simulated likelihood estimation
Multivariate Poisson regression model

ABSTRACT

This paper presents a simultaneous equations model of crash frequencies by severity level for freeway sections using five-year crash severity frequency data for 275 multilane freeway segments in the State of Washington. Crash severity is a subject of much interest in the context of freeway safety due to higher speeds of travel on freeways and the desire of transportation professionals to implement measures that could potentially reduce crash severity on such facilities. This paper applies a joint Poisson regression model with multivariate normal heterogeneities using the method of Maximum Simulated Likelihood Estimation (MSLE). MSLE serves as a computationally viable alternative to the Bayesian approach that has been adopted in the literature for estimating multivariate simultaneous equations models of crash frequencies. The empirical results presented in this paper suggest the presence of statistically significant error correlations across crash frequencies by severity level. The significant error correlations point to the presence of common unobserved factors related to driver behavior and roadway, traffic and environmental characteristics that influence crash frequencies of different severity levels. It is found that the joint Poisson regression model can improve the efficiency of most model coefficient estimators by reducing their standard deviations. In addition, the empirical results show that observed factors generally do not have the same impact on crash frequencies at different levels of severity.

1. Introduction

Recent work on the simultaneous evaluation of frequency and severity of crashes has indicated that the body of empirical evidence on influential factors is still developing. In particular, the effect of geometrics in the unconditional context (via the simultaneous accommodation of crash frequency and the severity of crashes) is a matter of substantial focus. The seminal papers by Lord and Mannering (2010) and Savolainen et al. (2011) point to the need for more empirical work, since the typology of models is potentially exhaustive. At the very least, the papers seem to suggest that joint estimation of severity and frequency would shed further light into the "efficiency" aspects of statistical parameters associated with such influential factors as geometry and traffic volume. The problem of loss of "efficiency" of parameters when correlations among unobserved factors are ignored in simultaneous equations model systems is well known in the statistical and econometric literature. Yet, the computational aspects of addressing this problem via full-information estimation methods present considerable challenges to the development of empirical models.

This paper is aimed at making a contribution in this area by presenting a model of freeway crash frequency by severity level for 275 freeway sections in the State of Washington. Five year freeway section crash counts (by severity level) are available for the 275 freeway sections and a modeling methodology that can simultaneously account for observed and unobserved factors contributing to crash frequencies by severity level is developed and presented in this paper. The work presented in this paper is motivated by recent studies in the literature along this line of inquiry. Wang et al. (2011) present a model aimed at ranking sites with respect to safety using a two-stage modeling approach. Their objective is to obtain insights into frequency and severity simultaneously by accounting for the effects of spatial correlations in crash occurrence, as well as heterogeneity in severity, on statistical parameters. This particular study is highly insightful in that it points to a ripe area in the domain of

[*] Corresponding author. Tel.: +1 909 869 3444; fax: +1 909 869 4342.
E-mail addresses: ye@csupomona.edu (X. Ye), ram.pendyala@asu.edu (R.M. Pendyala), shankarv@engr.psu.edu (V. Shankar), kkonduri@engr.uconn.edu (K.C. Konduri).
[1] Tel.: +1 480 727 9164; fax: +1 480 965 0557.
[2] Tel.: +1 814 865 9434; fax: +1 814 863 7304.
[3] Tel.: +1 860 486 2733; fax: +1 860 486 2298.

0001-4575/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.aap.2013.03.025

site ranking – an area that has historically been dominated by single equation methods primarily. Pei et al. (2011) adopted a Bayesian approach to the joint modeling of crash frequency and severity via a Poisson log-normal prediction. Their paper serves as a useful precedent to this study in the sense that it examines crash frequency and severity of signalized intersections as an example application. Quite appropriately, the authors explore the effect of correlations due to the multi-approach effects at signals; however, the insights from this study are limited to signalized intersection contexts. The examination of crash frequency by severity in other empirical contexts still requires attention. The nature of correlations across crash types in the signalized intersection context can be substantially different when compared with other highway contexts, due to the different degrees of exposure to crash types, and therefore, crash severities. El-Basouny and Sayed (2009) make the case for the significance of outcome correlations using a multivariate Poisson log-normal specification albeit in a Bayesian sense. The authors argue that correlations between property damage and injury rates can be significant, an issue that requires further exploration. The authors note that the degree of correlation does have the potential to substantially affect parameter estimates and model inferences, which would then affect the accuracy of hazardous location identification.

Modeling crash frequencies by type of crash (angle, head-on, rear-end, etc.), number of vehicles involved (single-vehicle, two-vehicle, multi-vehicle, etc.), and severity level (property damage only, possible injury, incapacitating injury, etc.) has been the subject of much research in the transportation safety arena. Researchers have employed a variety of single equation count data models including Poisson models, Negative Binomial models, and zero-inflated versions of these count models (to account for the presence of zero crash frequency counts in the data set that may be due to facilities being truly safe or simply due to the limited window of observation for which crash frequency data is collected) to model crash frequency. These single equation models have often been employed to identify factors contributing to total crash frequencies on highway facilities. These methods can also be employed to model crash frequencies by severity level (e.g., modeling the number of fatal crashes as a function of roadway and traffic characteristics, environmental conditions, etc.). Although modeling crash frequency by severity level using single-equation methods (sets of independent equations) can offer valuable insights into factors affecting crash frequencies, the fact that such model systems ignore the simultaneity that may be prevalent in the safety phenomenon under investigation is an issue that merits being addressed. A model system in this paper refers to a series of equations to model a number of dependent variables that are mutually interrelated. There may be a host of unobserved factors related to driver characteristics, vehicular characteristics, roadway and traffic characteristics that contribute to crash frequencies at various severity levels. In a single equation method, the random error component (and the constant term) may be viewed as capturing the effects of these unobserved factors. However, simultaneity may arise in crash frequency modeling due to the possibility that unobserved factors affecting crash frequency at one severity level may be correlated significantly with unobserved factors affecting crash frequency at another severity level. This possibility calls for the deployment of simultaneous equations modeling methodologies to effectively model crash frequencies at multiple severity levels. This paper aims to make a contribution in this area by applying a multivariate count data model that is capable of accounting for correlated unobserved factors across equations representing crash frequencies at different severity levels. The correlation is accommodated by allowing for the presence of error covariances across equations, thus contributing to the simultaneity in the phenomenon under investigation.

The need for modeling crash frequencies by severity level in a simultaneous equations framework has been recognized; however, the analytical and computational complexity associated with formulating and estimating such systems has hindered the development of these model systems, particularly in the count model (data) context. This paper applies the modeling estimation technique proposed in Ye et al. (2009), where an n-dimensional multivariate count data model (Poisson regression) is formulated and presented to account for error correlations through the incorporation of normally distributed heterogeneity terms. Model estimation is achieved through the use of maximum simulated likelihood estimation (MSLE) methods that provide consistent parameter estimates and valid statistics for hypotheses tests. This paper makes a contribution to the understanding of crash frequencies at various severity levels for freeway sections. By applying a simultaneous equations model system for freeway crash severities, the paper provides key insights into the factors that impact crash frequencies at various severity levels while accounting for the presence of error covariances (common unobserved factors).

This paper has two major objectives. The first objective is to explore the use of the normal distribution to represent the heterogeneity in Poisson regression models of traffic crash frequency, as opposed to the log-gamma distribution which has been widely used in the Negative Binomial model. The paper aims to show that researchers can take advantage of the multivariate normal distribution to accommodate correlations of heterogeneities among multiple interrelated traffic crash frequencies at different severity levels in order to improve efficiency of coefficient estimators. The second major objective of the paper is to offer insights into the effects of various roadway, geometric, and traffic volume factors on crash frequencies by severity for freeway sections, while explicitly accounting for correlations across unobserved attributes.

