Accident Analysis and Prevention 42 (2010) 1531–1537

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Identifying crash-prone locations with quantile regression

Xiao Qin[a,*], Marie Ng[b,1], Perla E. Reyes[c,2]

[a] Department of Civil and Environmental Engineering, South Dakota State University, CEH 148, Box 2219, Brookings, SD 57007, United States
[b] Institute for Health Metrics and Evaluation, University of Washington, 2301 5th Avenue, Suite 600, Seattle, WA 98121, United States
[c] Department of Statistics, 1300 University Ave., University of Wisconsin-Madison, Madison, WI 53706-1532, United States

ARTICLE INFO

Article history:
Received 30 March 2009
Received in revised form 17 February 2010
Accepted 16 March 2010

Keywords:
Quantile regression
Heterogeneity
Poisson-gamma
Confidence interval

ABSTRACT

Identifying locations that exhibit the greatest potential for safety improvements is becoming more and more important because of competing needs and a tightening safety improvement budget. Current crash modeling practices mainly target changes at the mean level. However, crash data often have skewed distributions and exhibit substantial heterogeneity. Changes at mean level do not adequately represent patterns present in the data. This study employs a regression technique known as the quantile regression. Quantile regression offers the flexibility of estimating trends at different quantiles. It is particularly useful for summarizing data with heterogeneity. Here, we consider its application for identifying intersections with severe safety issues. Several classic approaches for determining risk-prone intersections are also compared. Our findings suggest that relative to other methods, quantile regression yields a sensible and much more refined subset of risk-prone locations.

Published by Elsevier Ltd.

1. Background

Intersection safety has been one of the primary focuses for national, state and local traffic agencies. At-grade intersections alone account for over 39% of total crashes and 22% of fatal crashes (Potts et al., 2009). Numerous intersection improvement strategies and technologies, ranging from traffic control advancement to innovative geometric design, have been put forth by American Association of State Highway and Transportation Officials (AASHTO), state and local agencies. These strategies and measures vary considerably in terms of their costs. However, their relative efficacy remains unclear.

Given the tightening fiscal conditions, there is increasing urgency to implement safety measures in the most cost-effective manner. The general principle in implementation of safety measures is that safety improvement dollars should be spent on the sites that exhibit an inherently high risk of crash losses and that possess an economically justifiable opportunity for reducing the risky conditions. There are two classic approaches for determining crash-prone locations. One is based on the observed number of crashes, and the other is based on regression analyses. Crash history can be directly used to generate ranks via crash counts, rates, or other variations involving weighted average. On the other hand, regression analyses generate expected numbers of crashes for a site given certain characteristics of interest. One advantage of using regression-based methods is that these methods do not rely strictly on the crash frequencies as in count methods. They also take into consideration attributes associated with the number of crashes. Consequently, we can separate out variations in observations due to sampling error and identify crucial risk factors that predict outcome levels. Furthermore, by modeling the association between determinants and outcomes, we can make comparisons among sites and identify those with unusually high risk of crashes.

One of the most popular approaches for performing regression analysis is by means of general linear models (GLMs) or generalized linear models (GLMs) if the residual is not a multivariate normal distribution, which include ordinary least squares (OLS) as a special case. GLMs are appealing because they provide relatively simple solutions for modeling a wide spectrum of data including counts and proportions (rates). However, they have several limitations. First, they involve restrictive distributional assumptions. Specifically in the classic GLMs framework, the distribution of data must belong to the exponential family. In reality, however, few datasets follow exactly the distributional assumptions. Therefore, poor fit and erroneous parameter estimates are resulted. Second, in GLMs, link functions are required to be correctly specified. Misspecification of link functions can lead to the loss of efficiency in parameter estimates (e.g. Chiou and Muller, 1998). Third, homogeneity is often assumed when fitting GLMs. Violation of homogeneity assumption

* Corresponding author at: Department of Civil and Environmental Engineering, South Dakota State University, CEH 148, Box 2219, Brookings, SD 57006, United States. Tel.: +1 605 688 6355; fax: +1 605 688 6476.
E-mail addresses: xiao.qin@sdstate.edu (X. Qin), marieng@uwashington.edu (M. Ng), preyes@wisc.edu (P.E. Reyes).
[1] Tel.: +1 206 897 2871; fax: +1 206 897 2899.
[2] Tel.: +1 608 263 7329.

