Conditional Quantile Analysis for Crash Count Data

Xiao Qin, Ph.D., P.E., M.ASCE[1]; and Perla E. Reyes[2]

Abstract: Crashes are important evidence for identifying deficiencies existing in highway systems, but they are random and rare. The investigation of the nature of the problem normally draws on crashes collected over a multiyear period and from different locations to obtain a sizable sample. Hence, the issue of data heterogeneity arises because the pooled data originated from different sources. Data heterogeneity has to be addressed to obtain stable and meaningful estimates for variable coefficients. A desirable method of handling heterogeneous data is quantile regression (QR) because it focuses on depicting the relationship between a family of conditional quantiles of the crash distribution and the covariates. The QR method is appealing because it offers a complete view of how the covariates affect the response variable from the full range of the distribution, which is of particular use for distributions without symmetric or normal forms (i.e., heavy tails, heteroscedasticity, multimodality, etc.). Crash data possess some of the properties that quantile analysis can handle, as demonstrated in an intersection crash study. The compelling results illustrate that conditional quantile estimates are more informative than conditional means. The findings provide information relative to the effect of traffic volume, intersection layout, and traffic control on crash occurrence. **DOI: 10.1061/(ASCE) TE.1943-5436.0000247.** © *2011 American Society of Civil Engineers.*

CE Database subject headings: Data analysis; Data collection; Traffic accidents; Intersections.

Author keywords: Quantile regression; Data heterogeneity; Intersection crashes.

Introduction

Crash data are important and valuable source for traffic safety research because they measure the safety performance of an entity and disclose the relationship between a crash and its cause. Crashes are rare events, so often individual locations do not have adequate data for drawing a valid and explicit conclusion. To obtain a large sample, crash data are often pooled from a wide range of geographic locations and at different times in order to enhance the analysis. Data collected at the same time and location may exhibit similarities, whereas data collected at different times and locations may exhibit markedly different characteristics. As a result, although panel data have several advantages over the cross-sectional or time-series data, they may contain considerable heterogeneity. Heterogeneity means that the variance of the dependent variable changes from observation to observation; it may change as the independent variable changes. The variance of the observed dependent variable may be higher or lower than expected, indicating that the data are overdispersed or underdispersed. The accuracy of the estimates is compromised by not considering data heterogeneity. Furthermore, the statistics used to test the hypothesis under the Gauss-Markov assumption are no longer valid.

Data heterogeneity implies that data may originate or be collected from different sources. If so, generalized linear modeling relying on the conditional mean across different values of the independent variables to describe the data central tendency is less efficient because of the violation of homogeneity. The distribution of data therein becomes more important and informative than just the mean and the variance. Quantiles offer a more complete view of data from a broad spectrum. With the quantile regression (QR) methodology, it is possible to know the impact of the regressors on each quantile of the distribution. The major goal of this paper is to introduce an alternative modeling approach—quantile regression—to handle the heterogeneity issue in crash count data and to demonstrate the resulting disparity in crash trends.

Literature Review

Crash count data are usually overdispersed for a variety of reasons, including omission of important variables, link function misspecification, and structured error term. In fact, crash data overdispersion results from Bernoulli trials with unequal probability of independent events, or crashes (Lord et al. 2007). The different modeling alternatives suggested for accounting data over dispersions are as follows:

- Capturing heterogeneous crash count data by means of finite mixture regression models (Park and Lord 2009).
- Adopting a well-defined mean function (Mitra and Washington 2007) or improving the structure of the dispersion parameter ϕ by replacing a fixed value with a varying one (Miaou and Lord 2003; Miranda-Moreno et al. 2005; Geedipally and Lord 2008).

Most of the crash count models developed over the past two decades depend on generalized linear models (GLM). These are appealing because they provide relatively simple solutions for modeling a wide spectrum of data without the restriction of a multivariate normal distribution. The most representative GLMs in the discipline of traffic safety studies are Poisson and Poisson-gamma. These models have been used extensively for various data sets, many of which are imperfect in satisfying the modeling requirements. Several limitations impede GLMs from modeling complicated data in an effective fashion. First, GLMs involve restrictive distributional assumptions. Specifically in the classic

[1]Assistant Professor, CEH 148, Box 2219, Dept. of Civil and Environmental Engineering, South Dakota State Univ., Brookings, SD 57007 (corresponding author). E-mail: Xiao.Qin@sdstate.edu

[2]Dept. of Applied Mathematics and Statistics, Univ. of California, Santa Cruz, 1156 High Street M/S SOE2, Santa Cruz, CA 95064. E-mail: perla@soe.ucsc.edu

Note. This manuscript was submitted on February 16, 2010; approved on October 27, 2010; published online on August 15, 2011. Discussion period open until February 1, 2012; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Transportation Engineering***, Vol. 137, No. 9, September 1, 2011. ©ASCE, ISSN 0733-947X/2011/9-601–607/$25.00.