Following a brief review of the literature, the paper presents the modeling methodology adopted in this paper. This is followed by a description of the dataset. Model estimation results and key conclusions are presented in the final two sections of the paper.

2. Modeling crash frequency and severity

Previous research in crash severity analysis such as that undertaken by Shankar et al. (1996) and Shankar and Mannering (1996) has mostly involved the development of univariate models of severity, with specific focus on total severity of the crash. In such studies, the most severe outcome of the crash is modeled. While the most severe outcome approach is useful in terms of a methodology for identifying model specifications, it does not provide for a comprehensive analysis of the severity of a crash. For example, the specific injury levels of occupants, and the variation in property damage among vehicles in multiple vehicle crashes are not adequately modeled. When one starts to consider these aspects, the notion of multivariate severity modeling is certainly appealing, but makes the modeling task more complex. In addition to these issues, the consideration of multiple modes tends to complicate injury modeling. For example, when pedestrians are involved, it is very rare that property damage outcomes are observed. In this case, the severity distribution is near-truncated at the possible injury level on the lower bound, which then leads to additional modeling complexities when accounting for multiple modal characterizations in multivariate or univariate models.

There is undoubtedly a vast body of literature devoted to modeling crash severity outcomes as a function of crash type, driver characteristics, roadway and traffic characteristics, and environmental conditions. These papers have used a variety of discrete choice modeling approaches, most notably the ordered probit and multinomial logit modeling approaches, to model crash severity outcomes. Examples of ordered probit models of injury severity include Quddus et al. (2002), Kockelman and Kweon (2002), Zajac and Ivan (2003), and Ma and Kockelman (2006a). Examples of unordered logit models of injury severity include those by Chang and Mannering (1999), Shankar et al. (1996), Shankar and Mannering (1996), Khorashadi et al. (2005), and Milton et al. (2008). The last paper by Milton et al. (2008) deploys a mixed logit modeling methodology to account for variations in the effects of explanatory factors on injurity severity outcomes. Other methods to estimate injury severity include the logistic regression model (e.g., Ossenbruggen et al., 2001) and artificial neural networks (e.g., Abdelwahab and Abdel-Aty, 2002; Abdel-Aty and Abdelwahab, 2004).

While such models are very useful to model crash severity outcomes at the level of the individual crash, with the exception of the Milton et al. (2008) paper, they do not offer insights into crash frequencies. The Milton et al. (2008) paper examines crash severity by frequency at the unconditional level but does not do so in a multivariate context. Due to the non-negative nature of crash frequency data, count models have been used extensively to estimate crash frequency. Ivan et al. (2000) presented the use of Poison regression models to estimate single and multi-vehicle crash rates as a function of traffic density, land use, ambient light conditions and time of day. If there are excessive zeros among crash frequencies, it is considered that some of the zeros in the count data are generated by a process that is different from the rest of the counts. This has led to the use of Zero Inflated Regression Models (Lee and Mannering, 2002; Lord et al., 2005) in place of the traditional Poisson and Negative Binomial regression models. Crash frequency models using a variety of count data modeling approaches have also been presented by Ma and Kockelman (2006a), Milton and Mannering (1998), Shankar et al. (1997), Miaou (1994), and Miaou et al. (1992). This paper is intended to add to this body of literature by simultaneously modeling crash frequencies across severity levels for freeway sections through a multivariate count data model. Recent developments in the formulation of multivariate approaches offer promise (Anastasopoulos et al., 2012).

The development and application of multivariate frequency models has been attempted in the field of transportation before. Zhao and Kockelman (2001) developed and applied a Multivariate Negative Binomial Regression model to analyze household vehicle ownership by vehicle type. Ma and Kockelman (2006b) and Ma et al. (2007) continued efforts to develop a Poisson regression model with multivariate normal heterogeneities. In these two papers, Bayesian approach is developed to estimate the parameters of the Multivariate Poisson (MVP) Regression Model that jointly predicts crash frequencies at different severity levels. The analytical and computational complexity associated with using maximum likelihood estimation methods for model estimation is noted in their papers. However, Ye et al. (2009) has demonstrated that the maximum simulated likelihood estimation method can be applied for MVP model estimation if a special variance-covariance structure is specified for the random error terms.

Park and Lord (2007) make a significant contribution toward advancing the development of multivariate Poisson regression models for simultaneously modeling crash frequencies. They developed a multivariate Poisson regression – lognormal model for modeling crash frequencies by severity level at intersections using a Bayesian estimation approach. This paper attempts to further contribute to the area of crash analysis by modeling crash frequencies at different severity level on freeway segments using an MVP model that accommodates unobserved heterogeneity (overdispersion) and flexible error covariance structures. Unlike their estimation approach, however, the simulation-based maximum likelihood estimation method employed in this paper uses Halton sequence draws to accurately compute the log-likelihood function (evaluate multidimensional integrals of the Poisson distribution). This method has been used extensively in the travel behavior arena to model a variety of travel behavior choices while incorporating random taste variations (see Bhat, 2003 for a detailed description of this method). Also, the paper by Park and Lord (2007) focuses on crash severity frequencies for intersections while this paper focuses on modeling crash severity frequencies for freeway sections.

In the multivariate context, recent work (Valverde and Jovanis, 2009; Park and Lord, 2007; Ma and Kockelman, 2006a,b; El-Basouny and Sayed, 2009) suggests using a lognormal based approach for accommodating correlations across severities. There are limitations with these approaches in that the correlations do not explicitly account for multimodal effects. A recent study by Chiou and Fu (2013) offers considerable promise in this arena. They estimate a multinomial-generalized Poisson model with error components to simultaneously model crash frequency and severity. The formulation in this paper is intended to offer an alternative specification where multivariate error correlations are accommodated through the use of the multivariate normal distribution. Other approaches to severity modeling include the ordered probit approach where the typical setup involves a restricted single slope treatment for parametric effects across severity functions. This assumption can be relaxed to provide for multiple slopes, however, it remains to be seen if the ordered probit approach provides for stable behavior in frequency level models. The cited literature (Duncan et al., 1998) indicates the suitability of this approach in conditional severity modeling, a point also made in the recent study by Savolainen et al. (2011). The final issue that contributes to the complexity of multivariate modeling of crash severity is the roadway dimension. Recent work by Abdel-Aty and Keller (2005) indicates the depth of this issue when one examines intersections for instance.

To summarize, multivariate modeling has promise, as evidenced in this paper and recent work in this area in the cited literature. In comparison to univariate approaches, the main evidence that appears in the non-Bayesian body of work relates to improvements in efficiency in parameter estimation and therefore, more accurate identification of the effects of variables influencing severity. There is some claim to improved predictions in the Bayesian literature (see for example, Valverde and Jovanis, 2009), which appears to be an artifact of the Bayesian approach in general, and not necessarily the multivariate aspects specifically.

3. Modeling methodology

This section presents the modeling methodology adopted in this paper. First, the univariate Poisson regression model is presented using two different heterogeneity specifications – the log-gamma heterogeneity (i.e. the Negative Binomial regression model) and normal heterogeneity. Although the focus of this paper is to develop a joint multivariate model for crash frequencies at various severity levels, it will be insightful to present a univariate model of the total crash frequency for comparisons.

3.1. Negative Binomial (NB) model: univariate Poisson regression model with log-gamma heterogeneity

Count data models are most suited to modeling any dependent variable y_i that constitutes a frequency or "count". The dependent variable can only take non-negative integer values. In this paper, y_i represents crash frequency by severity (but the subscript representing severity level is suppressed without loss of generality) for road section i. The expectation of y_i is assumed to be λ_i and the count data model formulation is as follows:

$$\ln(\lambda_i) = x_i\beta + \varepsilon_i, \tag{1}$$

where x_i is a vector of explanatory variables indicating characteristics for road section i; β is a vector of coefficients associated with x_i.