0001-4575/$ – see front matter. Published by Elsevier Ltd.
doi:10.1016/j.aap.2010.03.009

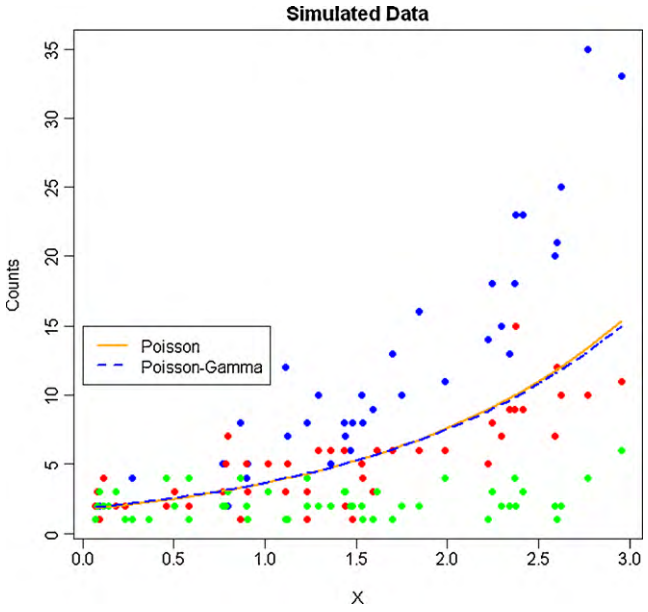

Fig. 1. Scatter plot of heterogeneous data.

can result in overdispersion, which refers to situations in which the observed variance is higher than the theoretical variance. Consequently, the accuracy of parameter and standard error estimates are undermined (e.g. Palmer et al., 2007). Although alternatives such as the quasi-likelihood methods have been proposed to tackle these problems, they only serve as partial solutions, and their finite sample performance can be unsatisfactory (Nelder and Lee, 1992).

In addition to the limitations aforementioned, another issue with GLMs is that they focus mainly on extracting a single trend to summarize the data. For example, when fitting count data, the results from the log-linear (Poisson) model represent the change in *conditional means* across different values of the explanatory variables. In situations where the assumption of homogeneity is met, conditional means generally serve as decent summary statistics of the central tendency. However, in situations where heterogeneity is present, conditional means give an incomplete summary. As an illustration, consider Fig. 1, heterogeneity is present in the data. Specifically, the variability of counts (y) increases as the value of x increases. Moreover, three different trends can be observed. As shown in the graph, neither the Poisson nor the Poisson-Gamma model captures the disparity in trends since both models are designed to capture the conditional means only. In this case, interesting aspects of the data are overlooked.

In fact, data similar to the above example is not uncommon in transportation safety research. In particular, data in many studies are obtained at different times across a wide range of geographical locations. Data collected at the same time or location may exhibit more similarities, whereas data collected at different times and locations may exhibit markedly different characteristics. As a result, the final dataset may have considerable heterogeneity. Recently, noticeable efforts have been made to address the crash data heterogeneity by adopting a random parameter model (Anastasopoulos and Mannering, 2009; El-Basyouny and Sayed, 2009). Instead of assuming that parameters are fixed across all the sites, some or all parameters are allowed to vary in order to account for heterogeneity across observations due to unobserved factors. Using rural interstate highway crash data in Indiana, substantial marginal effects and inferences were found between the models with and without accounting for the random parameters (Anastasopoulos and Mannering, 2009). Using more homogenous dataset grouped by 58 urban arterial corridors, El-Basyouny and Sayed expanded the random parameter models by considering variations through the variance and through the mean, respectively. The finding shows that the Poisson-lognormal models with the random corridor models significantly outperformed the ones without random parameters (El-Basyouny and Sayed, 2009). Both studies support that random parameter count model could be an effective approach to account for data heterogeneity.

In order to gain new insights regarding how covariates affect crash occurrence, it is better to capture the distribution of crash data given these factors. In other words, it is essential to look not only at the mean level but also at different quantiles. Quantiles are points at regular intervals in the cumulative distribution function (CDF) of a random variable. The median is probably the most well-known quantile measure that describes the value separating a population in half (i.e. 0.5 quantile). Both the median and the mean are often used to describe the central tendency of data. However, they differ quite substantially in terms of robustness. Specifically, mean is extremely sensitive to outliers and skewness in a distribution. Hence, it may not always be the best measure of central tendency.

The major goal of this paper is to provide an alternative regression analysis approach known as the quantile regression (QR). QR is a method for estimating how the quantile of an outcome variable changes with respect to the levels of the explanatory variables. It allows us to capture the disparity in trends caused by heterogeneity and offers a more complete picture of the data. This paper is organized as follows. In Section 2, we provide a brief literature review of QR and its application in various disciplines. In Section 3, we outline the basic mathematics underlying QR. In Section 4, we describe the application of QR for modeling crash counts at intersection. In Section 5, we repeat the analysis in Section 4 with the classic GLMs. In Section 6, we compare and contrast the findings in the previous two sections. Finally, in Section 7, we conclude by describing the implications of the findings.

