A multivariate random-parameters Tobit model for analyzing highway crash rates by injury severity

Qiang Zeng[a, *], Huiying Wen[a], Helai Huang[b], Xin Pei[c], S.C. Wong[d]

[a] School of Civil Engineering and Transportation, South China University of Technology, Guangzhou, Guangdong 510641, PR China

[b] Urban Transport Research Center, School of Traffic and Transportation Engineering, Central South University, Changsha, Hunan 410075, PR China

[c] Department of Automation, Tsinghua University, Beijing, PR China

[d] Department of Civil Engineering, The University of Hong Kong, Pokfulam Road, Hong Kong

ABSTRACT

In this study, a multivariate random-parameters Tobit model is proposed for the analysis of crash rates by injury severity. In the model, both correlation across injury severity and unobserved heterogeneity across road-segment observations are accommodated. The proposed model is compared with a multivariate (fixed-parameters) Tobit model in the Bayesian context, by using a crash dataset collected from the Traffic Information System of Hong Kong. The dataset contains crash, road geometric and traffic information on 224 directional road segments for a five-year period (2002-2006). The multivariate random-parameters Tobit model provides a much better fit than its fixed-parameters counterpart, according to the deviance information criteria and Bayesian R^2, while it reveals a higher correlation between crash rates at different severity levels. The parameter estimates show that a few risk factors (bus stop, lane changing opportunity and lane width) have heterogeneous effects on crash-injury-severity rates. For the other factors, the variances of their random parameters are insignificant at the 95 % credibility level, then the random parameters are set to be fixed across observations. Nevertheless, most of these fixed coefficients are estimated with higher precisions (i.e., smaller variances) in the random-parameters model. Thus, the random-parameters Tobit model, which provides a more comprehensive understanding of the factors' effects on crash rates by injury severity, is superior to the multivariate Tobit model and should be considered a good alternative for traffic safety analysis.

Keywords: Crash rate by severity; Random parameters; Multivariate Tobit model; Bayesian inference.

* Corresponding author
E-mail address: zengqiang@scut.edu.cn (Q. Zeng), hywen@scut.edu.cn (H. Wen), huanghelai@csu.edu.cn (H. Huang), peixin@mail.tsinghua.edu.cn (X. Pei), hhecwsc@hku.hk (S.C. Wong)

1. Introduction

In the field of traffic safety analysis, considerable research effort has been devoted to the application of innovative methodological approaches to model crash frequency. Because crash frequencies are non-negative integers, most of the advanced approaches are based on the basic statistical count model, Poisson regression, while addressing certain issues related to crash-frequency data (e.g., over-dispersion, under-dispersion, excess zero observations, spatiotemporal correlation, multilevel structure, and unobserved heterogeneity). Lord and Mannering (2010) and Mannering and Bhat (2014) presented overviews of these models.

Recently, as good alternatives to the traditional crash-frequency approaches, methods for analyzing crash rates (such as the number of crashes per 100 million vehicle miles traveled) have been developed in a number of studies (Anastasopoulos et al., 2008; Caliendo el al., 2015). Neutralizing the effects of crash exposure, the crash rates forms a standardized measure of the relative safety performance of roadway sites (e.g., roadway segments and intersections), which is more acceptable for the public than the crash frequency. Meanwhile, the crash rate is able to clearly reflect the risk of involving in a crash and thus may be a more effective criterion for the identification of hotspots (Ma et al., 2015b; Xu et al., 2014). Moreover, crash rates are commonly used in accident reporting systems, such as the annual crash reports of the National Highway Traffic Safety Administration (NHTSA, 2012). Therefore, the crash rate may sometimes be preferable to the crash frequency.

From the perspective of statistical modeling, crash rates are continuous, non-negative numbers, which differ substantially from (discrete integers) crash frequencies. Crash-rate data are usually left-censored at zero because no crashes may be observed at several sites during certain periods. Censoring refers to a limitation on data clustering which may result in a lower threshold (left-censored), an upper threshold (right-censored), or both. The censoring phenomenon in crash rates may appear for two distinct reasons, including a lack of crashes at the sites over the observation period and a failure to report crashes that occur (Anastasopoulos et al., 2012a). Generally, more severe crashes are more likely to be reported (Lord and Mannering, 2010). To deal with the censoring characteristic, Anastasopoulos et al. (2008) first introduced the Tobit model to analyze crash rates. Various forms of random-parameter Tobit models have since been proposed to account for unobserved heterogeneity across observations (Anastasopoulos et al., 2012a; Caliendo el al., 2015; Ma et al., 2015a; Yu et al., 2015). The Tobit-based (crash rate prediction) models avoid some shortcomings associated with Poisson-based (crash frequency prediction) models, such as the requirement of equal mean and variance of Poisson distribution and the hypothesized zero state in the zero-inflated count models (Lord et al., 2005, 2007), which may be a potentially technical advantage of crash rate modeling over crash frequency modeling.