ε_i is a random variable representing heterogeneity that accounts for unobserved factors and other random disturbances. Since y_i constitutes count data, the probability of y_i conditional on ε_i is given as:

$$\Pr(y_i|\varepsilon_i) = \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!}. \tag{2}$$

The Negative Binomial (NB) regression model is formulated based on the assumption that $\exp(\varepsilon_i) = t_i$ follows a gamma distribution, denoted as $\Gamma(1/\alpha^2, \alpha^2)$. The corresponding probability density function is:

$$f(t_i) = \frac{t_i^{1/\alpha^2-1}}{(\alpha^2)^{1/\alpha^2}\Gamma(1/\alpha^2)}\exp\left(-\frac{t_i}{\alpha^2}\right), \quad t_i > 0, \tag{3}$$

$$\text{where} \quad \Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt \tag{4}$$

The expectation and standard deviation of t are equal to 1 and α, respectively. By integrating t_i over its distributional domain, one may obtain the unconditional probability of y_i as:

$$\Pr\,(y_i) = \int_{-\infty}^{\infty} \Pr(y_i|t_i)f(t_i)dt_i = \frac{\Gamma(1/\alpha^2 + y_i)}{\Gamma(1+y_i)\Gamma(1/\alpha^2)} r_i^{y_i}(1-r_i)^{1/\alpha^2}, \tag{5}$$

$$\text{where} \quad r_i = \frac{\alpha^2\exp(x_i\beta)}{\alpha^2\exp(x_i\beta)+1} \tag{6}$$

Cameron and Trivedi (1986) proposed this unconditional probability function with a closed-form solution. This formulation has allowed the NB model to be widely applied for modeling count data in many different areas, including transportation.

It is to be noted that the true heterogeneity in the model is not t_i, but ε_i, which accounts for the presence of unobserved variables or factors excluded from the vector x_i. Since ε_i is equal to $\ln(t_i)$, the underlying distributional assumption on ε_i is the log-gamma distribution and the probability density function can be derived as:

$$f(\varepsilon_i) = \frac{1}{\Gamma(1/\alpha^2)}\exp\left\{\frac{1}{\alpha^2}[\varepsilon_i - \ln(\alpha^2)] - e^{[\varepsilon_i - \ln(\alpha^2)]}\right\}, \quad -\infty < \varepsilon_i < +\infty. \tag{7}$$

It is not a symmetric function with respect to the variable ε_i, indicating that the distribution of the random variable ε_i is asymmetric in nature (Lawless, 1980).

3.2. Univariate Poisson regression (UVP) model with normal heterogeneity

In view of the asymmetric nature of the log-gamma heterogeneity specification in the NB regression model, one may choose to use a normal distribution for representing heterogeneity ε_i. If ε_i is normally distributed with 0 expectation and standard deviation of σ, the probability density function is:

$$f(\varepsilon_i) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\varepsilon_i^2}{2\sigma^2}\right), \quad -\infty < \varepsilon_i < +\infty. \tag{8}$$

Unlike the log-gamma distribution, the normal distribution is symmetric and its expectation can be fixed at 0 regardless of the standard deviation of the distribution. Under the assumption of normality, one can integrate ε_i over its distributional domain and obtain the unconditional probability of y_i as:

$$\begin{aligned}\Pr(y_i) &= \int_{-\infty}^{\infty}[\Pr(y_i|\varepsilon_i)f(\varepsilon_i)]d\varepsilon_i \\ &= \int_{-\infty}^{\infty}\frac{\exp[-\exp(x_i\beta+\varepsilon_i)][\exp(x_i\beta+\varepsilon_i)]^{y_i}}{y_i!}\frac{1}{\sqrt{2\pi}\sigma} \\ &\quad \exp\left(-\frac{\varepsilon_i^2}{2\sigma^2}\right)d\varepsilon_i\end{aligned} \tag{9}$$

The key difference in comparison to the NB regression model presented in Eqs. (5) and (6) is that the unconditional probability function presented in Eq. (9) does not have a closed-form solution. The Monte Carlo integral method can be applied to approximate the unconditional probability as follows:

$$SP(y_i) \approx \frac{1}{R}\sum_{r=1}^{R}\frac{\exp[-\exp(x_i\beta+\sigma u_{ir})][\exp(x_i\beta+\sigma u_{ir})]^{y_i}}{y_i!} \tag{10}$$

where SP represents the simulated probability function and u_{ir} are random seeds drawn from a standard normal distribution, which can be converted to normal random seeds with standard deviation σ by multiplying them with a single factor σ. Then, the Maximum Simulated Likelihood Estimation (MSLE) method can be applied to estimate unknown parameters β and σ with the aid of quasi-random draws (Bhat, 2003). Ordinary Least Square (OLS) estimators may be used as starting values of β. The initial value for σ can be the standard deviation corresponding to the α value estimated from the NB regression model.

3.3. Three-dimensional multivariate Poisson (TVP) regression model

The greatest benefit of using a normal distribution to represent heterogeneity is that one can easily realize an n-dimensional Multivariate Poisson (MVP) regression model, where n ($n \geq 2$) dependent (count) variables can be jointly modeled. The correlation among the dependent variables can be naturally accommodated into the correlation between their heterogeneities, which turns out to be a multivariate normal distribution in the case of a simultaneous equations system. In this particular paper, crash frequencies are modeled jointly for three different severity levels. They are: Property Damage Only, Possible Injury, Injury and Fatality. Thus, we have three dependent count variables ($n = 3$) and the logarithms of expectations λ_1, λ_2, and λ_3 for the three count variables are formulated as:

$$\begin{cases}\ln(\lambda_{1i}) = x_{1i}\beta_1 + \varepsilon_{1i} = x_{1i}\beta_1 + f_1u_{1i} \\ \ln(\lambda_{2i}) = x_{2i}\beta_2 + \varepsilon_{2i} = x_{2i}\beta_2 + f_2u_{1i} + f_3u_{2i} \\ \ln(\lambda_{3i}) = x_{3i}\beta_3 + \varepsilon_{3i} = x_{3i}\beta_3 + f_4u_{1i} + f_5u_{2i} + f_6u_{3i}\end{cases} \tag{11}$$

where u_{1i}, u_{2i} and u_{3i} are three independent random variables, which are standard normally distributed and f_i are coefficients to be estimated; x_i are vectors of explanatory variables and β_i are associated coefficient vectors. It is preferable to use (f_1u_{1i}), $(f_2u_{1i} + f_3u_{2i})$, and $(f_4u_{1i} + f_5u_{2i} + f_6u_{3i})$ to represent the trivariate normally distributed heterogeneities ε_{1i}, ε_{2i} and ε_{3i}. Since u_{1i}, u_{2i} and u_{3i} are normally distributed, their linear combinations associated with parameters f_i are also normally distributed. Essentially, the parameters f_i can form a decomposed lower triangular matrix of the covariance-variance matrix of heterogeneities and uniquely correspond to the covariance-variance matrix. They can also be used to reconstruct the covariance-variance matrix of heterogeneities, as shown in Eq. (15) through 18. Coefficients in the lower triangular matrix of the covariance-variance matrix are specified and estimated in this formulation (instead of the variances and covariances directly). It has been shown that this replacement formulation can help circumvent computational challenges associated with implementing numerical procedures to maximize the simulated log-likelihood function (Ye et al., 2009).

