Multivariate Poisson–Lognormal Models for Jointly Modeling Crash Frequency by Severity

Eun Sug Park and Dominique Lord

A new multivariate approach is introduced for jointly modeling data on crash counts by severity on the basis of multivariate Poisson–lognormal models. Although the data on crash frequency by severity are multivariate in nature, they have often been analyzed by modeling each severity level separately, without taking into account correlations that exist among different severity levels. The new multivariate Poisson–lognormal regression approach can cope with both overdispersion and a fully general correlation structure in the data, as opposed to the recently suggested multivariate Poisson regression approach, which allows for neither overdispersion nor a general correlation structure in the data. The new method is applied to the multivariate crash counts obtained from intersections in California for 10 years. The results show promise toward the goal of obtaining more accurate estimates by accounting for correlations in the multivariate crash counts and overdispersion.

There has been considerable research on crash data analysis and statistical modeling (*1–12*). Crash data are often collected in terms of crash frequencies and severity levels. Examples of severity levels are fatal (K), incapacitating-injury (A), nonincapacitating injury (B), minor injury (C), and property damage only (PDO or O). Although the data on crash frequency by severity are multivariate in nature, they have often been analyzed by modeling each severity level separately, without taking into account correlations that exist among different severity levels. Usually, statistical models are produced for all crash severity levels (often referred to as KABCO) or for different crash severity levels, such as fatal and nonfatal crashes (e.g., KAB) or for PDO crashes.

Treating the correlated crash counts as independent and applying a univariate model to each count leads to less precise estimates for the effects of certain factors on crash risk. Unfortunately, there has not been much research on jointly modeling crash counts of different severity levels in highway safety. Notable exceptions include work by Tunaru (*13*), Bijleveld (*14*), Miaou and Song (*12*), Song et al. (*15*), and Ma and Kockelman (*16*). Ma and Kockelman (*16*) adapted a multivariate Poisson (MVP) regression approach developed by Tsionas (*17*) to assess the effects of various covariates on multivariate crash counts by severity. The MVP regression models, however, do not allow for overdispersion, which is often observed in crash data. In addition, the MVP regression models used by Ma and Kockelman (*16*) assume that the covariances for different severity levels are all identical [although the assumption of equal covariances has been relaxed in an extended MVP regression model developed by Karlis and Meligkotsidou (*18*)] and nonnegative, which is very restrictive. It is possible that different severity levels may have different covariances, and also the possibility of negative correlations cannot be entirely excluded.

An MVP–lognormal (MVPLN) regression approach developed by Chib and Winkelmann (*19*) can serve as a good alternative to a pure MVP regression approach for analysis of multivariate crash count data because the MVPLN approach can account for overdispersion and a fully general correlation structure, whereas the MVP model cannot. Also, MVPLN regression models are more general than multivariate negative binomial regression models in the sense that the former can account for negative correlations, whereas the latter cannot. Although these models already have been developed in statistics (*20*), there have been almost no attempts to employ those models in roadway safety to model data on multivariate crash frequency by severity. One exception is the work by Tunaru (*13*), who introduced an MVPLN model in the nonregression context (ranking the sites with accidents) to take into account general correlation structures. However, he did not consider any covariates.

Implementation of MVPLN models is not straightforward. It needs to be noted that no existing statistical software has the ability to estimate these models as built-in functions. As mentioned by Chib and Winkelmann (*19*), it is necessary to adapt simulation-based methods such as a Markov chain Monte Carlo (MCMC) simulation method (*21–23*) to cope with the multiple integral in the likelihood function. To estimate MVPLN models, the MATLAB (*24*) codes tailored to multivariate crash data modeling have been developed according to the MCMC algorithm of Chib and Winkelmann (*19*).

Analysis of multivariate crash count data by severity with the MVPLN regression models implemented by the MCMC method to assess the effects of covariates is presented. The models were developed by using crash data collected at three-leg unsignalized intersections in California.

MVPLN MODELS

The underlying model and the implementation algorithm on the basis of which the MCMC codes were developed are redescribed here in the context of the crash count data. Mathematical details can be found elsewhere (*19*). Because the dimensions of the matrices and vectors given by Chib and Winkelmann (*19*) were sometimes inconsistent, they are redefined here for clarification purposes.

Texas Transportation Institute, Texas A&M University System, 3135 TAMU, College Station, TX 77843-3135. Corresponding author: E. S. Park, e-park@tamu.edu.

