A Two-stage Bivariate Logistic-Tobit Model for the Safety Analysis of Signalized Intersections

Xuecai Xu[a,b,*], S.C. Wong[b], Keechoo Choi[c]

[a]School of Civil Engineering and Mechanics, Huazhong University of Science and Technology, Wuhan, China

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

[c]Department of Transportation Engineering, TOD-based Sustainable Urban Transportation Center, Ajou University, Korea

Abstract

Crash frequency and crash severity models have explored the factors that influence intersection safety. However, most of these models address the frequency and severity independently, and miss the correlations between crash frequency models at different crash severity levels. We develop a two-stage bivariate logistic-Tobit model of the crash severity and crash risk at different severity levels. The first stage uses a binary logistic model to determine the overall crash severity level. The second stage develops a bivariate Tobit model to simultaneously evaluate the risk of a crash resulting in a slight injury and the risk of a crash resulting in a kill or serious injury (KSI). The model uses 420 observations from 262 signalized intersections in the Hong Kong metropolitan area, integrated with information on the traffic flow, geometric road design, road environment, traffic control and any crashes that occurred during 2002 and 2003. The results obtained from the first-stage binary logistic model indicate that the overall crash severity level is significantly influenced by the annual average daily traffic and number of pedestrian crossings. The results obtained from the second-stage bivariate Tobit model indicate that the factor that significantly influences the numbers of both slight injury and KSI crashes is the proportion of commercial vehicles. The existence of four or more approaches, the reciprocal of the average turning radius and the presence of a turning pocket increase the likelihood of slight injury crashes. The average lane width and cycle time affect the likelihood of KSI crashes. A comparison with existing approaches suggests that the bivariate logistic-Tobit model provides a good statistical fit and offers an effective alternative method for evaluating the safety performance at signalized intersections.

Keywords: Signalized Intersection; Crash Severity; Crash Risk; Bivariate Logistic-Tobit Model

1. Introduction

A number of different approaches and perspectives have been used in crash prediction modeling. Lord and Mannering (2010) provided a comprehensive review of the different methodological approaches to crash frequency modeling, such as Poisson, negative-binomial, Poisson-lognormal, zero-inflated count, Conway-Maxwell-Poisson, gamma, generalized estimating equations, generalized additive, random-effects, negative-multinomial, random-parameter count, finite-mixture and Markov-switching models, and other intelligent algorithms. Savolainen et al. (2011) described modeling crash injury severity using artificial neural networks, Bayesian hierarchical binomial logit, Bayesian-ordered logistic, bivariate binary/ordered logistic, classification and regression tree, generalized ordered logit, Markov-switching multinomial logit, mixed generalized ordered logit, multivariate logit/logistic, nested logit and ordered logit/logistic models.

Models have been developed to assess intersection safety at signalized intersections in terms of either the crash frequency or crash severity. Wong et al. (2007), on whose research this study is based, used Poisson and negative-binomial regression models to quantify the influence of factors contributing to the incidence of slight injury crashes and the incidence of crashes resulting in a kill or serious injury (KSI) in Hong Kong. Liu (2007) generated a back-propagation neural network model using crash records from 62 signalized intersections in Taiwan and the characteristics of those intersections. The results indicated that the effects of the variables on the number of intersection-related crashes varied between intersections, leading to the proposal of a decision-making scheme to prevent erroneous investments. Ye et al. (2009) focused on collision types at rural intersections in Georgia and explored the crash frequency using multivariate Poisson models structured by simultaneous equations. This approach provided new insights into the crash frequency, although the effects of the risk factors that it addressed were found to be modest. Obeng (2011) analyzed the crash severity at signalized intersections using separate ordered logit models for females and males to investigate gender differences. The results indicated that the effects of driving conditions, type of crash, type of vehicle driven and vehicle-safety features on the risk of severe injury varied according to the gender of the driver. Haque et al. (2010) constructed Bayesian hierarchical models to examine motorcycle crashes at four-legged and “T” signalized intersections and found that the significant risk factors differed at the two intersections. In a similar study, Xie et al. (2013) used Bayesian hierarchical negative binomial models to evaluate the safety of signalized intersections in Shanghai at the intersection and corridor levels.

Several studies have investigated the heterogeneity of signalized intersections. For instance, Karlaftis and Tarko (1998) used negative binomial models and cluster analysis to explore the relationships between the crash frequency and possible influencing factors. However, the approach to accidents was quite general and the specific case of signalized intersections was not addressed. Chin and Quddus (2003) used the random-effects negative binomial model to identify the elements affecting intersection safety in Singapore, attempting to address the heterogeneity problem. Wang and Abdel-Aty (2006) and Wang et al. (2007) investigated the rear-end crash frequency at signalized intersections using the generalized estimating equations approach to account for temporal or spatial correlations within the dataset, which required that the same correlation matrix be used for different corridors. Guo et al. (2010) integrated the Poisson and negative binomial models with a Bayesian approach to evaluate the intersection safety with reference to corridor-level spatial correlations between 170

signalized intersections in Florida. The results indicated that the Poisson spatial model provided the best model fit and that its performance was related to the proximity function. However, the performance of alternative functions, such as the exponential function, should still be investigated. More recently, Castro et al. (2012) reformulated count modeling as a special case of generalized ordered-response modeling to address intersections. They presented a flexible count model, one accommodating temporal effects and the other accommodating both temporal and spatial effects. These models addressed temporal and spatial correlations and provided a fairly generalized method of crash analysis, which can be developed to accommodate more specific cases in future research.

However, these studies concentrated on either the crash frequency or crash severity at signalized intersections, and the possible correlations between their model estimates at different crash severity levels were not considered, which may have led to bias in the estimates (Lord and Mannering, 2010). Other studies have dealt simultaneously with the crash frequency and severity, using such methods as multi-level hierarchical structures (Kim et al., 2007), simultaneous equations (Kim and Washington, 2006) and multivariate analysis (Ma and Kockelman, 2006). These approaches either integrated crash frequency and crash severity models or involved a two-stage model.