The probability functions conditional on multivariate normal heterogeneities are given as:

$$\begin{cases} \Pr(y_{1i}|u_{1i}) = \dfrac{\exp(-\lambda_{1i})\lambda_{1i}^{y_{1i}}}{y_{1i}!} \\ \Pr(y_{2i}|u_{1i}, u_{2i}) = \dfrac{\exp(-\lambda_{2i})\lambda_{2i}^{y_{2i}}}{y_{2i}!} \\ Pr(y_{3i}|u_{1i}, u_{2i}, u_{3i}) = \dfrac{\exp(-\lambda_{3i})\lambda_{3i}^{y_{3i}}}{y_{3i}!} \end{cases} \tag{12}$$

Then, the unconditional probability can be obtained by integrating conditional probability functions over the distributional domain of random heterogeneities:

$$\Pr(y_{1i}, y_{2i}, y_{3i}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}[\Pr(y_{1i}|u_{1i})\Pr(y_{2i}|u_{1i}, u_{2i}) \Pr(y_{3i}|u_{1i}, u_{2i}, u_{3i})]d\Phi(u_{1i})d\Phi(u_{2i})d\Phi(u_{3i}) \tag{13}$$

Similar to the UVP-normal heterogeneity model, this expression does not have a computationally tractable closed-form solution. The unconditional probability function can be approximated by the simulated probability function given below:

$$SP(y_{1i}, y_{2i}, y_{3i}) = \frac{1}{R}\left[\sum_{r=1}^{R} Pr(y_{1i}|u_{1ir})Pr(y_{2i}|u_{1ir}, u_{2ir})Pr(y_{3i}|u_{1ir}, u_{2ir}, u_{3ir})\right] \tag{14}$$

where u_{1ir}, u_{2ir}, and u_{3ir} are three independent random draws from a standard normal distribution. These terms can be transformed as (f_1u_{1i}), $(f_2u_{1i}+f_3u_{2i})$, and $(f_4u_{1i}+f_5u_{2i}+f_6u_{3i})$ for realizing the trivariate normal heterogeneities that may be prevalent in the simultaneous equations system of crash severity frequencies. Since parameters f_i can form a decomposed lower triangular matrix of the covariance-variance matrix of heterogeneities, variance, covariance, standard deviation, and correlation terms associated with the error structure can be calculated based on parameter f_i as:

$$Var\ (\varepsilon_1) = f_1^2, \quad Var\ (\varepsilon_2) = f_2^2 + f_3^2 \quad \text{and} \quad Var\ (\varepsilon_3) = f_4^2 + f_5^2 + f_6^2 \tag{15}$$

$$Cov\ (\varepsilon_1, \varepsilon_2) = f_1f_2, \quad Cov(\varepsilon_1, \varepsilon_3) = f_1f_4 \quad \text{and} \quad Cov(\varepsilon_2, \varepsilon_3) = f_2f_4 + f_3f_5 \tag{16}$$

$$Std\,Dev(\varepsilon_1) = f_1, \quad Std\,Dev(\varepsilon_2) = \sqrt{f_2^2 + f_3^2} \quad \text{and} \quad Std\,Dev(\varepsilon_3) = \sqrt{f_4^2 + f_5^2 + f_6^2} \tag{17}$$

$$Corr(\varepsilon_1, \varepsilon_2) = \frac{f_1f_2}{\sqrt{f_1^2(f_2^2 + f_3^2)}}, \quad Corr(\varepsilon_1, \varepsilon_3) = \frac{f_1f_4}{\sqrt{f_1^2(f_4^2 + f_5^2 + f_6^2)}} \quad \text{and} \quad Corr(\varepsilon_2, \varepsilon_3) = \frac{f_2f_4 + f_3f_5}{\sqrt{(f_2^2 + f_3^2)(f_4^2 + f_5^2 + f_6^2)}}. \tag{18}$$

As in the case of the univariate model with normal heterogeneity, MSLE method can be applied to estimate unknown parameter β and f_i using reasonable starting values. Ordinary Least Squares (OLS) estimators may be used as starting values for β. The overall standard deviation of error terms ε_1, ε_2 and ε_3 can be approximated from the NB or UVP models. Then, these estimated standard deviations are taken as starting values for the f_i terms (i.e., f_1, f_3 and f_6 as shown in Eq. (11). The other f_i (i.e., f_2, f_4 and f_5) terms start at zero. 500 sets of random seeds are drawn from the Halton sequence (Bhat, 2003) to accomplish accurate approximation of the log-likelihood function.

The log-likelihood function and its first-order derivative were coded in Gauss (Aptech, 2006) and the default BFGS algorithm provided by the Maxlik module in Gauss was used for maximizing the log-likelihood function. One problem of the applied simulation method is that the variance-covariance matrix of the error structure needs to be calculated based on more than one estimated parameter. Therefore, it is not straightforward to draw statistical inferences on the elements of the variance-covariance matrix (as they are functions of estimated parameters, f_i). To overcome this problem, a simulation-based hypothesis test is applied to approximate statistical inferences for these calculated elements of interest. The idea is that all of the parameter estimators obtained through Maximum Likelihood Estimation (MLE) procedures are asymptotically multivariate normally distributed and their expectations are the estimated parameters themselves and the variance-covariance matrix is the negative inverse of the Hessian matrix at convergence. A few columns of random variables subject to the multivariate normal distribution of all the estimators can be drawn to implement the simulation-based hypothesis test. The relevant columns of random variables can be selected for calculating the elements of interest and the significance level can be approximated by counting the positive or negative numbers among the calculated elements. For example, suppose Cov(ε_1, ε_3) is of interest. Since Cov(ε_1, ε_3)=f_1f_4, one needs to pick out two columns of random variables corresponding to f_1 and f_4 and calculate the product f_1f_4 for each pair of random variables to generate a column of "randomized" Cov(ε_1, ε_3). The percent of positive or negative counts among the random seeds for Cov(ε_1,ε_3) can approximate the significance level of Cov(ε_1,ε_3), i.e., the probability that Cov(ε_1,ε_3) is positive or negative. In addition, one may approximate the expectation and standard deviation of the variances and covariances based on the mean value and standard deviation of these random elements. However, it should be noted that the elements of the variance-covariance matrix are no longer normally distributed and the traditional t-statistic, calculated as the ratio between the mean and the standard deviation of each estimator, cannot be used for statistical inference.

4. Data description

The data set used in this paper is derived from safety (crash frequency) data available for multilane divided highways in the State of Washington. These highways, which are part of the National Highway System, are considered critical routes because of their high economic importance – they are also known for their high speed of travel, significant traffic volumes, and congestion. To obtain roadway segments of sufficient length to allow for use in Washington State safety programming applications, segments along this multilane system are defined by median treatments (safety barriers, cables or landform barriers). A roadway segment's beginning point was identified where a previous run of a barrier terminated (or began) and ended where the next run of a barrier was encountered (or the current run ended). In all, the data consist of 275 roadway segments of varying lengths with a mean segment length of roughly 2.4 miles with a standard deviation of about 2.7 miles. Historical crash data were gathered for the 1990–1994 timeframe. For each roadway segment, crashes were sorted by year, and individual crash data reports on the roadway segments were aggregated based on the most severe person-injury in the crash. Thus, crash frequency counts by severity level were obtained for each freeway segment.

The crash data were combined with weather data from the Western Regional Climate Center which included total precipitation (all forms) and snowfall precipitation. These weather data were observed using permanent weather stations and were assigned based on proximity of the station to a roadway segment. Data from the Washington State Department of Transportation databases were used for geometric, pavement, roadside and traffic characteristics associated with roadway segments. Geometric data included number of lanes, width of lanes, shoulder widths, median width, minimum and maximum radii of horizontal curves, central angle of horizontal curves, grade, minimum grade, maximum grade, grade differential, tangent length, number of changes in grade, number of horizontal and vertical curves per mile, presence of interchanges and presence of exit and entrance ramps. Pavement data included roadway pavement type, shoulder pavement type and friction coefficients. Roadside information included slopes, presence of vegetation, ditch information, the presence of crossovers, and information on other fixed objects (such as trees and poles). Traffic operations data included speed limit, average annual daily traffic, average daily traffic per lane, single-unit truck traffic, combination truck traffic, large truck percentages, peak-hour factors and roadway access control.