Transportation Research Record: Journal of the Transportation Research Board, No. 2019, Transportation Research Board of the National Academies, Washington, D.C., 2007, pp. 1–6.
DOI: 10.3141/2019-01

Modeling Framework

Let **Y** denote an *n*-by-*J* matrix of the multivariate crash counts, where *n* corresponds to the number of intersections and *J* corresponds to the number of different severity types, and let **b** denote an *n*-by-*J* matrix of latent effects of which rows $b_i = (b_{i1}, \ldots, b_{iJ})$, $i = 1, \ldots, n$, correspond to a set of *J* intersection and outcome-specific latent effects. Let *k* be the number of covariates, and let *X* denote an *n*-by-*k* matrix of covariates of which rows $x_i = (x_{i1}, \ldots, x_{ik})$ are the *k*-dimensional row vectors corresponding to the *i*th intersection as follows:

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Let $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_J]$ denote a *k*-by-*J* matrix of the regression coefficients of which columns

$$\beta_j = \begin{pmatrix} \beta_{1j} \\ \vdots \\ \beta_{kj} \end{pmatrix}$$

are the *k*-dimensional column vectors consisting of parameters for the crash count of *j*th severity type. Suppose that, conditional on b_{ij} and parameters $\beta_j \in R^k$, the crash count of the *j*th severity type at the *i*th intersection, y_{ij}, follows a Poisson distribution with mean $\mu_{ij} = \exp(x_i \beta_j + b_{ij})$; that is,

$$y_{ij} | b_i, \beta_j \sim \text{Poisson}(\mu_{ij}) \tag{1}$$

where

$$\mu_{ij} = \exp(x_i \beta_j + b_{ij}) \tag{2}$$

for $j = 1, \ldots, J$ and $i = 1, \ldots, n$. The y_{ij}'s are independent given the μ_{ij}'s.

To model the correlations among the crash counts of *J* different severity types at an intersection, let

$$b_i | \Sigma \sim N_J(0, \Sigma) \qquad i = 1, \ldots, n \tag{3}$$

where Σ is an unrestricted covariance matrix, and N_J denotes a *J*-dimensional multivariate normal distribution. It was shown by Chib and Winkelmann (*19*) that the variance of y_{ij} is greater than the mean (allowing for overdispersion) as long as the diagonal elements of Σ are greater than 0, and the covariance between the counts, y_{ij} and y_{is}, can be positive or negative depending on the sign of the (*j*,*s*)th element of Σ. Thus, the correlation structure of the crash counts is unrestricted.

Estimation by Means of MCMC Method

As noted by Chib and Winkelmann (*19*), the marginal distribution of the counts $y_i = (y_{i1}, y_{i2}, \ldots, y_{iJ})$ cannot be obtained by direct computation because it requires the evaluation of a *J*-variate integral of the Poisson distribution with respect to the distribution of b_i.

The MCMC simulation is thus employed for parameter estimation under a Bayesian framework. For the prior on the parameters, it is assumed that $(\beta_1, \beta_2, \ldots, \beta_J, \Sigma)$ independently follow the distributions $\beta_j \sim N_k(\beta_0, B_0^{-1})$, $j = 1, \ldots, J$, $\Sigma^{-1} \sim \text{Wishart}(R_0, r_0)$, where (β_0, B_0, r_0, R_0) are known hyperparameters and Wishart $(\cdot,\cdot)$ is the Wishart distribution (*25*) with scale matrix R_0 and degrees-of-freedom parameter r_0. Then the joint posterior density is proportional to

$$\text{posterior} \propto \text{likelihood} \times \text{prior} = f_W\left(\sum^{-1} | r_0, R_0\right) \prod_{j=1}^{J} \phi_k\left(\beta_j | \beta_0, B_0^{-1}\right) \times \prod_{i=1}^{n} \left\{ \prod_{j=1}^{J} f\left(y_{ij} | \beta_j, b_{ij}\right) \right\} \phi_J\left(b_i | 0, \Sigma\right) \tag{4}$$

where f_W is the Wishart density and ϕ_k and ϕ_J are the *k*-variate and the *J*-variate normal density, respectively.

Three move types are used as by Chib and Winkelmann (*19*) in implementing the MCMC method: Sampling **b**, Sampling $\boldsymbol{\beta}$, and Sampling Σ^{-1}. Those three steps are presented briefly. Some notational errors and typographical errors found in the Chib and Winkelmann paper (*19*) have been corrected here.

Sampling ***b***

$$\text{Sampling } \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

where each b_i is a *J*-dimensional row vector.