GLM framework, the distribution of data must belong to the exponential family. In reality, few data sets follow the distributional assumptions exactly. Second, GLMs require correctly specified link function. Misspecification of link functions can lead to the loss of efficiency in parameter estimates (e.g., Chiou and Müller 1998; Mitra and Simon 2007). Third, homogeneity is often assumed when fitting GLMs. Violation of the homogeneity assumption can undermine the accuracy of parameter and standard error estimates (Palmer et al. 2007). The proposed alternatives, such as quasi-likelihood methods to tackle these problems, only serve as partial solutions, and their finite sample performance can be unsatisfactory (Nelder and Lee 1992). Additionally, GLMs focus mainly on extracting a single trend to summarize the data. They provide little information about a conditional distribution for which features depending on the independent variables other than location of the distribution (mean) are missing. 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.

The distribution of data can be reviewed by looking at quantiles, which 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; it describes the value separating a population in half (i.e., 0.5 quantile). Both the median and mean are often used to describe the central tendency of data; however, they differ substantially in terms of robustness. Specifically, the mean is extremely sensitive to outliers and skewness in a distribution. Hence, it may not always be the best measure of central tendency. Relating quantiles of the dependent variable with the independent covariate facilitates the origination of QR methodology. The groundbreaking work of estimating conditional quantile functions by Koenker and Bassett (1978) has inspired the prolific growth and development of QR in areas such as econometrics, social science, finance, and public health. In finance, QR has been applied to modeling value at risk (VAR) because market returns often follow a heavy-tailed distribution. The conditional means are often inadequate, but QR models are more appropriate for tracking the change in distribution over time (Taylor 1999). In social economics, especially studies of the factors affecting the wage structure, quantile regression has been widely used (Fizenberger et al. 2001; Gonzalez and Miles 2001; Garcia et al. 2001). For instance, Nielson and Rosholm (2001) studied 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. Machado and Mata (2001) conducted research on returns to education in Portugal. They found that returns to education were higher at high quantiles. In addition, the difference in the returns at the top and bottom of the wage distribution increased during the study period. Buchinsky (2001) discovered that returns to education for women varied across quantiles, age groups, and cohorts in the United States through a study over a period of two decades. Similar, in public health, 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 few pioneering studies include a paper by Hewson (2008), which explored the application of a quantile smoother for speed data, and a paper by Qin et al. (2010), which utilized QR to determine crash risk–prone locations. Given the data issues existing in transportation studies, QR can be a potentially useful tool.

Methodology

Before introducing the QR methodology, it is necessary to recall the state of the practice for crash count models. Crash frequency is often assumed to follow a Poisson distribution, as in Eq. (1):

$$N_i|\mu_i \sim \text{Poisson}\,(\mu_i) \quad \text{and} \quad \log(\mu_i|\mathbf{X_i} = \mathbf{X_i}\beta + \varepsilon_i \tag{1}$$

where N_i = number of crashes at site i; μ_i = expected number of crashes at site i; $\mathbf{X_i}$ = vector of the covariates of site i; β = vector of the unknown parameters for covariates; ε_i = random error; and $\exp(\varepsilon_i) \sim \text{Gamma}\,(\phi, \phi)$.

The exponential form of the random error follows a single-parameter gamma distribution, which accounts for the data heterogeneity across the sites. Hence, after considering the random error ε_i, the marginal distribution of N_i becomes a Poisson-gamma distribution, also a negative binomial distribution. Because a negative binomial distribution is not restricted by the assumption of equal variance and mean applied to Poisson, it is more appropriate for data with overdispersion. For this type of model, Eq. (1) clearly indicates that the relationship between the conditional mean crashes $\mu_i|\mathbf{X_i}$ and the covariates is the core interest of the regression model.

On the other hand, QR is interested in estimating conditional quantiles. It is extremely useful for the data that exemplify irregular distributions such as overdispersion, underdispersion, heavy tail or compressed tail of the distribution, or even multimodality. The explicit investigation of the stochastic relationship among the dependent variable and covariates regards QR as a more informative empirical analysis. The rest of this section introduces the concepts of quantiles, quantile regression and its estimation, and a jittering algorithm that converts step functions of CDF to a continuous CDF.