Information on a total of 22,619 individual crashes was included in this study. Due to the limited number of crashes that resulted in disabling injury and fatality, it was not possible to statistically differentiate among all five severity categories. Therefore, three severity categories are considered: property damage only, possible injury, and injury and fatality (with the third category encompassing evident injury, disabling injury, and fatality). With this definition, of the 22,619 individual crashes reported over the 5-year study period, 8367 resulted in property damage only, 6988 in possible injury, and 7264 in injury and fatality as the most severe outcome of the crash. Table 1 provides information on the mean, standard deviation, minimum and maximum of selected variables in the data set.

5. Model estimation results

This section presents model estimation results for various model forms considered in this paper. First, results are presented for univariate Poisson and Negative Binomial models of total crash frequency (sum of property damage only, possibly injury, injury and fatality crashes). Then, results are presented for the trivariate Poisson (TVP) regression model of crash frequency for the three severity levels under consideration.

5.1. Estimation results of UVP and NB models for total crash frequency

The univariate Poisson regression model (UVP with normal heterogeneity) and the Negative Binomial (NB) regression model of total crash frequency are presented on the right hand side of Table 2. The NB and UVP models are estimated using Gauss (Aptech, 2006) employing the Newton–Raphson optimization algorithm (the BFGS algorithm also yields the same solution, but requires more iterations to achieve convergence). As mentioned earlier, 500 sets of random seeds are drawn from the Halton sequence to accurately evaluate the log-likelihood function.

The value of α is estimated to be 0.6481 in the NB model. This parameter is statistically significant as evidenced by the large t-statistic value. In the UVP model, the estimate of the standard deviation of heterogeneity (ε_i) is 0.6752. The difference between normal distribution and log-gamma distribution is that the normal distribution is symmetric but the log-gamma distribution is asymmetric with a slight negative skew. It is found that the constant term and a few coefficients associated with explanatory variables in the UVP model are different in magnitude than those in the NB model. The marginal effect of explanatory variable x_i on dependent variable y can be expressed as:

$$\frac{\partial E(y|x)}{\partial x_i} = \frac{\partial E[\exp(x\beta+\varepsilon)]}{\partial x_i} = \frac{\partial \{E[\exp(\varepsilon)]\exp(x\beta)\}}{\partial x_i}. \qquad (19)$$

In the NB model, ε is log-gamma distributed and $\exp(\varepsilon)$ is gamma distributed with expectation equal to 1. Thus, the marginal effects in the NB model can be derived as:

$$\frac{\partial E(y|x)}{\partial x_i} = \beta_i \exp(x\beta). \qquad (20)$$

Table 1
Description of dependent and independent variables.

Variables	Min	Max	Mean	Std dev
Dependent variables				
Total frequency of crashes	0	182	16.45	21.87
Frequency of property-damage-only crashes	0	113	6.09	10.01
Frequency of possible injury crashes	0	96	5.08	8.49
Frequency of injury and fatal crashes	0	93	5.28	8.67
Independent variables				
AADT	3347	172,557	37,354.63	36,974.97
Logarithm of AADT	8.12	12.06	10.13	0.88
Section length (miles)	0.50	19.30	2.43	2.69
Logarithm of Section length	−0.69	2.96	0.51	0.82
Maximum grade in section	−5.50	6.72	−0.22	3.07
Maximum central angle in section	0.00	111.49	30.29	23.88
Scaled friction factor (scaled 0–100)	20.00	61.50	46.82	5.63
Number of grade breaks in section	0.00	28.00	3.87	4.09
Number of interchanges in section	0.00	4.00	0.85	0.83
Number of over-crossings in section	0.00	4.00	0.39	0.74
Number of ramp entries and exits in section	0.00	26.00	2.02	2.65
Number of median crossovers in section and the opposite direction	0.00	19.00	1.20	2.13
Average annual snowfall in inches	0.00	54.33	1.26	3.55

Table 2
Estimation results of recursive Poisson regression model with normal or log-gamma error structure.

Model type	Recursive Poisson model with normal error structure						Normal univariate Poisson model		Negative binomial model (log-gamma)	
Crash type	Property-damage-only crash freq.		Possible injury crash freq.		Injury and fatal crash freq.		Total crash freq.		Total crash freq.	
Variable	Coef.	S.D.	Coef.	S.D.	Coef.	S.D.	Coef.	S.D.	Coef.	S.D.
Constant	−5.7683	0.4023	−5.5019	0.4598	−8.3724	0.5396	−3.9295	0.3948	−4.9744	0.3688
Logarithm of AADT	0.7098	0.0294	0.6251	0.0334	0.8856	0.0365	0.6254	0.0300	0.7367	0.0278
Logarithm of section length (miles)	0.8428	0.0450	0.6281	0.0431	0.7584	0.0463	0.6944	0.0388	0.6624	0.0381
Maximum grade in section	0.0155	0.0084*	0.0354	0.0097	0.0242	0.0102	0.0294	0.0068	0.0220	0.0065
Maximum central angle in section/100	0.2083	0.1055	0.3267	0.1063	0.9549	0.1106	0.3260	0.0839	0.4382	0.0840
Friction factor/10	−0.2196	0.0462	−0.1309	0.0522	−0.1092	0.0536	−0.1968	0.0385	−0.1733	0.0375
Number of grade breaks in section	0.0117	0.0074**	0.0000	–	0.0371	0.0085	0.0162	0.0068	0.0137	0.0067
Number of interchanges in section	0.1575	0.0377	0.2201	0.0368	0.0000	–	0.1626	0.0289	0.1497	0.0280
Number of over-crossings in section	0.0000	–	0.0677	0.0345	0.0000	–	0.0780	0.0294	0.0613	0.0276
Number of ramp entries and exits in section	0.0000	–	0.0000	–	0.0000	–	0.0225	0.0088	0.0270	0.0088
Number of median crossovers in section and the opposite direction	0.0000	–	0.0166	0.0090*	−0.0381	0.0101	0.0000	–	0.0000	–
Average annual snowfall in inches	0.0000	–	0.0000	–	0.0204	0.0116*	0.0144	0.0073	0.0155	0.0070
Parameters in error structure										
$f_1, f_2, f_4, \sigma, \alpha$	0.9294	0.0232	0.0000	–	0.0000	–	0.6752	0.0173	0.6481	0.0159
f_3, f_5	–	–	0.9482	0.0250	0.0000	–	–	–	–	–
f_6	–	–	–	–	1.0458	0.0277	–	–	–	–
Goodness of fit measure										
LL(β)			−10,374.6				−4595.32		−4623.11	
LL(c)			−23,816.5				−15,847.5		−15,847.5	
Adj. $\rho^2(c)$			0.5633				0.7093		0.7076	
N			1375				1375		1375	

* Significance level of 0.10.
** Significance level of 0.15.
All the other coefficients are significant at the level of 0.05.

However, in the UVP model, ε is normally distributed and $\exp(\varepsilon)$ is log-normally distributed with expectation $\exp(\sigma^2/2)$. Then, the marginal effects in the UVP model can be derived to be:

$$\frac{\partial E(y|x)}{\partial x_i} = \exp\left(\frac{\sigma^2}{2}\right)\beta_i \exp(x\beta). \tag{21}$$

However, in both NB and UVP models, the coefficient β_i may be interpreted as the elasticity of the explanatory variable x_i because one can show that $\partial \ln[E(y|x)]/\partial x_i = \beta_i$ in both cases. In other words, one unit change in x_i will be associated with a $(100\cdot\beta_i)$% change in crash frequency.