2. Literature review

QR is originated from the econometric literature (Koenker and Bassett, 1978). It is closely related to a classic technique called the least absolute error (LAE). LAE estimates the regression slope by minimizing the sum of absolute residuals. The resulting best fit is the conditional median, which is the 0.5 quantile. (This is in contrast with the ordinary least squares (OLS) which estimates the regression slope by minimizing the sum of squared residuals. The resulting best fit is the conditional mean.) Note that, although LAE has better efficiency than OLS, it remains sensitive to outliers. Specifically, LAE only guards against unusual y values but not unusual x values (i.e. leverage points).

LAE and QR are closely related in the sense that they both involve optimizing certain function of absolute residuals. While LAE involves optimizing a symmetric piecewise linear absolute residuals function, QR involves optimizing an asymmetrically weighted absolute residuals function (see Section 3 for more details). Nonetheless, in contrast with LAE, QR offers the flexibility to capture the trend in data not only at the 0.5 quantile but also at other quantiles. This feature is particularly useful for analyzing dataset where clustering or heterogeneity is present. As a result, in fields like the social and biological sciences, QR is becoming increasingly popular.

In the area of finance, for instance, QR has been applied in value at risk (VAR) modeling. Since market returns often follow a heavy-tailed distribution, models focusing only the conditional means are often inadequate. QR is employed to understand the change in distribution over time (e.g. Taylor, 1999). Similarly, in economics, QR has been used to analyze earnings-related data, which often contain

outliers and the issue of heterogeneity is prevalent (e.g. Arias et al., 2001; Machado and Mata, 2001). As mentioned earlier, QR is particularly useful when data contain subgroups. In a study by Nielson and Rosholm (2001), QR was used to study the determinants of wages in Zambia with special emphasis on the public–private sector wage gap. Their paper nicely presented the differential trends in the entire wage distribution across education and age groups. Another research domain in which QR is becoming increasingly popular is public health. For example, recent studies have applied QR to examine the effect of health care reform (Winkelmann, 2006) and prenatal care utilization (Wehby et al., 2009).

Compared to these research areas, transportation research still has not fully embraced QR. Publication involving this methodology is sparse. The paper by Hewson (2008), which explored the application of quantile smoother for speed data, was perhaps one of the few pioneering studies. Given the data issues exist in transportation studies, QR can be a potentially useful tool in this field.

3. Methodology

3.1. Quantile definition and estimation

Let p be a number between 0 and 1, the $100p$ percentile of the distribution of a continuous random variable X denoted by $\eta(0.5)$ is defined by

$$p = F(\eta(p)) = \int_{-\infty}^{\eta(p)} f(y)dy \tag{1}$$

In general, the p-percentile of the distribution of any random variable X can be rewritten as the inverse function of its cumulative distribution function evaluated at p. Formally, the pth quantile of X with cumulative distribution function F on $\Re$ with $0 \le p \le 1$ is defined as

$$\eta(p) = F^{-1}(p) = \inf\{y : F(y) \ge p\} \text{ where } 0 < p < 1 \tag{2}$$

Note that $\eta(0.5)$ is the median, the 95th percentile is denoted as $\eta(0.95)$ and the commonly used 1st and 3rd quartiles are similarly represented as $\eta(0.25)$ and $\eta(0.75)$, respectively. $\eta(p)$ can be interpreted as the threshold that splits the possible values of X into two groups, such that $P(X \le \eta((p)) = p$ and $P(X > \eta((p)) = 1 - p$.

3.2. Quantile regression model

Similar to the sample mean that minimizes the sum of square errors, the sample median of a random sample $\{y_1, y_2, \ldots, y_n\}$ of a random variable Y is the minimal of the sum of absolute deviations. Therefore, the general pth sample statistics quantile $\eta(p)$ may be solved as a optimal solution to minimizing a weighted average of the samples whose values are larger or equal to $\eta(p)$ and the samples whose values are less or equal to $\eta(p)$ as Eq. (3) (Koenker, 1978).

$$\min_{\beta \in R^k} \left[\sum_{i \in \{i: y_i \ge \eta(p)\}} p \left| y_i - \eta(p) \right| + \sum_{i \in \{i: y_i < \eta(p)\}} (1-p) \left| y_i - \eta(p) \right| \right] \tag{3}$$

the pth sample statistics quantile $\eta(p)$ can be expressed as a linear function of the parameters of interest as

$$\eta(p) = X'\beta + \varepsilon$$

where p: $100p$ percentile such as 95 percentile, 50 percentile (median), etc; $\eta(p)$: the response variable corresponding to $100p$ percentile; $\boldsymbol{\beta}$: k-dimensional vector of unknown parameters of the covariates $\boldsymbol{X}$, and ε: random error.