Compared with analysis of the total crash rate, modeling the crash rate by injury severity can illustrate the effects of observed risk factors (such as the traffic, geometric, and environmental characteristics of sites) on the crash rate with a particular injury-severity outcome. The expected crash rates at each level of severity provide deeper insight into the safety situation of a certain road site. Therefore, although the overall crash rate may not reveal a site deficiency, overexposure of a specific crash severity may uncover otherwise undetected deficiencies. Moreover, models for the analysis of crash-injury-severity rates may be more appealing for ranking road sites that hold promise for safety improvement, because injury severity and the associated costs are the primary concerns in many programs (Miaou and Song, 2005). However, only two existing studies (Anastasopoulos et al., 2012b; Xu et al., 2014) have focused on jointly modeling crash rates by injury severity. Their model estimation results support the notion that a significant correlation exists between crash rates at various severity levels, which may be a result of common unobserved factors that affect crash rates across injury severity. Similarly, significant correlations among crash frequencies by injury severity or crash type have already been demonstrated in many multivariate count models (Aguero-Valverde and Jovanis, 2009; Barua et al., 2014; Bijleveld, 2005; El-Basyouny and Sayed, 2009; El-Basyouny et al., 2014; Ma and Kockelman, 2006; Ma et al., 2008; Park and Lord, 2007). Nonetheless, the fixed parameters in all these studies omit the potential heterogeneity in the effects of risk factors across observations, which have been found in the previous research using random-parameters models to analyze total crash rate (Anastasopoulos et al., 2012a), total crash frequency (Anastasopoulos and Mannering, 2009) and crash frequencies by severity (Barua et al., 2016). Ignoring the possible heterogeneity across observations and constraining the parameters to be constant, such fixed parameters models would lead to biased parameters and incorrect inferences (Washington et al., 2011). As noted by Anastasopoulos et al. (2012b), the inclusion of random parameters in the multivariate Tobit model would be able to capture the unobserved heterogeneity.

To this end, the main objective of this study is to develop a multivariate random-parameters Tobit model for the simultaneous analysis of crash rates by injury severity that accounts for both correlations between crash rates at different severity levels and the variations in the effects of risk factors across observations. To demonstrate the proposed model, it is compared to a multivariate (fix-parameters) Tobit model in the Bayesian context using crash-injury-severity-rate data on road segments in Hong Kong over a 5-year period. Accordingly, the remainder of this paper is organized as follows. The next section specifies the proposed models and criteria for model comparison. The collected data for model demonstration are described in Section 3. Section 4 introduces the detailed estimation of the proposed models and discusses the parameter estimation results. Finally, conclusions and recommendations for future research are presented in Section 5.

2. Methods

In this section, the formulations of the two candidate models for the simultaneous analysis of crash-injury-severity rates, multivariate Tobit and multivariate random-parameters Tobit regressions, are first specified explicitly under the Bayesian framework. Two criteria in the context of Bayesian inference, the deviance information criteria (DIC) and Bayesian R^2, are then introduced for the purpose of model comparison.

2.1. Model specification

2.1.1. Multivariate Tobit model

As mentioned, crash rates are generally left-censored at zero, because crashes may not be reported at some sites during the study period. The Tobit regression, first proposed by James Tobin (1958), is an appropriate method for the analysis of censored data (Anastasopoulos et al., 2008). To accommodate the possible correlation between crash rates at various severity levels, a multivariate Tobit model was advocated by Anastasopoulos et al. (2012b). Using a left-censored threshold of zero, the multivariate Tobit regression for the joint modeling of the crash rate by injury severity is expressed as follows:

$$Y_{it}^{k*} = \beta_{k0} + \sum_{m=1}^{M} \beta_{km} x_{it}^{m} + \varepsilon_{it}^{k}, \quad (1)$$

$$Y_{it}^{k} = \begin{cases} Y_{it}^{k*}, & if\ Y_{it}^{k*} > 0 \\ 0, & if\ Y_{it}^{k*} \le 0 \end{cases}, \quad i = 1,2,\cdots,N,\ t = 1,2,\cdots,T,\ k = 1,2,\cdots,K, \quad (2)$$

where Y_{it}^{k*} and Y_{it}^{k} represent the unobservable and observed crash rate at injury severity level k and site i during period t, respectively. N, T and K are the number of the observed sites, periods and categorized injury severity levels respectively. $x_{it}^{1}, x_{it}^{2}, \cdots x_{it}^{M}$ are the observed values of M risk factors at site i during period t, while $\beta_{k1}, \beta_{k2}, \cdots \beta_{kM}$ are the estimable coefficients corresponding to injury severity level k. β_{k0} is the constant, and ε_{it}^{k} denotes the random error term which is assumed to follow a multi-normal distribution with zero mean, that is,

$$\boldsymbol{\varepsilon}_{it} \sim N_K(\mathbf{0}, \boldsymbol{\Sigma}), \quad \boldsymbol{\varepsilon}_{it} = \begin{pmatrix} \varepsilon_{it}^{1} \\ \varepsilon_{it}^{2} \\ \cdots \\ \varepsilon_{it}^{K} \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1K} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2K} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{K1} & \sigma_{K2} & \cdots & \sigma_{KK} \end{pmatrix}. \quad (3)$$

In Eq. (3), σ_{kk} ($k=1,2,\cdots K$) represents the variance of error term ε_{it}^{k} and σ_{k_1,k_2} ($k_1 \neq k_2$) denotes the covariance between $\varepsilon_{it}^{k_1}$ and $\varepsilon_{it}^{k_2}$.

2.1.2. Multivariate random-parameters Tobit model

Studies have demonstrated that heterogeneous effects of certain factors may be present across observations of the total crash rate (Anastasopoulos et al., 2012a; Caliendo el al., 2015; Ma el al., 2015a; Yu et al., 2015). A random-parameters Tobit model is the prevalent approach applied to deal with this issue. It is reasonable to speculate that this phenomenon may also exist in crash-injury-severity rate analysis. Therefore, to accommodate the underlying unobserved heterogeneity in the multivariate Tobit model, the coefficients ($\beta_{k0},\beta_{k1},\cdots\beta_{kM}$) in Eq. (1) are set to be random parameters ($\beta_{it}^{k0},\beta_{it}^{k1},\cdots\beta_{it}^{kM}$):

$$Y_{it}^{k*}=\beta_{it}^{k0}+\sum_{m=1}^{M}\beta_{it}^{km}x_{it}^{m}+\varepsilon_{it}^{k}. \quad (4)$$

