A Two-Step Quantile Selection Model for the Safety Analysis at Signalized Intersections

Xuecai Xu[a], Y.C. Li[b], S.C. Wong[b], Feng Zhu[a*]

[a] *School of Civil and Environmental Engineering, Nanyang Technological University, Singapore*
[b] *Department of Civil Engineering, The University of Hong Kong, Pokfulam Road, Hong Kong, China*

Abstract: The simultaneous estimation of crash frequency and severity has been studied for years, but most of the existing methodologies adopt mean regression models to estimate the parameters. This study presents the quantile selection model as a methodological alternative in analyzing crash rate and severity at different levels, focusing on addressing the heterogeneity and endogeneity issues so as to identify the influencing factors at signalized intersections. A two-step estimation procedure is carried out, in which the Heckman selection framework accommodates the endogenous relationship between crash rate and crash severity at different levels, while the quantile regression estimates various quantiles of crash rate instead of the mean regression, and accounts for the heterogeneity attributed to unobserved factors. Compare to the general Heckman selection model, the quantile approach is able to provide more comprehensive information about the impact of the influencing factors on crash rate. The model uses 555 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 two years. The proposed model reveals more detailed information in terms of different quantiles and improves the prediction accuracy.

Keywords: Signalized Intersection; Crash Rate; Crash Severity; Quantile Selection Model; Heckman Selection Model

Introduction

The simultaneous estimation of crash frequency and severity at signalized intersections has attracted increasing attention in past decades. A variety of different approaches and perspectives (Park & Lord, 2007; Wong et al., 2007; Ye et al., 2009; Venkataraman et al., 2013; Agebelie & Roshandeh, 2015; Agebelie , 2016; Islam & Hernandez, 2016) have been applied in prediction modeling. Studies have suggested that bias estimations might result for separate crash frequency or separate crash severity levels at signalized intersections because possible correlations between crash frequency and severity levels are not considered (Lord & Mannering, 2010). Moreover, Xu et al. (2013) also stated that heterogeneity and endogeneity issues should be addressed in evaluating the influencing factors on the safety. For decades, various modeling approaches have been proposed to deal with these two issues, but there has been no uniform criterion and too many modeling assumptions, which may violate the natural attributes and lead to biased inferences.

Numerous modeling approaches have been adopted for the simultaneous estimation of

[*] Corresponding author. Email: zhufeng@ntu.edu.sg
Postal address: 50 Nanyang Avenue, Singapore, 639798; Tel: +65-6790-5267; Fax: +65-6791-0676

crash frequency and severity, for example multi-level hierarchical structures (Kim et al., 2007; Aguero-Valverde, 2014), simultaneous equations (Kim & Washington, 2006; Ye et al., 2009, 2013; Caliendo & Guida, 2014) and two-stage bivariate or multivariate analysis (Ma & Kockelman, 2006; Park & Lord, 2007; Xu et al., 2014). Primarily, these approaches can be considered either combined crash frequency or severity models, or two-stage models. From the perspective of combined crash frequency or severity models, the studies mainly used combined models, e.g. a multivariate Poisson-lognormal model (El-Basyouny and Sayed, 2011), a multinomial generalized Poisson structure (Chiou and Fu; 2013), a multivariate zero-inflated Poisson-lognormal regression model (Dong et al., 2014), or even Bayesian framework (Pei et al., 2011), and spatio—temporal analysis (Chiou and Fu, 2015) to address frequency and severity separately or simultaneously (Abdel-Aty and Keller, 2005).

Alternatively, some researchers focus on the two-stage bivariate or multivariate analysis. Most of the two-stage models integrated two different models e.g. the two-stage mixed multivariate model (Wang et al., 2011), the outcome model with a multinomial probit selection model (Bhat et al., 2014), the two-stage bivariate logistic-Tobit model (Xu et al., 2014), the two-stage model with binary probit model and switching regression model (Ding et al., 2015), and the two-stage bivariate binary probit and bivariate ordered probit model (Li et al, 2017), to simultaneously accommodate frequency and severity. All of these studies verified that two-stage analysis provides potential for future study.

Fundamentally, the aforementioned models belong to the mean regression, in which the model assumptions cannot be easily extended to non-central locations, and are not always satisfactory with the nature of real-world data, especially in the case of homoscedasticity assumption (Qin, 2012). In light, quantile regression (QR) approach was proposed to specify conditional quantiles as functions of predictors (Koenker & Bassett, 1978). Currently, QR has been widely used in many fields and areas, such as example sociology, economics, finance and medical science (Qin, 2012; Wang et al., 2016), but the application in transportation research is still at the initial stage (Qin et al. 2010; Kwon et al., 2011; Qin, 2012; Wu et al., 2014; Washington et al., 2014). A pioneering study by Hewson (2008) examined the potential role of QR for modeling the speed data, and demonstrated the potential benefits of using QR methods, providing more interest than the conditional mean. Subsequently, QR was introduced into safety area to identify the crash frequency (Qin et al., 2010; Qin, 2012; Wu et al., 2014) and crash severity (Liu et al., 2013; Washington et al. 2014).

Compared with the mean regression, QR can estimate different effects at different quantiles of the response variable, in which a specific distribution is not required. Moreover, QR is more robust against outliers because the estimation results are less sensitive to outliers and multi-modality (Liu et al., 2013). In particular, QR can process the heterogeneity issue for data collected from different sources at different locations and times without many assumptions (Qin et al. 2010; Qin, 2012), which addresses the relationship between safety and influencing factors more appropriately. Besides, the Heckman selection model (Heckman, 1979) is suitable to identify the relationship between crash rate and severity levels as it helps to correct the selection bias, providing a direct test for endogeneity from the estimation. Again, the Heckman selection model has been commonly adopted in the economics field, but application in transportation research was unfamiliar (Zhou & Kockelman, 2008; Kaplan et al., 2016; Xu et al. 2017).

Although QR approach and Heckman selection model are popular in different areas, these two modeling approaches have been rarely combined. In this study, the QR and Heckman selection model are integrated to establish a sophisticated Quantile Selection Model (QSM). In this QSM, the crash rate is realized by QR approach that is more robust to non-normal errors, while the crash injury severity is addressed by sample selection model, so that both the heterogeneity and endogeneity issues are addressed simultaneously. The model provides a two-step analysis and deals with the zero-sample issue, based on which it can accommodate the heterogeneity (i.e., shared unobserved factors) between signalized intersections and then address the endogeneity (between crash rate and severity levels) at signalized intersections. An illustrative example using a crash dataset from signalized intersections in Hong Kong is used to evaluate the suitability of the proposed model.