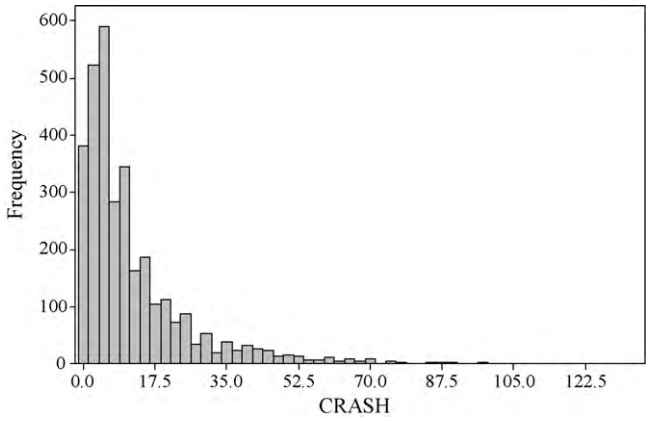

Fig. 2. Crash data histogram.

Hence, the optimization problems become solving the estimates for βs

$$\hat{\beta}(p) = \arg\min_{\beta \in R^k} \left[\sum_{i \in \{i: y_i \ge X'\beta)\}} p \left| y_i - X'\beta \right| + \sum_{i \in \{i: y_i < X'\beta\}} (1-p) \left| y_i - X'\beta \right| \right] \tag{4}$$

For any quantile p between 0 and 1, $\hat{\boldsymbol{\beta}}(p)$ is called the pth regression quantile which minimizes the sum of weighted absolute residuals. As a special case, the sample median minimizes the sum of the absolute errors of the sample set when p is equal to 0.5, which is also called the L1 regression.

In general, QR can be considered as a linear programming problem and can be solved efficiently with various optimization methods such as simplex algorithm, interior point method, smoothing algorithm and their derivations (Chen, 2005). The simplex algorithm is the most popular algorithm, but it is computationally demanding. Its processing time increases considerably as the size of data increases. The interior algorithm has been developed as an alternative for handling large datasets and has been proven to be superior to the simplex algorithm. The smoothing algorithm, on the other hand, is a heuristic approach which aims to improve the estimate through numerous iterations. All three algorithms are implemented in the SAS QUANTREG procedure and details are described in the procedure document (Chen, 2005).

4. Quantile regression application in intersection safety

Identification of crash-prone locations is salient in the assessment of intersection designs and the implementation of safety measures. In this section, we explore the use of QR for modeling intersection crash data, with the specific objective to determine locations with high risk of crashes. QR serves as a desirable option for modeling intersection crash data because the distribution of crash data is often skewed (see Fig. 2). Moreover, due to the data collection process, crash data often show substantial heterogeneity.

Data were obtained from the Wisconsin intersection crash summary statistics (Knapp et al., 2005). Crash data for 1710 intersections in the state of Wisconsin were collected along with other features. We modeled the number of crashes at the intersections based on various attributes including crash characteristics, area type, traffic volume, traffic control, and the geometric features of the intersection. These geometric features included the number of

Table 1
Description of variables.

Variables	Description
TOT0103	Total number of crashes between 01 and 03
AREATYPE	Types of area (rural or urban)
LEG	Number of intersection approach legs (3 or 4)
ENTVEH	Million of annual entering vehicles (MEV)
DIVIDED	Existence of major roadway median
TRFCNTL	Types of traffic controls (all-way, side, other, signal)
LANE	Number of major roadway lanes (2 or 4)
LEFTTURN	Existence of left-turn lane(s)

intersection approach legs (legs), the number of major roadway lanes (lane), the presence of median in major roadway, and the presence of left-turn lane(s) (left turn). Table 1 describes the covariates considered in the study, and Table 2 summarizes the statistics for key variables.

Let y_i denote the number of crash counts at the intersection i. Unlike GLMs, QR does not rely on any distribution assumption. Estimates of slopes are derived by minimizing the optimization function. The logarithm of the response variable y_i (crash count) was used to normalize the data and to achieve more accurate fitting.

$$\min\left[\sum_{i\in\{i:y_i\geq\eta(p)} p|y_i-\eta(p)| + \sum_{i\in\{i:y_i\leq\eta(p)} (1-p)|y_i-\eta(p)|\right] \quad (5)$$

and the link function is

$$\eta(p) = V^{\beta_v}\, exp(\beta_0 + \alpha Z) \text{ and define } \beta = (\beta_0, \beta_v, \alpha')'$$

where p:*100p* percentile such as 95 percentile, 50 percentile (median), etc $\eta(p)$: the expected number of crashes corresponding to p percentile, V: million annual daily entering volume, $\boldsymbol{Z}$: the matrix of categorical variables such as traffic control, median, left turn, area type, $\boldsymbol{\alpha}$: regression coefficient vector.