Because $\beta_{it}^{1m},\beta_{it}^{2m},\cdots\beta_{it}^{Km}$ may be also correlated, suggested by Barua et al. (2016), they are assumed to be multi-normally distributed as:

$$\boldsymbol{\beta}_{it}^{m}\sim N_K(\boldsymbol{\beta}_m,\boldsymbol{\Sigma}_m),\quad \boldsymbol{\beta}_{it}^{m}=\begin{pmatrix}\beta_{it}^{1m}\\ \beta_{it}^{2m}\\ \cdots\\ \beta_{it}^{Km}\end{pmatrix},\quad \boldsymbol{\Sigma}_m=\begin{pmatrix}\sigma_{11}^{m} & \sigma_{12}^{m} & \cdots & \sigma_{1K}^{m}\\ \sigma_{21}^{m} & \sigma_{22}^{m} & \cdots & \sigma_{2K}^{m}\\ \cdots & \cdots & \cdots & \cdots\\ \sigma_{K1}^{m} & \sigma_{K2}^{m} & \cdots & \sigma_{KK}^{m}\end{pmatrix},\quad m=0,1\cdots,M, \quad (5)$$

in which $\boldsymbol{\beta}_m$ and $\boldsymbol{\Sigma}_m$ are the mean vector and variance-covariance matrix of $\boldsymbol{\beta}_{it}^{m}$ respectively. Noticeably, if the covariance of two random parameters is statistically insignificant (say, at the 95 % credibility level), they are assumed to follow independent normal distributions (Barua et al., 2016). If the variance of a certain random parameter is not statistically significant (say, at the 95 % credibility level), the random parameter is simplified to be fixed across observations (Anastasopoulos et al., 2012a).

2.2. Model comparison

As in many other studies that included modeling under the Bayesian framework (Dong et al., 2014; Huang et al., 2016a; Zeng and Huang, 2014), the DIC and Bayesian R^2 are used to compare the above candidate models.

The DIC is intended as a Bayesian generalization of Akaike's information criteria, which penalizes larger-parameter models. Specifically, it provides a Bayesian measure of model complexity and fitting and is defined as (Spiegelhalter et al., 2002):

$$DIC = \overline{D(\theta)} + pD, \tag{6}$$

where $\overline{D(\theta)}$ is the posterior mean deviance that can be taken as a Bayesian measure of fitting, and pD is a complexity measure for the effective number of parameters. Generally, models with lower DIC values are preferred. However, it is worth noting that determination of a critical difference in DIC is very difficult. According to Spiegelhalter et al. (2005), very roughly, more than 10 differences may rule out the model with the higher DIC; differences between 5 and 10 are considered substantial; and if the DIC difference is less than 5, and the model inferences are significantly different, it could be misleading to simply report the model with the lowest DIC.

The Bayesian R^2 measure, which could be viewed as a global model-fit measurement, is proposed to estimate the ratio of the explained sum of squares to the total sum of squares (Ahmed et al., 2011; Zeng and Huang, 2014). To evaluate the model fit comprehensively, the Bayesian R^2 values of crash rates at each injury severity k and all observations, represented by R_k^2 and R_T^2 respectively, are all calculated:

$$R_k^2 = 1 - \frac{\sum_{t=1}^{T}\sum_{i=1}^{N}\left(Y_{it}^k - \lambda_{it}^k\right)^2}{\sum_{t=1}^{T}\sum_{i=1}^{N}\left(Y_{it}^k - \bar{Y}_k\right)^2}, \quad k = 1, 2, \cdots, K, \tag{7}$$

$$R_T^2 = 1 - \frac{\sum_{k=1}^{K}\sum_{t=1}^{T}\sum_{i=1}^{N}\left(Y_{it}^k - \lambda_{it}^k\right)^2}{\sum_{k=1}^{K}\sum_{t=1}^{T}\sum_{i=1}^{N}\left(Y_{it}^k - \bar{Y}\right)^2}, \tag{8}$$

$$\lambda_{it}^k = \begin{cases} \beta_{k0} + \sum_{m=1}^{M}\beta_{km}x_{it}^m, & if\ \beta_{k0} + \sum_{m=1}^{M}\beta_{km}x_{it}^m > 0 \\ 0, & if\ \beta_{k0} + \sum_{m=1}^{M}\beta_{km}x_{it}^m \le 0 \end{cases}, \tag{9}$$

$$\bar{Y}_k = \frac{1}{T \times N}\sum_{t=1}^{T}\sum_{i=1}^{N} Y_{it}^k, \quad k = 1, 2, \cdots, K, \tag{10}$$

$$\bar{Y} = \frac{1}{K \times T \times N}\sum_{k=1}^{K}\sum_{t=1}^{T}\sum_{i=1}^{N} Y_{it}^k. \tag{11}$$

In the above equations, λ_{it}^k is the expected crash rates at injury severity level k and site i during period t. $\bar{Y}_k$ and $\bar{Y}$ are the mean of crash rates at injury severity k and all observations, respectively.

3. Data preparation and preliminary analysis

A crash dataset obtained from the Traffic Information System (TIS) maintained by the Transport Department of Hong Kong is used to demonstrate the proposed multivariate random-parameters Tobit model and to compare it with the multivariate Tobit model. This dataset has already been employed in the previous studies on crash frequency prediction (Huang et al., 2016b; Zeng et al., 2016). Therefore, their estimation results could be referred to in this research. It contains 112 road segments with two end points defined by the Hong Kong Annual Traffic Census (ATC) system (as shown in Fig. 1). Their traffic volumes are continuously measured by the 112 core stations of the ATC system. The directional average annual daily traffic (AADT) for each road segment adjacent to the 112 core stations ($N = 224$) from 2002 to 2006 ($T = 5$) is derived from the ATC system for the analysis.

Geographical information system techniques are used to map crashes to these directional segments, based on the movement attributes of the vehicles involved and a detailed description of each crash recorded in the TIS system. With respect to crash severity, the TIS classifies crashes into three categories – fatal, serious injury, and slight injury – according to the severity of injury among the casualties. Due to the rareness of fatality, it is combined with serious injury to form the category of killed and seriously injured (KSI) crashes ($K = 2$). The road geometric and traffic information is also obtained from the TIS system.

