Comparative Analysis of Bayesian Quantile Regression Models for Pedestrian Injury Severity at Signalized Intersections

Xuecai Xu[a], Xiangjian Luo[b], Daiquan Xiao[a*], S. C. Wong[c]

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

[b]*Shenzhen Urban Transportation Planning Center, Shenzhen, China*

[c]*Department of Civil Engineering, The University of Hong Kong, Hong Kong, China*

[*]Corresponding Author

Abstract: This study intended to (1) investigate the pedestrian injury severity involved in traffic crashes; and (2) address the heterogeneity issue at signalized intersections. To achieve the objectives, Bayesian binary and ordinal quantile regression models were proposed to address the pedestrian injury severity at signalized intersections. The suitability of the proposed methods was illustrated with the Hong Kong dataset from 2008 to 2012 and 376 signalized intersections involving 2090 pedestrian-related crashes are selected. It's found that age, injury location, pedestrian special circumstance, pedestrian contributory and presence of Tram/LRT stops and right turning pocket are significant variables. The results indicated that both Bayesian binary and ordinal quantile regression models not only provide a more comprehensive and in-depth understanding of the relationship between pedestrian injury severity and the explanatory variables, but highlight the heterogeneity issue for the data collected at different locations and different times without many assumptions. The goodness-of-fit of the proposed models outperforms existing mean models, while the Bayesian binary quantile model provides a better fit than the Bayesian quantile regression for ordinal model. The results can benefit the pedestrian facilities improvement/management and guide a much safer pedestrian environment.

Keywords: Pedestrian Injury Severity; Bayesian Binary Quantile Regression; Bayesian Quantile Regression for Ordinal Model; Signalized Intersection

Background and Motivation

Each year about 1.24 million people are killed by traffic crashes, and more than one fifth of these deaths occur among pedestrians all over the world, while in some countries or areas the proportion reaches as high as two thirds. Among these, millions of people are injured in traffic-related crashes while walking. As reported from the Annual Traffic Census, Hong Kong Transport Department in 2017, although the total pedestrian casualties have been reduced during the past ten years,

pedestrian traffic fatalities are on the rise, increasing by about 10 percent each year, and account for about 15 percent of all motor vehicle deaths.

Among the pedestrian casualties, intersections accounts for major proportions instead of mid-block segments (Xu et al., 2016a), thus the good design and capacity of the intersection is an important component of preventing traffic injuries and improving pedestrian safety. Usually a well-designed intersection can help increase the traffic capacity and travel speed, while decreasing the traffic conflicts and the pedestrian injuries. Moreover, most of pedestrian injuries can be controlled and prevented by considering the pedestrian facilities and influencing factors, so it is of significance to identify them and utilize them in practice, especially at well-designed signalized intersections. Therefore, the intention of this study is to identify the influencing factors of pedestrian injury severity at different signalized intersections so as to control the predictable and preventable risks to pedestrians.

Various approaches and perspectives (Savolainen, et al. 2011; Mannering and Bhat, 2014) have been proposed to evaluate the roadway safety in recent years. For the pedestrian-related crashes, different methods, such as on-site investigation, mathematical modeling, simulation, etc., have been employed to evaluate the pedestrian injury severity. Among all these, econometric modeling approaches, which specifically focus on the analysis of injury severity from the perspective of overall safety and its economic implications, reveals considerable promise. Currently, most of the studies about pedestrian injury severity (Zajac and Ivan, 2003; Sze and Wong, 2007; Kim et al., 2008; Eluru et al. 2008; Clifton et al., 2009; Kim et al., 2010; Mohamed et al., 2013; Sasidharan and Menendez, 2014; Yasmin et al; 2014; Xu et al., 2016b; Bhat et al., 2017; Prato et al., 2018; Zhao et al., 2019) belong to mean regression, in which the model assumptions cannot be easily extended to non-central locations and do not always complement the nature real-world data, especially in the case of homoscedasticity (Qin, 2012). A more appropriate and more complete view is required to capture the distributional properties with a broader spectrum than only mean and variance. Quantile regression (QR), a very different method from mean regression, provides a more in-depth understanding of the relationship between the outcome and the explanatory variables.

QR, initially proposed by Koenker and Bassett (1978), has attracted increasing attention in various fields, such as sociology, economics, finance, medical science, etc. (Qin, 2012; Wang et al., 2016; Koenker, 2017). Quantile regression is a powerful tool, more thoroughly than the mean regression, for comparing various aspects (location, scale, and shape) of any kind of distribution of the outcome across different covariate patterns (Orsini and Bottai, 2011).The main advantage of quantile regression is that it does not require the data to follow a specific distribution, but provide multiple variations from several regression curves for different percentage points of the distribution. This would reveal different effects at different quantiles of the response variable. Furthermore, quantile regression is more robust against outliers because the estimation results may be less sensitive to outliers and multi-modality (Liu et al., 2013). More importantly, quantile

regression can address the heterogeneity issue for the data collected from different sources at different locations and different times without many assumptions (Qin et al. 2010; Qin and Reyes, 2011; Qin, 2012), which is beneficial to deal with the unobserved factors at different signalized intersections more accurately.

During the past thirty years, QR has been utilized in various fields and areas(Qin et al. 2010; Qin and Reyes, 2011; Qin, 2012; Wu et al., 2014; Washington et al., 2014), whereas the application in transportation field remains sparse, and one of high-related studies is by Xu et al., (2018). The initial study by Hewson (2008) proposed the potential role of quantile regression for modeling the speed data, and demonstrated the potential benefits of using quantile regression methods, which provided more interest than the conditional mean regression approaches. From the view point of discrete variables, Qin et al. (2010) determined crash-prone locations with quantile regression. The flexibility of estimating trends at different quantiles was provided, and the data with heterogeneity issue was handled. The results indicated that quantile regression can offer a sensible and much more refined subset of risk-prone locations. Continuously, Qin and Reyes (2011) and Qin (2012) analyzed crash frequencies with quantile regression. The heterogeneous crash data were addressed, and a complete view of how the covariates affected the responsible variable from the full range of the distribution was provided, which benefits for the data with heavy tails, heteroscedasticity and multi-modality. The findings suggest that quantile regression estimates can be more informative than conditional mean regression. Similar study by Wu et al. (2014) investigated crash data using quantile regression for counts. The results displayed more detailed information on the marginal effect of covariates change across the conditional distribution of the response variable, and revealed more robust and accurate predictions on crash counts. From the perspective of railway transportation mode, Liu et al. (2013) analyzed the train derailment severity using zero-truncated negative binomial regression and quantile regression, and provided insights for train derailment severity under various operational conditions and by different accident causes. By identifying accident blackspots in a transportation network, Washington et al. (2014) employed quantile regression to model equivalent property damage only (PDO) crashes. The proposed method provided covariate effects on various quantiles of the population and performed better than traditional Negative Binomial (NB) model.