The primary goal was to use the QR model to identify intersections which exhibit the most severe safety problems. Therefore, we focused our attention on the 95th percentile regression. The QUANTREG procedure was used to estimate the coefficients for the covariates and to perform statistical inferences on the estimated coefficients (Chen, 2005). Table 3 presents the results of the estimated coefficients and 95% confidence intervals. Standard errors and confidence intervals for the QR coefficients were obtained via asymptotic and bootstrap methods respectively. Except for area type (urban or rural) and median (yes or no), all the coefficients are statistically significant at $\alpha = 0.05$ level. The results also suggest that the daily entering volume is a significant predictor of number of crashes. The negative coefficient associated with LEG indicates that 3-legged intersections in general have lower crash frequency than 4-legged intersections. This could be attributed to the fact that 3-legged intersections have fewer conflicting points. Holding all else constant, the results suggest that 3-legged intersections generally have approximately a quarter ($e^{-0.2851}$) fewer number of crashes than 4-legged intersections. Similarly, the coefficient associated with LANE suggests that 2-lane intersections tend to have lower number of crashes compared to 4-lane intersections holding all else constant. As for traffic control type, the results show that relative to signalized intersections both stop-controlled (all-way or side) intersections have significantly fewer crashes than signalized intersections.

Based on the model, intersections with particularly high incidence of crash were identified by comparing the observed counts with the predicted values. Specifically, a site was classified as crash-prone if the observed number of crashes exceeded the prediction. Using this procedure 86 out of 1710 intersections were identified as extreme sites, which constituted approximately 5% of the sample size

Table 2
Summary statistics of variables.

TOT0103		AREATYPE	Sites	LEG	Sites
Min	1	Urban	1151 (67.3%)	3	361 (21.1%)
Mean	12	Rural	559 (22.7%)	4	1349 (78.9%)
Max	134				
SD	14.17				
ENTVEH (MEV)		**DIVIDED**	**Sites**	**TRFCNTL**	**Sites**
Min	0.53	Yes	788 (46.1%)	All-way	25 (1.475%)
Mean	14.73	No	922 (53.9%)	Side	931 (54.45%)
Max	89.87			Other	3 (0.175%)
SD	12.9			Signal	751 (43.9%)
LANE	**Sites**	**LEFTTURN**	**Sites**		
2	697 (40.8%)	Yes	1023 (59.8%)		
4	1013 (59.2%)	No	687 (40.2%)	Total	1710

Table 3
95th regression quantile estimates.

Parameter estimates								
Parameter[a]		DF	Estimate	Standard error	95% confidence limits		t value	$P>\|t\|$
Intercept		1	2.7938	0.1110	2.5760	3.0116	25.16	<0.0001
ENTVEH		1	0.5386	0.0455	0.4494	0.6278	11.84	<0.0001
LEG	3-Legged	1	−0.2827	0.0611	−0.4025	−0.1629	−4.63	<0.0001
LANE	2-Lane	1	−0.1909	0.0603	−0.3092	−0.0725	−3.16	0.0016
TRFCNTL	All-way	1	−0.4680	0.2039	−0.8679	−0.0680	−2.29	0.0219
TRFCNTL	Other	1	−0.8271	0.5728	−1.9506	0.2964	−1.44	0.1490
TRFCNTL	Side	1	−0.3744	0.0584	−0.4888	−0.2599	−6.42	<0.0001

[a] Covariates such as area type, median and left turn lane are removed from the model because they are not statistically significant.

5. Mean regression model estimates

In this section, we repeat the same analysis using the classic GLMs approach. The Poisson-Gamma model was used in this case for the non-negative count data and to tackle the problem of overdispersion. Unlike the QR model, the GLMs estimate the changes of conditional means given the various contributing factors.

5.1. Point estimated parameters of Poisson-gamma model

Let y_i denote the number of crashes at intersection i and the distribution of y_i conditional on its mean λ_i is assumed to follow a Poisson distribution independently over sites.

$$y_i|\lambda_i \sim \text{Poisson}(\lambda_i), \quad i = 1, 2, \ldots, n \tag{6}$$

The log function used to link the mean number of crash counts with all possible covariates and unstructured errors is defined as

$$\log(\lambda_i) = \beta_v \ \log V + \beta_0 + Z\alpha + \varepsilon_i, \quad t = 1, 2, \ldots, k \tag{7}$$

where V: million annual daily entering volume, $\boldsymbol{Z}$: the matrix of categorical variables such as traffic control, median, left turn, area type, α: regression coefficient vector, and ε_i: an unstructured random effect independent of $\boldsymbol{Z}$.