Quantiles

Let p be a number between 0 and 1. The $100p$ percentile of the distribution of a continuous random variable X denoted by $\eta(p)$ is defined in Eq. (2).

$$p = F(\eta(p)) = \int_{-\infty}^{\eta(p)} f(y)dy \tag{2}$$

In general, the p-percentile of the distribution of any random variable X can be written as the inverse function of its CDF evaluated at p. Formally, the pth quantile of X with cumulative distribution function F on $\Re$ with $0 \leq p \leq 1$ is defined as Eq. (3):

$$\eta(p) = F^{-1}(p) = \inf\{y : F(y) \geq p\} \quad \text{where } 0 < p < 1 \tag{3}$$

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

Quantile Regression Model

Like the mean that minimizes the sum of square errors, the 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 given X, $\eta_Y(p|\mathbf{X})$ may be solved as an optimal solution to minimizing a weighted average of the samples whose values are larger or equal to $\eta_Y(p|\mathbf{X})$ and the samples whose values are less than or equal to $\eta_Y(p|\mathbf{X})$ (Koenker 1978), as formulated in Eq. (4).

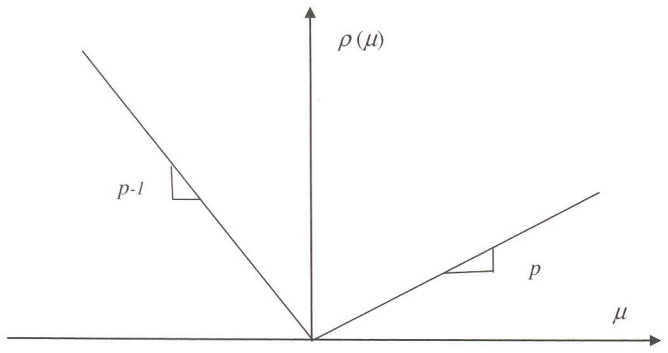

Fig. 1. Quantile regression function

$$\min_{\beta \epsilon R^k} \left[\sum_{i \epsilon \{i: y_i \geq \eta_Y(p|\mathbf{X})\}} p|y_i - \eta_Y(p|\mathbf{X})| + \sum_{i \epsilon \{i: y_i < \eta_Y(p|\mathbf{X})\}} (1-p)|y_i - \eta_Y(p|\mathbf{X})| \right] \quad (4)$$

The graphic presentation of Eq. (4) is shown in Fig. 1.

Following Fig. 1, the bracket in Eq. (4) can be simplified as $\sum \rho_p(\mu) = \sum \mu(p - I_{(\mu<0)})$, where $\mu = y_i - \eta_Y(p|\mathbf{X})$. If the pth sample statistics quantile $\eta_Y(p|\mathbf{X})$ is a linear function of the parameters of interest, it can be solved efficiently by linear programming methods. Because crash count has to be nonnegative, a logarithmic transformation is imposed on the quantiles. The log transformation allows comparison of the QR estimates with the conditional mean estimates obtained in Eq. (1). Note that the quantiles are equivalent with respect to any monotone increasing transformation such as logarithmic transformation, so the transformed random variable $\eta_{\log(Y)}(p|\mathbf{X})$ is equal to $\log[\eta_Y(p|\mathbf{X})]$ (Koenker and Hallock 2001) in Eq. (5).

$$\log[\boldsymbol{\eta}\mathbf{Y}(\mathbf{p}|\mathbf{X})] = \mathbf{X_i}\beta + \boldsymbol{\epsilon}\mathbf{i} \quad (5)$$

where $p = 100p$ percentile, such as 95th percentile, 50th percentile (median), etc.; $\eta_Y(p|\mathbf{X_i})$ = response variable corresponding to $100p$ percentile; β = k-dimensional vector of unknown parameters of the covariates $\mathbf{X_i}$; and ε_i = random error term.

Hence, the optimization problem becomes solving the estimates for βs in Eq. (6).

$$\hat{\beta}(p) = \arg\min_{\beta \epsilon R^k} \left[\sum \rho_p(y_i - \mathbf{X_i'}\beta) \right] \quad (6)$$

For any quantile p between 0 and 1, $\hat{\beta}(p)$ 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 $p = 0.5$.

Smooth Quantile for Counts with Jittering Algorithm

Although generation of quantiles is possible from either discrete or continuous data, a potential issue arises when estimating conditional quantiles with count data. The problem originates from the conjunction of a nondifferentiable objective function and a discrete dependent variable (Machado and Santos Silva 2005). To apply QR to count data, the proposal for smoothing approaches can be roughly categorized into discretization and jittering. Introduction of a latent variable classifies a discrete variable between two continuous values defined by the latent variable. This approach is similar to the ordinal logistic regression model. In this model, the order of a multilevel outcome is determined by an underlying latent variable to be estimated. One obvious shortcoming of this approach is that it introduces a new parameter for each observation, which makes it computationally inefficient. This study employs the jittering method, which constructs a continuous variable whose conditional quantiles have a one-to-one relationship with the conditional quantiles of the counts of interest. This requires imposing an artificial relationship on the data.

