Quantile Effects of Causal Factors on Crash Distributions

Xiao Qin

Crash data are heterogeneous because they are collected from different sources and locations at different times. This data heterogeneity may cause a significant bias in the estimation of standard errors for the coefficients as well as the coefficients' statistical inferences. In the past decade, several promising modeling strategies have been proposed to handle overdispersed crash data, most of which have focused on estimating the conditional mean crash count. This paper applies an alternative crash modeling approach: quantile regression (QR) in the context of a count data model. The application of QR to model crash frequency is illustrated, and empirical results are interpreted. Poisson gamma, the benchmark statistical model for crash counts, is referenced to estimate the covariate coefficients for the mean crash count. Focusing on the mean may result in important aspects of the data being missed. A more detailed analysis, using a QR model for crash count data, confirms that crash predictors have varying impacts on the different areas of the crash distribution. Moreover, the marginal effects of covariates provide a more direct observation of changes in the quantity, rather than the percentage, of crash frequency when responding to one-unit changes in regressors.

Crash modeling is an effective approach for exploring the relationship between crash frequency and a set of predictors from the statistical perspective. Once the relationship is established, the mean crash count can be estimated with the values of a set of regressors. It is anticipated that the regressors will not only be statistically correlated but also logically related to crash occurrences. Such a method for conditional mean regression assumes the error to be random noise, and the mean can be represented as the true value around which observations fluctuate. The location (mean) of the conditional distribution is of more concern than its scale or shape.

Decades of experience with crash data distribution and modeling have shown that crash count data are usually overdispersed for a variety of reasons, including the omission of important variables, a misspecification of the link function, or a structured error term. Under these circumstances, it is no longer appropriate to treat the error term as random noise for the sake of modeling convenience. Hence, several modeling approaches have been proposed to improve the model assumptions. Representative work includes

- Generalizing the parametric count model through hurdle models, such as a zero-inflated Poisson or a zero-inflated negative binomial (*1–3*);
- Describing heterogeneous crash count data through finite mixture regression models (*4*);
- Adopting a well-specified mean function (*5*);
- Improving the structure of the dispersion parameter ϕ by replacing a fixed value with a varying one (*6–8*); and
- Addressing the crash data heterogeneity by specifying a random parameter model (*9, 10*), in which some or all of the parameters are allowed to vary to account for the heterogeneity across locations caused by unobserved factors.

These approaches have explicitly considered the data heterogeneity and significantly improved the unbiased estimation of standard errors for the coefficients as well as the coefficients' statistical inferences.

Data heterogeneity and, more specifically, data overdispersion may originate from the fact that crash data are collected from different sources at different locations at different times. At a disaggregate level, crashes may result from many Bernoulli trials with an unequal probability of independent individual crashes (*3*). If so, a generalized linear model that relies on the conditional mean across different values of predictors to describe the data central tendency is less efficient because of its inherent limitations. The mean is more sensitive to the influence of data outliers. The model assumptions cannot be easily extended to noncentral locations and are not always met with real-world data, especially in the case of the homoscedasticity assumption. Therefore, other measures of data central tendency or the shape of the distribution become more relevant, more appropriate, and more informative than merely the mean and the variance. For instance, the median (50th percentile) is known for its robustness of data central tendency when the data distribution is highly skewed (compressed or stretched) with heavy tails. If other quantiles besides the median are used, distributional properties, such as skew and modality, can be described in greater detail. Similarly to mean regression, the relationship between the quantile locations of the response variable and a set of covariates can be established. It should be possible to find out how the covariates affect the response variable. This methodology is called quantile regression (QR) and is commonly used when an understanding of the effect of covariates in all aspects of the response distribution is desired.

The origin of QR can be traced to early work related to the development of median regression, when conditional median modeling was developed with the least absolute error (*11*). The groundbreaking introduction of QR in 1978 by Koenker and Bassett specified conditional quantiles as functions of predictors (*12*). The least absolute error and QR are closely related in the sense that they both optimize certain functions of absolute residuals. The least absolute error optimizes a symmetric piecewise linear absolute error, whereas QR optimizes an asymmetrically weighted residual function. In 2005, Machado and Santos Silva proposed QR for count data and demonstrated its implementation (*13*). Since then, QR has been used to

Box 2219, Department of Civil and Environmental Engineering, South Dakota State University, CEH 148, Brookings, SD 57007. Xiao.Qin@sdstate.edu.

Transportation Research Record: Journal of the Transportation Research Board, No. 2279, Transportation Research Board of the National Academies, Washington, D.C., 2012, pp. 40–46.
DOI: 10.3141/2279-05

model not only continuous variables but also discrete variables, which grants it immediate applicability in many disciplines, including highway safety. Recent developments in QR have been prolific in many areas, such as nonparametric models, multivariate quantile regression, nonlinear models, and Bayesian models, most of which can be found in Koenker and Hallock's review article (*14*).

The application of QR is growing rapidly in research communities, such as sociology, economics, finance, medicine, and public health (*15–18*). In contrast, the development of QR in transportation research is still in its infancy. Publications involving this methodology are scarce. A few pioneering studies include the paper by Hewson, who applied a quantile smoother for speed data (*19*), and the papers by Qin et al. (*20*) and Qin and Reyes (*21*), who utilized QR to determine crash-prone locations and modeled crash frequencies at intersections. Progressing along with QR theory development is the availability of statistical tools. To date, QR is provided in several statistical software packages, including QREG and QCOUNT in STATA (*22, 23*), and QUANTREG in SAS and R (*24, 25*). Given the data issues existing in transportation studies, QR can be a useful method.