As shown in the right two blocks of Table 2, UVP model coefficients are somewhat different from NB model coefficients. Out of ten coefficients for explanatory variables, six coefficients in UVP model are greater than those of the NB model in magnitude. As for the sign of the coefficients, both UVP and NB models yield coefficients with the same sign. It is also noted that the goodness-of-fit measure for the UVP model is seemingly better than that of the NB model (0.7093 versus 0.7076), indicating that the normal distribution serves as an acceptable alternative to a log-gamma distribution. A statistical test is needed to check whether the UVP model is significantly better than the NB model. The UVP model is non-nested with the NB model because one model specification cannot be achieved by restricting parameter(s) of the other model. In this case, a non-nested test can be conducted to compare these two models. Horowitz (1983) initially proposed a statistical test to compare non-tested models estimated by maximum likelihood method. Ben-Akiva and Lerman (1985) modified the Horowitz' test into a form represented by Akaike Information Criterion (AIC). The test is provided under the null hypothesis that model 1 is the true specification and the following holds asymptotically:

$$P(\bar{\rho}_2^2 - \bar{\rho}_1^2 > z) \le \Phi\{-[-2zL(0) + K_2 - K_1]^{0.5}\}, \tag{22}$$

where, $\bar{\rho}_i^2$: the adjusted likelihood ratio index at zero for model i = 1, 2, K_i: the number of parameters in model i, $\Phi(\ )$: the standard normal cumulative distribution function, $L(0)$ = log-likelihood value at zero.

The probability that the adjusted likelihood ratio index of model 2 is greater by some $z > 0$ than that of model 1, given that the latter is the true model, is asymptotically bounded by the right-hand side of Eq. (22) above. If the model with the greater $\bar{\rho}^2$ is selected, then this bounds the probability of erroneously choosing the incorrect model. Using this procedure, the models with alternative distributions for heterogeneities can be compared against one another. For Poisson regression models, one may use L(c) in place of L(0) in Eq. (22), where L(c) refers to the log-likelihood value of the model with the constant term only. In this paper, the non-nested test is used to compare Poisson regression models with log-gamma and normal heterogeneities. The right-hand side in Eq. (22) is calculated to be 1.065×10^{-13}, indicating that the total crash frequency model with normal heterogeneity is significantly better than the NB model.

As the focus of this research effort is on explaining crash frequencies by severity level, the total crash frequency models are not discussed here in detail. A more detailed discussion of the empirical results is presented next in the context of the trivariate Poisson regression model with multivariate normal heterogeneity.

5.2. Estimation results of TVP model for crash frequencies at three severity levels

Estimation results for the trivariate Poisson (TVP) regression model with multivariate normal heterogeneity are presented in Table 3. The estimation results provide parameter estimates for three simultaneous equations included in the model system, where one equation for each severity level is considered. Table 4 presents the expectation and standard deviation of elements in the error variance-covariance matrix and the error correlation matrix approximated by drawing 1,000,000 pseudo-random seeds. Among the random seeds drawn for covariances and correlations, no negative seeds are observed. These results indicate that all correlations

Table 3
Estimation results of joint Poisson regression model with trivariate normal error structure.

Model type	Joint Poisson model with trivariate normal error structure					
Crash type	Property-damage-only crash freq.		Possible injury crash freq.		Injury and fatal crash freq.	
Variable	Coef.	S.D.	Coef.	S.D.	Coef.	S.D.
Constant	−5.6645	0.3917	−4.9408	0.4355	−5.9409	0.5373
Logarithm of AADT	0.6758	0.0305	0.5941	0.0309	0.7116	0.0384
Logarithm of section length (miles)	0.8491	0.0429	0.6521	0.0426	0.7280	0.0414
Maximum grade in section	0.0145	0.0078*	0.0308	0.0090	0.0218	0.0094
Maximum central angle in section/100	0.3610	0.1015	0.4508	0.1051	0.7947	0.1045
Friction factor/10	−0.1797	0.0482	−0.1937	0.0519	−0.2398	0.0513
Number of grade breaks in section	0.0115	0.0067*	0.0000	–	0.0315	0.0076
Number of interchanges in section	0.1439	0.0350	0.1840	0.0361	0.0000	–
Number of over-crossings in section	0.0000	–	0.0546	0.0353**	0.0000	–
Number of ramp entries and exits in section	0.0000	–	0.0000	–	0.0000	–
Number of median crossovers in section and the opposite direction	0.0000	–	0.0223	0.0082	−0.0173	0.0097*
Average annual snowfall in inches	0.0000	–	0.0000	–	0.0179	0.0107*
Parameters in error structure						
$f_1, f_2, f_4, \sigma, \alpha$	0.9345	0.0239	0.2482	0.0307	0.2901	0.0298
f_3, f_5	–	–	0.9537	0.0272	0.1784	0.0274
f_6	–	–	–	–	0.9262	0.0248
Goodness of fit measure						
LL(β)			−10,300.8			
LL(c)			−23,816.5			
Adj. $\rho^2(c)$			0.5663			
N			1375			

* Significance level of 0.10.
** Significance level of 0.15.
All the other coefficients are significant at the level of 0.05.

between each pair of error terms are highly statistically significant (p-value < 1 × 10^{-6}). This finding of significant error correlations will be discussed in greater detail later in this section. Similar model estimation results were obtained when f_2, f_4 and f_5 in the error structure are fixed at zero as shown in the left block of Table 2. In this case, the TVP model reduces to three recursive UVP models which ignore error covariances. Relative to the TVP model, the UVP model's goodness-of-fit measure appears comparable (0.5633 vs. 0.5663). However, a likelihood ratio test can be conducted to compare the joint and recursive models since the recursive model can be achieved by restricting three parameters (i.e., f_2, f_4 and f_5) at zero in the joint model. The value of the test statistic can be calculated as 147.6 (=2 × [−10300.8 − (−10374.6)]), which is much greater than the critical value of Chi-square distribution (16.266 at three degrees of freedom for a probability level of 0.999). The likelihood ratio test indicates the three-equation recursive model (UVP) can be rejected in favor of the TVP model with error covariances. Meanwhile, among 26 explanatory variables (with the exception of those in the error structure), 22 of them have coefficient estimators with smaller variances, which shows the improvement in model efficiency through the joint TVP model.

Table 4
Estimated expectation and standard deviation of elements in error covariance matrix.

Elements	Mean	Std. dev.
Var1	0.8737	0.0446
Var2	0.9729	0.0530
Var3	0.9760	0.0476
Cov12	0.2320	0.0302
Cov13	0.2711	0.0291
Cov23	0.2422	0.0294
Std1	0.9344	0.0239
Std2	0.9860	0.0269
Std3	0.9876	0.0241
Corr12	0.2517	0.0305
Corr13	0.2938	0.0294
Corr23	0.2486	0.0281

The models of crash severity frequencies used almost the same set of variables as those used in the univariate models of total crash frequency, except the variable indicating the number of ramp entries and exits in the freeway section (due to statistical insignificance in the TVP model). The values of coefficients are generally found to be quite different across the three severity frequency equations. For example, the variable representing the maximum central angle in the section (divided by 100) takes a coefficient value in the model for injury and fatality frequency (0.7947) much greater than that in the possible injury frequency model (0.4508) and the property damage only frequency model (0.3610). It infers that the 1-degree increment in the central angle contributes to 0.3610% more property-damage-only crashes, 0.4508% more possible injured crashes, but 0.7947% more injury and fatal crashes. Therefore, the factor of central angle contributes more to severe traffic crashes than mild traffic crashes. If a single model of total crash frequency were specified (instead of a simultaneous equations model system), such differences will be masked leading to a biased inference regarding the impacts of roadway geometrics on safety. Based on this research, highway safety officials may consider new policies regarding restrictions on central angles of horizontal curves with a view to reducing severe traffic crashes. It should be noted that some variables are found to have coefficients with differing signs in the equations for crash frequencies at different severity levels. For example, the variable indicating the number of median crossovers has a negative impact on injury and fatal crash frequencies, but a positive impact (reversal of sign) on possible injury crash frequency.