Data Description

In this study, traffic crash information extracted from the Traffic Accident Database System (TRADS) of 262 signalized intersections in Hong Kong during the periods between 2002 and 2004 were adopted. The crash information of an intersection in a particular year is considered as an independent observation, and hence a total of 555 observations distributed across Hong Kong Island (HKI), Kowloon (KLN) and the New Territories (NT) were obtained (Xu et al., 2014). The count data included intersections without crashes or with one or multiple crashes in two severity levels: 1) slight injury, or 2) killed or serious injured (KSI). To conform to the proposed model selection procedure, intersections zero crashes are included. Among the 555 observations, 135 (24%) exhibited zero crashes and 420 (76%) exhibited one or multiple crashes. Among the 76% crashes, 60% of 420 crash cases belonged to the slight injury category, and the remaining 16% belonged to the KSI category.

To review the performance of crash data, the distributions on the crash rate of slight injury and KSI are plotted in **Figure 1**. The crash rate of slight injury (SCrRt) or KSI (KCrRt) are defined as the numbers of million crashes per year divided by the annual exposure, which is calculated by multiplying the annual average daily traffic (AADT) by 365 days. Considering that skewed distribution is observed, the QR approach is therefore preferred for modeling crash rate in this study. Furthermore, due to the data collection process, substantial heterogeneity may exist in the crash data among the signalized intersections.

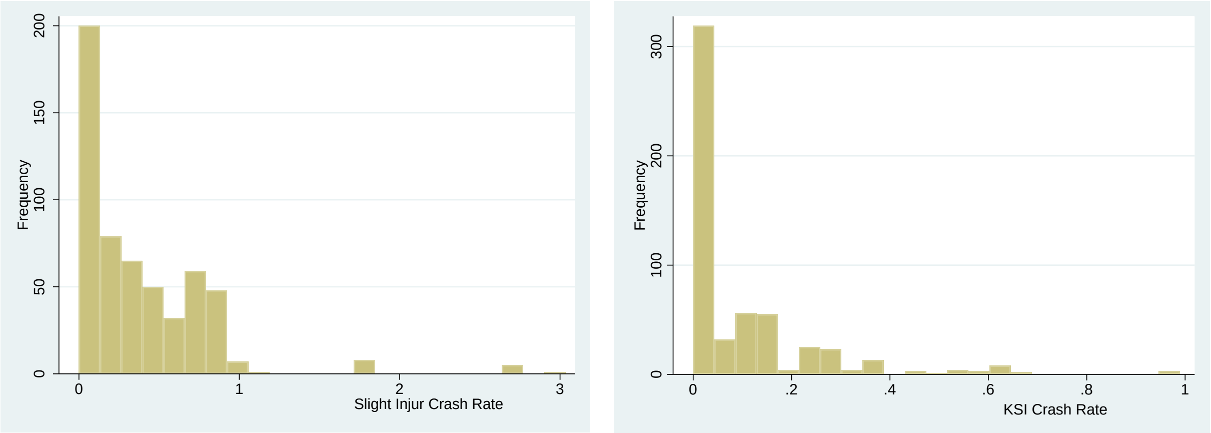


Figure 1 Crash Rate Histogram of Slight Injury and KSI

Other contributory factors, including roadway characteristics (number of approach lanes, number of conflict points, number of turning movements required, average lane width and reciprocal of the turning radius), traffic characteristics (proportion of commercial vehicles and speed limit), signal-phasing scheme (number of signal phases, signal cycle time and number of pedestrian crossings), geometric characteristics (number of approaches, presence of tram stops and light rail transit [LRT] stops), road environments on HKI and in KLN and presence of turning pockets were also observed (Wong et al., 2007; Xu et al., 2014). The available characteristics of the data sample are presented in **Table 1**.

Table 1 Sample characteristics for the selected signalized intersections (Sample size = 555)

Variable	**Description**	**Categories**			
i) Dependent variables					
Slight	Slight injury	Yes: 60%		No: 40%	
KSI	Killed and serious injury (KSI)	Yes: 15%		No: 85%	
		Mean	**S.D.**	**Min.**	**Max.**
SCrRt	Crash rate of slight injury	0.62	0.50	0.39	3.03
KCrRt	Crash rate of KSI	0.60	0.47	0.68	2.45
ii) Exposure					
AADT	AADT	35,934.16	23,219.35	903	121,221
iii) Continuous variables					
Roadway characteristics					
Nolanes	Number of approach lanes	8.49	3.52	2	18
Noconflict	Number of conflict points	8.74	8.53	0	30
Notrnstream	Number of turning movements required	6.32	2.70	1	12
Lanewidth	Average lane width (m)	3.31	0.31	2.7	5.5
Reciprad	Reciprocal of the turning radius	0.09	0.03	0	0.2
Traffic characteristics					
Comveh	Proportion of commercial vehicles	0.21	0.10	0.01	0.66
Speed	Speed limit (km/h)	50.04	0.85	50	70
Signal-phasing scheme					
Nostages	Number of signal stages	3.14	0.78	2	7
Cycletime	Cycle time (s)	98.31	18.30	44	140
Pedcrossing	Number of pedestrian crossings	4.06	2.21	0	8
iv) Indicator variables (Yes=1, No=0)					
Geometrical characteristics					
2 Appr.	Two approaches	0.16		0	1
3 Appr.	Three approaches	0.30		0	1
4 Appr.	Four or more approaches	0.69		0	1
Tramstop	Presence of tram stops	0.06		0	1
Lrtstop	Presence of LRT stops	0.02		0	1
Road environment					
HKI	Hong Kong Island	0.23		0	1
KLN	Kowloon	0.58		0	1
Signal-phasing scheme					
Turningpock	Presence of turning pocket	0.08		0	1

S.D. = Standard deviation

Method

Quantile regression (QR) model

Quantiles are cut points dividing the range of a probability distribution into contiguous intervals with equal probabilities. Let *p* be a number between zero and one, then the 100*p* percentile of the distribution for a continuous random variable *y*, denoted by *Q(p)*, can be expressed as follows:

$$p = P\left(y \le Q(p)\right) = F\left(Q(p)\right) = \int_{-\infty}^{Q(p)} f(y)dy \tag{1}$$

where *f(y)* is the density distribution function. Converted from Equation (1), *Q(p)* with $0 \le p \le 1$ is defined as follows:

$$Q(p) = F^{-1}(p) = \inf\left\{y : F(y) \ge p\right\} \qquad 0 \le p \le 1 \tag{2}$$

where F^{-1} denotes the inverse function of the cumulative distribution function, and *inf* denotes the greatest lower bound. It is noted that *Q*(0.5) is the median. The first and third quartiles are *Q*(0.25) and *Q*(0.75), and the 95th percentile is expressed as *Q*(0.95), respectively.