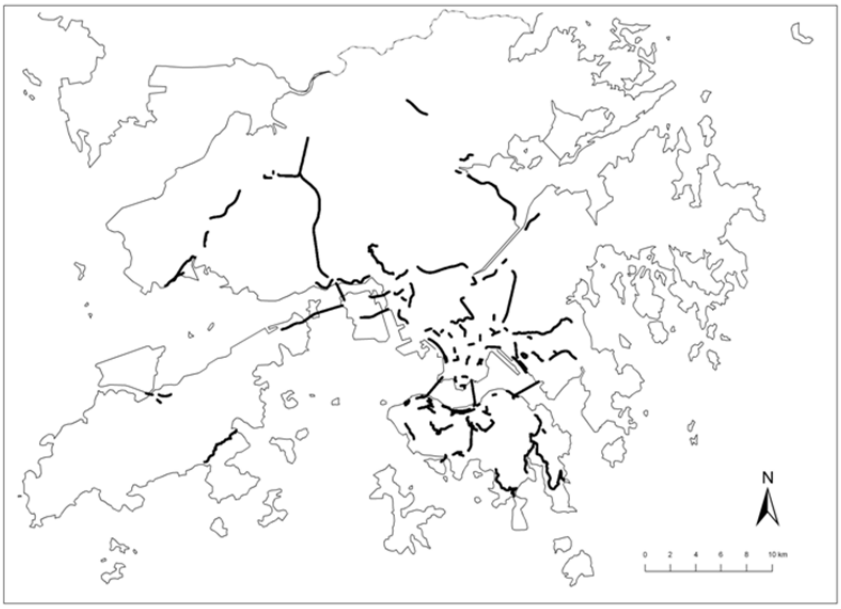

Fig. 1. Selected roadway segments in Hong Kong for the analysis.

The yearly crash rate (number of crashes per million vehicle-kilometers traveled) by injury severity, CR_{it}^{k}, which is used as the dependent variable in this study, is calculated as:

$$CR_{it}^{k} = \frac{No\_crash_{it}^{k}}{AADT_{i}^{t} \times L_{i} \times 365/1000,000}, \quad (14)$$
$$i = 1, 2, \cdots 224,\ k = 1, 2,\ t = 2002, 2003, 2004, 2005, 2006$$

in which $No\_crash_{it}^{k}$ is the number of crashes at injury severity degree k that occurred on road segment i in year t; $AADT_{i}^{t}$ is the AADT on road segment i in year t; and L_i is the length of segment i, ranging from 0.15 km to 9.07 km with mean 1.47 km. Among the total observations, 76 slight injury crash rates and 349 KSI crash rates are 0. Table 1 illustrates the definitions and descriptive statistics of the variables used in the development of the model. The results of the correlation tests and multi-collinearity diagnoses indicate that there is no significant correlation or collinearity among these factors.

Table 1. Descriptive statistics of the variables.

Variable	Description	Mean	SD	Min.	Max.
Response variable					
Slight	Slightly injured crash count per million vehicle-kilometers traveled	1.70	2.21	0	24.35
KSI	Killed and seriously injured crash count per million vehicle-kilometers traveled	0.46	0.98	0	9.86
Risk factors					
AADT	Average annual daily traffic (10^3 vehicles)	22.08	19.94	1.16	101.63
Width	Average width of each lane (m)	3.63	0.64	2.40	7.30
SL	Posted speed limit (km/h)	60.3	14.7	50	110
Merge	Number of merging ramps	0.84	1.00	0	4
Diverge	Number of diverging ramps	1.75	2.27	0	17
Inter	Number of intersections	1.90	2.37	0	16
Gradient	Average segment gradient (10^{-2})	0.04	2.74	-11	11
Curvature	Average segment curvature	21.9	17.5	0	85
LCO	Lane changing opportunity	2.43	1.61	0	7.85
Median	Presence of median barrier: yes = 1, no = 0	0.70	0.46	0	1
BS	Presence of bus stop: yes = 1, no = 0	0.64	0.48	0	1
Rainfall	Annual precipitation (m)	2.28	5.65	0.76	3.22

The lane changing opportunity (LCO) variable refers to the length-weighted average number of eligible opportunities to change lanes in a subsegment with identical lane markings. No lane changing is allowed in road sections with double continuous lines (as shown in Fig. 2(a)), thus $LCO = 0$. In sections with one continuous line and one broken line, lane changing is only allowed from the side of the broken line to the side of the continuous line (shown in Fig. 2(b)), thus $LCO = 1$. In sections with a single broken line, lane changing is allowed between both adjacent lanes (as shown in Fig. 2(c)), thus $LCO = 2$. Pei et al. (2012) provided a more detailed description of LCOs.

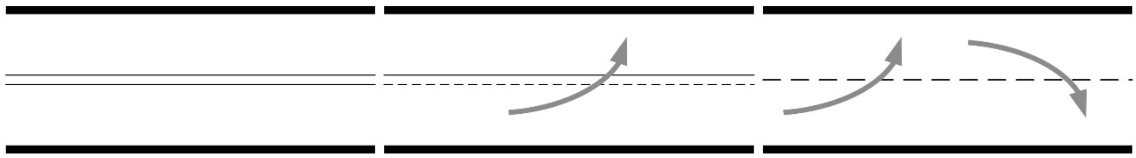

Fig. 2. Lane changing opportunities for different road section configurations.