Because observed count data Y have a discrete distribution, $\eta_Y(p|\mathbf{X})$ is not a continuous function of the parameters of interest. By adding a uniformly distributed random variable U between 0 and 1 to Y, a new variable $Z = U + Y$ can be created, resulting in a conditional quantile function that is continuous in p. More importantly, Z's conditional quantiles $\eta_Z(p|\mathbf{X})$ have the one-to-one relationship with $\eta_Y(p|\mathbf{X})$, the conditional quantiles of the counts Y described in Eq. (7). The reason for adding a uniform distribution to the count data is that the new variable Z has a boundary. Moreover, compared to other distributions and jittering algorithms, the uniform distribution is more computationally efficient.

$$\eta_Y(p|\mathbf{X}) = \lceil \eta_Z(p|\mathbf{X}) + 1 \rceil \quad (7)$$

where $\lceil a \rceil$represents the ceiling function, meaning the next largest integer.

Estimation of Coefficients and Confidence Intervals

In general, considering QR as a linear programming problem can help it solve efficiently with various optimization methods such as simplex algorithm, interior point method, smoothing algorithm, and their derivations (Chen 2005). Simplex algorithm is the most popular algorithm, but it is computationally demanding. The processing time increases considerably as the size of data increases. The interior algorithm developed as an alternative for handling large data sets (i.e., $n > 10^5$) has proved superior compared to the simplex algorithm. The smoothing algorithm, on the other hand, is a heuristic approach that aims to improve the estimate through numerous iterations. The SAS QUANTREG procedure implements all three algorithms, and QUANTREG also provides three methods for estimating confidence intervals for the coefficients: sparsity, rank, and resampling. The sparsity method is the most direct and fastest method, but it can have problems with data that are not independently and identically distributed. The rank method uses the simplex algorithm and is computationally expensive. The resampling method uses bootstrap but is not suitable for small data sets. Koenker and Hallock (2001) proved that the discrepancies between these competing methods are slight. The SAS procedure document (Chen 2005) has details of these algorithms.

Data Description

This paper reconsiders a study conducted by Knapp et al. (2005) of the impact of intersection geometric design and traffic control on intersection crashes . Crash data for 1,770 intersections in Wisconsin were collected along with other features. This paper models the total number of crashes between 2001 and 2003 at the intersections based on various intersection attributes. Intersections are categorized by area type (AREATYPE) including rural and urban; traffic control (TRFCNTL) including four-way stop, two-way stop, signalized, and other; and geometric features of the intersection. The geometric features include the number of intersection approach legs (LEGS), the number of through lanes on the major roadway approach (LANE), the presence of a median on the major roadway (DIVIDED), and the presence of left-turn lanes (LEFTTURN). Millions of annual entering vehicles (ENTVEH) for the traffic

Table 1. Descriptive Statistics for Variables

Continuous variables	Description	Mean	S. dev.	Range
CRASHES	Total 3-year number of crashes	16.05	16.69	[0, 134]
ENTVEH	Million of annual entering vehicles	6.96	4.77	[0, 29.88]
Categorical variables	Description	Values	Frequency	Proportion (%)
AREATYPE	Types of area	Rural	581	32.82
		Urban	1189	67.18
LEGS	Number of intersection approach legs	3	380	21.47
		4	1390	78.53
DIVIDED	Existence of major roadway median	Yes	819	46.27
		No	951	53.73
TRFCNTL	Types of traffic controls (all-way, two-way, other, signal)	All-way	40	2.26
		Side	947	53.50
		Signal	780	44.70
		Other	3	0.17
LANE	Number of major roadway lanes	2	722	40.79
		4	1048	59.21
LEFT-TURN	Existence of left-turn lanes	Yes	1054	59.55
		No	716	40.45