However, when some outcome variables take on non-continuous values, the conventional QR may be inadequate. Manski (1975) presented the general semi-parametric binary quantile regression estimator, and Kordas (2006) explored the binary quantile regression models and verified that the approach can lead to a much richer view of how covariates influence the response variable. Nevertheless, the traditional frequentist approach to binary quantile regression has difficulty optimizing the regression parameters and the confidence intervals around the estimates are problematic. To overcome the drawbacks, a Bayesian approach to binary quantile regression was proposed to handle it appropriately. Yu and Moyeed (2001) presented the general Bayesian quantile regression employing a likelihood function based on the asymmetric Laplace distribution

(ALD), and demonstrated ALD is an effective way for modeling Bayesian quantile regression. Specifically, Benoit and Van den Poel (2012) developed a Bayesian binary quantile regression based on ALD, and the applicability was examined with Monte Carlo experiments. Then Benoit et al. (2013) extended to Bayesian lasso binary quantile regression and other fields. Miguéis et al. (2013) applied Bayesian binary quantile regression to evaluate credit risk and demonstrated that the methodology can be an important tool for credit companies in making decision about credit scoring; Lavín et al. (2016) used Bayesian quantile binary regression approach to estimate payments for environmental services, and Mollica and Petrella (2017) analyzed the Bachelor-Master transition phenomenon with the Bayesian binary quantile regression, whereas Rahaman (2016) introduced a Bayesian estimation method for quantile regression in univariate ordinal models. All the studies characterize the non-continuous features with the Bayesian quantile regression models, which provides the foundation for pedestrian injury severity analysis.

In sum, the attempt of quantile regression for binary and ordinal models within the Bayesian framework to analyze pedestrian injury severity is a pioneer study. Thus, the objective of this study is to investigate the severity of pedestrian injury at signalized intersections, as well as addressing the heterogeneity due to unobserved factors at different signalized intersections.

Data Description

The dataset was integrated from the Traffic Accident Database System (TRADS) with the geo-database of the Traffic Information System (TIS) maintained by the Hong Kong Department of Transport from 2008 to 2012. As described in detail by Sze and Wong (2007), three components from TRADS were included: the crash environment, casualty injuries, and vehicle involvement profiles. All three were converted into a geo-database and displayed in ArcGIS.

To investigate the factors that contribute to pedestrian injury severity, 376 signalized intersections were selected from three areas, Hong Kong Island, Kowloon, and New Territory, involving 2,090 pedestrian-related crashes as shown in Figure 1. In Hong Kong, injury severity is divided into three levels: fatal, serious injury, and slight injury. As required by TRADS, the number of pedestrians who sustained fatal injuries could be merged in the dataset (i.e. fatal cases accounted for only 6.8%), and as the two adjacent injury categories were quite similar, merging the fatal and serious injury categories was not expected to substantially affect the inferences. Consequently, in binary quantile regression model the dependent variable in the proposed model was a dichotomous injury outcome, in which the response of interest referred to killed and serious injury (i.e. KSI), and slight injury was treated as the contrast; For the quantile regression in ordinal model, the injury severity is ordered as 1 for slight injury, 2 for KSI.

To identify the factors that influence pedestrian injury severity and accommodate the unobserved heterogeneity issue among signalized intersections, the dataset was integrated according to the unique intersection IDs in the Arc GIS following the time series from 2008 to 2012, which reflects the demographic characteristics of the pedestrian and traffic characteristics, environmental features,

and geometric design data. The variables included are displayed in Table 1, in which the upper part indicates the proportions of categorical variables and the lower part provides the descriptive statistics of the continuous variables.

Figure 1 *Selected Signalized Intersections in Hong Kong*

Table 1 Summary of the parameters in the pedestrian injury model

Factor	Attribute	Count(proportion)
Year	2008	479(22.9%)
	2009	438(20.9%)
	2010	420(20.1%)
	2011	384(18.4%)
	2012	369(17.7%)

Injury severity	Killed or severe injury	586(28.0%)
	Slight injury	1504(72.0%)
Sex	Male(1)	1091(52.2%)
	Female(0)	999(47.8%)
Age (years)	Under 15(1)	183(8.7%)
	15–65(2)	1504(71.9%)
	Above 65(3)	403(19.4%)
Injury location	Head injury(2)	641(30.7%)
	Others(1)	1449(69.3%)
Pedestrian location	On the crossing(2)	570(27.3%)
	Within 15m of the crossing(3)	1306(62.5%)
	Others(1)	214(10.2%)
Pedestrian action	Crossing road or junction(2)	1116(53.4%)
	Walking along footpath(3)	184(8.8%)
	Others(1)	790(37.8%)
Pedestrian special circumstance	Overcrowded footpath(2)	300(14.4%)
	Obstructed footpath(3)	245(11.7%)
	Others(4)	915(43.8%)
	None(1)	630(30.1%)
Pedestrian contributory	Heedless crossing(2)	417(20.0%)
	Inattentive(3)	250(12.0%)
	Others(4)	698(33.4%)
	None(1)	725(34.6%)
Day of week	Weekday (Monday-Friday)(1)	1553(74.3%)
	Weekend(Saturday-Sunday)(0)	537(25.7%)
Time of day	7:00–9:59AM(1)	295(14.1%)
	10:00AM–3:59PM(2)	751(35.9%)
	4:00–6:59PM(3)	452(21.6%)
	7:00PM–6:59AM(4)	592(28.4%)
Speed limit	Below 50km/h(1)	30(1.4%)
	50km/h(2)	2041(97.7%)
	Above 50km/h(3)	19(0.9%)
Traffic aids	Poor aids(0)	205(9.8%)
	Normal(1)	1885(90.2%)
Traffic congestion	Severe congestion(2)	375(17.9%)
	Moderate congestion(3)	505(24.2%)
	No congestion(1)	1210(57.9%)
Obstruction	At or near obstruction(0)	388(18.6%)