Virtually all other explanatory variables have coefficient estimates with intuitively reasonable values and signs. The logarithms of AADT and the length of freeway section, both of which may be considered to serve as exposure measures, significantly contribute to crash severity frequencies (for all severity levels). The maximum grade and maximum central angle in the section, representing adverse geometric characteristics of the roadway section, also have positive coefficients indicating that they contribute to crash severity frequencies for all three severity levels. The friction factor shows a negative coefficient suggesting that higher friction factors are associated with lower crash frequencies for all severity levels. This is an interesting finding because one would expect some degree of compensation in driving behavior, as in choice of speed due to better friction.

The number of interchanges in the section appears with statistically significant positive coefficients in property-damage-only and possible injury crash frequency models. The possible injury effect appears to be larger presumably due to the fact that weaving, slower speeds, and the proximity of an interchange affect vehicle interactions. The number of over-crossings in the section takes a positive coefficient in possible injury crash frequency model although it is not highly significant. All of these factors presumably contribute to additional traffic conflicts and driving maneuvers that, in turn, lead to more crashes at all severity levels. Grade breaks appear to have a non-ignorable effect on injury and fatality occurrences, in part due to speed deviations and substantial changes in speeds of interacting vehicles.

The average annual snowfall in inches is the only weather-related variable specified in the models and it appears with a positive coefficient in injury and fatality crash frequency models. The t-test shows it is statistically significant at a level of 0.10. This finding suggests that snowfall contributes positively to injury/fatality crash frequencies. The positive effect of snowfall should again be viewed in the context of compensating driver behavior. As such, the positive effect can be viewed as net detrimental effect due to adverse weather and potentially resulting from vehicles leaving the roadway due to lack of control. An interesting variable for further exploration in the unconditional context would be the “excessive speed for driving conditions” variable, represented as a probability measure to account for severity related effects. Such a variable has been found to be significant in prior studies in conditional severity models (see for example, Shankar et al., 1996).

Returning to the point made earlier, it is found that all error correlations (presented in Table 4) are highly significant with very small standard deviations relative to the magnitude of correlations. It is clear that a simultaneous equations modeling methodology that accommodates cross-equation error covariances is appropriate and should be preferred for modeling safety phenomena such as that considered in this paper. The standard deviations of heterogeneities in the joint model are estimated at 0.9344, 0.9860 and 0.9876, which are comparable with those from the recursive model (0.9294, 0.9482 and 1.0458). The correlations among three heterogeneities range between 0.20 and 0.30, indicating the presence of common unobserved factors that affect crash frequency at different severity levels. It is likely that there are driver, vehicular, roadway and traffic characteristics that are unobserved and simultaneously contributing to crash frequencies of various severity levels. The correlated unobserved factors contributing to different crash frequencies clearly call for the modeling of crash frequencies jointly using simultaneous equation systems. Such systems are also able to capture the differential impacts of variables on crash frequencies at various severity levels. While such differential impacts could potentially be captured in independent (non-joint) single-equation model systems, the estimates of such impacts are likely to be inaccurate if one were to ignore the presence of correlated error terms (i.e., the presence of correlated unobserved factors).

6. Conclusions

This paper focuses on modeling crash frequencies for freeway sections and identifying the factors and their relative contribution to crash frequencies by severity level. Freeway safety is a topic of much interest to transportation professionals for several reasons. First, as speeds on freeways are higher than on other roadways, the occurrence of a crash on a freeway section is likely to result in a more severe outcome than on other roadways. Second, freeways are of vital importance for meeting travel needs of businesses and people and the occurrence of crashes on freeways results in severe economic losses through delays in travel time and costs associated with emergency response services, not to mention the enormous personal toll and cost associated with severe crashes. Enhancing safety on the nation’s freeways could play a significant part in reducing fatalities and saving precious public and private resources.

Research in the transportation safety arena has focused on the development of models of crash frequency using a variety of count data modeling methodologies including Poisson regression, Negative Binomial regression, and zero-inflated versions of these models. However, most research efforts in the past have treated crash frequency models as single equation models where a single crash frequency (either total or of a certain type or severity) is modeled as a function of roadway characteristics, traffic characteristics, roadway conditions, and environmental conditions. While such single equation systems offer very useful insights, there is the potential to significantly enhance such models by simultaneously modeling crash frequencies of different severity levels while explicitly recognizing that there may be common unobserved factors affecting crash frequencies at various severity levels. To this end, this paper presents a simultaneous equations model for several count dependent variables while explicitly recognizing the multivariate nature of the error structure across equations. The unobserved heterogeneity terms (or random error components) in the different equations are potentially correlated with one another leading to the simultaneity in the phenomenon under study and the multivariate nature of the error structure.

In this paper, a simultaneous multivariate Poisson (MVP) regression model that accounts for correlated error structures is formulated, developed, and estimated to analyze crash frequencies for three different severity levels, i.e., property damage only, possible injury, injury and fatal. Five-year crash frequency data for 275 freeway sections in the State of Washington are used in this study. The study involved formulating the MVP model and adopting a multivariate normal heterogeneity specification that allows the accommodation of correlated error structures while facilitating efficient model estimation using maximum simulated likelihood estimation (MSLE) procedures. In addition, the study involved the development of a simulation method to compute error variances and covariances in a computationally practical way.

The model estimation results are intuitively reasonable and offer valuable insights into the factors that affect crash frequencies at various severity levels. The key findings in this paper are summarized as below:

(1) The replacement of log-gamma distribution with the normal distribution for representing heterogeneity in the Poisson regression model of traffic crash frequency can improve the goodness-of-fit of the model.
(2) The trivariate error correlations estimated in the empirical context of this paper are highly statistically significant, strongly supporting the notion that crash frequencies (by severity) should be modeled in a rigorous simultaneous equations modeling framework such as that presented in this paper.
(3) The joint Poisson regression model can improve the efficiency of most model coefficient estimators by reducing their standard deviations.
(4) The joint model can help differentiate the impact of the same freeway attribute on the crash frequency at different severity levels. For example, it is found that a freeway section with

greater maximum central angle tends to have more severe traffic crashes than mild ones.

Limitations of the current study offer directions for future research in this domain. Data reliability is an issue that is frequently encountered in the modeling of crashes, particularly when crash data is obtained from police reports (which is the case here). It is likely that under-reporting of crashes could have an impact on the model coefficient estimates and the consequent inferences drawn in this paper. Examining the potential impact of under-reporting on correlations across crash severities is a fruitful area for future research. In this paper, an attempt has been made to include a number of significant variables to control for factors contributing to crash frequencies, but it is inevitable that important factors (e.g., driver-related, environment-related, and vehicle-related) remain unobserved and are therefore omitted from the model. As is likely in any modeling effort, the existence of correlation between omitted exogenous variables and random error terms introduces the potential for bias. In this particular study, the impact of important geometric effects on freeway segment crash propensities has been modeled. As many of these variables have been tested in prior research (see for example, Milton et al., 2008), it is possible to draw inferences based on the body of evidence in the literature despite the potential presence of omitted variable effects. While it cannot be said with certainty that the omitted variable effect is not introducing bias, this study has shown that the statistical efficiency of parameter estimates can be substantially improved with a multivariate treatment of crash frequencies by severity that explicitly accounts for correlations across unobserved heterogeneity terms. Another limitation of this study is that severe injury and fatality crashes were combined into a single count. Datasets that separate out severe injury and fatal crash frequencies may offer richer insights into the differential impacts of various factors on crash frequencies at different severity levels. Finally, it would be valuable to pursue further research that explores the robustness of the findings to model specification and geographical contexts.