Table 2. Quantile Regression Coefficient Estimates

Quantile				0.25			
Parameter		Estimate	Std. error		95% CI	t value	Pr > |t|
INTERCEPT		1.0236	0.084	0.8588	1.1884	12.18	< 0.0001
ENTVEH		0.7785	0.0431	0.694	0.8629	18.08	< 0.0001
LEG	3-LEGGED	−0.4332	0.0683	−0.5671	−0.2992	−6.34	< 0.0001
TRFCNTL	4-WAY	−1.5097	0.3896	−2.2737	−0.7456	−3.88	0.0001
	OTHER	−0.173	3.1534	−6.3577	6.0118	−0.05	0.9563
	SIDE	−0.496	0.0546	−0.603	−0.3889	−9.08	< 0.0001
Quantile				0.5			
Parameter		Estimate	Std. error		95% CI	t value	Pr > |t|
INTERCEPT		1.346	0.0696	1.2096	1.4825	19.35	< 0.0001
ENTVEH		0.7857	0.0352	0.7167	0.8546	22.35	< 0.0001
LEG	3-LEGGED	−0.2957	0.0546	−0.4028	−0.1886	−5.41	< 0.0001
TRFCNTL	4-WAY	−0.5717	0.2752	−1.1115	−0.0319	−2.08	0.0379
	OTHER	−0.2745	2.3237	−4.8319	4.2829	−0.12	0.906
	SIDE	−0.3646	0.0404	−0.4439	−0.2852	−9.01	< 0.0001
Quantile				0.75			
Parameter		Estimate	Std. error		95% CI	t value	Pr > |t|
INTERCEPT		1.871	0.0828	1.7086	2.0334	22.6	< 0.0001
ENTVEH		0.7035	0.0348	0.6352	0.7718	20.2	< 0.0001
LEG	3-LEGGED	−0.2562	0.0508	−0.3557	−0.1566	−5.05	< 0.0001
TRFCNTL	4-WAY	−0.5482	0.1891	−0.919	−0.1774	−2.9	0.0038
	OTHER	−0.1345	3.847	−7.6798	7.4107	−0.03	0.9721
	SIDE	−0.3821	0.0511	−0.4823	−0.2819	−7.48	< 0.0001
Quantile				0.95			
Parameter		Estimate	Std. error		95% CI	t value	Pr > |t|
INTERCEPT		2.6406	0.1589	2.329	2.9523	16.62	< 0.0001
ENTVEH		0.5483	0.0643	0.4223	0.6744	8.53	< 0.0001
LEG	3-LEGGED	−0.2983	0.0712	−0.438	−0.1586	−4.19	< 0.0001
MEDIAN	DIVIDED	0.1418	0.0649	0.0146	0.2691	2.19	0.029
TRFCNTL	4-WAY	−0.6026	0.4024	−1.3917	0.1866	−1.5	0.1344
	OTHER	−0.8286	20.6819	−41.3922	39.7351	−0.04	0.968
	SIDE	−0.3677	0.0738	−0.5124	−0.223	−4.99	< 0.0001

Note: Baseline is signalized, undivided four-legged intersections, and the coefficients for this type of intersections are zero.

exposure is the only continuous variable. Table 1 summarizes the statistics for key variables.

Results and Discussion

The relationship between crash count quantiles and explanatory variables was estimated by using a simplex algorithm, and the confidence intervals were computed for each estimated coefficient using the resampling method in SAS QUANTREG because of the moderate sample size. Table 2 shows the estimated coefficients and 95% confidence intervals for statistically significant variables (5% level of significance) at the 25th, 50th, 75th, and 95th percentiles of crash count distribution. Specifically, quantile $\eta(0.25)$ is the predicted 25th percentile of crash frequency, conditional upon the variables and their given values. Therefore, it presents a broader view of the variables related to intersections with low, intermediate, and high numbers of crashes. In other words, instead of assuming the coefficients are fixed across all the sites, some or all of them are allowed to vary to account for heterogeneity attributable to unobserved factors.

The variables of DIVIDED and LEFTTURN are omitted from the table because they are not statistically significant. Baseline data are signalized, and undivided four-legged intersections and the coefficients for this type of intersection are zero. In general, the overall patterns of the estimated coefficients at each quantile are similar in terms of the number of statistically significant variables and the signs of individual parameters. The overall results are rather consistent with previous studies: the number of crashes increases with the increase in traffic exposure (ENTVEH); three-legged intersections are safer than four-legged intersections, provided other variables remain equal; and unsignalized intersections, including four-way stop and two-way stop controls, have fewer crashes than their signalized counterparts if other variables remain equal. A closer examination of the magnitude of the estimated coefficients reveals similarities and differences among quantiles. First, ENTVEH is less likely to affect safety in the high tail than in the low tail. A unit increase of a million annual entering vehicles at an intersection may lead to 2.17 crashes at the 25th percentile or 1.73 crashes at the 95th percentile. This suggests that any significant traffic volume increase at locations with a historically low number of crashes may trigger a considerable surge in crashes