4. Model estimation and result analysis

4.1. Model estimation

Without the requirement of the traditional maximum likelihood estimation for closed-form likelihood functions, Bayesian inference is able to handle very complex models (such as the random-parameters model in this study) (Lord and Mannering, 2010). Moreover, Freeware WinBUGS, which is a popular platform for Bayesian inference, can be used to construct a flexible programming environment. As a consequence, both of the candidate models are programmed, estimated, and evaluated in WinBUGS, which is much more easily implemented than other alternatives, such as the maximum simulated likelihood estimation and copula methods (Anastasopoulos et al., 2012a, b).

In the absence of sufficient knowledge, non-informative priors are specified for the parameters and the hyper-parameters. Specifically, a diffused normal distribution $N(0,10^4)$ is used as the priors of β_{km} and all of the elements of $\boldsymbol{\beta}_m$ $(m = 0,1,\cdots 12,\ k = 1,2)$, and a Wishart prior $W(\mathbf{P}, r)$ is used for $\boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\Sigma}_m^{-1}$, where $\mathbf{P} = \begin{bmatrix} 1, & 0 \\ 0, & 1 \end{bmatrix}$ represents the scale matrix and $r = 2$ is the degrees of freedom (El-Basyouny and Sayed, 2009; Zeng et al., 2016). For each model, a chain of 500,000 iterations of the Markov chain Monte Carlo (MCMC) simulation are made, with the first 4000 iterations acting as burn-ins. The Gelman-Rubin statistics available in WinBUGS is used to evaluate the MCMC convergence.

4.2. Model comparison

The results of DIC and a number of hyper-parameters for model comparison are summarized in Table 2. According to the results, it can be seen that the random effects of slight injury crash rates (σ_{11}) and KSI crash rates (σ_{22}), and their covariance σ_{21}/σ_{12} in the multivariate random-parameters Tobit model, are all lower than their respective counterparts in the multivariate Tobit model, possibly because a portion of the random effects derived from the unobserved heterogeneity across observations is accounted for by the random parameters. In contrast to the random effects, the correlation coefficient ρ ($=\sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$) is increased dramatically from 0.491 in the fixed-parameters model to 0.817 in the random-parameters model. It reveals that the crash rates at the two severity levels have a high positive correlation, which is reasonable as they are likely to rise due to the same deficiencies in roadway design and/or other unobserved factors (El-Basyouny and Sayed, 2009).

Table 2. Model comparison results.

	Multivariate Tobit			Multivariate random-parameters Tobit		
	Mean	SD	95 % Credible interval	Mean	SD	95 % Credible interval
σ_{11}	**3.442**	**0.154**	**(3.152, 3.753)**[a]	**1.387**	**0.329**	**(0.809, 1.983)**
$\sigma_{21}(=\sigma_{12})$	**0.823**	**0.059**	**(0.712, 0.943)**	**0.465**	**0.081**	**(0.307, 0.617)**
σ_{22}	**0.816**	**0.036**	**(0.747, 0.890)**	**0.238**	**0.041**	**(0.165, 0.323)**
ρ [b]	**0.491**	**0.024**	**(0.443, 0.538)**	**0.817**	**0.048**	**(0.711, 0.895)**
R_1^2	**0.305**	**0.005**	**(0.295, 0.314)**	**0.740**	**0.064**	**(0.616, 0.851)**
R_2^2	**0.158**	**0.006**	**(0.146, 0.168)**	**0.785**	**0.039**	**(0.703, 0.854)**
R_T^2	**0.356**	**0.003**	**(0.349, 0.360)**	**0.754**	**0.050**	**(0.659, 0.844)**
$\overline{D(\theta)}$	6559			3440		
pD	29			1388		
DIC	6588			4828		

[a] Boldface indicates statistical significance at the 95 % credibility level.

[b] $\rho=\sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$.

Moreover, the $\overline{D(\theta)}$ value of the multivariate random-parameters Tobit model (=3440) is much smaller than that of the multivariate Tobit model (=6559), which suggests that the random-parameters model fits the crash-rate data much better than the fixed-parameters model. It can be further confirmed by the Bayesian R^2 measure results that the values of R_1^2, R_2^2 and R_T^2 of the multivariate random-parameters Tobit model are all greater than those of the multivariate Tobit model. These results are consistent with those in the previous research which shows that accommodating unobserved heterogeneity could significantly improve model fit (Anastasopoulos and

Mannering, 2009; Anastasopoulos et al., 2012a). Although there are more effective parameters (as reflected by *pD*) in the multivariate random-parameters Tobit model, which increase the complexity, its much lower DIC indicates that it substantially outperforms the multivariate Tobit model.

4.3 Interpretation of parameter estimation

The parameter estimation results in the multivariate Tobit and multivariate random-parameters Tobit models are presented in Tables 3 and 4, respectively. From the results in Table 4, we can see that the random-parameters' standard deviations of three factors (bus stop, lane changing opportunity and lane width)[1] are significant at the 95 % credibility level for both slight injury and KSI crash rates. They demonstrate the heterogeneous effects of these risk factors on the slight injury and KSI crash rates.

Table 3. Parameter estimation in the multivariate Tobit model[a].

Variable	Slight injury			KSI		
	Mean	SD	95% Credible interval	Mean	SD	95% Credible interval
Constant	**3.656**	**0.574**	**(2.657, 4.725)**[b]	0.397	0.284	(-0.121, 0.920)
AADT	**-0.024**	**0.004**	**(-0.033, -0.015)**	**-0.010**	**0.002**	**(-0.017, -0.003)**
SL	**-0.042**	**0.005**	**(-0.053, -0.030)**	**-0.009**	**0.003**	**(-0.016, -0.002)**
Merge	**-0.234**	**0.069**	**(-0.370, -0.100)**	**-0.093**	**0.033**	**(-0.158, -0.027)**
Diverge	0.034	0.035	(-0.035, 0.104)	**0.070**	**0.017**	**(0.036, 0.104)**
Inter	**-0.076**	**0.034**	**(-0.149, -0.002)**	**-0.086**	**0.019**	**(-0.122, -0.050)**
Median	**-0.438**	**0.182**	**(-0.787, -0.080)**	-0.160	0.090	(-0.333, 0.017)
BS	**0.601**	**0.150**	**(0.306, 0.895)**	**0.327**	**0.072**	**(0.186, 0.469)**
Gradient	-3.072	2.125	(-7.266, 1.104)	**-2.136**	**1.036**	**(-4.162, -0.114)**
Curvature	**-0.018**	**0.004**	**(-0.027, -0.009)**	-0.004	0.002	(-0.011, 0.002)
LCO	**0.254**	**0.045**	**(0.165, 0.342)**	**0.052**	**0.022**	**(0.009, 0.095)**
Width	**0.256**	**0.097**	**(0.072, 0.444)**	**0.197**	**0.048**	**(0.108, 0.289)**