The Poisson-gamma model is specified by assuming that $\exp(\varepsilon_i)$ follows a gamma distribution independently. In most crash prediction literature, it is widely accepted that its mean is 1 and variance is $1/\phi$ for some positive quantity (or parameter) ϕ. In other words,

$$\exp(\varepsilon_i) \sim \text{Gamma} \ (\phi, \phi) \tag{8}$$

and ϕ is usually called an inverse dispersion parameter. Let $\boldsymbol{\beta} = (\beta_0, \beta_v, \boldsymbol{\alpha}')'$, based on this particular parameterization, y_i follows a negative binomial distribution with mean $\exp(\boldsymbol{x}\beta)$ and variance $\exp(\boldsymbol{x}\beta)(1 + \exp(\boldsymbol{x}\beta)/\phi)$. Here, $\boldsymbol{\beta}$ is estimated via SAS GENMOD procedure (SAS 9.1, 2003) and the results are listed in Table 4.

A comparison of Tables 3 and 4 indicates that the estimated coefficients yielded by the QR and GLMs are identical in signs but vary slightly in magnitude. This suggests that the relationship between crash counts and the covariates follows the same pattern at the 95% percentile and the mean level. For instance, high daily entering volume is associated with more crashes at both the 95% percentile and mean levels. Nonetheless, the difference in the values of the coefficients implies that the magnitude of change in crash incidence per unit increase in traffic volume is slightly less at the 95% percentile than at the mean level.

5.2. Confidence intervals for means

To identify locations with high incidence of crash, the observed counts were compared with the 90% confidence intervals for the predicted value. Confidence intervals are computed as follows:

$$X'\beta \pm t_{n-p}^{\alpha/2} S \nu_0^{1/2} \tag{9}$$

where $\nu_0 = x_0'(X'X)^{-1}x_0$ and $\alpha = 0.1$. This limit indicates the range of possible true values for the conditional mean crash incidence given X. A site was classified as crash-prone if the observed number of crashes exceeded the *upper* 90% confidence limits. For the situation at hand, 621 out of 1710 intersections were classified as extreme sites, which constituted approximately 64% of the sample size.

The procedure described is commonly used for detecting outliers in the dependent variable (Laughlin et al., 1975; Agrawal and Lord, 2006; Geedipally and Lord, 2008; Lord, 2008; Wood, 2005). However, caution must be taken when applying this strategy. First, the presence of extreme observations can inflate the estimated variance (S), resulted in masking (i.e. the detection of outliers being hindered by their very presence). Second, this procedure offers little control over the number of observations declared as extreme values. Therefore, situations may occur in which more observations than desired are classified as extreme cases.

6. Comparative analysis and discussion

In this section, we compare the rankings yielded by four methods, namely ranks based on observed crash counts only (henceforth count method), ranks based on predicted mean crash counts from Poisson-Gamma GLMs (henceforth mean method), ranks based on the differences between observed crashes and the 90% upper confidence limits of the predicted values derived from GLMs (henceforth CI limit method), and ranks based on the differences between observed crashes and predicted crashes derived from 95th quantile QR (henceforth QR method).

All 1710 intersections were ranked. Fifteen sites were selected for illustration purpose. Table 5 shows the crash rankings yielded by the methods. The selected sites are ranked as the top 15 crash-prone locations based on the QR method. Similar ranking is obtained by the CI limit method. Overall, more than half of the sites categorized as top 15 by the QR method are similarly identified by the CI limit approach. Among which five of the sites are given the same ranks by both methods. However, some discrepancies exist. For instance, Site N, which is ranked 14th by the QR method is ranked 71st by the CI limit approach.

On the other hand, remarkably different rankings are obtained by the count and the mean methods. For example, Site E, which is ranked 5th by the QR method is ranked 131st and 1171st by the count and the mean methods respectively. It is also interesting to note that among the four methods, ranks yielded by the mean method generally tend to be low. For instance, Site A, which is ranked 1st unanimously by the count, CI limit and QR methods, is ranked 16th based on the mean method. One possible explanation for the unsatisfactory performance of the mean method is that it failed to account for the presence of heterogeneity in the data. The

Table 4
Regression parameter estimates.