Similar to the mean of a random sample leading to the minimal of the sum of square errors, the median of a random sample $\left\{y_1, y_2, ... y_n\right\}$ for a random variable *y* results in the minimal of the sum of absolute deviations. Consequently, the general *Q(p)* can be interpreted as an optimal solution to minimize the weighted average of the samples whose values are larger or equal to *Q(p)*, and the samples whose values are less or equal to *Q(p)* as follows:

$$\min\left[\sum_{i\in\left\{i:y_i \le Q(p)\right\}} p\left|y_i - Q(p)\right| + \sum_{i\in\left\{i:y_i > Q(p)\right\}} (1-p)\left|y_i - Q(p)\right|\right] \tag{3}$$

Assume *y* is a linear function of the variables as the following:

$$y = X'\beta + \varepsilon \tag{4}$$

where *y* denotes the response variable, β is the vector of unknown parameters of the covariates *X*, and ε is random error. Therefore, the optimization problems can be converted into solving the estimation for β s:

$$\hat{\beta}(p) = \underset{\beta\in R^k}{\arg\min}\left[\sum_{i\in\left\{i:y_i \ge X'\beta\right\}} p\left|y_i - X'\beta\right| + \sum_{i\in\left\{i:y_i < X'\beta\right\}} (1-p)\left|y_i - X'\beta\right|\right] \tag{5}$$

As for any quantile *p* between 0 and 1, $\hat{\beta}(p)$ can be regarded as the *p*th regression quantile, which minimizes the sum of weighted absolute residuals (Qin et al. 2010; Qin & Reyes, 2011; Qin, 2012).

Heckman selection model

The Heckman selection model formulates from two equations, including a **regression equation**:

$$y_i = X_i\beta + \mu_1 \tag{6}$$

and a **selection equation**

$$Z_i\gamma + \mu_2 > 0 \tag{7}$$

with the following holds:

$$\begin{aligned} &\mu_1 \sim N(0,\sigma) \\ &\mu_2 \sim N(0,1) \\ &corr(\mu_1,\mu_2) = \rho \end{aligned} \tag{8}$$

where y_i denotes the dependent variables, X_i denotes the observable features of the independent variables, β denotes the parameters to be estimated and μ_1 is a normally distributed error term with a mean of zero and a standard deviation σ to be estimated. Z_i denotes observable features including the overlapping variables with X_i, and γ denotes the vectors of parameters to be estimated. μ_2 is a distributed error term with a mean of zero and a standard deviation equal to one. ρ represents the correlation between the two error terms to be estimated. Using these two equations, samples larger than zero can be selected and estimated based on various modeling methods, through which the Heckman selection model provides consistent, asymptotically efficient estimates for all of the parameters. More detailed estimation procedures can be referred to Leung and Yu (1996), Schwiebert (2015) and Xu et al. (2017).

Quantile Selection Model

Different from Equation (4), Equation (6) is a mean-type regression model. However, if a more comprehensive view of the relationship between dependent and independent variables at various response levels is required, the estimates by Equation (6) may be inadequate. Moreover, the random errors in Equation (6) may not follow the normal distribution, thus the assumptions will be violated. Therefore, when the effect of explanatory variables on the entire distribution of dependent variables is required, and the random errors do not conform to the normal distribution hypothesis, the quantile selection model is therefore proposed as supplement for the conventional Heckman selection model.

In this study, it is assumed that a **quantile regression model** can be used to explain the crash rate for slight injury/KSI:

$$y_i = X_{1i}\beta_1 + C_i\beta_2 + \mu_i = X_i\beta + \mu_i \tag{9}$$

where y_i denotes the crash rate for slight injury/KSI, not mean type but quantile; X_i is a vector of observable features related to slight injury/KSI, in which X_{1i} represents the endogenous variables; C_i stands for the exogenous variables; β_1, β_2 and β are vectors of parameters to be estimated; and μ_i is error term to be estimated. Here, the dependent variable y_i may not always be observed, and it is specially observed only when the crashes actually belong to the slight injury/KSI categories. Therefore, in the **selection model**, the dependent variable is observed if:

$$Z_{1i}\gamma_1 + C_i\gamma_2 + \upsilon_i = Z_i\gamma + \upsilon_i > 0 \tag{10}$$

where Z_i is a vector of observable features related to slight injury/KSI, which includes the overlapping variables with X_i; Z_{1i} represents the endogenous variables that may or may not be the same as X_{1i}; γ_1, γ_2 and γ are vectors of the parameters to be estimated; and υ_i is a distributed error term with a mean of zero and a standard deviation equal to one. This equation describes the probability that slight injury/KSI is greater than zero.

The error terms hold the following distribution:

$$\begin{gathered} \upsilon_i \sim N(0,1) \\ corr(\mu_i, \upsilon_i) = \rho \end{gathered} \tag{11}$$

where ρ represents the correlation between the two error terms to be estimated. The parameter $\lambda = \sigma\rho$, known as the inverse Mills ratio, is the estimated selection coefficient.