The full conditional posterior density for b_i, $\pi(b_i | \ldots)$, is not given by any known density [see paper by Chib and Winkelmann (*19*)] and requires the Metropolis–Hastings (M-H) algorithm (*26*). A multivariate-*t* distribution with degrees of freedom v_1, the location parameter $\hat{b}_i$, and the scale parameter $V_{\hat{b}_i}$, $f_T(b_i | \hat{b}_i, V_{\hat{b}_i}, v_1)$, is used as a proposal density for b_i Here, v_1 is a tuning parameter and $\hat{b}_i$ and $V_{\hat{b}_i}$ are the mode and the inverse of minus the Hessian matrix of $\log \pi^+(b_i | y_i, \boldsymbol{\beta}, \Sigma)$ at the mode $\hat{b}_i$, where log denotes a natural log and

$$\log \pi^+\left(b_i | y_i, \boldsymbol{\beta}, \Sigma\right) = -0.5 \ln |2\pi\Sigma| - 0.5\left(b_i \Sigma^{-1} b_i'\right) + \sum_{j=1}^{J} \left[-\exp\left(x_i \beta_j + b_{ij}\right) + y_{ij}\left(x_i \beta_j + b_{ij}\right) \right] \tag{5}$$

To find $\hat{b}_i$, and $V_{\hat{b}_i} = (-H_{\hat{b}_i})^{-1}$, the Newton–Raphson algorithm with the gradient vector $g_{b_i} = -b_i \Sigma^{-1} + [y_i - \exp(x_i \boldsymbol{\beta} + b_i)]$ and Hessian matrix $H_{b_i} = -\Sigma^{-1} - \text{diag}\{\exp(x_i \boldsymbol{\beta} + b_i)\}$ can be used. A proposal b_i^* drawn from $f_T(b_i | \hat{b}_i, V_{\hat{b}_i}, v_1)$ is then accepted with probability

$$\min\left\{ \frac{\pi^+\left(b_i^* | y_i, \beta, \Sigma\right) f_T\left(b_i | \hat{b}_i, V_{\hat{b}_i}, v_1\right)}{\pi^+\left(b_i | y_i, \beta, \Sigma\right) f_T\left(b_i^* | \hat{b}_i, V_{\hat{b}_i}, v_1\right)}, 1 \right\}$$

Sampling $\boldsymbol{\beta}$

$$\text{Sampling } \boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_J]$$

where each β_j is a *k*-dimensional vector.

The full conditional posterior density for $\boldsymbol{\beta}$ is not given by any known density either, and it requires the M-H algorithm. A "block-at-a-time" M–H algorithm is used to sample β_J ($j = 1, \ldots, J$) one

at a time. A multivariate-t distribution with degrees of freedom v_2, the location parameter $\hat{\beta}_j$, and the scale parameter $V_{\hat{\beta}_j}$, $f_T(\beta_j | \hat{\beta}_j, V_{\hat{\beta}_j}, v_2)$, can be used as a proposal density. Here, v_2 is a tuning parameter and $\hat{\beta}_j$ and $V_{\hat{\beta}_j}$ are the mode and the inverse of minus the Hessian matrix of $\log \pi^+(\beta_j | y_j, b_j)$ at the mode $\hat{\beta}_j$ where log denotes a natural log, y_j and b_j denote the jth columns of the matrices **Y** and **b,** respectively, and

$$\log \pi^+(\beta_j | y_j, b_j) = -0.5 \log |2\pi B_{0j}^{-1}| - 0.5\left((\beta_j - \beta_{0j})' B_{0j}(\beta_j - \beta_{0j})\right) + \sum_{i=1}^{n}\left[-\exp(x_i\beta_j + b_{ij}) + y_{ij}(x_i\beta_j + b_{ij})\right] \quad (6)$$

To find $\hat{\beta}_j$ and $V_{\hat{\beta}_j} = (-H_{\hat{\beta}_j})^{-1}$, the Newton–Raphson algorithm with the gradient vector

$$-B_{0j}(\beta_j - \beta_{0j}) + \sum_{i=1}^{n}\left[y_{ij} - \exp(x_i\beta_j + b_{ij})\right]x_i'$$

and Hessian matrix

$$H_{\beta_j} = -B_{0j} - \sum_{i=1}^{n}\left[\exp(x_i\beta_j + b_{ij})\right]x_i'x_i$$

can be used. A proposal β_j^* drawn from $f_T(\beta_j | \hat{\beta}_j, V_{\hat{\beta}_j}, v_2)$ is then accepted with probability

$$\min\left\{\frac{\pi^+(\beta_j^* | y_j, b_j)\, f_T(\beta_j | \hat{\beta}_j, V_{\hat{\beta}_j}, v_2)}{\pi^+(\beta_j | y_j, b_j)\, f_T(\beta_j^* | \hat{\beta}_j, V_{\hat{\beta}_j}, v_2)}, 1\right\}$$

Sampling Σ^{-1}

The Gibbs sampling algorithm is used to sample Σ^{-1} because the full conditional posterior distribution of Σ^{-1} is given by

$$\sum^{-1} | \mathbf{b} \sim \text{Wishart}\left(n + \upsilon_0, \left[R_0^{-1} + \sum_{i=1}^{n} b_i' b_i\right]^{-1}\right) \quad (7)$$

Inferences Based on MCMC Samples

Thousands (or millions, if necessary) of samples can be simulated indirectly from the joint posterior distribution by using the MCMC algorithm. Here the samples represent the values of the parameters, and inferences (point estimates, uncertainty estimates, or interval estimates, or all three) on the parameters can be directly made on the basis of those samples (often called posterior samples). For instance, the sample mean and sample standard deviation of the posterior samples of β can be used as the point estimate and the corresponding uncertainty estimate (standard error) for β. Also, the 2.5th percentile and the 97.5th percentile can be used to construct the 95% credible interval for the elements of β, which is another useful way of representing uncertainty. It needs to be emphasized that convergence of the chain has to be ensured before one makes any inferences.