	No obstruction nearby(1)	1702(81.4%)
Junction type	T-junction(2)	819(39.2%)
	Y-junction(3)	34(1.6%)
	Cross-roads(4)	347(16.6%)
	Others(1)	890(42.6%)
Road type	Single-way carriageway(1)	951(45.5%)
	Two-way carriageway(2)	543(26.0%)
	Multi-/dual carriageway(3)	596(28.5%)
Environmental contributory	Pedestrian negligence(2)	25(1.2%)
	Other factors(3)	47(2.2%)
	None(1)	2018(96.6%)
Weather	Clear(2)	1927(92.2%)
	Dull(3)	108(5.2%)
	Fog/mist(4)	38(1.8%)
	Strong wind and unknown(1)	17(0.8%)
Rain	Not raining(2)	1821(87.1%)
	Light rain(3)	215(10.3%)
	Heavy rain(4)	43(2.1%)
	Unknown(1)	11(0.5%)
Natural light	Daylight(1)	1443(69.0%)
	Dawn/dusk(2)	65(3.1%)
	Dark(3)	582(27.9%)
Street light	Good(2)	850(40.7%)
	Poor(3)	12(0.6%)
	Obscured and others(1)	1228(58.7%)
Road surface	Wet(2)	282(13.5%)
	Dry(3)	1800(86.1%)
	Unknown(1)	8(0.4%)
Crossing facility	Traffic signal(2)	846(40.5%)
	Others(3)	1148(54.9%)
	None(1)	96(4.6%)
Presence of tram/LRT stops	Yes(1)	272(13.0%)
	No(0)	1818(87.0%)
Presence of bus stops	Yes(1)	723(34.6%)
	No(0)	1367(65.4%)
Presence of right turning pocket	Yes(1)	218(10.4%)
	No(0)	1872(89.6%)

	Range	Mean	S.D.
Geometric design			
Number of approaches	Min:1; Max:4	3.04	0.79
Number of approach lanes	Min:1; Max:20	7.73	3.57
Number of traffic streams	Min:1; Max:12	4.55	2.04
Average lane width (m)	Min:2.47; Max:6.85	3.41	0.49
Number of pedestrian streams	Min: 0; Max: 10	3.25	1.83
Number of conflict points	Min: 0; Max:46	11.36	7.93
Number of conflict locations	Min: 0; Max:46	10.24	7.43
Signal phasing scheme			
Cycle time	Min:30; Max: 150	103.46	19.69
Number of stages	Min:1; Max:5	2.80	0.91

Note: The numbers in brackets represent the codification system in the dataset when input into the software.

TRADS provide the following variables: the injury characteristics include the time, date, year, severity levels, and injury location; the crash environment involves speed limit, traffic aids, traffic congestion, obstruction, junction/road type, weather, light conditions, road surface, crossing facility, presence of tram/LRT stops, and bus stops, geometric design and signal phasing scheme; TIS gives the pedestrian features as the main variables, including the gender, age, location, action, special circumstance, contributory and the number of pedestrian streams.

Methodology

In this study, although pedestrian injury severity can be considered as binary variable (slight injury or KSI (killed and serious injury), the responses in such a case still have the ordinal meaning (1 for slight injury, 2 for KSI), so when the dependent variable is discrete and outcomes still have the inherent binary or ordinal features, both binary quantile regression model and quantile regression in ordinal model are proposed to find out whether the pedestrian injury severity is better reflected in binary or ordinal features.

Binary Quantile Regression Model

Let $y = (y_1, y_2, \cdots y_n)$ be the vector of the observed binary pedestrian injury severity, and propose the binary quantile regression model originally from Manski (1975) with an arbitrary quantile level $p \in (0,1)$:

$$y_i = I_{[y_i^* \geq 0]} \tag{1}$$

$$y_i^* = x_i \beta_p + \varepsilon_i \tag{2}$$

where I_E denotes the indicator function of the event E, x_i is the vector of explanatory variables, β_p is the vector of regression coefficients, ε_i is a random error and $i=1,\ldots, n$. The quantile regression model for quantitative responses can be described as:

$$Q_{y_i^*|x_i,\beta_p}(p) = x_i\beta_p \tag{3}$$

where $Q_{y_i^*|x_i,\beta_p}(\cdot)$ represents the conditional quantile function y_i^*. If p=0.5, $Q_y(0.5)$ is the conditional median, the value that splits the conditional distribution of the outcome variable into two parts with equal probability.

In the Bayesian framework, the inference is relied on the asymmetric Laplace distribution (ALD), and by using the data augmentation method, quantile regression modeling for the continuous responses can be extended for the treatment of binary response variables (Benoit and Van den Poel, 2017; Mollica and Petrella,2017).

The conventional estimates, also called frequentist $\hat{\beta}_p$, for the quantile regression model can be expressed as the following optimization problem:

$$\hat{\beta}_p = \arg\min \sum_{i=1}^{n} \rho_p(y_i^* - x_i\beta_p) \tag{4}$$

where ρ_p is called loss function. It can be found out that the *p*-th regression quantile coincides with the maximum likelihood estimate under independent ALD for the unobserved error terms, which is needed for the specification of the likelihood in the Bayesian framework. To implement the Bayesian inference, the adoption of Exponential-Gaussian mixture representation of the ALD can be considered as a convenient option. Particularly, let the error ε_i follows a skewed ALD, i.e. $\varepsilon_i \sim ALD(0,1,k)$ where the location, scale and skewness parameters represent 0, 1, and *k*, respectively.

Given the likelihood based on ALD, the posterior distribution is proportional to the product of the likelihood and the prior distribution of the parameters. Since the joint posterior distribution does not have a known tractable form, Markov Chain Monte Carlo (MCMC) method is required for posterior inferences. In this study, Gibbs sampling is employed so that the ALD is represented in terms of location-scale mixture of normal-exponential distribution, i.e.