The estimation of the quantile selection regression consists of two steps, starting from the selection model. In the first step, the **probit regression** is used to model the sample selection process in Equation (10), and then the inverse Mills ratio λ, the error from the probit equation explaining selection, is calculated based on the probit regression results. This step can be done using maximum likelihood. In the second step, the inverse Mills ratio is added to **quantile regression analysis** as an independent variable, and ordinary least squares regression is used to provide the consistent parameter estimates in quantiles of Equation (9). The estimation can be computed as follows:

$$\hat{\beta}_\tau = \arg\min_{\beta \in R^k} \left[\sum_{i=1}^{N} \left[\tau(Y_i^* - X_i^{'}\beta) \right] + \sum_{i=1}^{N} \left[(1-\tau)(Y_i^* - X_i^{'}\beta) \right] \right] \tag{12}$$

where $\hat{\beta}$ is a consistent estimator of the τ th quantile regression coefficient for any given $\tau \in (0,1)$, Y_i^* is latent outcome from probit regression.

In this study, the inverse Mills ratio term includes two parts: a selection effect and an effect due to the endogeneity. If the endogeneity between crash rate and injury severity levels is absent, the endogeneity effect is zero, and the model is reduced to the general two-step selection model. The selection effect gives the expected outcome of the fully observed sample while holding the entire explanatory variables constant (including the endogenous variable), and the sign of the selection effect with the endogenous variable is determined by the correlation coefficient ρ. By estimating the preceding equations, the crash rate and severity at different levels can be simultaneously and respectively calculated, and the full distribution of all quantiles is addressed along with the heterogeneity and endogeneity at the signalized intersections. More estimation methods related to the quantile selection models are given in Alhamzawi (2015; 2016), Arellano and Bonhomme (2017).

Results and Discussion

To avoid correlations between the independent variables, a correlation test is conducted to identify the variables without co-linearity to be included in the model, allowing the final results to be obtained. Based on the correlation matrix, the number of approach lanes, conflict points, turning movements and signal stages are highly correlated. Therefore, these variables are not included in the model at the same time.

A two-step quantile selection regression models is developed in this study. The first-step model is a selection model that determines the presence of injury cases; whereas the second-step model examines the effect of the influencing variables on the quantiles of crash rate. To test for bias, we examine the relationship between the residuals in two steps. If the unobserved factors in the first-step selection model are correlated with the unobservable variables in the second step of the model, the correlation is not zero, which implies that unobservable variables in the crash injury selection model also affect the second step of the model and heterogeneity issue exists; If the unobserved factors in the first step are unrelated to the unobservable variables in the second step, that the results of the first step do not affect the results of the second step, and the residuals are not correlated..

The results are presented for the best model specification of slight injury and KSI as quantile selection models in **Tables 2** and **3**, respectively, which gives the estimated coefficients and 95% confidence intervals for statistically significant variables at the 25th, 50th, 75th, 90th, 95th percentile of crash rate distribution. Consequently, it presents a broader and complete view of the variables with different crash rates, that is to say, rather than assuming the coefficient are fixed across all the signalized intersections, some or all of them are allowed to vary to account for heterogeneity attributed to unobserved factors. Moreover, the error terms for ρ at different quantiles are positive for slight injury and negative for KSI, meaning that the unobserved factors that cause slight injury/KSI are positively and negatively correlated with one another, respectively.

Table 2 Estimated results of the quantile selection model for slight injury

Variables	**Coefficient (t-statistics)**				
	0.25	**0.50**	**0.75**	**0.90**	**0.95**
Reciprad	1.722* (6.40)	2.722* (9.49)	4.224* (24.29)	3.556* (11.39)	1.908(0.44)
Cycletime	0.001* (2.82)	0.004* (9.50)	0.008* (17.94)	0.005* (17.49)	0.008(0.68)
Tramstop	0.143* (5.26)	0.230* (6.68)	0.273* (9.43)	0.411* (26.58)	0.293(0.42)
KLN	0.166* (9.68)	0.218* (11.72)	0.137* (6.89)	0.264* (23.47)	0.431(1.03)
Cons	-0.261* (-6.14)	-0.610* (-11.42)	-0.657* (-11.45)	-0.246* (-8.99)	-0.363(-0.48)
	Coefficient	**Std. Err.**	**Z-statistic**		
Slight injury model					
LnAADT	0.482*	0.074	6.49		
Reciprad	4.572*	2.086	2.19		
Tramstop	0.812*	0.240	3.39		
KLN	0.680*	0.121	5.62		
Cons	-5.504*	0.783	-7.02		
Goodness-of-fit					
ρ	0.778	0.783	0.786	0.773	0.757
σ	0.486	0.480	0.487	0.481	0.461
λ	0.378	0.376	0.383	0.372	0.349
MAD	0.297	0.249	0.324	0.390	0.558
RMSE	0.469	0.399	0.450	0.503	0.646

Note: * Significant at the 5% level.

Mean absolute deviation (MAD) = $\frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \hat{Y}_i\right|$, 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. MAD and RMSE are used to compute the errors and reflect the goodness-of-fit of the proposed model.

Table 3 Estimated results of the quantile selection model for KSI

Variables	Coefficient (t-statistics)				
	0.25	**0.50**	**0.75**	**0.90**	**0.95**
Lanewidth	-0.063* (-3.17)	0.054* (2.78)	0.311* (4.30)	0.832* (63.41)	0.651(0.90)
Cycletime	0.001* (2.10)	0.006* (5.01)	0.012* (8.60)	0.005* (23.83)	0.003(0.27)
KLN	0.157* (12.43)	0.188* (4.10)	0.215* (473)	0.329* (35.86)	0.304(0.70)
Cons	0.160 (1.91)	-0.532 (-1.93)	-1.573* (-5.04)	-2.556* (-44.66)	-1.710(-0.53)
	Coefficient	Std. Err.	Z-statistic		
KSI model					
Comveh	0.482*	0.074	6.49		
Tramstop	0.812*	0.240	3.39		
Cons	-5.504*	0.783	-7.02		
Goodness-of-fit					
ρ	-0.856	-0.909	-0.853	-0.836	-0.835
σ	0.636	0.638	0.632	0.590	0.507
λ	-0.544	-0.580	-0.539	-0.493	-0.423
MAD	0.393	0.326	0.427	0.539	0.582
RMSE	0.602	0.489	0.572	0.693	0.712

Note: * Significant at the 5% level.

To demonstrate the effectiveness of the proposed model, **Tables 4** and **5** give the results obtained from the Heckman selection model, with which the proposed model shares some similar features. The correlations between the two severity levels, ρ in Tables 2 and 3 before 75th percentile are stronger than those in Tables 4 and 5, which corresponds to the dataset distribution in Figure 1. Furthermore, MAD and RMSE values at 50th quantile in Tables 2 and 3 are smaller than that in Tables 4 and 5. The results of the proposed quantile selection models provide more holistic and accurate results, revealing both the crash rates at all quantiles for slight injury and KSI severity simultaneously.