References

Abdelwahab, H.T., Abdel-Aty, M.A., 2002. Investigating driver injury severity in traffic accidents using fuzzy ARTMAP. Computer Aided Civil and Infrastructure Engineering 17, 396–408.

Abdel-Aty, M.A., Abdelwahab, H.T., 2004. Predicting injury severity levels in traffic crashes: a modeling comparison. Journal of Transportation Engineering 130 (2), 204–210.

Abdel-Aty, M., Keller, J., 2005. Exploring the overall and specific crash severity levels at signalized intersections. Accident Analysis and Prevention 37 (3), 417–425.

Aguero-Valverde, J., Jovanis, P.P., 2009. Bayesian multivariate Poisson lognormal models for crash severity modeling and site ranking journal. transportation research record. Journal of the Transportation Research Board 2136, 82–91.

Anastasopoulos, P., Shankar, Ch., Haddock, V.N., Mannering, J.E.F.L., 2012. A Multivariate tobit analysis of highway accident-injury-severity rates. Accident Analysis and Prevention 45 (1), 110–119.

Aptech, 2006. GAUSS 8.0 Aptech Systems. Maple Valley, Washington.

Bhat, C.R., 2003. Simulation estimation of mixed discrete choice models using randomized and scrambled halton sequences. Transportation Research Part B 37 (9), 837–855.

Ben-Akiva, M., Lerman, S.R., 1985. Discrete Choice Analysis: Theory and Application to Travel Demand. The MIT Press, Cambridge.

Cameron, C., Trivedi, P., 1986. Econometric models based on count data: comparisons and applications of some estimators and tests. Journal of Applied Econometrics 46, 347–364.

Chang, L., Mannering, F.L., 1999. Analysis of injury severity and vehicle occupancy in truck- and non-truck-involved accidents. Accident Analysis and Prevention 31, 579–592.

Chiou, Y.-C., Fu, C., 2013. Modeling crash frequency and severity using multinomial-generalized Poisson model with error components. Accident Analysis and Prevention 50 (1), 73–82.

Duncan, C.S., Khattak, A.J., Council, F.M., 1998. Applying the ordered probit model to injury severity in truck-passenger car rear-end collisions. Transportation Research Record: Journal of the Transportation Research Board 1635, 63–71.

El-Basouny, K., Sayed, T., 2009. Collision prediction models using multivariate Possion log-normal regression. Accident Analysis and Prevention 41 (4), 820–828.

Horowitz, J.L., 1983. Statistical comparison of non-nested probabilistic discrete choice models. Transportation Science 17 (3), 319–350.

Ivan, J.N., Wang, C., Bernardo, N.R., 2000. Exploring two-lane highway crash rates using land use and hourly exposure. Accident Analysis and Prevention 32, 787–795.

Khorashadi, A., Niemeier, D., Shankar, V., Mannering, F.L., 2005. Differences in rural and urban driver-injury severities in accidents involving large-trucks: an exploratory analysis. Accident Analysis and Prevention 37, 910–921.

Kockelman, K.M., Kweon, Y., 2002. Driver injury severity: an application of ordered probit models. Accident Analysis and Prevention 34, 313–321.

Lawless, J.F., 1980. Inference in the generalized gamma and log gamma distributions. Technometrics 22 (3), 409–419.

Lee, J., Mannering, F.L., 2002. Impact of roadside feature on the frequency and severity of run-off-roadway accidents: an empirical analysis. Accident Analysis and Prevention 34, 149–161.

Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transportation Research Part A: Policy and Practice 44 (5), 291–305.

Lord, D., Washington, S.P., Ivan, J.N., 2005. Poisson, Poisson-gamma and zero inflated regression models of motor vehicle crashes: balancing statistical fit and theory. Accident Analysis and Prevention 37, 35–46.

Ma, J., Kockelman, K.M., 2006a. Crash frequency and severity modeling using clustered data from Washington State. In: Proceedings of the IEEE Intelligent Transportation Systems Conference 2006, Toronto, Canada September 17–20, p. 2006.

Ma, J., Kockelman, K.M., 2006b. Bayesian multivariate Poisson regression for models injury count, by severity. Transportation Research Record: Journal of the Transportation Research Board 1950, 24–34.

Ma, J., Kockelman, K.M., Damien, P., 2007. A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using bayesian methods. In: Presented at the 86th Annual Meeting of the Transportation Research Board, National Research Council, Washington, D.C.

Miaou, S.-P., 1994. The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. Accident Analysis and Prevention 26, 471–482.

Miaou, S.-P., Hu, P.S., Wright, T., Rathi, A.K., Davis, S.C., 1992. Relationship between truck accidents and geometric design: a Poisson regression approach. Transportation Research Record 1376, 10–18.

Milton, J., Mannering, F.L., 1998. The relationship among highway geometrics traffic-related elements, and motor-vehicle accident frequencies. Transportation 25 (4), 395–413.

Milton, J.C., Shankar, V., Mannering, F.L., 2008. Highway accident severities and the mixed logit model: an exploratory empirical analysis. Accident Analysis and Prevention 40 (1), 260–266.

Ossenbruggen, P.J., Pendharkar, J., Ivan, J.N., 2001. Roadway safety in rural and small urbanized areas. Accident Analysis and Prevention 33, 485–498.

Park, E.S., Lord, D., 2007. Multivariate Poisson-lognormal models for jointly modeling crash frequency by severity. Transportation Research Record: Journal of the Transportation Research Board 2019, 1–6 (the National Research Council, Washington, D.C.).

Pei, X., Wong, S.C., Sze, S.S., 2011. A joint-probability approach to crash prediction models. Accident Analysis and Prevention 43 (3), 1160–1166.

Quddus, M.A., Noland, R.B., Chin, H.C., 2002. An analysis of motorcycle injury and vehicle damage severity using ordered probit models. Accident Analysis and Prevention 33, 445–462.

Shankar, V., Mannering, F.L., Barfield, W., 1996. Statistical analysis of accident severity on rural freeways. Accident Analysis and Prevention 28, 391–401.

Shankar, V., Mannering, F., 1996. An exploratory multinomial logit analysis of single-vehicle motorcycle accident severity. Journal of Safety Research 27 (3), 183–194.

Shankar, V.N., Milton, J.C., Mannering, F.L., 1997. Modeling statewide accident frequencies as zero-inflated probability processes: an empirical inquiry. Accident Analysis and Prevention 29, 829–837.

Savolainen, P., Mannering, F., Lord, D., Quddus, M., 2011. The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. Accident Analysis and Prevention 43 (5), 1666–1676.

Wang, C., Quddus, M., Ison, S., 2011. Predicting accident frequency at their severity levels and its application in site ranking using a two-stage mixed multivariate model. Accident Analysis and Prevention 43 (6), 1979–1990.

Ye, X., Pendyala, R.M., Washington, S.P., Konduri, K., Oh, J., 2009. A simultaneous equations model of crash frequency by collision type for rural intersections. Safety Science 47 (3), 443–452.

Zajac, S.S., Ivan, J.N., 2003. Factors influencing injury severity of motor vehicle-crossing pedestrian crashes in rural Connecticut. Accident Analysis and Prevention 35, 369–379.

Zhao, Y., Kockelman, K.M., 2001. Household vehicle ownership by vehicle type: application of a multivariate negative binomial model. In: Presented at the 81st Annual Meeting of the Transportation Research Board, National Research Council, Washington, D.C., 2001.