For instance, Abdel-Aty and Keller (2005) explored overall crash severity levels using an ordered logistic model and specific crash severity levels using a hierarchical tree-based regression model. Their results showed that the aggregation of crash types was a less effective method than the development of separate models for each level of collision, an insight that has informed the design of this study. However, it should be noted that the two models presented were kept relatively separate, with no interaction permitted. Pei et al. (2011, 2012) developed a joint-probability model to integrate crash occurrence prediction and crash severity prediction within a single framework and used the Markov-chain Monte Carlo approach to establish a full Bayesian estimate of the effects of the explanatory factors. The results indicated that the proposed model was appropriate for signalized intersections and roadway safety, but only the binary approach to crash severity was provided as an illustrative example. El-Basyouny and Sayed (2011) used a multivariate Poisson-lognormal intervention model for the analysis of crash counts by severity level, and extended the model to incorporate random parameters to account for the correlation between sites. Chiou and Fu (2013) addressed the crash frequency and severity simultaneously in an integrated model with a multinomial generalized Poisson structure. The proposed covariance structure was shown to enhance the model's performance.

Wang et al. (2011) used the less-common two-stage model approach to model the crash frequency at different severity levels. They proposed a two-stage mixed multivariate model and showed how disaggregated data at the level of individual accident could be used to predict a certain type of low-frequency accident. Bhat et al. (2014) formulated a count outcome model with multinomial probit selection that accommodates unobserved heterogeneity and endogeneity issues at intersections. Their results showed that the model can be used for intersection crash analysis.

Crashes do not occur at every roadway segment and intersection during a particular observation period. Crash or crash-rate data can therefore be considered left-censored at zero, in accordance with the requirements of the Tobit model, which has been used by previous scholars to address the issue of safety. For example, Obeng and Burkey (2006) used a Tobit model to measure the property damage costs resulting from crashes at signalized intersections.

They found that driver characteristics, type of vehicle, vehicle speed, presence of a median barrier, amber light time and type of crash all had a significant influence on the property damage costs arising from crashes at signalized intersections. Anastasopoulos et al. (2008) explored the use of Tobit regression to address the censoring problem, offering new insights into the factors that significantly influenced the accident rates on interstate highways. Anastasopoulos et al. (2012b) then used a multivariate Tobit regression model to investigate accident-injury severity rates and found that the multivariate Tobit model helped with the analysis of the factors that determine the accident-injury severity on roadway segments.

However, Anastasopoulos et al. (2012b) estimated parameters that were fixed across observations and did not consider the possibility that the unobserved heterogeneity may be present across observations (which, if present, would suggest a random-parameters modeling approach).. El-Basyouny and Sayed (2009) found that the effects of covariates on accident frequency varied significantly across corridors, and that a Poisson-lognormal model with random parameters for each corridor improved the goodness of fit and accounted for the heterogeneity issues among different corridors. Anastasopoulos and Mannering (2009) and Anastasopoulos et al. (2012a) found that the random parameters approach permitted some or all of the parameters to vary randomly across observations. Anastasopoulos and Mannering (2011) explored fixed and random parameter logit models using crash-specific and non-crash-specific injury data. Their results verified that the models based on individual crash-data provided a better overall fit, and that random parameter models using less detailed data can still provide a reasonable level of accuracy. Likewise Venkataraman et al. (2013) verified that a random parameters negative binomial model showed significant improvement compared with a fixed parameter negative binomial model in relation to severity, the number of vehicles involved and the collision and location type. Chen and Tarko (2014) used random parameters and random effects models to investigate traffic safety in highway work zones. Their results showed that the marginal effects on crash frequency from the random effects model were similar to those from the random parameter model, and that the negative binomial model with random effects is a useful programming tool for police enforcement in highway work zones. Russo et al. (2014) used a random parameters bivariate ordered probit model to consider the fault status and examine the factors that affect the degree of injury sustained by drivers involved in angled collisions. They investigated concerns relating to within-crash correlation and the heterogeneity issues, and the results showed that the random parameters bivariate ordered probit models provided significant flexibility, allowing a more careful assessment of the effects of the influencing factors.

The aim of this paper is to develop a two-stage bivariate logistic-Tobit model capable of simultaneously modeling the crash severity and crash frequency at different severity levels. The model accommodates possible correlations (i.e., shared unobserved factors) between signalized intersections and deals with the left-censored issue (a predominance of zero or low crashes) at signalized intersections. An illustrative example composed of crash data from signalized intersections in Hong Kong is used to evaluate the suitability of the proposed model.

2. Data Description

This study uses 420 observations, of which 133 are of zero crashes, from 262 signalized intersections in the Hong Kong metropolitan area, with particular reference to Hong Kong Island, Kowloon and the New Territories, to evaluate the safety performance (see Wong et al.,

2007).

The crash dataset is obtained from the Traffic Accident Database System (TRADS) maintained by the Hong Kong Transport Department and the Hong Kong Police Force. TRADS categorizes crashes as of slight, serious and fatal severity. As there are few fatal crashes and both serious and fatal crashes lead to very serious damage, we consider crashes that result in death and in serious injury as a single category, KSI. Therefore, during the first stage of the proposed model, the crash severity is evaluated using a binary logistic model for slight injury and KSI levels. In the second stage, bivariate Tobit models are developed that correspond to the two severity levels: one for slight injury crashes and the other for KSI crashes.

Traffic volume significantly influences crash occurrences, and various studies have demonstrated a non-linear relationship between the crash incidence and exposure (Wong et al., 2007). The annual average daily traffic (AADT) is therefore quantified using a logarithmic transformation and is expected to reveal the proportionality of the relationship between the crash risk and traffic volume.