Analysis of parameter estimates								
Parameter[a]		DF	Estimate	Standard error	Wald 95% confidence limits		Chi-square	$P >$ ChiSq
Intercept		1	1.7008	0.0696	1.5644	1.8373	596.99	<0.0001
ENTVEH		1	0.7005	0.0284	0.6449	0.7561	610.27	<0.0001
LEG	3-Legged	1	−0.2851	0.0389	−0.3613	−0.2089	53.76	<0.0001
LANE	2-Lane	1	−0.1143	0.0369	−0.1866	−0.0420	9.61	0.0019
TRFCNTL	All-way	1	−0.2936	0.1278	−0.5440	−0.0432	5.28	0.0216
TRFCNTL	Other	1	−0.4518	0.4328	−1.3001	0.3965	1.09	0.2966
TRFCNTL	Side	1	−0.4255	0.0356	−0.4954	−0.3557	142.72	<0.0001
Dispersion		1	0.2646	0.0120	0.2411	0.2880		

[a] Covariates such as area type, median and left turn lane are removed from the model because they are not statistically significant.

Table 5
Ranking comparisons.

Site	Observed crashes	Rank by count	Predicted mean crash counts from GLM	Rank by mean	Difference from upper 90% CI limit	Rank by CI limit	Difference from the predicted 95th quantile	Rank by QR
A	134	1	47.65	16	83.46	1	47.78	1
B	130	2	50.90	11	75.81	2	39.30	2
C	87	11	26.11	313	59.77	3	32.71	3
D	97	7	38.22	93	56.90	5	24.23	4
E	42	131	8.72	1177	32.76	27	21.89	5
F	117	3	54.21	4	59.06	4	21.79	6
G	39	162	6.81	1377	31.63	28	21.71	7
H	86	13	32.83	171	51.71	8	21.25	8
I	88	10	34.76	134	51.64	9	20.34	9
J	98	6	43.15	48	52.47	7	18.11	10
K	36	190	7.55	1305	28.02	37	18.01	11
L	110	4	52.11	7	54.44	6	17.64	12
M	27	292	4.17	1604	22.51	60	16.30	13
N	24	361	3.56	1640	20.14	71	13.90	14
O	53	65	17.09	669	34.90	20	13.81	15

predicted value substantially underestimated the true number of crashes, resulted in questionable rankings.

Although the CI limit approach yields more plausible rankings, it has the tendency to select too many sites as high-risk locations. As mentioned earlier, 621 out of 1710 intersections were classified as extreme sites, which constituted approximately 64% of the sample size. Among which 220 were sites with relatively low observed number of crashes (less than or equal to 15 crashes). This reflects the lack of specificity by the CI limit method. In other words, this method may not adequately provide a clear indication of *the most* problematic sites.

Conversely, the QR method offers a more specific selection of crash-prone sites. Using the QR method, 86 out of 1710 intersections were classified as extreme sites, which constituted approximately 5% of the sample size. Although minor discrepancies are present between the QR ranking and observed number of crashes, the QR method appears to provide a more specific and discriminative selection of sites.

7. Conclusions

Identifying the locations with high risks of accidents is crucial for the planning of transportation policies and the implementation of safety measures. Many methods are available for determining crash-prone sites. Nevertheless, given that many factors contribute to crash incidences, regression-based approaches offer a means to predict and classify crash-prone locations based on the relevant factors. Unfortunately, the current practices in road safety modeling lack the flexibility and capability to handle heterogeneity and other data issues. In this paper, we illustrated the use of quantile regression for identifying crash-prone intersections. QR is more favorable to traditional regression approaches in a sense that it does not involve any distributional assumption concerning the error and is less sensitive to violation of distributional assumptions. Moreover, by providing estimates at different quantile levels, QR has the capacity to capture heterogeneity in data and present a more well-rounded description of the trends.

Here, we applied QR on the 95th quantile of the intersection crash data. More specifically, we estimated the association between the 95th quantile of the intersection crash data and various factors. These factors included daily entering volume, number of legs and number of lanes on the intersection main approaches, as well as traffic control types. The predicted values from the QR model were used to identify locations which exhibited severe safety issues. We compared the ranking derived from this method with the count method and two GLMs-based approaches including the mean method and the CI limit method.

The number of locations identified by the QR method accounted for approximately 5% of the total sample. These outcomes were fairly consistent with the count and CI limit methods. However, we believe that QR is preferable to the other methods in practice for at least two reasons. First, unlike the count method, which focuses only on crash frequencies, QR offers insights into the risk of crash given certain location-specific characteristics. In other words, it not only enables us to identify sites which exhibit an overall high number of crashes, it also allows us to distinguish sites which exhibit unusually high number of crashes *within its class*. Second, compared with the CI limit method, which lacks specificity in sites selection, QR provides a more discriminative set of crash-prone locations. Having a highly specific set of crash-prone locations can greatly enhance the process of priority settings.