APPLICATION TO CALIFORNIA INTERSECTION CRASH DATA

The MVPLN models are applied to crash count data of five different severity levels—Sev1: fatal (K), Sev2: incapacitating-injury (A), Sev3: nonincapacitating injury (B), Sev4: minor injury (C), Sev5: property damage only (PDO or O)—collected from 451 three-leg unsignalized intersections in California obtained through the Highway Safety Information System (HSIS). Although the original data contained the crash counts from 537 intersections, only the intersections having 10 years of crash data history were retained (resulting in 451 intersections). There were 77 fatal injuries, 202 Sev2 accidents, 738 Sev3 accidents, 865 Sev4 accidents, and 2,857 PDO accidents at those 451 intersections for 10 years. Table 1 contains summary statistics of the variables of interest, where the unit of crash frequency is the number of crashes per intersection for 10 years. The major and minor roads of the intersection are defined as a function of the entering traffic flow. The legs with the highest entering flows are defined as major annual average daily traffic (AADT).

Table 2 gives the estimates (posterior means and standard deviations) of the regression coefficients **β** based on an MVPLN model implemented by the MCMC method with the MATLAB (*24*) codes developed specifically for this research. The functional form used for the models is described in Equations 1 through 3. The dependent variable is defined as the number of crashes per 10 years. To ensure that the chain has converged to the posterior distribution by the end of the burn-in period, trace plots and the autocorrelation function plots of posterior sample values were inspected, although those plots are not presented here because of space limitations.

For comparison purposes, Table 2 reports the estimates obtained by applying the univariate Poisson regression model and the univariate negative binomial regression model implemented in SAS (*27*) as well. For an objective comparison, the prior distributions of the parameters and the starting values in MCMC implementation were obtained independently of the SAS results. Here, vague priors not requiring much prior knowledge of the parameters are used to illustrate that the suggested MVPLN models can be applied even without precise prior knowledge. When good prior information on the parameters exists, however, it can be incorporated by the use of more informative (precise) prior distribution, and it may further improve the precision of the MVPLN models. A good discussion on elicitation of priors in crash data analysis can be found in a paper by Schluter et al. (*28*), for example. Finally, all the variables described

TABLE 1 Summary Statistics of Variables for California Intersection Data

Variable Name	Mean	SD	Min.	Max.
Dependent variables				
Sev1	0.1707	0.5204	0	5
Sev2	0.4479	0.9609	0	6
Sev3	1.6364	2.5159	0	20
Sev4	1.9180	3.5571	0	28
Sev5	6.3348	9.9493	0	88
Independent variables				
Lighting (1 = yes)	0.3525	0.4783	0	1
Painted left turn (1 = yes)	0.3925	0.4888	0	1
Curb med left turn (1 = yes)	0.1330	0.3340	0	1
Right turn channel (1 = yes)	0.1397	0.3470	0	1
ML lanes (no. of main lanes)	3.6851	0.7292	2	4
Mountain terrain (1 = yes)	0.1397	0.3470	0	1
Rolling terrain (1 = yes)	0.3570	0.4796	0	1
Logmaj (logarithm of major AADT)	9.4195	0.7514	7.7956	11.2683
Logmin (logarithm of minor AADT)	4.9193	1.5148	2.3026	10.0481

TABLE 2 Estimates of Regression Coefficients Obtained by Applying MVPLN Model, Univariate Poisson Regression Model, and Univariate Negative Binomial Regression Model