Data on other risk factors, such as the geometric road design, traffic characteristics, road environment and signal phasing, are collected from traffic impact assessment reports made in 2002 and 2003. These reports were produced for planning and design purposes. The safety performance and crash records of the intersections have not previously been investigated. Our sampling process should therefore not show a marked bias. We used the number of approaches, number of approach lanes, number of conflict points, number of turning movements required, average lane width, reciprocal of the turning radius, proportion of commercial vehicles, number of signal phases, signal cycle time, number of pedestrian crossings, presence of tram stops and light rail transit (LRT) stops and presence of turning pockets as the variables. Wong et al. (2007) provided a detailed description of these variables. Descriptive statistics for the selected signalized intersections are given in Table 1.

Table 1 Descriptive Statistics for the Selected Signalized Intersections

Variable	Description	Mean	Std. dev.	Min.	Max.
Dependent variables					
Logit	0=slight injury, 1=KSI	0.20	0.40	0	1
Srisk	Slight injury crash risk	0.49	0.43	0	3.03
Krisk	KSI crash risk	0.12	0.16	0	0.99
Exposure					
LnAADT	Ln(AADT)	10.36	0.77	6.81	11.42
Numeral variables					
nolanes	Number of approach lanes	9.05	3.41	2	16
noconflict	Number of conflict points	8.84	8.37	0	30
notrnstream	Number of turning movements required	6.40	2.59	2	12
lanewidth	Average lane width (m)	3.27	0.27	2.7	4.6
reciprad	Reciprocal of the turning radius	0.09	0.03	0	0.13
Traffic characteristics					

comveh	Proportion of commercial vehicles	0.23	0.10	0.01	0.66
Signal-phasing scheme					
nostages	Number of signal stages	3.14	0.70	2	5
cycletime	Cycle time	100.51	17.05	60	130
pedcrossing	Number of pedestrian crossings	4.26	2.20	0	8
Indicator variables					
Geometrical characteristics					
2 Appr.	**Two approaches (Yes=1, No=0)**	0.01		0	1
3 Appr.	**Three approaches (Yes=1, No=0)**	0.25		0	1
4 Appr.	**Four or more approaches (Yes=1, No=0)**	0.74		0	1
tramstop	Presence of tram stops (Yes=1, No=0)	0.08		0	1
lrtstop	Presence of LRT stops (Yes=1, No=0)	0.01		0	1
Road environment					
HKI	Hong Kong Island (Yes=1, No=0)	0.19		0	1
KLN	Kowloon (Yes=1, No=0)	0.66		0	1
Signal-phasing scheme					
turningpock	Presence of a turning pocket (Yes=1, No=0)	0.07		0	1

Number of observations=420.

3. Methodology

In this section, a two-stage bivariate logistic-Tobit model is developed. The crash severity level and crash risk are first addressed sequentially using a binary logistic model, and then simultaneously using a bivariate Tobit model that enables the simultaneous investigation of the slight injury and KSI risks.

The rationale for this model design lies in the prospective sequential nature and jointness of the crash severity levels and overall crash data. Crucially, although the crash severity is a continuum spanning different levels, we identify the two discrete categories of slight injury and KSI. The use of binary categories allows for an initial estimate of the crash severity using the binary logistic method, which can be followed by the generation of censored models for both slight injury and KSI crashes. By estimating the bivariate logistic-Tobit model, we can address the heterogeneity and left-censored issues at signalized intersections.

We first specify the two-stage bivariate logistic-Tobit model. The crash severity is regarded as a binary variable and is expected to be determined by the intersection characteristics and other influencing factors. The binary logistic stage of the model is thus expressed as follows:

$$Z_i = \log it[p_i] = \log\left[\frac{p_i}{1-p_i}\right] = \beta_0 + \beta X_i, \quad (1)$$

where p_i represents the crash severity probability at the signalized intersection i, $\log it[p_i]$ is the log of the odds ratio or likelihood ratio that the dependent variable is 1, X_i is a vector of the influencing variables, β is a vector of the regression coefficients and β_0 is the constant of the equation.

The probability of an observed class y_i with a features vector X_i is p_i if y_i =1 or $1-p_i$ if y_i =0. The likelihood function is described as:

$$L(\beta_0, \beta) = \prod_{i=1}^{n} p_i^{y_i} (1-p_i^{1-y_i}), \tag{2}$$

where *n* is the number of observations. Further details of the model can be found in Washington et al. (2011).

The Tobit model, first presented by James Tobin in 1958, was originally developed to explain the range of dependent variables in regression models censored at either a lower threshold (left-censored), an upper threshold (right-censored) or both. Truncated data only provide non-limited values, and censored data also provide limited data information (Anastasopoulos et al., 2008). The data on crash rates can be considered left-censored at zero, as not all of the signalized intersections experienced a crash during the observation period.

In the second stage of the proposed model, a bivariate Tobit approach is used to simultaneously evaluate the slight injury and KSI risks, thus explaining the heterogeneity of the unobserved factors affecting safety at signalized intersections. The way that the Tobit model is used is because the slight injury and KSI risks are continuous dependent variables. The slight injury risk and KSI risk are respectively defined as the numbers of slight injury crashes and KSI crashes per year divided by the annual exposure. The annual exposure is calculated by multiplying the AADT by 365. The crash rate is preferred over the count, because the crash rate can neutralize the effect of "exposure," which can be used to identify hazardous locations more effectively. Crash frequency models often need to convert the results into crash risk by setting the exposure as an offset variable. Therefore, the crash rate is more directly useable by traffic agencies to reflect safety risk. These risks are dependent on the crash severity level and the relevant influencing factors. The structural equations can be expressed as follows:

$$\begin{cases} Y_i^{S^*} = \beta^S X_i + \gamma^S Z_i + \varepsilon_i^S \\ Y_i^S = \begin{cases} Y_i^{S^*} if Y_i^{S^*} > 0 \\ 0 if Y_i^{S^*} \le 0 \end{cases} \\ Y_i^{K^*} = \beta^K X_i + \gamma^K Z_i + \varepsilon_i^K \\ Y_i^K = \begin{cases} Y_i^{K^*} if Y_i^{K^*} > 0 \\ 0 if Y_i^{K^*} \le 0 \end{cases} \end{cases}, \tag{3}$$

where $Y_i^{S^*}$ and $Y_i^{K^*}$ represent the unobservable slight injury risk and KSI risk at the signalized intersection *i*, respectively, and Y_i^S and Y_i^K indicate the slight injury risk and KSI risk at the signalized intersection *i*, respectively. X_i is the vector of the variables derived from the characteristics of the signalized intersections and other influencing factors; β^S and β^K are the vectors of the estimable parameters associated with the characteristics of the signalized intersections; γ^S and γ^K are the estimable parameters associated with the severity level; ε_i^S and ε_i^K are the random error terms; and Z_i is the predicted value from the first stage. The Tobit model's likelihood function for zero observations (0) and positive observations (1) can be expressed as follows:

$$L = \prod_{0}[1-\Phi(\beta X/\sigma)]\prod_{1}\sigma^{-1}\phi[(Y_i-\beta X/\sigma)] \quad , \tag{4}$$

where L is the likelihood estimate; $X=\{X_i, \forall_i\}$; $Y_i=\{Y_i^S, Y_i^K\}$; $\beta=\{\beta^S, \beta^K, \gamma^S, \gamma^K\}$; σ is the standard deviation of the normally distributed latent variable Y_i^*, where $Y_i^*=\{Y_i^{S^*}, Y_i^{K^*}\}$; Φ is the standard normal-distribution function; and ϕ is the standard normal-density function. More details of the bivariate Tobit model can be found in Chen and Zhou (2011).

The model described above can be considered an extension of a bivariate model combining the binary logistic and bivariate Tobit approaches. The method used to estimate the models is similar to the two-stage least-squares approach used for simultaneous equation models. The predicted values for the crash severity level are obtained from the binary logistic model and included as regressors in the respective bivariate Tobit components. The bivariate Tobit models incorporate the predicted crash severity level and all of the influencing factors as regressors to predict the slight injury and KSI risks. The results obtained from the bivariate Tobit model are used to simultaneously estimate the number of slight injury and KSI crashes.

The accuracy of the two-stage model is determined with the statistical methods used to evaluate goodness of fit. Information criteria such as Akaike's information criterion (AIC) and the Bayesian information criterion (BIC) are applied, in which the effects of the number of parameters and sample size are considered. The lower the value of the AIC and BIC, the better the statistical fit of the model. The AIC and BIC can be estimated by:

$$\text{AIC} = -2\log(L) + 2m, \tag{5}$$

$$\text{BIC} = -2\log(L) + m\log n, \tag{6}$$

where L is the likelihood of the data given the proposed model, m is the number of parameters and n is the number of observations.

The two-stage model has received less attention in the transportation literature. Greene (2003) developed a two-stage model similar to ours using an ordered logistic approach as its first stage and a linear regression as its second stage. Although Greene's second stage consisted of only one linear regression, it offers possible future applications, especially within economics (see, for example, Bellemare and Barrett, 2006). It can also easily be used in conjunction with a statistical package for maximum-likelihood estimation.

4. Results and Discussion

Although data for a number of independent variables are collected, only those variables that are significant are included in the final model. All of the predictor variables must be verified as statistically independent with no co-linearity before the final model is estimated. STATA 11.0 (StataCorp LP, 2009) is used to conduct the relevant analysis and estimates.

We avoid correlations between the variables by conducting a correlation test to identify the variables to be included in the model. As in Wong et al. (2007), the number of turning movements, number of approach lanes, number of conflict points and number of signal stages are highly correlated with one another. Therefore, these variables are not included in the model at the same time.

In the first stage of the proposed model, the classical maximum likelihood method is used to

estimate the parameters. The crash severity results obtained from the binary logistic model are listed in Table 2 and show that the overall crash severity is significantly influenced by the AADT and the number of pedestrian crossings.

Table 2 Results of the First Stage of the Bivariate Logit-Tobit Model

Variables	Estimated coefficient	Std. Err.	Z-statistic
LnAADT	0.58*	0.16	3.55
Pedcrossing	0.13*	0.06	2.04
Const	4.08*	1.59	2.56
Goodness-of-fit assessment			
Number of observations	420		
Log likelihood at zero	-211.55		
Log likelihood at convergence	-204.74		
LR chi-square	13.62		

* 5% level of significance.

It has been found that logit models can have misspecification issues if the effects of the parameters that vary across the observations are not allowed for (Enberg, 1990; Yau and Ma, 1999; Chen and Kuo, 2001; Malchow-Moller and Svarer, 2003; Wang and Tsodikov, 2010; Roy 2012). For traffic safety problems, although random-effects and binary logit models have been investigated separately (Mannering and Bhat, 2014), their applications were mainly for injury severity analyses (Haleem and Gan, 2013; Pai et al., 2013; Yu and Abdel-Aty, 2014). There have been few applications of the random parameter binary logit model to crash counts. Nevertheless, to avoid the possible misspecification issue, the random parameters were introduced into the binary logit model in the first stage and found the results to be very similar to those of the binary logit model without random parameters. After the predicted value was incorporated into the second stage, the goodness-of-fit values are comparable to those of the proposed model. Therefore, the incorporation of random parameters does not seriously affect the results for this dataset. However, it would be desirable to adopt the random parameters model for other datasets when heterogeneity is a concern, which could be a useful future study.

As shown in Table 2, the LnAADT is positively related to the overall severity level, implying that a higher AADT increases the likelihood of the overall crash severity. This is consistent with the findings of Poch and Mannering (1996), Chin and Quddus (2003) and Wong et al. (2007). For every one-unit increase in the LnAADT, we expect a 0.58 increase in the log-odds of the overall crash severity, provided that the other variables are kept constant.