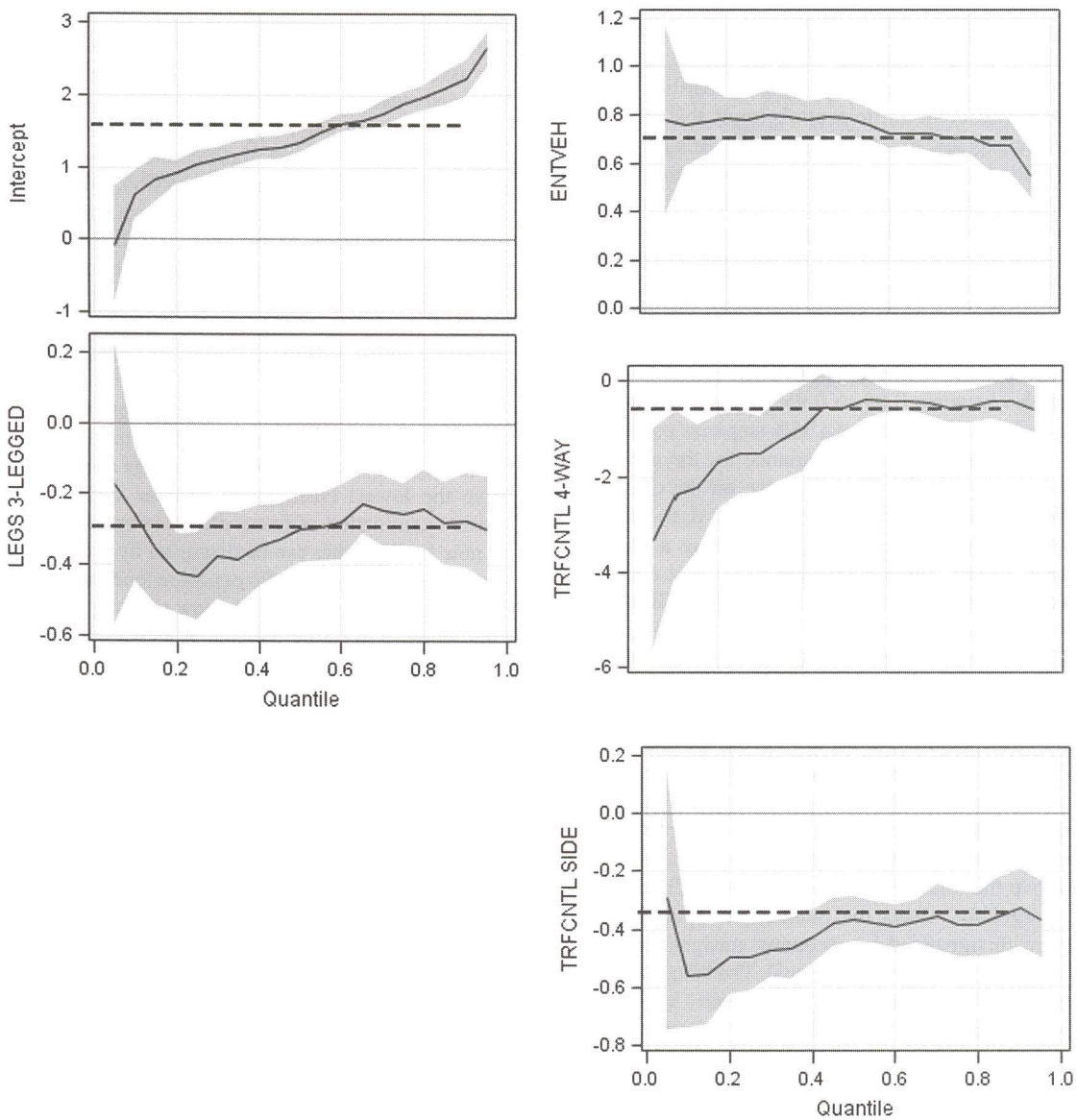

Fig. 2. Quantile plots for variable coefficients

Table 3. Negative Binomial Regression Coefficient Estimates

Parameter		Estimate	Std. error	95% CI		t value	Pr > |t|
INTERCEPT		1.5297	0.0654	1.4015	1.6578	547.35	< 0.0001
ENTVEH		0.7379	0.0311	0.6769	0.799	561.57	< 0.0001
LEG	3-LEGGED	−0.3144	0.0426	−0.398	−0.2309	54.43	< 0.0001
TRFCNTL	4-WAY	−0.6508	0.1233	−0.8925	−0.409	27.84	< 0.0001
	OTHER	−0.3907	0.4685	−1.309	0.5276	0.7	0.4043
	SIDE	−0.3867	0.0397	−0.4645	−0.3089	94.9	< 0.0001