$$\varepsilon_i = \theta w_i + \tau\sqrt{w_i}\,\mu_i \tag{5}$$

where w_i and μ_i are mutually independent, $w_i \sim Exp(1)$, $\mu_i \sim N(0,1)$, and $Exp(\cdot)$ and $N(\cdot)$ represent the Exponential and Gaussian distribution respectively. The constants (θ, τ) are defined as the followings:

$$\theta = \frac{1-2p}{p(1-p)} \text{ and } \tau^2 = \frac{2}{p(1-p)}$$

After substitution into Equation (2), it can be expressed as:

$$y_i^* = x_i\beta_p + \theta w_i + \tau\sqrt{w_i}\,\mu_i \tag{6}$$

which implies that $y_i^* \left| w_i, x_i, \beta_p \right. \sim N(x_i\beta_p + \theta w_i, \tau^2 w_i)$, so Equation (6) expands the likelihood specification into a hierarchical structure, which is to transfer the normal linear model framework into the quantile regression approach (Mollica and Petrella, 2017). In the discrete case, the observed binary data y with both the vectors $y^* = (y_1^*, ..., y_n^*)$ and $w = (w_1, ...w_n)$ can be augmented and the complete-data likelihood can be described as follows:

$$L_c(\beta_p, y^*, w) = \prod_{i=1}^{n} (y_i I_{[y_i^* \geq 0]} + (1 - y_i) I_{[y^* < 0]}) f(y_i^* \left| w_i, x_i, \beta_p \right.) f(w_i \left| x_i, \beta_p \right.) \tag{7}$$

The model estimation can be conducted in R package bayesQR developed by Benoit and Van den Poel (2017). More details about the Bayesian binary quantile regression can be referred to Benoit and Van den Poel (2017), and Mollica and Petrella (2017).

Quantile Regression in Ordinal Model

The quantile regression in ordinal model can be represented with a continuous latent random variable z_i as follows:

$$z_i = x_i^{'}\beta_p + \varepsilon_i \tag{8}$$

where x_i denotes a *k×1* vector of covariates, β_p denotes *k×1* vector of unknown parameters at the *p*th quantile, ε_i follows an ALD, and *i=1…n*, in which *n* denotes the number of observations. Here the variable z_i is unobserved and concerned with the observed discrete response y_i, which has *J* categories or outcomes, through the cut-point vector γ_p as follows:

$$\gamma_{p,j-1} < z_i \leq \gamma_{p,j} \Rightarrow y_i = j \tag{9}$$

where $\gamma_{p,0} = -\infty$ and $\gamma_{p,J} = \infty$. Here $\gamma_{p,1}$ is set to 0, which anchors the location of the distribution required for parameter identification (Rahman, 2016). For the data vector $\mathrm{y} = (\mathrm{y}_1, \mathrm{y}_2, \cdots \mathrm{y}_\mathrm{n})'$, the likelihood for the model can be described as the function of unknown parameters (β_p, γ_p) as follows:

$$f(\beta_p, \gamma_p; y) = \prod_{i=1}^{n}\prod_{j=1}^{J} P(y_i = j \left| \beta_p, \gamma_p \right.)^{I(y_i=j)} = \prod_{i=1}^{n}\prod_{j=1}^{J} [F_{AL}(\gamma_{p,j} - x_i^{'}\beta_p) - F_{AL}(\gamma_{p,j-1} - x_i^{'}\beta_p)]^{I(y_i=j)} \tag{10}$$

where $F_{AL}(\bullet)$ represents the cumulative distribution function of an AL distribution and $I(y_i = j)$ is an indicator function equal to 1 if $y_i = j$ and 0 otherwise.

Similar to binary quantile regression, the Bayesian method of estimating quantile regression in ordinal model employs the latent variable of Equation (8) combined with Equation (6), so the p^{th} quantile regression in ordinal model can be described as:

$$z_i = x_{i_i}^{'}\beta_p + \theta w_i + \tau\sqrt{w_i}\,\mu_i \tag{11}$$

where w_i and μ_i are the same meaning as in binary quantile regression model, and here the latent variable follows $z_i \left| \beta_p, w_i \sim N(x_i^{'}\beta_p + \theta w_i, \tau^2 w_i) \right.$ with the convenient properties of normal distribution in the estimation procedure. In this study, the estimation is conducted with a simpler algorithm that relies on Gibbs sampling. More estimation methods can be referred to Rahman (2016).

For model comparison, as provided by many other studies under the Bayesian (Haque et al., 2010; Zeng and Huang, 2014; Xu et al., 2016a; Zeng et al. 2017), the Deviance Information Criterion (DIC) is used to compare the models abovementioned:

$$DIC = D(\bar{\theta}) + 2p_D = \bar{D} + p_D \tag{12}$$

where $D(\bar{\theta})$ is the deviance evaluated at $\bar{\theta}$, the posterior mean of the parameter of interest, p_D is the effective number of parameter in the model, and $\bar{D}$ is the posterior mean of the deviance statistic $D(\bar{\theta})$. The lower the DIC, the better the model fits. Generally speaking, differences in DIC of more than 10 definitely rule out the model with the higher DIC; differences between 5 and 10 are considered substantial, while the difference less than 5 indicates that the models are not statistically different.

Results

By applying R package and Stata 15 respectively, both quantile and mean regression parameters are estimated using Bayesian approach to calculate the fitted probabilities for each variable. Before estimating the regression models, the correlation test was conducted to avoid the multicollinearity issues. There are numerous methods that can be used to deal with multicollinearity, among which the Variance Inflation Factor (VIF) is a popular choice (Miguéis et al., 2013). It is generally considered that a strong multicollinearity may exist when a value of VIF is greater than 10, and this can cause projection bias. By applying R package, values of VIF can be calculated. Most variables have a VIF ranging from 1.02 to 5.13 and this is within the acceptable range. However, the VIF for the number of conflict points and conflict locations are 20.5 and 21.6 respectively, which are greater than 10, hence multicollinearity is a considered issue between those two variables.

By the correlation test, the linear coefficient obtained for them is 0.969, and this indicates that there is indeed a collinearity issue between them. Similarly, a high correlation existed between

the time of day and natural light and street light, implying that those three variables should not be included together in the model. In addition, weather, rain, and road surface; the number of approaches, approach lanes, and traffic streams; and the number of pedestrian streams, conflict points, and conflict locations, respectively, were all correlated. Obstruction and traffic aids were also highly correlated, indicating that only one of the two should be included in the model. After the variables were put into the software, the insignificant variables were removed step by step in terms of critical values of 95% confidence interval. Other variables didn't show up in the results because they are not significant for the pedestrian injury severity.