The number of pedestrian crossings is positively related to the overall crash severity at signalized intersections. This is probably because such crossings increase the likelihood of conflict between vehicles and pedestrians, particularly at times when the number of pedestrians using the crossings is elevated. Vehicle-pedestrian crashes often lead to severe injuries, as pedestrians are more vulnerable in a crash situation. Increasing the number of pedestrian crossings by one point causes the overall crash severity level to rise by 0.13 in the log-odds, provided that the other variables remain constant.

In the second stage of the proposed model, with the error distribution under the joint normality assumption, the classical maximum likelihood method provides an asymptotically efficient estimator for the regression parameters (Chen and Zhou, 2011). The results of the bivariate Tobit models for slight injury and KSI crashes are shown in Table 3.

From Table 3 it is clear that the proportion of commercial vehicles is another significant factor that influences both the slight injury and KSI crashes in the bivariate Tobit model. An increase in the probability of the crash severity level of one point decreases the risk of a slight injury crash by 3.08 points and increases the risk of a KSI crash by 1.09 points, provided that the other variables are held constant. Therefore, if the crash severity probability increases, the risk of a slight injury crash decreases, whereas that of a KSI crash increases.

Table 3 Results of the Second Stage of the Bivariate Logit-Tobit Model

Variables	Coefficient	Std. Err.	Z-statistic
Slight injury model			
Prob_sev	-3.08*	0.25	-12.59
4Appr.	0.48*	0.04	11.50
Reciprad	3.64*	0.67	5.47
Comveh	-0.50*	0.17	-2.92
Turningpock	0.19*	0.07	2.82
Cons	-0.69*	0.09	-7.64
KSI model			
Prob_sev	1.09*	0.11	9.98
Lanewidth	-0.02*	0.01	-6.05
Comveh	0.33*	0.08	4.10
Cycletime	0.001*	0.001	2.80
Cons	-0.14*	0.05	-2.64
Goodness-of-fit assessment			
Sigma1	0.34		
Sigma2	0.18		
Rho	0.26		
Number of observations	420		
Log likelihood at zero	-484.17		
Log likelihood at convergence	-107.77		
Chi-square		290.52	
Degrees of freedom	20		
AIC	38.45		
BIC	95.01		
MAD	0.324		
MAPE	10.494		
RMSE	0.602		

Note: prob_sev is the predicted probability of the crash severity level from the first stage of the model; * indicates a 5% level of significance; the coefficients of all of the variables, sigma 1 and sigma 2, are scaled up to a million vehicles; the mean absolute deviation (MAD) = $\frac{1}{n}\sum_{i=1}^{n}\left|Y_i-\hat{Y}_i\right|$, mean absolute percentage error (MAPE) = $\frac{100}{n}\sum_{i=1}^{n}\left|Y_i-\hat{Y}_i\right|/\left|Y_i\right|$ and root mean square error (RMSE) = $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i-\hat{Y})^2}$ where Y_i is the observed value, $\hat{Y}_i$ is the predicted value and n is the number of observations.

The slight injury risk is negatively influenced by the existence of four or more approaches. Larger signalized intersections may therefore decrease the likelihood of crashes. Small intersections may generate more crashes in the Hong Kong environment due to lower design standards. If one more approach is provided, the slight injury risk reduces by 0.48.

The third significant factor in the slight injury model is the reciprocal of the average turning radius, which is positively correlated with slight injury crashes. Fewer lanes and a smaller lane width hinder the maneuvers of drivers turning at a signalized intersection, increasing the likelihood of conflict and collisions. This is especially significant for large trucks and

double-decker buses. A one unit increase in the reciprocal of the average turning radius increases the risk of a slight injury crash by 3.64, provided that the other variables remain constant.

The slight injury crash risk at signalized intersections is negatively sensitive to the proportion of commercial vehicles (mainly heavy trucks and buses): increasing the proportion of these vehicles decreases the risk of a slight injury crash. Conversely, the KSI risk is positively sensitive to the proportion of commercial vehicles, as increasing the proportion of these vehicles increases the risk of a KSI crash. Collisions with or between commercial vehicles usually have a greater force of impact and involve more people than collisions with or between non-commercial vehicles. A higher proportion of commercial vehicles means a higher proportion of heavy vehicles, thus in the event of a crash, the likelihood of a KSI is higher. There is apparently a migration effect from slight to KSI injury crashes, resulting in this opposite trend. Emphasizing safety education to reduce the aggressive behavior of commercial vehicle drivers is a more effective means of reducing the risk of injury to other road users than limiting the number of commercial vehicles allowed on the road. A one unit increase in the proportion of commercial vehicles results in a 0.50 variation in the risk of slight injury crashes and a 0.33 variation in the risk of KSI crashes, provided that the other variables are held constant.

Similarly, turning pockets must be used sparingly, because the presence of turning pockets is positively correlated with the risk of slight injury crashes. An additional left-turning or right-turning pocket will increase the traffic volume of an intersection, because the turning pocket is separated from the through lane. However, if the pocket is not appropriately designed, conflicts between through and turning vehicles may increase the risk of slight injury crashes.

The average lane width is negatively related to the KSI risk: an increase in the average lane width reduces the number of KSI crashes at signalized intersections. A wider lane gives drivers—especially aggressive drivers—more space to maneuver, thus reducing the crash risk. An increase in the average lane width of one unit causes a decrease in the KSI crash risk of 0.02, provided that the other variables are held constant.

Another significant factor in the KSI model is the cycle time at signalized intersections. The model shows a positive correlation with the cycle time: a longer cycle time increases the KSI crash risk. Red-light jumping may increase if aggressive drivers know that they will have to wait for a long red light if they miss the last seconds of amber light, which is a dangerous maneuver that leads to more serous crashes.