Table 4 Estimated results of the Heckman selection model for slight injury

Variables	Coefficient	Std. Err.	Z-statistic
Crash rate of slight injury			
Reciprad	4.923*	1.005	4.90
Cycletime	0.004*	0.001	2.63
Tramstop	0.244*	0.112	2.19
KLN	0.279*	0.068	4.12
Cons	-0.714*	0.215	-3.32
Slight injury model			
AADT	0.001*	0.001	7.43
Reciprad	4.256*	2.094	2.03
Tramstop	0.872*	0.240	3.63
KLN	0.698*	0.123	5.68
Cons	10.267*	0.010	3.68
Goodness-of-fit			
ρ	0.778		
σ	0.486		
λ	0.378		
Number of observations		555	
Wald Chi-square		45.85	
MAD		0.257	
RMSE		0.409	

Note: * Significant at the 5% level.

Table 5 Estimated results of the Heckman selection model for KSI

Variables	**Coefficient**	**Std. Err.**	**Z-statistic**
Crash rate of KSI			
Lanewidth	-0.039*	0.014	-2.81
Cycletime	0.007*	0.003	2.24
KLN	0.330*	0.113	2.91
Cons	0.854*	0.561	1.52
KSI model			
Comveh	1.767*	0.628	2.81
Tramstop	0.585*	0.228	2.57
Speed	0.229*	0.003	71.98
Cons	9.979*	0.004	2.58
Goodness-of-fit			
ρ	-0.856		
σ	0.636		
λ	-0.545		
Number of observations	555		
Wald Chi-square	21.90		
MAD	0.385		
RMSE	0.635		

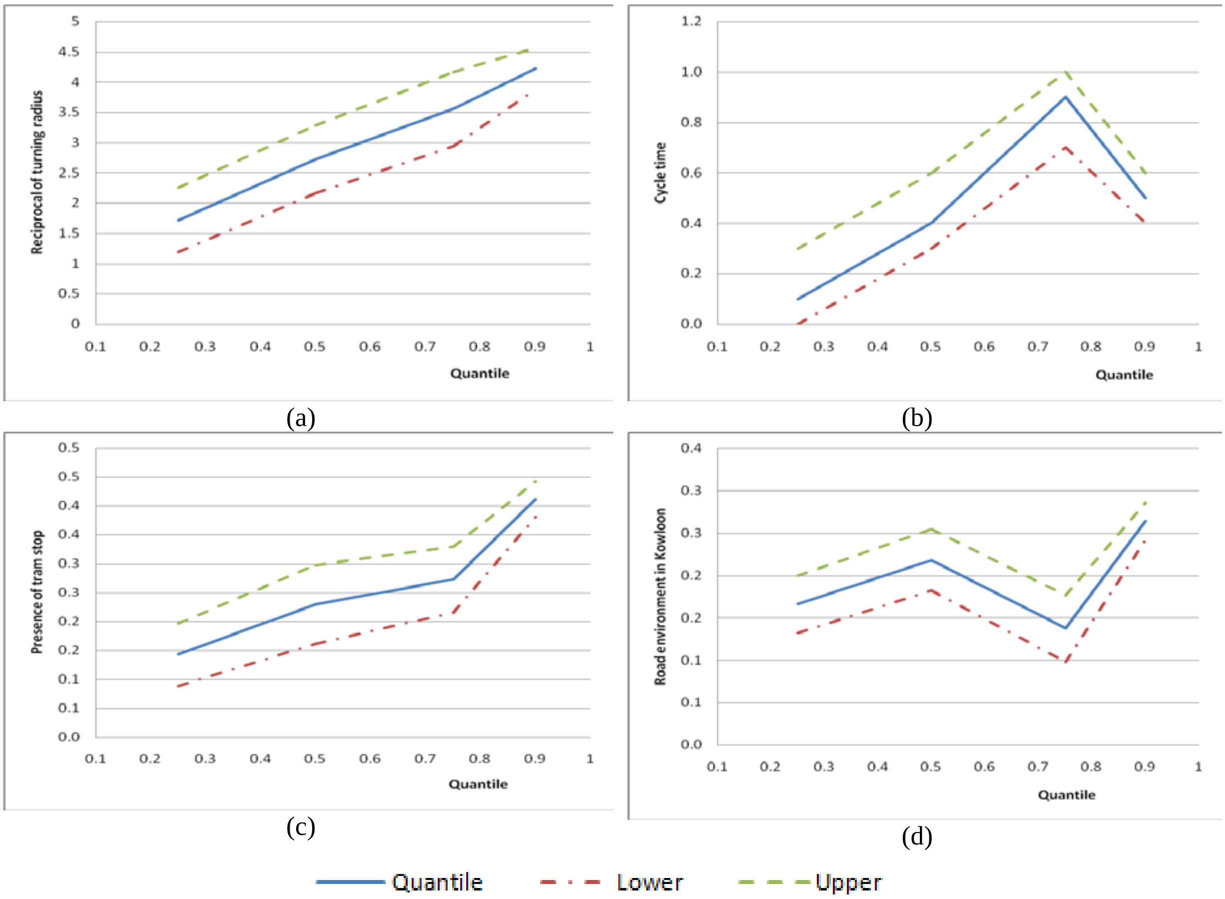
Note: * Significant at the 5% level.

Table 6 Elasticity of Crash Rate Varying with Influencing Variables

Elasticity	**Crash Rate for Slight Injury**			
	0.25	**0.50**	**0.75**	**0.90**
Reciprad	159%	81%	71%	43%
Cycletime	1%	1.5%	0.2%	0.6%
Tramstop	146%	76%	49%	51%
KLN	170%	72%	24.5%	33%
Elasticity	**Crash Rate for KSI**			
	0.25	**0.50**	**0.75**	**0.90**
Lanewidth	16.2%	15.6%	41.4%	94.7%
Cycletime	0.2%	2.1%	1.6%	3.9%
KLN	40%	55%	27%	37.5%

Table 6 shows how the elasticity of crash rate varies with the significant variables in each quantile. The elasticity of crash rate can be calculated from one-unit increase or decrease in one variable while holding the mean values of other influencing variables in each quantile constant. The elasticity of crash rate in the slight injury model shows similar trends for all of the variables except cycle time, whereas no obvious trends in the elasticity of crash rate in the KSI model are observed for the variables except the lane width variation.