Severity	Variable	Multivariate Poisson–Lognormal Model	Univariate Poisson Regression Model	Univariate Negative Binomial Regression Model
Sev1	Constant	−13.0261 (1.6854)	−15.2279 (2.0375)	−14.9638 (2.2200)
	Lighting	−0.5544 (0.3229)	−0.5955 (0.3204)	−0.5704 (0.3470)
	Painted left turn	0.5349 (0.2859)	0.5032 (0.2886)	0.5158 (0.3138)
	Curb med. left turn	0.4994 (0.3534)	0.6221 (0.3446)	0.6228 (0.3884)
	Right turn channel	0.2777 (0.3156)	0.3752 (0.2870)	0.2991 (0.3356)
	ML lanes	0.2934 (0.2764)	0.2815 (0.3045)	0.2714 (0.3152)
	Mountain	−0.1367 (0.3720)	−0.3431 (0.3764)	−0.1864 (0.4232)
	Rolling	−0.3916 (0.2733)	−0.5641 (0.2689)	−0.5400 (0.3005)
	Logmaj ADT	0.8818 (0.1698)	1.1537 (0.1894)	1.1188 (0.2088)
	Logmin ADT	0.2069 (0.0873)	0.2052 (0.0810)	0.2223 (0.0921)
				Dispersion: 0.7059
	Pearson chi-square/DF		1.2232	1.0667
Sev2	Constant	−12.5689 (1.2596)	−13.2302 (1.1873)	−13.4023 (1.4116)
	Lighting	0.2345 (0.1993)	0.2997 (0.1733)	0.2844 (0.2072)
	Painted left turn	0.5569 (0.2031)	0.4796 (0.1706)	0.5572 (0.2023)
	Curb med. left turn	0.1780 (0.2856)	0.2229 (0.2431)	0.2290 (0.2882)
	Right turn channel	0.2285 (0.2379)	0.3425 (0.1821)	0.2686 (0.2408)
	ML lanes	0.1625 (0.1737)	0.1571 (0.1563)	0.1490 (0.1703)
	Mountain	0.3866 (0.2667)	0.3106 (0.2294)	0.4187 (0.2762)
	Rolling	0.4564 (0.1918)	0.4112 (0.1611)	0.4710 (0.1910)
	Logmaj	0.9097 (0.1336)	1.0435 (0.1186)	1.0548 (0.1422)
	Logmin	0.2331 (0.0612)	0.1899 (0.0492)	0.1952 (0.0619)
				Dispersion: 0.6070
	Pearson chi-square/DF		1.2699	1.0042
Sev3	Constant	−9.8505 (0.8479)	−9.9059 (0.5815)	−10.1854 (0.8482)
	Lighting	0.2081 (0.1360)	0.2315 (0.0907)	0.2025 (0.1321)
	Painted left turn	0.1088 (0.1388)	0.0648 (0.0844)	0.1206 (0.1271)
	Curb med. left turn	0.0560 (0.1875)	0.0780 (0.1188)	0.0896 (0.1811)
	Right turn channel	0.0793 (0.1619)	0.2511 (0.1002)	0.0499 (0.1655)
	ML lanes	0.0417 (0.0995)	0.0404 (0.0692)	0.0491 (0.0911)
	Mountain	0.4458 (0.1650)	0.3636 (0.1074)	0.5708 (0.1691)
	Rolling	0.0734 (0.1329)	0.0447 (0.0846)	0.0885 (0.1257)
	Logmaj	0.8936 (0.0907)	0.9463 (0.0604)	0.9645 (0.0881)
	Logmin	0.1789 (0.0419)	0.1608 (0.0262)	0.1670 (0.0393)
				Dispersion: 0.6048
	Pearson chi-square/DF		2.0799	1.0555
Sev4	Constant	−11.9536 (0.8721)	−12.4660 (0.5726)	−11.4316 (0.8863)
	Lighting	0.5212 (0.1409)	0.5264 (0.0845)	0.5422 (0.1394)
	Painted left turn	0.0119 (0.1485)	−0.0357 (0.0774)	0.0169 (0.1354)
	Curb med. left turn	−0.1958 (0.1990)	−0.1487 (0.1172)	−0.1396 (0.1984)
	Right turn channel	0.2490 (0.1789)	0.2908 (0.0917)	0.3392 (0.1743)
	ML lanes	0.0134 (0.1007)	0.0140 (0.0649)	0.0093 (0.0966)
	Mountain	0.4015 (0.1790)	0.3253 (0.1007)	0.4683 (0.1837)
	Rolling	0.0518 (0.1451)	0.0569 (0.0787)	0.0536 (0.1353)
	Logmaj	1.0857 (0.0926)	1.2034 (0.0593)	1.0921 (0.0938)
	Logmin	0.2317 (0.0442)	0.1982 (0.0240)	0.1997 (0.0417)
				Dispersion: 0.8015
	Pearson chi-square/DF		2.8881	1.2074
Sev5	Constant	−9.9596 (0.6670)	−10.1806 (0.3065)	−9.6546 (0.6358)
	Lighting	0.4203 (0.1051)	0.3544 (0.0465)	0.4881 (0.1049)
	Painted left turn	−0.2159 (0.1127)	−0.2326 (0.0420)	−0.2327 (0.1027)
	Curb med. left turn	−0.1494 (0.1482)	−0.1836 (0.0611)	−0.2024 (0.1471)
	Right turn channel	0.0715 (0.1263)	0.1864 (0.0525)	0.1016 (0.1311)
	ML lanes	0.1257 (0.0723)	0.1041 (0.0373)	0.1423 (0.0692)
	Mountain	0.5337 (0.1347)	0.5352 (0.0533)	0.5966 (0.1376)
	Rolling	0.1260 (0.1046)	0.1403 (0.0437)	0.0699 (0.1004)
	Logmaj	0.9777 (0.0717)	1.0593 (0.0315)	0.9829 (0.0676)
	Logmin	0.2493 (0.0333)	0.2193 (0.0132)	0.2291 (0.0321)
				Dispersion: 0.6225
	Pearson chi-square/DF		5.4932	1.1823