To analyze the regression parameters, 95% Bayesian credible interval (BCI) from the marginal posterior distributions of each parameter is estimated, i.e. a 95% BCI contains the true parameter value with ~95% certainty. If the 95% BCI of the posterior mean does not include 0, it implies that this effect is statistically significant at the 95% level. Table 2 and Table 3 give the estimation results for both models from 12,500 iterations after a burn-in of 2,500 iterations. Shown from Table 2, most variables are significant at 95% BCI for both models, except that obstruction and presence of bus stops are significant for binary regression model while presence of right turning pocket is significant for binary quantile regression model. Furthermore, the DIC values at the 25^{th}, 50^{th}, 75^{th}, 90^{th}, 95^{th} quantile in the binary quantile regression model are much smaller than that in the binary regression model, which indicates that binary quantile regression model is better fit than binary regression model. Amongst the quantile models, the 75^{th} quantile model provides the best fit.

Similarly, in Table 3 the significant variables are the same as in Table 2, and DIC values of quantile models are smaller than that of ordinal probit model. Moreover, the 75^{th} quantile model provides the best fit, which is correct since the distribution of the continuous variable z is skewed and so is the ALD for p=0.75.

From vertical perspective of Table 2 and Table 3, in comparison with the two mean models, the DICs of Bayesian binary probit and ordinal probit models are almost equal to each other, indicating that two models are comparable and there is difference about whether the pedestrian injury severity levels are considered either as binary or ordinal. On the other hand, both Bayesian binary and ordinal quantile regression models are comparable to each other from DIC values, and both the 75^{th} quantile models provide the best fit among all. However, the DIC values in Bayesian binary quantile model are smaller than those in ordinal one, implying that Bayesian binary quantile models fits better, and performs better for pedestrian injury severity.

Table 2 Estimation Results for Bayesian binary quantile and mean regression models

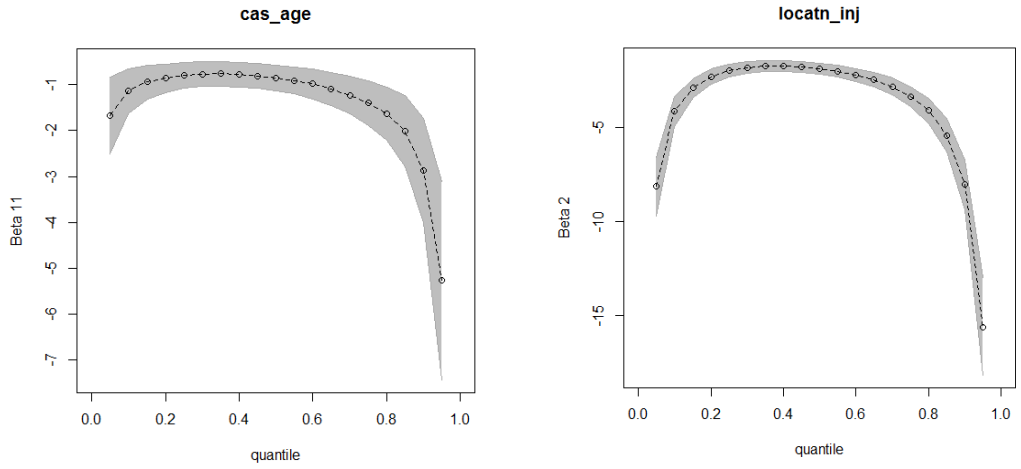
Variable	Binary quantile estimates (95% BCI)					Binary Probit (95% BCI)
	0.25	0.50	0.75	0.90	0.95	
Age	-0.799* (-1.082,-0.524)	-0.848* (-1.135,-0.566)	-1.397* (-1.883,-0.921)	-2.868* (-4.007,-1.724)	-5.263* (-7.432, -3.114)	-0.323* (-0.438,-0.209)
Injury location	-1.970* (-2.317,-1.630)	-1.871* (-2.177,-1.571)	-3.347* (-3.882,-2.829)	-8.039* (-9.374,-6.748)	-15.590* (-18.139,-12.978)	-0.821* (-0.949,-0.694)
Pedestrian special circumstance	-0.290* (-0.407,0.180)	-0.272* (-0.390, -0.158)	-0.458* (-0.667,-0.258)	-1.032* (-1.575,-0.504)	-1.853* (-2.779,-0.857)	-0.120* (-0.171,-0.697)
Pedestrian contributory	-0.234* (-0.352,-0.119)	-0.251* (-0.373,-0.131)	-0.404* (-0.614,-0.193)	-0.861* (-1.353,-0.344)	-1.496* (-2.525,-0.493)	-0.101* (-0.152,-0.052)
Obstruction						0.247* (0.089,0.406)
Presence of tram/LRT stops	-0.748* (-1.194,-0.317)	-0.831* (-1.240,-0.414)	-1.301 * (-1.973, -0.590)	-2.518* (-4.014,-1.005)	-4.753* (-7.673,-1.899)	-0.289* (-0.485,-0.091)
Presence of bus stops						-0.141* (-0.285,-0.001)
Presence of right turning pocket	-0.501* (-0.974,-0.042)	-0.531* (-0.993,-0.058)	-0.937* (-1.679,-0.162)	-2.283* (-4.083,-0.330)	-4.363* (-7.834,-0.470)	
Intercept	2.848* (2.118,3.616)	4.513* (3.707,5.371)	9.829* (8.383,11.432)	24.280* (21.114, 27.73)	45.310* (39.843,50.818)	1.645* (1.441,1.848)
No. of observation			2090			2090
DIC 2062.346		2045.360	2036.859	2053.850	2056.389	2210.567

Table 3 Estimation Results for Bayesian ordinal quantile regression and ordinal probit models

Variable	Ordinal quantile estimates (95% BCI)					Ordinal Probit (95% BCI)
	0.25	0.50	0.75	0.90	0.95	
Age	0.778* (0.534, 1.092)	0.839* (0.568,1.253)	1.379* (0.912,1.388)	2.638* (1.235,4.070)	5.043* (3.041,7.276)	0.329* (0.218,0.439)
Injury location	1.937* (1.643,2.348)	1.834* (1.580, 2.189)	3.343* (2.289,3.828,)	8.004* (6.748,9.374)	15.072* (12.134,18.989)	0.823* (0.715,0.934)
Pedestrian special circumstance	0.283* (0.178,0.482)	0.272* (0.153,0.358)	0.453* (0.253, 0.676,)	1.027* (0.537,1.598)	1.842* (0.837,2.783)	0.116* (0.066,0.164)
Pedestrian contributory	0.234* (0.123, 0.386)	0.216* (0.123,0.381)	0.410* (0.139, 0.641,)	0.873* (0.336,1.367)	1.473* (0.487,2.436)	0.104* (0.052,0.153)
Obstruction						-0.257* (-0.422,-0.092)
Presence of tram/LRT stops	0.775* (0.308,1.176)	0.825* (0.456,1.432)	1.300* (0.593,1.937)	2.504* (1.014,4.083)	4.435* (1.808,7.738)	0.286* (0.117,0.481)
Presence of bus stops						0.141* (0.026,0.260)
Presence of right turning pocket	0.550* (0.046, 0.987)	0.586* (0.052,0.989)	0.935* (0.152,1.683)	2.227* (0.362,4.095)	4.328* (0.483,7.497)	
$\sigma(\mathbf{Std.Dev.})$	4.804* (0.060)	3.452* (0.043)	1.282* (0.015)	3.287* (0.041)	4.319* (0.058)	1.650* (0.099)
No. of observation			2090			2090
DIC 2064.736		2053.421	2045.593	2057.546	2058.354	2209.518