As the results in Table 2 and 3, crash rate is positively influenced by the reciprocal of turning radius, cycle time, presence of tram stop and road environment in Kowloon. However, the closer examination of the magnitude of the estimated coefficients reveals some differences and similarities between the quantiles. First, the difference is that significant variables may reveal different impacts on crash rate at different percentiles, e.g. the reciprocal of turning radius is more likely to influence crash rate in the low tail than in the high tail. A unit increase of the reciprocal of turning radius may increase the crash rate 159% at the 25th percentile or crash rate 43% at the 90th percentile. This indicates that the significant reciprocal of turning radius variation at signalized intersections with low tail may trigger a sharp crash explosion easily, thus more attentions should be paid to the turning radius located at those intersections by roadway designers or planners.



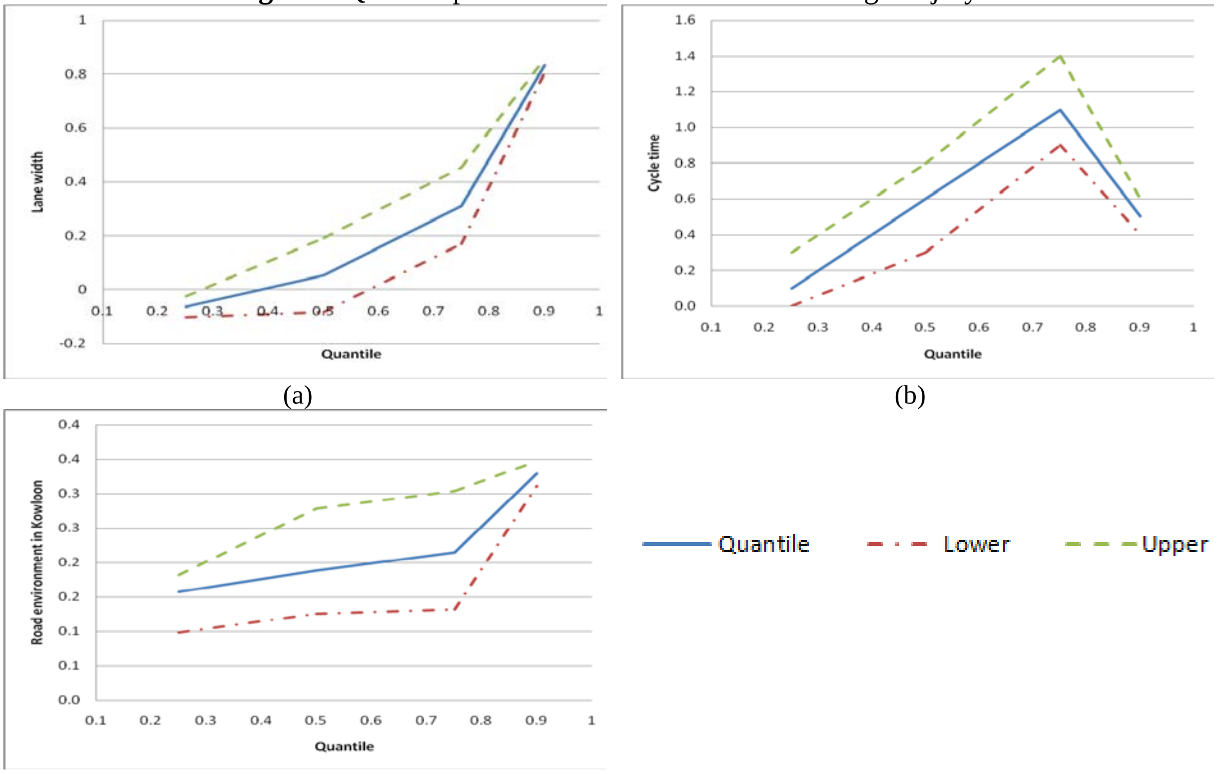
(a)

(b)

(c)

(d)

Figure 2 Quantile plots for variable coefficients of slight injury



(a)

(b)

(c)

Figure 3 Quantile plots for variable coefficients of KSI

Secondly, consistent with the results of the Heckman selection model, the similarity is that all influencing variables are of significance to the 90th percentile, while most of the variables are not significant at the 95th percentile. As shown from Figure 1, the distribution of crash rate concentrates before 75% correspondingly, and begins to weaken from the 75th percentile to the 90th percentile. The reason that all the variables are not statistically significant at the 95th percentile is not only due to the shortage of crash data, but also because of other influencing factors besides the listed variables, such as vehicle problems, drivers' issues, and even environmental conditions, etc. This suggests that the diversity of data sets may need to be considered to evaluate the safety impact at signalized intersections.

Figures 2 and **3** illustrate the estimation results of the coefficients for all the variables. The solid line denotes the coefficients for the 25, 50, 65, 75, and 90 percentiles, which are enveloped by two dashed lines representing a 95% confidence interval.

Slight injury model

Various studies have demonstrated that AADT is significantly related to crash rate (Abdel-Aty & Radwan, 2000; Wong et al., 2007; Milton et al., 2008). Likewise, as shown in Table 2, it is demonstrated that AADT is positively influence the slight injury, implying that traffic volume increases the possibility of slight injury. The more traffic volume on the roadway, the higher the probability that conflicts will be generated. However, because there is greater traffic volume, the vehicles may not speed much, thus the severity mostly resides in reasonable slight injury.

As demonstrated in Figure 2(a), the reciprocal of the turning radius is significantly and positively correlated with the crash rate at all quantiles. A larger turning radius, i.e. smaller reciprocal of the turning radius, is accompanied by better sight distance, such that the severity of crashes decreases. However, for slight injury, the positive relation to crash rate and slight injury indicates that the possibility of crashes and slight injury still increases with the traffic flow, even under the large turning radius condition.