Note: Baseline is signalized, undivided four-legged intersections, and the coefficients for this type of intersections are zero.

compared with the sites already experiencing a high number of crashes. Highway designers or planners may need to act more cautiously and proactively with projects that intend to increase the capacity of the roadway or intersection. Secondly, for the number of intersection approaches, three-legged intersections have fewer crashes than four-legged intersections at any quantile level. The difference, however, displays a concave shape, with the 25th percentile being the lowest. Various intersection traffic controls demonstrate substantial variation. Intersections controlled by four-way stops have the lowest number of crashes, but the effect dwindles and eventually reaches a plateau near the 50th percentile. Intersections controlled by two-way stops are relatively stable, and their value fluctuates between −0.4 and −0.6. The disparities among different quantiles for different traffic controls imply that these factors are not equally correlated to a low, a moderate, or a high crash count. When evaluating traffic control design at an intersection, its crash history has to be considered along with the projected traffic demand. For example, when weighing a signalization project for an intersection, the site crash history should be reviewed in detail because more crashes may be expected when converting an intersection with a low crash history than a high crash history.

Fig. 2 depicts the quantile regression results of the parameters in quantile plots. The solid line represents the estimates of the coefficient for percentiles between 0.05 to 0.95, and the shaded area between the two dashed lines describes a 95% stepwise confidence band. A narrow band suggests a small variance for the estimated coefficient, which displays a stronger influence on the crash frequency. Superimposed on each plot is a horizontal dash line that represents the mean estimate of the coefficient using a negative binomial regression model. Table 3 lists the coefficient estimates along with statistical inferences obtained by the SAS GENMOD procedure.

The position of the mean estimate in relation to the quantile estimate unambiguously indicates that the mean, the location parameter, is insufficient to explain the variation of the response variable. As can be found in the quantile plot for ENTVEH, the conditional mean estimate intersects the quantile plot around the 80th percentile, suggesting an overestimate of the number of crashes. For three-legged intersections, a clear concave shape can be observed below the 50th percentile, and the quantile estimates fluctuate around the mean estimate above the 50th percentile of the crash distribution. Traffic control inarguably affects the intersection operation and safety. Compared with signalized intersections, unsignalized ones have fewer crashes, other variables being equal. The values of mean coefficients for four-way stop and two-way stop control are near the median, but they are unable to illustrate the changes in the low tail of the distribution. This implies that warranting a signal control to an unsignalized intersection already experiencing many crashes may not drastically increase crash count. Although the quantile plot is more indicative compared to a mean estimate, like any statistical model and its results, it needs to be combined with other data collection and engineering studies for appropriate decision-making. For example, the decision of warranting a signalized intersection or changing signalized intersections to other control types needs to include crash severity and prevailing traffic conditions, such as traffic mix, turning movements, pedestrian activities, etc.

Conclusions

A common practice in highway safety studies is to combine data from different locations and at different times to increase sample size for a more statistically valid analysis. This, however, raises the issue of data heterogeneity that may potentially make the parameter estimates unstable and less efficient. The QR method provides an alternative approach to cope with heterogeneity data. Compared to the current crash count models using GLM, such as Poisson or negative binomial regression, QR can effectively depict the varying effects of covariates on crash frequency at different levels of its distribution. In this study, 1,770 intersections with a three-year crash history and corresponding geometric characteristics and traffic controls were analyzed using the QR models. The effort was focused on understanding how the crash contributing factors influenced results at various quantiles of crash distribution, from low to high.

The overall results are rather consistent with previous studies in which the traffic exposure (ENTVEH) has a positive effect on crash frequency; intersections with three legs and four-way stop control have the lowest crash count if other variables remain equal. A closer examination of factor effects at various quantile levels leads to new findings that can be disguised by a conditional mean-based regression analysis. Traffic exposure tends to be less likely to affect safety in the high tail than the low tail, suggesting that any significant increase in traffic demand at locations with low crash history may experience a surge in crashes. Three-legged intersections appear to be safer than four-legged intersections, but the difference seems to follow a concave shape at the low tail, with the 25th percentile as the lowest. Similarly, unsignalized intersections have fewer crashes than signalized intersections when other variables are equal. The disparities, however, are more appreciable at the low tail than the high tail of the crash distribution. This implies that warranting a signal control to an unsignalized intersection already experiencing many crashes may not drastically increase crash count.

In summary, quantile-based regression analysis seeks to extend the ideas of estimating the conditional mean to estimating conditional quantiles of the response variable, which is expressed as a link function of a series of covariates. Covariates can influence the conditional distribution of the response in many ways. Explicit investigation of these effects using QR can provide a much more complete view of the stochastic relationship between variables and, therefore, a more indicative empirical analysis. Furthermore, given the flexibility of QR, efforts can be directed to crash diagnosis on

the high quantile of the distribution if the locations with an abnormally high number of crashes are of particular interest to safety stakeholders.