[a] Rainfall is excluded, because neither of its effects on crash rates at the two severity degrees is significant at the 95 % credibility level.

[b] Boldface indicates statistical significance at the 95 % credibility level.

Specifically, the presence of a bus stop results in normally distributed parameters, with means of 0.616 and 0.300 and standard deviations of 0.963 and 0.747 for slight injury and KSI crash rates, respectively. That is, the presence of a bus stop on most road segments (73.9 % and 65.6 %) increases the slight injury and KSI crash rates, probably because of the increased interaction between buses and other vehicles when they enter or leave bus bays (Pei et al., 2012; Zeng et al., 2016); however, for the minority of road segments (26.1 % and 34.4 %), the presence of a bus stop actually

[1] The covariance of the three pair of random parameters is not significant at the 95 % credibility level. Therefore, they are independently and normally distributed in the multivariate random-parameters Tobit model.

decreases the slight injury and KSI crash rates.

The effects of LCOs on the slight injury and KSI crash rates are found to follow two independent normal distributions, with means of 0.235 and 0.052 and standard deviations of 0.167 and 0.058, such that more LCOs would lead to a higher slight injury (KSI) crash rate for 92 % (81.4 %) of roadway segments but a lower slight injury (KSI) crash rate for the other 8 % (18.6 %). The general finding of LCOs (i.e., increasing slight injury and KSI crash rates) could be the result of increased vehicle interaction caused by lane changing maneuvers (such as overtaking), which may increase the incidence of traffic conflicts (Pei et al., 2012; Zeng et al., 2016).

Table 4. Parameter estimation in the multivariate random-parameters Tobit model[a].

Variable	Slight injury			KSI		
	Mean	SD	95 % Credible interval	Mean	SD	95 % Credible interval
Constant	**3.508**	**0.429**	**(2.945, 4.238)**[b]	**0.752**	**0.209**	**(0.401, 1.131)**
AADT	**-0.023**	**0.004**	**(-0.033, -0.014)**	**-0.008**	**0.002**	**(-0.015, -0.002)**
SL	**-0.040**	**0.005**	**(-0.051, -0.029)**	**-0.008**	**0.002**	**(-0.014, -0.001)**
Merge	**-0.205**	**0.065**	**(-0.331, -0.077)**	**-0.057**	**0.028**	**(-0.113, -0.001)**
Diverge	0.021	0.035	(-0.046, 0.090)	**0.049**	**0.016**	**(0.016, 0.081)**
Inter	**-0.073**	**0.037**	**(-0.141, -0.001)**	**-0.069**	**0.017**	**(-0.104, -0.035)**
Median	**-0.457**	**0.173**	**(-0.794, -0.147)**	**-0.194**	**0.080**	**(-0.350, -0.048)**
BS	**0.616**	**0.141**	**(0.341, 0.879)**	**0.300**	**0.063**	**(0.178, 0.423)**
SD of BS	**0.963**	**0.233**	**(0.515, 1.320)**	**0.747**	**0.055**	**(0.645, 0.852)**
Curvature	**-0.018**	**0.004**	**(-0.026, -0.009)**	-0.004	0.002	(-0.010, 0.002)
LCO	**0.235**	**0.045**	**(0.150, 0.318)**	**0.052**	**0.017**	**(0.020, 0.084)**
SD of LCO	**0.167**	**0.062**	**(0.070, 0.286)**	**0.058**	**0.018**	**(0.023, 0.090)**
Width	**0.320**	**0.097**	**(0.160, 0.486)**	**0.098**	**0.051**	**(0.009, 0.158)**
SD of Width	**0.256**	**0.039**	**(0.182, 0.321)**	**0.084**	**0.018**	**(0.046, 0.113)**

[a] Gradient and Rainfall are excluded, because neither of their effects on crash rates at the two severity degrees is significant at the 95 % credibility level.

[b] Boldface indicates statistical significance at the 95 % credibility level.

The mean and standard deviation of the random parameter of lane width for the slight injury crash rates are 0.320 and 0.256, respectively, whereas those for the KSI crash rates are 0.098 and 0.084, respectively. Given these distributional parameters with their 95 % credible intervals away from zero, widening the lanes in 89.4 % and 87.8 % of roadway segments would increase the slight injury and KSI crash rates, respectively, whereas widening the lanes of the remaining 10.6 % and 12.2 % would have the opposite effect. Gross and Jovanis (2007) found a U-shaped relationship between lane width and crash risk and speculated that drivers may respond to narrow lanes with more-cautious behavior, thereby decreasing the likelihood of a crash.

With respect to the coefficients of the other factors, the positive or negative signs are consistent and the magnitude is comparable in the two multivariate models.

Furthermore, the signs of each factor's coefficients for both slight injury and KSI crash rates are identical, which means that they have consistent effects on the crash rates for the two injury severity degrees. Nevertheless, the fixed parameters of Constant and Median are changed to be significant but the effect of Gradient on the KSI crash rates is changed to be insignificant at the 95 % credibility level in the multivariate random-parameters Tobit regression. In addition, the standard deviations of most coefficients are smaller in the random-parameters model, which indicates that it leads to more precise parameter estimation.