Basically, cycle time is positively associated with the crash rate of slight injury. A longer cycle time is usually accompanied by longer vehicle queues and delays, implying that more vehicles arrive at the intersections during the signal cycle, and more chances conflict with each other and run into slight injury. Plus, longer cycle times may arouse some drivers' emotions, leading to aggressive driving that ends in injuries. However, as shown in Figure 2(b) and Table 6, the crash rate of slight injury is decreased at 75th percentile when the cycle time reaches certain limit. The turning curve is such that the slight injury data mostly concentrate before 75th percentile, implying that the vehicles queues are either adequately long or congested, thus leading to the slight injury down.

The presence of tram stops is positively related to the slight injury and crash rate at all quantiles as shown in Figure 2(c), and Table 6 shows that the influence of presence of tram stops is larger at lower tail than that at higher tail. In Hong Kong Island the tram stops are located in the center of the road, and next to the signalized intersections. Reasonably, more tram stops may possibly increase the conflicts between passengers and automobiles, thus leading to more crashes. However, because the travel speed of tram is relatively low and passengers get aboard and alight at a raised area of tram stops. And therefore, the conflicts between passengers and the automobiles are more commonly attributed to slight injury.

Although the road environment in KLN is significantly and positively related to the crash rate and slight injury in Table 2, clear trend of crash rate is not observed for all the quantiles as shown in Figure 2(d) because the curve fluctuates up and down at 50th and 75th percentiles, respectively. This indicates that the impact of road environment is not stable in Kowloon. The complicated traffic conditions in Hong Kong create more chances for crashes, and the complicated road environment there causes more trouble whether the slight injury or KSI.

KSI model

The average lane width is negatively linked to a higher risk of KSI severity at 25th percentile, which is uniform with previous findings from Xu et al. (2014); however, it becomes positively related to crash rate of KSI beyond the 25th percentile, as shown in the curve in Figure 3(a), which depicts the rising trend. Currently, there is controversy about whether wider or narrower lanes are safer in roadway design, and the turning curve in the figure demonstrates the dispute, but this study suggests that wider lanes are safer, especially for KSI. Thus, it is recommended that the lane width at signalized intersections should be carefully designed to minimize the crash rate of KSI.

Similar to the results of quantile selection model for slight injury, cycle time is positively correlated with crash rate of KSI at signalized intersections before 75th percentile, such that a longer cycle time increases the crash risk of KSI, while the curve begins to reduce after 75th percentile in Figure 3(b), corresponding to the data distribution in Figure 1. For those aggressive drivers, if they know that they will have to wait for a long red light, red-light jumping may increase if they miss the last seconds of the amber light, which is a dangerous maneuver that leads to more serious crashes.

Shown from Figure 3(c), the traffic in Kowloon significantly and positively affects the KSI crash rate for all quantiles, and the curve begins to increase significantly after the 75th percentile. Thousands of people walk Kowloon’s streets every day, increasing the probability of crashes, some of which may be attributed to KSI if pedestrians and the drivers are aggressive. Most of Kowloon is flat and world famous stores are gathered here, attracting tourists from around the world. In contrast, foothills occupy a large proportion of the geography of Hong Kong Island and other areas, and although these areas have some attractions, they attract fewer pedestrians than those in Kowloon.

The KSI severity at signalized intersections is positively sensitive to the proportion of commercial vehicles, in agreement with Xu et al. (2014). As the collisions with or between commercial vehicles usually have a greater force of impact and involve more people than those 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.

The presence of tram stops is positively related to the KSI severity. Similar to the quantile selection model for slight injury, more tram stops may increase the conflicts between passengers and vehicles, leading to more crashes.

To sum up, crash rate can be a predictor in the quantile selection model for slight injury and KSI, and vice versa. In other words, the increase in the crash severity can be reflected from the decrease in crash rate since crash severity is considered as one independent variable in the crash rate function, while the increase of crash rate implies that the crash severity may be reduced, thus the endogeneity can be characterized and verified by the quantile selection model directly.

Conclusions

In this paper, the crash rate and crash severity are modeled to evaluate the safety performance at signalized intersections in Hong Kong, while taking into account the heterogeneity and simultaneity issues. The quantile selection model, to the authors’ knowledge, is by far the first attempt in the literature on the crash rate and crash severity to simultaneously model the safety at signalized intersections. A two-step procedure is used to assess the crash rate and crash severity simultaneously and address the slight injury and KSI separately, and the quantile regression accommodates the heterogeneity (i.e., shared unobserved factors) between signalized intersections, and the Heckman selection framework tackles the endogeneity (between crash rate and crash severity) at signalized intersection. Compared with the Heckman selection model, the proposed quantile selection model offers a more complete view and a highly comprehensive analysis of the relationship between variables, which reflects different effects at different quantiles of the response variable, and fewer assumptions were made for quantile regression, which is helpful to describe the relationship between crash rate and crash severity more naturally.

Generally, the results of the quantile selection model for slight injury indicate that the crash rate is positively correlated with the reciprocal of the turning radius, cycle time, the presence of tram stops and road environment in Kowloon, whereas the slight injury severity is significantly influenced by the AADT, the reciprocal of the turning radius, the presence of tram stops, and the road environment in Kowloon, but show some difference at different quantiles. Regarding the results of the quantile selection model for KSI, cycle times and road environment in Kowloon increase the likelihood of crash rate for KSI whereas the average lane width shows some variation at different quantiles. The proportion of commercial vehicles and the presence of tram stops increase the likelihood of KSI severity. The quantile selection model also addresses the correlation between the crash rate for slight injury/KSI and slight injury/KSI severity respectively, which implies that the unobserved variables are heterogeneous between the signalized intersections in Hong Kong.

Nevertheless, the dataset has its limitations. First, about 555 observations are included and are very limited, thus the estimation results may be more accurate if more observations are involved. The second limitation concerns the explanatory variables. A broader range of explanatory variables could result in statistically significant coefficient estimates, thus more variables should be collected. The third limitation is that the temporal effect at signalized intersections is not addressed, and the model performance may be improved if the temporal effect is integrated into the future dataset for typical signalized intersections.

Acknowledgements

The study is supported by Singapore Ministry of Education Academic Research Fund Tier 1 MOE2017-T1-001-069 and the National Natural Science Foundation of China (No. 71561167001). The third author was also supported by the Francis S Y Bong Professorship in Engineering.

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.