Consequently, the results of Bayesian binary quantile regression model in Table 2 will be enumerated. The variables pedestrian action, traffic aids, number of conflict points, and presence of turning pocket are neglected from the table because their credible intervals overlap the value of zero on several quantile levels. Hence, these variables are not statistically significant for the analysis. The casualty age, injury location, pedestrian special circumstance, pedestrian contributory, presence of tram/LRT stops and presence of right turning pocket are significant variables. As can been seen from Table 2, the coefficients of significant variables are reduced from 25^{th} to 95^{th}, implying that the trend of injury severity is going worse. Each variable's variation can be reflected from Figure 2.

Figure 2 gives the estimates (dotted lines) and the 95 percent confidence bands (shaded gray areas) for the regression coefficients associated with the significant variables for a dense set of quantiles, whereas the mean regression models can't reflect the variables visually. The horizontal line at zero is marked for reference. The figure depicts the information shown numerically in Table 2 for five quantiles and extends it to a larger set of quantiles. The confidence bands allow visual inspection of the import of the sampling error.



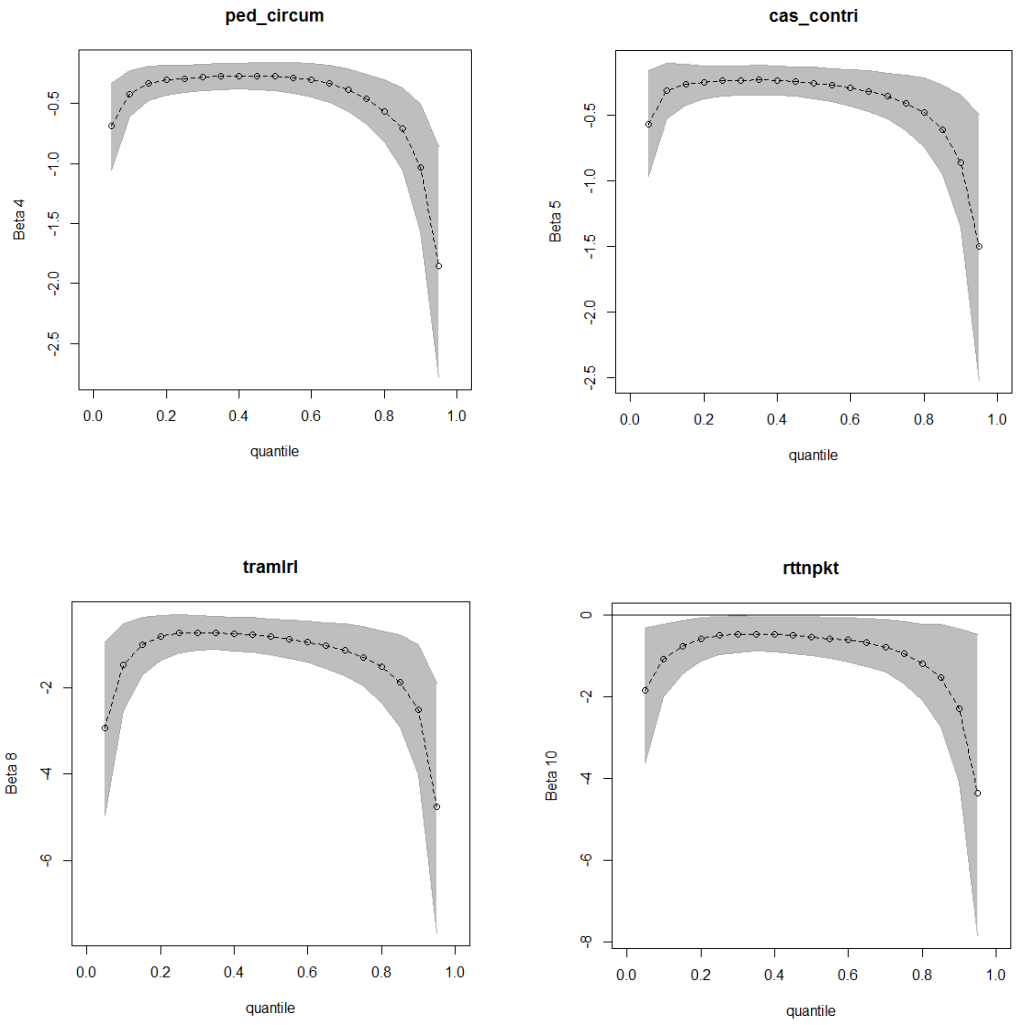


Figure 2 *Regression parameters*

Discussions

Shown from Table 2 and Table 3, the closer examination of the magnitude of the estimated coefficients reveals some similarities and differences between quantiles. First, the similarity is that all the influencing variables are of significance and the coefficients are reduced from low quantile to high quantile, that is to say, the impact of all the variables may not be even, and some are more likely to influence pedestrian injury severity in the low tails than in the high tails. This indicates that certain influencing factors would lead to the specific injury severity trend and need to paid more attention. Secondly, the difference is that significant variables may reveal different impacts on injury severity at different percentiles; e.g., the injury location is significant all through the quantiles, while obstruction is only significant at mean regression models. This implies that from the variables considered in the model, 95% confidence interval may be not

suitable for quantile regression models; thus next step 97% or 99% confidence interval may reveal different patterns.