References

Buchinsky, M. (2001). "Quantile regression with sample selection: Estimating women's return to education in the U. S." *Empir. Econ.*, 26, 87–113.

Chen, C. (2005). *An introduction to quantile regression and the QUANTREG procedure*, SAS Institute Inc., Gary, NC.

Chiou, J. M., and Muller, H. G. (1998). "Quasi-likelihood regression with unknown link and variance functions." *J. Am. Stat. Assoc.*, 93, 1376–1387.

Fizenberger, B., Hujur, R., MaCurdy, T. E., and Schnabel, R. (2001). "Testing for uniform wage trends in West Germany: A cohort analysis using quantile regression for censored data." *Empir. Econ.*, 26, 41–86.

Garcia, J., Hernandez, P. J., and Lopez-Nicolas, A. (2001). "How wide is the gap? An investigation of gender wage differences using quantile regression." *Empir. Econ.*, 26, 149–167.

Geedipally, S. R., and Lord, D. (2008). "Effects of the varying dispersion parameter of Poisson-models on the estimation of confidence interval of crash prediction models." *Transportation Research Record 2061*, 46–54.

Gonzalez, X., and Miles, D. (2001). "Wage inequality in a developing country: Decrease in minimum wage or increase in education returns." *Empir. Econ.*, 26, 135–148.

Hewson, P. (2008). "Quantile regression provides a fuller analysis of speed data." *Accid. Anal. Prev.*, 40, 502–510.

Knapp, K. K., Campbell, J., and Kienert, C. (2005). "Intersection crash summary statistics for Wisconsin." *Final Report*, University of Wisconsin–Madison.

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

Koenker, R., and Hallock, K. F. (2001). "Quantile regression: An introduction." *J. Econ. Perspect.*, 15, 143–156.

Lord, D., Washington, S. P., and Ivan, J. N. (2007). "Further notes on the application of zero inflated models in highway safety." *Accid. Anal. Prev.*, 39(1), 53–57.

Machado, J. A. F., and Mata, J. (2001). "Earning functions in Portugal 1982–1994: Evidence from quantile regressions." *Empir. Econ.*, 26, 115–134.

Machado, J. A. F., and Santos Silva, J. M. C. (2005). "Quantiles for count." *J. Am. Stat. Assoc.*, 100(472), 1226–1237.

Miaou, S.-P., and Lord, D. (2003). "Modeling traffic crash-flow relationships for intersections: Dispersion parameter, functional form, and Bayes versus empirical Bayes methods." *Transp. Res. Rec.*, 1840, 31–40.

Miranda-Moreno, L. F., Fu, F. F., Saccomanno, L., and Labbe, A. (2005). "Alternative risk models for ranking locations for safety improvement." *Transp. Res. Rec.*, 1908, 1–8.

Mitra, S., and Washington, S. (2007). "On the nature of over-dispersion in motor vehicle crash prediction models." *Accid. Anal. Prev.*, 39(3), 459–468.

Nelder, J. A., and Lee, Y. (1992). "Likelihood, quasi-likelihood and pseudo-likelihood: Some comparisons." *J. R. Stat. Soc. Ser. B*, 54(1), 273–284.

Nielson, H. S., and Rosholm, M. (2001). "The public-private sector wage gap in Zambia in the 1990s: A quantile regression approach." *Empir. Econ.*, 26, 169–182.

Palmer, A., Losilla, J. M., Vives, J., and Jimenez, R. (2007). "Overdispersion in the Poisson regression model." *Methodology*, 3(3), 89–99.

Park, B. J., and Lord, D. (2009). "Application of finite mixture models for vehicle crash data analysis." *Accid. Anal. Prev.*, 41(4), 683–691.

Qin, X., Ng, M., and Reyes, P. (2010). "Identifying crash-prone locations with quantile regress." *Accid. Anal. Prev.*, 42(6), 1531–1537.

Taylor, J. (1999). "A quantile regression approach to estimating the distribution of multiperiod returns." *J. Derivatives*, 7, 64–78.

Wehby, G. L., Murray, J. C., Castilla, E. E., Lopez-Camelo, J. S., and Ohsfeld, R. L. (2009). "Quantile effects of prenatal care utilization on birth weight in Argentina." *Health Econ.*, 18(11), 1307–1321.

Winkelmann, R. (2006). "Reforming health care: Evidence from quantile regressions for counts." *Health Econom.*, 25, 131–145.