NOTES: 1. Multivariate Poisson–lognormal model was implemented by MCMC coded in MATLAB (*24*).
2. Univariate Poisson regression and univariate negative binomial regression were implemented in SAS (*27*).
3. Numbers in parentheses represent uncertainty estimates; posterior standard deviations under multivariate Poisson–lognormal model and standard errors under univariate Poisson regression model and univariate negative binomidal regression model, respectively.
4. Significant (at $\alpha = 0.05$) effects are shown in bold.

in Table 1 were included in the models to facilitate the comparison between the multivariate and univariate models.

In Table 2 it can be observed that for Sev1 and Sev2 all three models give similar results in terms of the estimated model coefficients and their significance except for the variable "rolling" of Sev1 (which was significant only under the univariate Poisson regression model). For Sev3 to Sev5, however, univariate Poisson regression models give significantly different results (in terms of both point estimates and uncertainty estimates) from those of the MVPLN models or univariate negative binomial regression models. For Sev3 to Sev5, it appears that under the univariate Poisson regression model the standard errors are seriously underestimated and as a result many of the covariates are incorrectly declared to be significant. The values of Pearson's chi-square divided by degrees of freedom for univariate Poisson regression models are considerably greater than 1 for Sev3 to Sev5, which indicates an apparent overdispersion problem. It is well known that the more overdispersion there is, the more seriously standard errors are underestimated, in which case those standard errors are not correct estimates of true uncertainties, and the corresponding interval estimates will not be able to capture the true parameter values. This problem cannot be overcome with MVP regression models either because overdispersion is not accounted for by those models. For the unbiased estimates, the small standard errors (more precise estimates) lead to more accurate parameter estimates only when they are not underestimated.

MVPLN models and univariate negative binomial models give, in general, consistent results in terms of significance of model coefficients. For Sev1, however, the uncertainty estimates from the MVPLN model are noticeably smaller than those from the univariate negative binomial model. Unlike univariate Poisson regression models, both MVPLN models and univariate negative binomial regression models are able to account for overdispersion, and their standard errors can serve as good estimates of true uncertainties. This finding supports the fact that by accounting for correlation in multivariate crash frequency by severity, an MVPLN model leads to more precise parameter estimates than a univariate negative binomial model does.

With regard to interpretation of the models' output, Table 2 shows that the coefficients sometimes change from positive to negative values, and vice versa, for different crash severity levels (e.g., "painted left-turn" bay). This characteristic can indicate that some variables may have a different effect given the severity outcome of the crash. However, some coefficients appear to be counterintuitive. For example, the presence of lighting is associated with more crashes for Sev4 and Sev5 (for all three models). Given the limited number of observations available in this study, it is possible that confounding factors, such as the location where lighting is used, may explain this outcome. A speed-related factor such as the speed limit could have been an influencing factor on severity. It is regrettable that the speed limit data were not available for this analysis, which might also have caused confounding of the "lighting" variable's effect.

Although the effect of confounding factors such as speed limit may also exist for other severity levels, the effect could be different for different severity levels unless the correlations among different severity levels are very high. It is expected that Sev5 crash counts would be more closely correlated with Sev4 crash counts than with Sev1 crash counts because Sev4 and Sev5 are more similar in nature. This similarity explains why the effect of "lighting" is significant only for Sev4 and Sev5. Obviously, these outcomes need to be further investigated. Although the limitations of this database are recognized, the purpose of this study is to present a general methodology that can account for correlations in the multivariate crash counts and to illustrate the method on real crash counts for different severity types.

TABLE 3 Posterior Means of Covariance Matrix (Σ) of Latent Effects

	Sev1	Sev2	Sev3	Sev4	Sev5
Sev1	0.6592	0.4884	0.4743	0.5487	0.4638
Sev2	0.4884	0.7408	0.5251	0.6054	0.4998
Sev3	0.4743	0.5251	0.7224	0.6595	0.5424
Sev4	0.5487	0.6054	0.6595	0.9357	0.6760
Sev5	0.4638	0.4998	0.5424	0.6760	0.6651

Tables 3 and 4 contain the MCMC estimates of the covariance matrix and correlation matrix of the latent effects (generating the correlation structure in the multivariate crash counts) of the MVPLN model, respectively.