In Table 3, sigma 1 and sigma 2 denote the two estimated standard errors obtained from the bivariate Tobit regression model. Rho is the disturbance correlation between the slight injury model and the KSI model and takes a value of 0.26, indicating that a correlation exists between the two parts of the bivariate Tobit model. The test statistic for the first stage of the proposed model is 13.62 and for the second stage of the model is 290.52, each with a distributed χ^2. These values are significant at the 95% level, which provides strong evidence for the sequential theoretical formulation of the crash severity level and the resulting empirical specifications obtained from the Tobit model.

Table 4 Results of the Bivariate Poisson-lognormal Model and Bivariate Tobit Model

	Bivariate Poisson-lognormal Model			Bivariate Tobit Model		
Variables	**Coefficient**	**Std. Err.**	**Z-statistic**	**Coefficient**	**Std. Err.**	**Z-statistic**
Slight injury model						
Ln AADT	0.29*	0.11	2.82	0.42*	0.03	14.89
4Appr.	-0.43*	0.21	-2.08	0.46*	0.04	11.44
Lanewidth	1.40*	0.48	2.89	0.05*	0.01	5.27
Noconflict				0.02*	0.003	7.82
Reciprad				2.12*	0.61	3.50
Comveh				-0.72*	0.17	-4.26
Cycletime				0.01*	0.001	3.00
Pedcrossing				0.02*	0.01	2.08
Tram	-0.69*	0.27	-2.58			
KLN	0.69*	0.025	2.82	0.17*	0.04	4.36
Cons	-7.83	1.89	-4.14	3.69*	0.27	13.82
KSI model						
Ln AADT				0.18*	0.02	9.83
4Appr	-0.34*	0.11	-2.13	0.08*	0.03	2.53
Noconflict				0.01*	0.002	4.45
Notrnstream	-0.08*	0.31	-2.70			
Lanewidth	-0.75*	0.26	-2.92			
Comveh				0.34*	0.12	2.79
Cycletime	0.02*	0.004	4.34	0.004*	0.001	4.86
Pedecrossing	0.29*	0.04	6.97			
Lrtstop						
HKI				-0.30*	0.04	-8.47
Cons	-0.58	1.06	-0.54	1.78*	0.21	8.66
Goodness-of-fit assessment						
Sigma1	0.53			0.30		
Sigma2	0.28			0.18		
Rho	-1.00			0.23		
Number of observations	420			420		
Log likelihood at zero	-257.38			-396.69		
Log likelihood at convergence	-94.21			-118.14		
Chi-square	337.26			434.64		
Degree of freedom	13			17		
AIC	540.76			270.28		
BIC	520.01			280.88		
MAD	0.324			0.292		
MAPE	11.538			11.865		
RMSE	0.605			0.608		

Note: The coefficients of all of the variables, sigma 1 and sigma 2, are scaled up to a million vehicles; * 5% level of significance.

We demonstrate the effectiveness of the proposed model by comparing the results with those obtained from the multivariate (bivariate) Poisson-lognormal regression model (Park and Lord, 2007) and the single bivariate Tobit model for crashes at different crash severity levels (Yoo, 2005; Anastasopoulos et al., 2012b), as shown in Table 4. The bivariate Poisson-lognormal model has the strongest correlation between the two severity levels, as it has the greatest Rho value (-1.00) at the 5% significance level. However, the AIC and BIC values of the Poisson-lognormal model are the largest of the three models. The simple bivariate Tobit model has the weakest correlation between the two severity levels, as it has the lowest Rho value (0.23) at the 5% significance level. The AIC and BIC values of the simple bivariate Tobit model are smaller than those of the bivariate Poisson-lognormal model. The proposed model and the bivariate Poisson-lognormal model have the same MAD values, which are larger than those of the simple bivariate Tobit model. However, the likelihood

ratios of the bivariate Tobit model at zero and convergence are lower than those of the bivariate Poisson-lognormal model and comparable with those of the proposed model. Although the correlation in the proposed model is not the strongest (Rho value = 0.26) at the 5% significance level, the AIC, BIC, MAPE and RMSE values are the lowest, indicating that its performance is the best of the three models. It should be noted that the proposed model has the most degrees of freedom and that it includes the variables from both stages. As defined, both MAPE and RMSE do not show signs of errors. However, MAPE takes percentages of actual values and does not penalize extreme deviations or cancel offsetting errors, whereas RMSE penalizes extreme errors and does not offset the errors. Both values are smaller in the proposed model, indicating that the errors for the proposed model are offset and implying that if adequate geometric and traffic data are available, a fully specified model may be better than the proposed model.

5. Conclusions

In this paper, a two-stage bivariate logistic-Tobit model is developed to evaluate the safety performance at signalized intersections in Hong Kong. A binary logistic model is used to assess the crash severity level and a bivariate Tobit model is used to simultaneously address the slight injury and KSI risks.

The results of the binary logistic model indicate that the crash severity level is positively correlated with the AADT and number of pedestrian crossings. The results of the bivariate Tobit model suggest that the proportion of commercial vehicles is most likely to influence both slight injury and KSI crashes. The overall severity level, existence of four or more approaches, reciprocal of the average turning radius and presence of a turning pocket increase the likelihood of slight injury crashes. The cycle time increases the likelihood of KSI crashes, whereas the average lane width reduces the likelihood of KSI crashes. The bivariate Tobit model also addresses the correlation between the risk of slight injury crashes and the risk of serious or fatal injury crashes, which implies that the unobserved variables are heterogeneous between the signalized intersections in Hong Kong.

The proposed two-stage model has several advantages over the conventional crash prediction analysis method, which uses separate crash frequency and crash severity models.

Previous studies have demonstrated that crash severity types may be correlated (i.e., that they share unobserved effects) (Milton et al., 2008). The proposed model circumvents this limitation by using bivariate models in its second stage, making it a useful tool.