The variable casualty age (cas_age) shows that the impact is higher in the extreme high quantiles than in the middle and low quantiles, especially after 80^{th} quantile. It can be inferred that pedestrians whose age is over 65 are more vulnerable to serious severity compared to younger pedestrians (the category under 15 is the base), due to their weakness in the field of physiological conditions, perception of safety, and reaction in hazardous situations. The finding is in line with the studies by Pour-Rouholamin and Zhou (2016), and Xu et al. (2016a).

The variable injury location (locatn_inj) has the similar trend on the severity of pedestrian injuries as the age, i.e. the impact is extremely higher in the high quantiles than in the middle and low quantiles. It is interesting to note that the effect is downward after 60^{th} quantile significantly. This indicates that there is a clear correlation between the severity of pedestrian injuries and the injury location, and it can be inferred that head injuries can easily cause severe injuries compared to other locations. This is in line with the actual situation and results in Xu et al. (2016a).

As for the variable pedestrian special circumstance (ped_circum), it can be seen from Figure 1 that the variable has a higher impact in the high quantiles and after 60^{th} quantile the effect begin to be downward obviously, indicating that pedestrian special circumstance in addition to overcrowded and obstructed footpath, other circumstances can easily contribute to severe injuries. Because at signalized intersections, there exists a variety of unknown situations, such as driving through the red light, bicyclists going on the rampage, drunk drivers, etc., all of these may cause more severe injuries, especially in Hong Kong due to the narrow streets and crowded traffic.

Regarding the variable pedestrian contributory (cas_contri), the tendency is more evident for high quantiles by considering the category none as the base. When it comes to the 60^{th} quantile, the effect begins to be downward. This implies that people who are inattentive or heedless when crossing the road can easily suffer from severe injuries.

Concerning the variable the presence of tram/LRT stops (tramlrl), it is obvious that the impact is higher in the extreme high quantiles than in the low and middle quantiles, implying that the presentence of a tram or LRT stops facilities can easily lead to serious traffic injuries. Since most of the tram stops are located in the center of arterials in Hong Kong, the interactions and conflicts between the passengers and vehicles are increased, thus generating more potential probabilities of injury severity, which is consistent with Xu et al. (2016a). Hence, in practice it is suggested to design more alternative facilities connecting the tram/LRT stops with the pedestrianisation.

Different from Xu et al. (2016a), the variable presence of right turning pocket (rttnpkt) is statistically significant and its impact is higher in the high quantiles than in the low and middle

quantiles and after 60th quantile the effect is downward. This indicates that whether there exists right turning pocket or not makes a difference in pedestrian injuries. Compared to absence of right turning pocket, the presentence of right turning pocket are more likely to be related to severe pedestrian injuries when the conflict between turning vehicles and the pedestrians occur.

According to the results obtained, from an empirical point of view, for the pedestrian over the age 65, certain facilities or devices, e.g. barriers, underpass or overpass, and refuge island where necessary, should be set up to help them cross the intersections safely, whether the injury location, special circumstances or inattentive or heedless all should be paid more attention with flashing sign or electronic screens; more alternative facilities should be designed to connect the tram/LRT stops and pedestrianisation; the presence of right turning pocket increases the pedestrian injury severity, thus one way of increasing the safety is to optimize the signal phase to avoid the conflict between right turning and the corresponding pedestrians so that the pedestrian injury severity levels may be reduced.

To measure the performance of the Bayesian binary quantile regression model proposed, we compute the receiver operating characteristic curve (ROC) and analyze the area under that curve (AUC) (Miguéis et al., 2013). An AUC value close to 1 indicates that the model has perfect discrimination, while an AUC value close to 0.5 means that the model has poor discrimination. Figure 3 depicts the ROC and AUC obtained in this study. It can be seen that the AUC value is 0.737. This value is dramatically greater than the null-model benchmark of 0.5, implying that our model performs well in terms of discrimination.

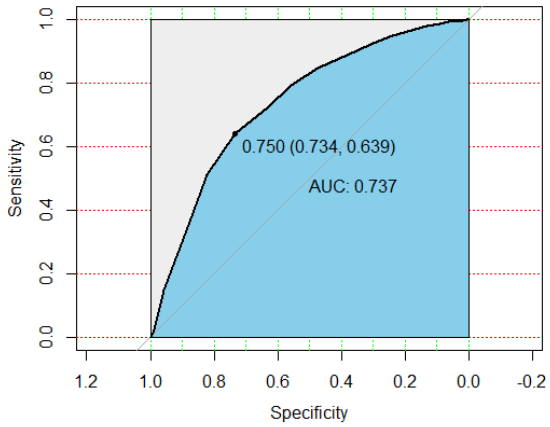


Figure 3 *Sensitivity Analysis Result*

Conclusions

A variety of studies have concerned the pedestrian safety problem at signalized intersections, but quantile regression model has not been widely employed. In this paper we proposed binary and ordinal quantile regression models within Bayesian framework to address the pedestrian injury severity at signalized intersections. This method permits to highlight the heterogeneity issue due to unobserved factors for the data collected at different locations (376 intersections) and different times (5 years) without many assumptions by QR model. From the Bayesian point of view, the inference is addressed in a straightforward manner without depending on the asymptotic properties and computational demanding methods. Moreover, the Bayesian framework allows for an easy derivation of the posterior credible intervals, which provides a clear measure of the uncertainty related to the estimates. The suitability of the method is illustrated with the Hong Kong dataset from 2008 to 2012.

This study adds to the pedestrian injury severity in three aspects. First, by introducing the QR into pedestrian injury severity analysis at the first attempt, both Bayesian binary and ordinal quantile regression models can reveal a more comprehensive and in-depth understanding of the relationship between the outcome and the explanatory variables. Second, the heterogeneity issue due to unobserved factors is addressed at different locations and different times without many assumptions, which avoids the complicated modeling process. Finally, the goodness-of-fit of the proposed models outperforms existing mean models, while the Bayesian binary quantile model provides a better fit than the Bayesian ordinal quantile regression model, both of which provide additional significant variables that are missed with other mean models such as conventional binary probit and ordinal probit regression models.

One concern is that if the killed proportion is separated from KSI, i.e. ordinal models is extended to three levels, whether the Bayesian ordinal quantile regression model reveals the same or different results needs further verification, which may better reflect the pedestrian injury severity. An extension of the present pedestrian injury severity problem could be dealt with by multilevel structure combing with QR model within Bayesian framework, in this way the spatial heterogeneity can be addressed (Fan et al., 2019), which is our next-step work. Additionally, the time series data in this study were not actually utilized to accommodate the temporal instability issue (Mannering, 2018) of pedestrian injuries. After the spatial and temporal issues are integrated, this will broaden the scope of pedestrian injury severity at signalized intersections as well as arterials or corridors, and can guide a much safer pedestrian environment.