To be comprehensive, we shall note that QR is not without its limitations. For example, for certain discrete dataset, such as binary data, in which the sample objective function is non-differentiable, estimation of conditional quantiles may become problematic. A recent paper by Machado and Santos Silva (2005) proposed the use of jittering to impose smoothness to the data. They offered a framework for estimating conditional quantiles for a range of GLMs, which could be potentially useful in transportation research. Further research is needed to investigate the application of the new method.

In summary, given the various data issues in transportation studies, quantile regression serves as a useful tool for attaining more accurate estimation. Using the intersection crash data, we have shown that QR allows us to determine crash-prone locations with more specificity. Through this brief introduction and demonstration of the application of QR, we hope to encourage researchers to explore the use of this modern statistical method.

Acknowledgement

The authors would like to thank Maria (Fenghuan) Wang for assisting with the literature review.

References

Agrawal, R., Lord, D., 2006. Effects of sample size on the goodness-of-fit statistic and confidence intervals of crash prediction models subjected to low sample mean values. Transportation Research Record 1950, 35–43.

Anastasopoulos, P., Mannering, F., 2009. A note on modeling vehicle-accident frequencies with random-parameters count models. Accident Analysis and Prevention 41 (1), 153–159.

Arias, O., Hallock, K.F., Sosa-Escudero, W., 2001. Individual heterogeneity in the returns to schooling: instrumental variables quantile regression using twin data. Empirical Economics 26, 7–40.
Chen, C., 2005. An Introduction to Quantile Regression and the QUANTREG Procedure. SAS Institute Inc, Cary, NC.
Chiou, J.M., Muller, H.G., 1998. Quasi-likelihood regression with unknown link and variance functions. Journal of American Statistical Association 93, 1376–1387.
El-Basyouny, K., Sayed, T., 2009. Accident prediction models with random corridor parameters. Accident Analysis and Prevention 41 (5), 1118–1123.
Geedipally, S.R., Lord, D., 2008. Effects of the varying dispersion parameter of Poisson-gamma models on the estimation of confidence intervals of crash prediction models. Transportation Research Record 2061, 46–54.
Hewson, P., 2008. Quantile regression provides a fuller analysis of speed data. Accident Analysis and Prevention 40, 502–510.
Knapp, K., Campbell, J., Kienert, C., 2005. Intersection Crash Summary Statistics for Wisconsin. Traffic Operations and Safety Laboratory, University of Wisconsin-Madison, Department of Civil and Environmental Engineering.
Koenker, R., Bassett, G., 1978. Regression quantiles. Econometrica 46, 33–50.
Laughlin, J.C., Hauer, L.E., Hall, J.W., Clough, D.R., 1975. NCHRP Report 162: methods for evaluating highway safety improvements. National Research Council, Washington, DC.
Lord, D., 2008. Methodology for estimating the variance and confidence intervals for the estimate of the product of baseline models and AMFs. Accident Analysis and Prevention 40 (3), 1013–1017.
Machado, J.A.F., Mata, J., 2001. Earning functions in Portugal 1982–1994: evidence from quantile regressions. Empirical Economics 26, 115–134.
Machado, J.A.F., Santos Silva, J.M.C., 2005. Quantiles for counts. Journal of American Statistical Association 100 (472), 1226–1237.
Nelder, J.A., Lee, Y., 1992. Likelihood, quasi-likelihood and pseudolikelihood: some comparisons. Journal of the Royal Statistical Society, Series B (Methodological) 54 (1), 273–284.
Nielson, H.S., Rosholm, M., 2001. The public-private sector wage gap in Zambia in the 1990s: a quantile regression approach. Empirical Economics 26, 169–182.
Palmer, A., Losilla, J.M., Vives, J., Jimenez, R., 2007. Overdispersion in the Poisson regression model. Methodology 3 (3), 89–99.
Potts, I.B., Hutton, J.M., Harwood, D.W., 2009. Strategic Intersection Safety Program Guide. Publication No. FHWA-SA-09-004.
SAS Institute Inc, 2003. GENMOD Procedure. SAS Institute Inc, Cary, NC.
Taylor, J., 1999. A quantile regression approach to estimating the distribution of multiperiod returns. Journal of Derivatives. 24, 64–78.
Wehby, G.L., Murray, J.C., Castilla, E.E., Lopez-Camelo, J.S., Ohsfeld, R.L., 2009. Quantile effects of prenatal care utilization on birth weight in Argentina. Health Economics 18 (11), 1307–1321.
Winkelmann, R., 2006. Reforming health care: evidence from quantile regressions for counts. Journal of Health Economics 25, 131–145.
Wood, G.R., 2005. Confidence and prediction intervals for generalized linear accident models. Accident Analysis and Prevention 37 (2), 267–273.