Abdel-Aty, M., Radwan, A.E., 2000. Modeling traffic accident occurrence and involvement. *Accident Analysis and Prevention* 32(5), 633-642.

Agbelie, B.R.D.K., 2016. The effect of gender on two-passenger vehicle highway crash-injury severity: A mixed logit empirical analysis. *Journal of Transportation Safety & Security* 8(3), 280-291.

Agbelie, B.R.D.K., Roshandeh, A.M., 2015. Impacts of signal-related characteristics on crash frequency at urban signalized intersections. *Journal of Transportation Safety & Security* 7(3), 199-207.Aguero-Valverde, J., 2014. Direct spatial correlation in crash frequency models: Estimation of the effective range. *Journal of Transportation Safety & Security* 6(1), 21-33.

Alhamzawi, R., 2015. Model selection in quantile regression models. *Journal of Applied Statistics* 42(2), 445-458.

Alhamzawi, R., 2016. Bayesian model selection in ordinal quantile regression. *Computational Statistics and Data Analysis* 103, 68-78.

Arellano, M., Bonhomme, S., 2017. Quantile selection models with an application to understanding changes in wage inequality. *Econometrica* 85(1), 1-28.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.

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.

Caliendo, C., Guida, M., 2014. A new bivariate regression model for the simultaneous analysis of total and severe crashes occurrence. *Journal of Transportation Safety & Security* 6(1), 78-92.

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.

Chiou, Y. C., Fu, C., 2015. Modeling crash frequency and severity with spatiotemporal dependence. *Analytic Methods in Accident Research* 5, 43–58.

Ding, C., Ma, X., Wang, Y., Wang, Y., 2015. Exploring the influential factors in incident clearance time: disentangling causation from self-selection bias. *Accident Analysis and Prevention* 85, 58-65.

Dong, C., Richards, S.H., Clarker, D.B., Zhou, X., Ma, Z., 2014. Examining signalized intersection crash frequency using multivariate zero-inflated Poisson regression. *Safety Science* 70, 63–69.

El-Basyouny, K., Sayed, T., 2011. A full Bayes multivariate intervention model with random parameters among matched pairs for before-after safety evaluation, *Accident Analysis and Prevention* 43(1), 87–94.

Heckman, J.J., 1979. Sample selection bias as a specification error. *Econometrica* 47,153–161.

Hewson, P., 2008. Quantile regression provides a fuller analysis of speed data. *Accident Analysis and Prevention* 40, 502–510.

Islam, M.B., Hernandez, S., 2016. Fatality rates for crashes involving heavy vehicles on highways: A random parameter tobit regression approach. *Journal of Transportation Safety & Security* 8(3), 247-265.

Kaplan, S., Nielsen, T.A.S., Prato, C.G., 2016. Walking, cycling and the urban form: A Heckman selection model of active travel mode and distance by young adolescents. *Transportation Research Part D-Transport and Environment* 44, 55-65.

Kwon, J., Barkley, T., Hranac, R., Petty, K., Compin, K., 2011. Decomposition of Travel Time Reliability into Various Sources Incidents, Weather, Work Zones, Special Events, and Base Capacity. *Transportation Research Record* 2229, 28-33.

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.

Koenker, R., Bassett, G., 1978. Regression quantiles. *Econometrica* 46, 33-50.

Li, L., Hasnine, M.S., Habib, K.M.N., Persaud, B., Shalaby, A., 2017. Investigating the interplay between the attributes of at-fault and not-at-fault drivers and the associated impacts on crash injury occurrence and severity level. *Journal of Transportation Safety & Security* 9(4), 439-456.

Liu, X., Saat, M.R., Qin, X., Barkan, C.P.L., 2013. Analysis of U.S. freight-train derailment severity using zero-truncated negative binomial regression and quantile regression. *Accident Analysis and Prevention* 59, 87-93.

Leung, S.F., Yu, S., 1996. On the Choice between Sample Selection and Two-part Models. *Journal of Econometrics* 72, 197–229.

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

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

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., 2011. A joint-probability approach to crash prediction models. *Accident Analysis and Prevention* 43(3), 1160–1166.

Qin, X., 2012. Quantile effects of casual factors on crash distributions. *Transportation Research Record* 2219, 40-46.

Qin, X., Ng, M., Reyes, P. E., 2010. Identifying crash-prone locations with quantile regression. *Accident Analysis and Prevention* 42(6), 1531-1537.

Qin, X., Reyes, P. E., 2011. Conditional Quantile Analysis for Crash Count Data. *Journal of Transportation Engineering* 137(9), 601–607.

Schwiebert, J., 2015. Estimation and Interpretation of a Heckman Selection Model with Endogenous Covariates. *Empirical Economics* 49 (2), 675–703.

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 Analysis and Prevention* 43(6), 1979–1990.

Wang, Y., Feng, X., Song, X., 2016. Bayesian quantile structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal* 23(2), 246-258.

Washington, S., Haque, M.M., Oh, J., Lee, D., 2014. Applying quantile regression for modeling equivalent property damage only crashes to identify accident blackspots. *Accident Analysis and Prevention* 66, 136-146.

Wu, H., Gao, L. Zhang, Z., 2014. Analysis of crash data using quantile regression for counts. *Journal of Transportation Engineering* 140(4). DOI: 10.1061/(ASCE)TE.1943-5436.0000650.

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, 1107-1113.

Xu, X., Kwigizile, V., Teng, H., 2013. Identifying access management factors associated with safety of urban arterials mid-blocks: A panel data simultaneous equation models approach. *Traffic Injury Prevention* 14(7), 734-742.

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.

Xu, X., Wong, S.C., Zhu, F., Pei, X., Huang, H., Liu,Y., 2017. A Heckman Selection Model for the Safety Analysis of Signalized Intersections. *PLoS ONE,* 12(7).

https://doi.org/10.1371/journal.pone.0181544.

Ye, X., Pendyala, R.M., Shankar, V., Konduri, K.C., 2013. A simultaneous equations model of crash frequency by severity level for freeway sections. *Accident Analysis and Prevention* 57,140–149.

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, 443–452.

Zhou, B., Kockelman, K.M., 2008. Self-Selection in Home Choice Use of Treatment Effects in Evaluating Relationship between Built Environment and Travel Behavior. *Transportation Research Record* 2077, 54-61.