Some signalized intersections have low or zero crash counts, especially for fatal crashes, and cannot therefore be easily analyzed using crash frequency models at different severity levels. Zero-inflated models can account for the excessive number of zeros in the count models and can also be implemented in our two-stage analysis framework. The Tobit approach can be replaced with a zero-inflated approach if the crash frequency is a concern rather than the crash rate. The second stage of the bivariate Tobit model is capable of predicting the number of crashes at different crash severity levels even when the signalized intersection under study has a low or zero crash count.

The proposed model offers flexible, convenient specifications and estimation procedures. The two stages of the proposed model are not limited to the specific methods used here and the most suitable model can be used instead at each stage. For instance, if more than three

severity levels are considered, the ordered logistic/probit model can be substituted with the binary logistic model, another non-regression model or an artificial intelligence algorithm (e.g., a genetic algorithm or neural network), all of which offer appropriate tools for a severity analysis. The same is true of the second-stage bivariate Tobit model, which can accommodate different correlation patterns between the crash severity outcomes and unobserved heterogeneity according to various requirements. It can also be extended to multivariate models if required.

This method involves a less complex estimation procedure than other models. Researchers with less mathematical expertise will find it convenient to estimate the model using the associated statistical package. This may benefit practitioners and facilitate the validation process.

In sum, compared with a simple bivariate Tobit model, the proposed two-stage approach has the following advantages: (1) the two-stage approach handles endogenous and heterogeneous effects by incorporating the crash severity and crash frequency into the forecast; (2) by incorporating crash severity and crash frequency, the two-stage approach reduces the effects of the overly complicated single level modeling structure and the effects of complex modeling estimation; (3) the two-stage approach retains all of the benefits of a single level model; and (4) the two-stage model is easy to implement.

However, as Mannering and Bhat (2014) have stated, all methodological approaches have inherent limitations. Accordingly, one weakness of the proposed model is that in the first stage only two variables are statistically significant, which may generate biased results. Accordingly, one weakness of the proposed model is that in the first stage only two variables are statistically significant. Having data that includes a broader range of explanatory variables could result in additional variables producing statistically significant coefficient estimates. In addition, the endogeneity issue is also only partially addressed with this two-stage approach. More fully addressing endogeneity issues is promising direction for future research.

Future research in this area should establish a more comprehensive dataset by integrating the time-series cross-sectional information available for typical signalized intersections and then using the panel data model to simultaneously investigate the temporal and spatial effects at signalized intersections. The model performance for signalized intersections may be improved by incorporating data on pedestrian flow and approach speed, addressing the endogeneity between safety and mobility.

Acknowledgements
This study was jointly supported by the Fundamental Research Fund for the Central Universities (HUST: 2013QN031), the National Natural Science Foundation of China (NSFC) (No: 51208222), the Scientific Research Foundation for Returned Overseas Chinese Scholars, the State Education Ministry of China, the University Research Committee of the University of Hong Kong (201109176069), the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. HKU 7175/12E) and a grant from the National Research Foundation of Korea funded by the Korea government (MEST) (NRF-2010-0029446).

References

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.

Anastasopoulos, P., Mannering, F., 2009. A note on modeling vehicle-accident frequencies with random-parameters count models. Accident Analysis and Prevention, 41(1), 153–159.
Anastasopoulos, P., Mannering, F., 2011. An empirical assessment of fixed and random parameter logit models using crash-and non-crash-specific injury data. Accident Analysis and Prevention, 43(3), 1140-1147.
Anastasopoulos, P., Mannering, F., Shanker, V., Haddock, J., 2012a. A study of factors affecting highway accident rates using the random-parameters Tobit model. Accident Analysis and Prevention, 45(1), 628–633.
Anastasopoulos, P., Shankar, V., Haddock, J., Mannering, F., 2012b. A multivariate Tobit analysis of highway accident-injury-severity rates. Accident Analysis and Prevention, 45(1), 110–119.
Anastasopoulos, P., Tarko, A. P., Mannering, F. L., 2008. Tobit analysis of vehicle accident rates on interstate highways. Accident Analysis and Prevention, 40(2), 768–775.
Bellemare, M. F., Barrett, C. B., 2006. An ordered Tobit model of market participation: evidence from Kenya and Ethiopia. American Journal of Agricultural Economic, 88(2), 324–337.
Bhat, C.R., Born, K., Sidharthan, R., Bhat, P.C., 2014. A count data model with endogenous covariates: formulation and application to roadway crash frequency at intersections. Analytic Methods in Accident Research, 1, 53-71.
Castro, M., Paleti, R., Bhat, C. R., 2012. A latent variable representation of count data models to accommodate spatial and temporal dependence: application to prediction crash frequency at intersections. Transportation Research Part B: Methodological, 46(1), 253–272.
Chen, E., Tarko, A.P., 2014. Modeling safety of highway work zones with random parameters and random effects model. Analytic Methods in Accident Research, 1, 86-95.
Chen, S., Zhou, X., 2011. Semi-parametric estimation of a bivariate Tobit model. Journal of Econometric, 165(2), 266–274.
Chen, Z., Kuo, L. 2001. A note on the estimation of the multinomial logit model with random effects. American Statistician, 55(2), 89-95.
Chin, H. C., Quddus, M. A., 2003. Applying the random effect negative binomial model to examine traffic accident occurrence at signalized intersections. Accident Analysis and Prevention, 35(2), 253–259.
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.
El-Basyouny, K., Sayed, T., 2009. Accident prediction models with random corridor parameters. Accident Analysis and Prevetion, 41(5), 1118-1123.
El-Basyouny, K., Sayed, T., 2011. A full Bayes multivariate intervention model with random parameters among matched pairs for before-after safety evaluation, 43(1), 87-94.
Enberg, J. 1990. A random-effects logit model of work-welfare transitions. Journal of Econometrics, 43(2), 63-75.
Greene, W., 2003. LIMDEP 8.0 Reference Guide. Plainview NY: Econometric Software, Inc.
Guo, F., Wang, X., Abdel-Aty, M., 2010. Modeling signalized intersection safety with corridor-level spatial correlations. Accident Analysis and Prevention, 42(1), 84–92.
Haque, M. M., Chin, H. C., Huang, H., 2010. Applying Bayesian hierarchical models to examine motorcycle crashes at signalized intersections. Accident Analysis and Prevention, 42(1), 203–212.
Haleem, K., Gan, A. 2013. Effect of driver's age and side of impact on crash severity along urban freeways: A mixed logit approach. Journal of Safety Research, 46, 67-76.
Karlaftis, M., Tarko, A. P., 1998. Heterogeneity considerations in accident modeling. Accident Analysis and Prevention, 30(4), 425–433.
Kim, D. G., Lee, Y., Washington, S. P., Choi, K., 2007. Modeling crash outcome probabilities at rural intersections: application of hierarchical binomial logistic models. Accident Analysis and Prevention, 39(1), 125–134.
Kim, D. G., Washington, S. P., 2006. The significance of endogeneity problems in crash models: an examination of left-turn lanes in intersection crash models. Accident Analysis and Prevention, 38(6), 1094–1100.
Liu, P., 2007. A neural network approach on analyzing and reducing signalized intersection crashes. Third International Conference on Natural Computation (ICNC 2007), IEEE Computer Society.
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.
Ma, J., Kockelman, K. M., 2006. Bayesian multivariate Poisson regression for models injury count, by severity. Transportation Research Record, 1950, 24–34.
Malchow-Moller, N, Svarer, M. 2003. Estimation of the multinomial logit model with random effects. Applied Economics Letters, 10(7), 389-392.
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.
Milton, J., Shankar, V., Mannering, F., 2008. Highway accident severities and the mixed logit model: an exploratory empirical analysis. Accident Analysis and Prevention, 40(1), 260–266.