Acknowledgement

This study was jointly supported by Fundamental Research Fund for the Central Universities [HUST：2018KFYYXJJ001].

References

Benoit, D.F., Van den Poel, D. (2012). Binary quantile regression: A Bayesian approach based on the asymmetric Laplace distribution. *Journal of Applied Econometrics* 27, 1174-1188.

Benoit, D.F., Van den Poel, D. (2017). bayesQR: A Bayesian approach to quantile regression. *Journal of Statistical Software* 76(7). DOI:10.18637/jss.v076.i07.

Bhat, C.R., Astroza, S., Lavieri, P.S. (2017). A new spatial and flexible multivariate random-coefficients model for the analysis of pedestrian injury counts by severity level. *Analytic Methods in Accident Research*, 16, 1-22.

Clifton, K.J., Burnier, C.V., Akar, G. (2009). Severity of injury resulting from pedestrian-vehicle crashes: What can we learn from examining the built environment? *Transportation Research Part D-Transport and Environment* 14(6), 425-436.

Eluru, N., Bhat, C.R., Hensher, D.A. (2008). A mixed generalized ordered response model for examining pedestrian and bicyclist injury severity level in traffic crashes. *Accident Analysis and Prevention* 40(3), 1033-1054.

Fan, Y., Zhang, G., Ma, J., Lee, J., Meng, T., Zhang, X., Jiang, X. (2019). Comprehensive evaluation of signal-coordinated arterials on traffic safety. *Analytic Methods in Accident Research* 21, 32-43.

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.

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

Kim, J.K., Ulfarsson, G.F., Shankar, V.N., Kim, S. (2008). Age and pedestrian injury severity in motor-vehicle crashes: A heteroskedastic logit analysis. *Accident Analysis and Prevention* 40(5), 1695-1702.

Kim, J.K., Ulfarsson, G.F., Shankar, V.N., Mannering, F.L. (2010). A note on modeling pedestrian-injury severity in motor-vehicle crashes with the mixed logit model. *Accident Analysis and Prevention* 42(6), 1751-1758.

Koenker, R. (2017).Quantile regression: 40 years on. *Annual Reviews of Economics* 9, 155-176.

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

Kordas, G. (2006). Smoothed binary regression quantiles. *Journal of Applied Econometrics* 21(3), 387–407.

Lavín, F.V., Flores, R., Ibarnegaray, V. (2016). A Bayesian quantile binary regression approach to estimate payment for environmental services. *Environment and Development Economics* 22, 156-176.

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.

Mannering, F. (2018). Temporal instability and the analysis of highway accident data. *Analytic Methods in Accident Research* 17, 1-13.

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.

Manski, F.F. (1975). Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics* 3(3), 205-228.

Miguéis, V.L., Benoit, D.F., Van den Poel, D. (2013). Enhanced decision support in credit scoring using Bayesian binary quantile regression. *Journal of Operation Research Society* 64, 1374-1383.

Mohamed, M.G., Saunier, N., Miranda-Moreno, L.F., Ukkusuri, S.V. (2013). A clustering regression approach: A comprehensive injury severity analysis of pedestrian-vehicle crashes in New York, US and Montreal, Canada. *Safety Science* 54, 27-37.

Mollica, C., Petrella, L. (2017). Bayesian binary quantile regression for the analysis of Bachelor-Master transition. *Journal of Applied Statistics* 44(15), 2791-2812.

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.

Orsini, N., Bottai, M. (2011). Logistic quantile regression in Stata. *The Stata Journal* 11(3), 327-344.

Pour-Rouholamin, M., Zhou, H. (2016). Investigating the risk factors associated with pedestrian injury severity in Illinois. *Journal of Safety Research*, 57, 9-17.

Prato, C.G., Kaplan, S., Patrier, A., Rasmussen, T.K. (2018). Considering built environment and spatial correlation in modeling pedestrian injury severity. *Traffic Injury Prevention*, 19(1), 88-93.

Rahman, M.A. (2016). Bayesian quantile regression for ordinal models. *Bayesian Analysis* 11(1), 1-24.

Sasidharan, L., Menendez, M. (2014). Partial proportional odds model-an alternate choice for analyzing pedestrian crash injury severities. *Accident Analysis and Prevention*, 72, 330-340.

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

Sze, N.N., Wong, S.C. (2007). Diagnostic analysis of the logistic model for pedestrian injury severity in traffic crashes. *Accident Analysis and Prevention* 39(6), 1267-1278.

Wang, Y., Feng, X., and 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.

Yu, K., Moyeed, R.A. (2001). Bayesian quantile regression. *Statistics & Probability Letters* 54, 437-447.

Xu, X., Li, Y., Wong, S.C., Zhu, F. (2018). A two-step quantile selection model for safety analysis at signalized intersections. *Journal of Transportation Safety & Security*. https://doi.org/10.1080/19439962.2018.1509919.

Xu, X., Wong, S.C., Choi, K. (2016b). Quasi-induced exposure method for pedestrian safety at signalized intersections. *Journal of Transportation Safety & Security* 8(2), 129-147.

Xu, X., Xie, S., Wong, S.C., Xu, P., Huang, H., Pei, X. (2016a). Severity of pedestrian injuries due to traffic crashes at signalized intersections in Hong Kong: a Bayesian spatial logit model. *Journal of Advanced Transportation* 50, 2015-2028.

Yasmin, S., Eluru, N., Ukkusuri, S.V. (2014). Alternative ordered response framework for examining pedestrian injury severity in New York City. *Journal of Transportation Safety & Security* 6(4), 275-300.

Zajac, S.S., Ivan, J.N. (2003). Factors influencing injury severity of motor vehicle-crossing pedestrian crashes in rural Connecticut. *Accident Analysis and Prevention* 35(3), 369-379.

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

Zeng, Q., Sun, J., Wen, H. (2017). Bayesian hierarchical modeling monthly crash counts on freeway segments with temporal correlation. *Journal of Advanced Transportation.* DOI: 10.1155/2017/5391054.

Zhao, S., Iranitalab, A., Khattak, A.J. (2019). A clustering approach to injury severity in pedestrian-train crashes at highway-rail grade crossings. *Journal of Transportation Safety & Security* 11(3), 305-322.