MVP regression models suggested by other researchers, such as Ma and Kockelman (*16*), are very restrictive in the sense that they assume the covariances for different severity levels to be identical and nonnegative as well as lacking in overdispersion. However, the new MVPLN regression models that can be implemented by the MCMC method allow for a fully general correlation structure as well as overdispersion in the crash data. From Tables 3 and 4, it can be observed that there is a positive correlation between each of the latent effects in the crash counts of the five severity types, but the correlations for the different severity levels are not identical. Thus, as with any statistical models, these correlations need to be incorporated into the estimation of the model.

There are a few important avenues for further work. First, the predicted values obtained from the MVPLN model can be compared with the values estimated from a univariate negative binomial regression model estimated for all crash severities combined (e.g., KABCO) with the output of an ordered-logit crash severity model, as proposed by Miaou et al. (*29*). Traditional tools and the ones proposed by Oh et al. (*10*) could be used for this comparison analysis. Second, the stability of the MVPLN models subjected to low sample mean values and small sample size should be investigated along with a sensitivity analysis for different hyperprior specifications; crash data often exhibit these two unique properties. As noted by Lord (*30*) and by Lord and Miranda-Moreno (*31*), statistical models have the potential to become unstable when they are estimated with this kind of data. Third, the changes in the signs of the coefficients between different crash severity levels need to be investigated further. It also would be desirable to investigate if there is any confounding between the variable "lighting" and the variables that are not included in the data that causes the derivation of counterintuitive signs for the coefficient of the former variable for Sev4 and Sev5 crashes.

TABLE 4 Posterior Means of Correlation Matrix of Latent Effects

	Sev1	Sev2	Sev3	Sev4	Sev5
Sev1	1.0000				
Sev2	0.7035	1.0000			
Sev3	0.6904	0.7203	1.0000		
Sev4	0.7030	0.7297	0.8035	1.0000	
Sev5	0.7043	0.7152	0.7834	0.8575	1.0000

SUMMARY AND CONCLUSIONS

A new multivariate approach for modeling data on crash counts by severity is presented based on MVPLN models employing the MCMC algorithm as a computational engine. The method was applied to the multivariate crash counts from 451 intersections in California obtained for 10 years. It turns out that not only are there correlations across severity levels but also the correlations are not identical. Neither the univariate modeling approach nor the previously suggested MVP regression approach (*16*) would have revealed these findings. Overdispersion in the data was apparent, which again cannot be handled by an MVP approach.

The new MVPLN regression approach can cope with both overdispersion and a fully general correlation structure in the data. It also was observed that for fatal crashes (Sev1) the uncertainty estimates from the MVPLN model are noticeably smaller than those from the univariate negative binomial model, which suggests that by accounting for correlation in the multivariate crash counts, an MVPLN model leads to more precise parameter estimates than a univariate negative binomial model does.

ACKNOWLEDGMENTS

The authors thank Srinivas Geedipally and Craig Lyon for assisting in collecting and assembling the HSIS data used in this study. The work presented was carried out initially as part of NCHRP Project 17-29, and the authors thank NCHRP and TRB for funding the study.