Obeng, K., 2011. Gender differences in injury severity risks in crashes at signalized intersections. Accident Analysis and Prevention, 43(4), 1521–1531.
Obeng, K., Burkey, M., 2006. Explaining property damage from crashes at signalized intersections. Transportation Planning and Technology, 29(3), 217–231.
Park, E. S., Lord, D., 2007. Multivariate Poisson-lognormal models for jointly modeling crash frequency by severity. Transportation Research Record, 2019, 1–6.
Pai, C.W., Hsu, J.J., Chang, J.L., Kuo, M.S. 2013. Motorcyclists violating hook-turn area at intersections in Taiwan: An observational study. Accident Analysis and Prevention, 59, 1-8.
Pei, X., Wong, S. C., Sze, N. N., 2011. A joint-probability approach to crash prediction models. Accident Analysis and Prevention, 43(3), 1160–1166.
Pei X., Wong, S. C., Sze, N. N., 2012. The roles of exposure and speed in road safety analysis. Accident Analysis and Prevention, 48(1), 464–471.
Poch, M., Mannering, F. L., 1996. Negative binomial analysis of intersection-accident frequencies. Journal of Transportation Engineering, 122(2), 105–113.
Roy, S. 2012. Accounting for response misclassification and covariate measurement error using a random effect logit model. Communication in Statistics-Simulation and Computation, 41(9), 1623-1636.
Russo, B.J., Savolainen, P.T., Schneider IV, W.H., Anastasopoulos, P., 2014. Comparison of factors affecting injury severity in angle collisions by fault status using a random parameters bivariate ordered probit model. Analytic Methods in Accident Research, 2, 21-29.
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.
Venkataraman, N., Ulfarsson, G.F., Shankar, V.N., 2013. Random parameter models of interstate crash frequencies by severity, number of vehicles involved, collision and location type. Accident Analysis and Prevention, 59, 309-318.
Wang, C., Quddus, M. A., Ison, S. G., 2011. Predicting accident frequency at their severity levels and its application in site ranking using a two-stage mixed multivariate model. Accident and Prevention, 43(6), 1979–1990.
Wang, S., Tsodikov, A. 2010. A self-consistency approach to multinomial logit model with random effects. Journal of Statistical Planning and Inference, 140, 1939-1947.
Wang, X., Abdel-Aty, M., 2006. Temporal and spatial analyses of rear-end crashes at signalized intersections. Accident and Prevention, 38(6), 1137–1150.
Wang, X., Abdel-Aty, M., Chen, X., 2007. Safety analysis at signalized intersections: research strategies, modeling techniques, and significant factors. Plan, Build, and Manage Transportation Infrastructure in China Congress 2007 (ISSTP), 389–400.
Washington, S. P., Karlaftis, M. G., Mannering, F. L., 2011. Statistical and Econometric Methods for Transportation Data Analysis, Second Edition. Boca Raton, FL: Chapman and Hall/CRC.
Wong, S. C., Sze, N. N., Li, Y. C., 2007. Contributory factors to traffic crashes at signalized intersections in Hong Kong. Accident Analysis and Prevention, 39 (6), 1107–1113.
Xie, K., Wang, X., Huang, H., Chen, X., 2013. Corridor-level signalized intersection safety analysis in Shanghai, China using Bayesian hierarchical models. Accident Analysis and Prevention, 50(1), 25–33.
Yau, K., Ma, P. 1999. A simulation study for the binomial-logit model with correlated random effects. Journal of Statistical Computation and Simulation, 63(2), 169-186.
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.
Yoo, S., 2005. Analyzing household bottled water and water purifier expenditures: simultaneous equation bivariate Tobit model. Applied Economics Letters, 12(5), 197–301.
Yu, R.J., Abdel-Aty, M. 2014. Analyzing crash injury severity for a mountainous freeway incorporating real-time traffic and weather data. Safety Science, 63, 50-56.