Specifically, the average annual daily traffic is found to have significant negative effects on both slight injury and KSI crash rates, which is consistent with the findings in many previous studies (Anastasopoulos et al., 2012a, b; Huang et al., 2016b; Qi et al., 2007). These findings may be attributed to the reduced travel speed caused by increasing traffic volume, which may significantly decrease the likelihood of (slight injury and SKI) crash occurrence. It is interesting to find that crash rates at the two injury severity levels are both lower on roadway segments with higher speed limits, which may contradict engineering intuition (Aguero-Valverde and Jovanis, 2008). However, some researchers have argued that road segments designed for higher speeds are usually well-planned, constructed, and managed (Milton and Mannering, 1998; Zeng et al., 2016) and that these features may improve highway safety performance.

The significantly negative coefficients of Merge and Inter indicate that a greater number of merging ramps and intersections on roadway segments are associated with lower slight injury and KSI crash rates. These results could be explained by the risk compensation theory, which suggests that drivers may adapt to an adverse driving environment (more merging ramps and intersections) by altering their driving behavior (such as being more careful or slowing down) (Mannering and Bhat, 2014). Some drivers may overcompensate for the adverse conditions, leading to a lower crash risk. Conversely, the positive coefficients of Diverge suggest that an increase in the number of diverging ramps results in higher slight injury and KSI crash rates, which conforms to engineering intuition and the findings of previous studies (Zeng et al., 2016), because the sites that approach diverging ramps can be hazardous.

The estimation results show that the presence of a median barrier on a road segment results in significantly lower slight injury and KSI crash rates. Studies have suggested that median barriers can effectively prevent cross-median crashes (Donnell and Mason, 2006; Zeng et al., 2016). Curvature is found to significantly decrease slight injury crash rates, which may be somewhat counterintuitive. However, stronger centrifugal force derived from greater road curvature tend to increase the injury severity once a crash happens, that is, it is less likely to be a slight injury crash. This may partially explain the significantly negative coefficient of Curvature on slight injury crash rates.

5. Conclusions and future research

This study advocates a multivariate random-parameters Tobit model for the simultaneous analysis of crash-injury-severity rates, which accommodates both correlation between crash rates at different severity levels and unobserved heterogeneity across observations. A crash dataset obtained from the Traffic Information System maintained by the Transport Department of Hong Kong, in which crashes are classified into slight injury and KSI degrees, is used to demonstrate the proposed model and to compare it with a multivariate fixed-parameters Tobit model. The models are estimated and evaluated in the Bayesian context via programming in the freeware WinBUGS.

The results show that bus stop, lance changing opportunity and lane width present heterogeneous effects on both slight injury and KSI crash rates. After accounting for the heterogeneity of these factors' effects, the random effects and covariance are all decreased to certain degrees while the correlation coefficient reaches 0.817, which suggests a high positive correlation between the crash rates at the two severity levels. The multivariate random-parameters Tobit model fits the collected crash data much better than the multivariate Tobit model, as reflected by a dramatic decrease in the DIC value (1760) and significant increases in the Bayesian R^2 values. What's more, compared with the fixed-parameters model, the random-parameters model produces consistent signs and more precise estimates for the fixed coefficients of the other factors. The average annual daily traffic, the speed limit, the number of merging ramps and intersections, and the presence of median barriers are found to have significant negative effects on slight injury and KSI crash rates, while the number of diverging ramps is found to have a significant positive association with KSI crash rates.

In summary, the empirical analysis demonstrates the superiority of the multivariate random-parameters Tobit model and the significance of correlation between crash rates at various severity degrees and the heterogeneous effects of certain risk factors in the crash-injury-severity-rate data. Noticeably, like other random-parameters models, the proposed multivariate random-parameters Tobit model may suffer from transferability issues since the individual parameter vector associated with each observation is unique. However, if significant random parameters are found in crash data, the fixed-parameters model will be estimated with a persistent bias and transferability will be problematic since this bias will be a function of unobserved heterogeneity (Mannering et al., 2016). Thus, we still think that the proposed model has considerable potential in analysis of crash rates by severity. Because significant spatial correlation always exists among crash rates of adjacent road sites, further research effort could be devoted to incorporate spatial correlation into the multivariate random-parameters Tobit model. It could also be applied to rank sites that hold promise for safety improvement.

Acknowledgements

This research was jointly supported by the Natural Science Foundation of China (Nos. 51378222, 51578247, 71371192, and 71301083), and a grant from the Joint Research Scheme of National Natural Science Foundation of China/Research Grants Council of Hong Kong (Nos. 71561167001 and N_HKU707/15).

References

Aguero-Valverde, J., Jovanis, P., 2008. Analysis of road crash frequency with spatial models. Transportation Research Record 2061, 55-63.

Aguero-Valverde, J., Jovanis, P., 2009. Bayesian multivariate Poisson lognormal models for crash severity modeling and site ranking. Transportation Research Record 2136, 82-91.

Ahmed, M., Huang, H., Abdel-Aty, M., Guevara, B., 2011. Exploring a Bayesian hierarchical approach for developing safety performance functions for a mountainous freeway. Accident Analysis and Prevention 43 (4), 1581-1589.

Anastasopoulos, P. C., Mannering, F. L., 2009. A note on modeling vehicle accident frequencies with random-parameters count models. Accident Analysis and Prevention 41, 153-159.

Anastasopoulos, P. C., Mannering, F. L., Shankar, V. N., Haddock, J. E., 2012a. A study of factors affecting highway accident rates using the random-parameters tobit model. Accident Analysis and Prevention 45, 628-633.