REFERENCES

1. Abbess, C., D. Jarett, and C. C. Wright. Accidents at Blackspots: Estimating the Effectiveness of Remedial Treatment, with Special Reference to the "Regression-to-Mean" Effect. *Traffic Engineering and Control,* Vol. 22, No. 10, 1981, pp. 535–542.
2. Hauer, E., J. C. N. Ng, and J. Lovell. Estimation of Safety at Signalized Intersections. In *Transportation Research Record 1185,* TRB, National Research Council, Washington, D.C., 1988, pp. 48–61.
3. Persaud, B. N., and L. Dzbik. Accident Prediction Models for Freeways. In *Transportation Research Record 1401,* TRB, National Research Council, Washington, D.C., 1993, pp. 55–60.
4. Kulmala, R. *Safety at Rural Three- and Four-Arm Junctions: Development and Applications of Accident Prediction Models.* VTT Publications 233. Technical Research Centre of Finland, Espoo, 1995.
5. Poch, M., and F. L. Mannering. Negative Binomial Analysis of Intersection-Accident Frequencies. *Journal of Transportation Engineering,* ASCE, Vol. 122, No. 2, 1996, pp. 105–113.
6. Lord, D. *The Prediction of Accidents on Digital Networks: Characteristics and Issues Related to the Application of Accident Prediction Models.* PhD dissertation. Department of Civil Engineering, University of Toronto, Ontario, Canada, 2000.
7. Ivan, J. N., C. Wang, and N. R. Bernardo. Explaining Two-Lane Highway Crash Rates Using Land Use and Hourly Exposure. *Accident Analysis and Prevention,* Vol. 32, No. 6, 2000, pp. 787–795.
8. Lyon, C., J. Oh, B. N. Persaud, S. P. Washington, and J. Bared. Empirical Investigation of the IHSDM Accident Prediction Algorithm for Rural Intersections. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1840,* Transportation Research Board of the National Academies, Washington, D.C., 2003, pp. 78–86.
9. Miaou, S.-P., and D. Lord. Modeling Traffic Crash-Flow Relationships for Intersections: Dispersion Parameter, Functional Form, and Bayes Versus Empirical Bayes. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1840,* Transportation Research Board of the National Academies, Washington, D.C., 2003, pp. 31–40.
10. Oh, J., C. Lyon, S. P. Washington, B. N. Persaud, and J. Bared. Validation of the FHWA Crash Models for Rural Intersections: Lessons Learned. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1840,* Transportation Research Board of the National Academies, Washington, D.C., 2003, pp. 41–49.
11. Lord, D., A. Manar, and A. Vizioli. Modeling Crash–Flow–Density and Crash–Flow–V/C Ratio for Rural and Urban Freeway Segments. *Accident Analysis and Prevention,* Vol. 37, No. 1, 2005, pp. 185–199.
12. Miaou, S.-P., and J. J. Song. Bayesian Ranking of Sites for Engineering Safety Improvements: Decision Parameter, Treatability Concept, Statistical Criterion and Spatial Dependence. *Accident Analysis and Prevention,* Vol. 37, No. 4, 2005, pp. 699–720.
13. Tunaru, R. Hierarchical Bayesian Models for Multiple Count Data. *Austrian Journal of Statistics,* Vol. 31, Nos. 2 and 3, 2002, pp. 221–229.
14. Bijleveld, F. D. The Covariance Between the Number of Accidents and the Number of Victims in Multivariate Analysis of Accident Related Outcomes. *Accident Analysis and Prevention,* Vol. 37, No. 4, 2005, pp. 591–600.
15. Song, J. J., M. Ghosh, S. Miaou, and B. Mallick. Bayesian Multivariate Spatial Models for Roadway Traffic Crash Mapping. *Journal of Multivariate Analysis,* Vol. 97, 2006, pp. 246–273.
16. Ma, J., and K. M. Kockelman. Bayesian Multivariate Poisson Regression for Models of Injury Count by Severity. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1950,* Transportation Research Board of the National Academies, Washington, D.C., 2006, pp. 24–34.
17. Tsionas, E. G. Bayesian Multivariate Poisson Regression. *Communications in Statistics—Theory and Methods,* Vol. 30, No. 2, 2001, pp. 243–255.
18. Karlis, D., and L. Meligkotsidou. Multivariate Poisson Regression with Covariance Structure. *Statistics and Computing,* Vol. 15, 2005, pp. 255–265.
19. Chib, S., and R. Winkelmann. Markov Chain Monte Carlo Analysis of Correlated Count Data. *Journal of Business and Economic Statistics,* Vol. 19, 2001, pp. 428–435.
20. Winkelman, R. *Econometric Analysis of Count Data,* 4th ed. Springer, New York, 2003.
21. Tierney, L. Markov Chains for Exploring Posterior Distributions. *Annals of Statistics,* Vol. 22, No. 4, 1994, pp. 1701–1762.
22. Gilks, W. R., S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice.* Chapman & Hall, London, 1996.
23. Liu, J. S. *Monte Carlo Strategies in Scientific Computing.* Springer, New York, 2001.
24. *MATLAB Neural Network Toolbox 5.* MathWorks, Inc., Natick, Mass., 2006.
25. Anderson, T. W. *An Introduction to Multivariate Statistical Analysis,* 2nd ed. Wiley, New York, 1984.
26. Chib, S., and E. Greenberg. Understanding the Metropolis-Hastings Algorithm. *American Statistician,* Vol. 49, No. 4, 1995, pp. 327–335.
27. *SAS: Version 9 of the SAS System for Windows.* SAS Institute Inc., Cary, N.C., 2002.
28. Schluter, P. J., J. J. Deely, and A. J. Nicholson. Ranking and Selecting Motor Vehicle Accident Sites by Using a Hierarchical Bayesian Model. *The Statistician,* Vol. 46, No. 3, 1997, pp. 293–316.
29. Miaou, S.-P., R. P. Bligh, and D. Lord. Developing Median Barrier Installation Guidelines: A Benefit/Cost Analysis Using Texas Data. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1904,* Transportation Research Board of the National Academies, Washington, D.C., 2005, pp. 3–19.
30. Lord, D. Modeling Motor Vehicle Crashes Using Poisson-Gamma Models: Examining the Effects of Low Sample Mean Values and Small Sample Size on the Estimation of the Fixed Dispersion Parameter. *Accident Analysis & Prevention,* Vol. 38, No. 4, 2006, pp. 751–766.
31. Lord, D., and L. F. Miranda-Moreno. Effects of Low Sample Mean Values and Small Sample Size on Estimation of Fixed Dispersion Parameter of Poisson-Gamma Models for Modeling Motor Vehicle Crashes: Bayesian Perspective. Presented at 86th Annual Meeting of the Transportation Research Board, Washington, D.C., 2006.