Anastasopoulos, P. C., Shankar, V. N., Haddock, J. E., Mannering, F. L., 2012b. A multivariate tobit analysis of highway accident-injury-severity rates. Accident Analysis and Prevention 45, 110-119.

Anastasopoulos, P. C., Tarko, A. P., Mannering, F. L., 2008. Tobit analysis of vehicle accident rates on interstate highways. Accident Analysis and Prevention 40 (2), 768-775.

Barua, S., El-Basyouny, K., Islam, M. T., 2014. A full Bayesian multivariate count data model of collision severity with spatial correlation. Analytic Methods in Accident Research 3, 28-43.

Barua, S., El-Basyouny, K., Islam, M. T., 2016. Multivariate random parameters collision count data models with spatial heterogeneity. Analytic Methods in Accident Research 9, 1-15.

Bijleveld, F. D. 2005. The covariance between the number of accidents and the number of victims in multivariate analysis of accident related outcomes. Accident Analysis and Prevention 37 (4), 591-600.

Caliendo, C., De Gugliemo, M. L., Guida, M., 2015. Comparison and analysis of road tunnel traffic accident frequencies and rates using random-parameter models. Journal of Transportation Safety and Security 8 (2), 177-195.

Dong N., Huang H., Xu P., Ding Z., Wang. D., 2014. Evaluating spatial proximity structures in TAZ-level crash prediction models. Transportation Research Record

2432, 46-52.

Donnell, E.T., Mason Jr, J.M., 2006. Predicting the frequency of median barrier crashes on Pennsylvania interstate highways. Accident Analysis and Prevention 38 (3), 590-599.

El-Basyouny, K., Sayed, T., 2009. Collision prediction models using multivariate Poisson-lognormal regression. Accident Analysis and Prevention 41(4), 820-828.

El-Basyouny, K., Barua, S., Islam, M. T., 2014. Investigation of time and weather effects on crash types using full Bayesian multivariate Poisson lognormal models. Accident Analysis and Prevention 73, 91-99.

Huang, H., Song, B., Xu, P., Zeng, Q., Lee, J., Abdel-Aty, M., 2016a. Macro and micro models for zonal crash prediction with application in hot zones identification. Journal of Transport Geography 54, 248-256.

Huang, H., Zeng, Q., Pei, X., Wong, S.C., Xu, P., 2016b. Predicting crash frequency using an optimized radial basis function neural network model. Transportmetrica A, 12, 330-345.

Gross, F., Jovanis, P. P., 2007. Estimation of the safety effectiveness of lane and shoulder width: Case-control approach. Journal of Transportation Engineering 133 (6), 362-369.

Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. Transportation Research Part A 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.

Lord, D., Washington, S., Ivan, J. N., 2007. Further notes on the application of zero-inflated models in highway safety. Accident Analysis and Prevention 39, 53-57.

Ma, J., Kockelman, K.M., 2006. Bayesian multivariate Poisson regression for models of injury count, by severity. Transportation Research Record 1950, 24-34.

Ma, J., Kockelman, K.M., Damien, P., 2008. A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods. Accident Analysis and Prevention 40 (3), 964-975.

Ma, X., Chen, F., Chen, S., 2015a. Modeling crash rates for a mountainous highway by using refined-scale panel data. Transportation Research Record 2515, 10-16.

Ma, L., Yan, X., Weng, J., 2015b. Modeling traffic crash rates of road segments through a lognormal hurdle framework with flexible scale parameter. Journal of Advanced Transportation 49 (8), 928-940.

Mannering, F.L., Bhat, C.R., 2014. Analytic methods in accident research: methodological frontier and future directions. Analytic Methods in Accident Research, 1, 1-22.

Mannering, F. L., Shankar, V., Bhat, C. R., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. Analytic methods in accident research, 11, 1-16.

Miaou, S. P., Song, J. J., 2005. Bayesian ranking of sites for engineering safety improvements: decision parameter, treatability concept, statistical criterion, and spatial dependence. Accident Analysis and Prevention 37(4), 699-720.

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

NHTSA, 2012. 2010 motor vehicle crashes: Overview.

Park, E.S., Lord, D., 2007. Multivariate Poisson-lognormal models for jointly modeling crash frequency by severity. Transportation Research Record 2019, 1-6.

Pei, X., Wong, S.C., Sze, N.N., 2012. The roles of exposure and speed in road safety analysis. Accident Analysis and Prevention 48, 464-471.

Qi, Y., Smith, B. L., Guo, J., 2007. Freeway accident likelihood prediction using a panel data analysis approach. Journal of transportation engineering 133 (3), 149-156.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., Van Der Linde, A., 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society 64 (4), 583-639.

Spiegelhalter, D., Thomas, A., Best, N., Lunn, D., 2005. WinBUGS user manual. MRC Biostatistics Unit, Cambridge, United Kingdom.

Tobin, J., 1958. Estimation of relationships for limited dependent variables. Econometrica 26, 24-36.

Washington, S.P., Karlaftis, M.G., Mannering, F.L., 2011. Statistical and econometric methods for transportation data analysis, 2nd edition. CRC press, New York.

Xu, X., Wong, S.C., Choi, K., 2014. A two-stage bivariate logistic-Tobit model for the safety analysis of signalized intersections. Analytic Methods in Accident Research 3-4, 1-10.

Yu, R., Xiong, Y., Abdel-Aty, M., 2015. A correlated random parameter approach to investigate the effects of weather conditions on crash risk for a mountainous freeway. Transportation Research Part C 50, 68-77.

Zeng, Q., Huang, H., 2014. Bayesian spatial joint modeling of traffic crashes on an urban road network. Accident Analysis and Prevention 67, 105-112.

Zeng, Q., Huang, H., Pei, X., Wong S.C., 2016. Modeling nonlinear relationship between crash frequency by severity and contributing factors by neural networks. Analytic Methods in Accident Research 10, 12-25.