QR models conditional quantiles as functions of predictors that specify changes in the response variable as the changes in predictors. Because any quantile can be used, QR is flexible enough to model any position of a probabilistic distribution. QR offers a direct observation of the impact of covariates on the response variable, yields rich information on different parts of the response distribution, and provides a complete understanding of the marginal effects of predictors. This analysis will be particularly useful for safety researchers and practitioners who are interested in investigating the factors that affect locations with a very low or an extremely high number of crashes (i.e., the information indicated in the low tail and the high tail of the crash frequency distribution). This paper introduces QR as an alternative crash modeling approach in the context of a count data model, provides a practical example of modeling crash frequency, and interprets the empirical results.

QR FOR COUNT DATA

Quantiles are points taken at regular intervals from the cumulative distribution function of a random variable. Special quantiles are named after the length of the interval (e.g., the 2-quantile is called the median, the 4-quantiles are called quartiles, the 10-quantiles are called deciles, and the 100-quantiles are called percentiles). Let *p* be a number between zero and one, and the 100*p* percentile of the distribution of a continuous random variable *Y,* denoted by *Q*(*p*), be defined in Equation 1:

$$p = F(Y) = F(Q(p)) = \int_{-\infty}^{Q(p)} f(y)\,dy \qquad (1)$$

Formally, the *p*-quantile of *Y* with cumulative distribution function *F* on $\Re$ with $0 \le p \le 1$ is defined in Equation 2:

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

where F^{-1} represents the inverse function of the cumulative distribution function, and inf represents the greatest lower bound. Note that *Q*(0.5) is the median, and the commonly used first and third quartiles are denoted as *Q*(0.25) and *Q*(0.75), respectively. *Q*(*p*) can be interpreted as the threshold that splits the possible values of *Y* into two sets, such that $P(Y \le Q(p)) = p$ and $P(Y > Q(p)) = 1 - p$.

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, which is analogous to the mean of a random sample as the minimal of the sum of square errors. Koenker suggested that the *p*th sample statistics quantile, given by **X**, $Q_Y(p|\mathbf{X})$, may be solved as an optimal solution to minimize a weighted average of the samples whose values are larger than or equal to $Q_Y(p|\mathbf{X})$ and the samples whose values are less than or equal to $Q_Y(p|\mathbf{X})$, as formulated in Equation 3 (*12*):

$$\sum_{i\in\{i:y_i \ge Q_Y(p|\mathbf{X})\}} p|y_i - Q_Y(p|\mathbf{X})| + \sum_{i\in\{i:y_i \ge Q_Y(p|\mathbf{X})\}} (1-p)|y_i - Q_Y(p|\mathbf{X})| \qquad (3)$$

The bracket in Equation 3 can be simplified as $\sum\rho_p(\mu) = \sum\mu(p - \mathbf{1}_{(\mu<0)})$, where $\mu = y_i - Q_Y(p|\mathbf{X})$ and $\mathbf{1}_A$ is an indicator function of a subset *A* of a set *U*. If the *p*th sample quantile $Q_Y(p|\mathbf{X})$ is a linear function of the parameters of interest $\mathbf{X}\beta$, it can be solved efficiently by linear programming methods. The process of estimating the vector of unknown coefficients β for the vector of variables **X** is called QR.

QR is a standard application for a continuous variable but not for a discrete variable. Because the distribution of a discrete variable is not continuous, neither are its quantiles. Hence, quantiles cannot be modeled directly as a continuous function of the regressors. In other words, the objective function in the optimization is not differentiable. Machado and Santos Silva have proposed a few alternatives that extend the quantile regression to count data (*13*). The proposal to smooth approaches can be roughly categorized into discretization and jittering. Discretization introduces a latent variable and classifies a discrete variable between two continuous values that are defined by the latent variable. One obvious shortcoming of this approach is that it introduces a new parameter for each observation, making it computationally inefficient. The other popular alternative in QR for count is the jittering method. The jittering method constructs a continuous variable whose conditional quantiles have a one-to-one relationship with the conditional quantiles of the discrete variable. Adding a uniformly distributed random variable *U* between [0, 1] to a discrete variable *Y* creates a new variable $Z = U + Y,$ and results in a conditional quantile function that is continuous in *p*. Similarly, a continuous variable can be generated with the jittering method from the nonnegative crash count.

The quantile of the new continuous variable *Z* has two important features: (*a*) according to its distribution, the *p*th quantiles of *Z* can never be smaller than *p,* and (*b*) the quantiles of *Z* can never be negative because *Z* is the sum of the crash count *Y* and a uniform variable *U* between zero and one (*26*). Therefore, the *p*th quantile of *Z* can be specified in Equation 4 as

$$Q_Z(p|\mathbf{X}) = p + \exp(\mathbf{X}\boldsymbol{\beta}(p)) \qquad (4)$$

To estimate the unknown coefficients, a natural logarithm is applied to obtain a linear function of βs in Equation 5:

$$T(Z;p) = \begin{cases} \log(Z-p) & Z > p \\ \log\zeta & Z \le p \end{cases} \qquad (5)$$

where *T* means a transformed function and ζ is a very small positive number. The transformed quantile function in Equation 6 is linear in its parameters:

$$T(Q_Z(p|\mathbf{X})) = \mathbf{X}\boldsymbol{\beta}(p) \qquad (6)$$

The last effort is to prove that the quantiles of transformed *Z*, $Q_{T(Z)}$, are the same as the quantiles of *Z*, Q_Z. One of the properties of quantiles is that they are invariant to monotonic transformations and are also invariant to censoring from below up to the quantile of interest (*14*). Hence, the optimization problem becomes to solve the estimates for βs in Equation 7:

$$\hat{\boldsymbol{\beta}}(p) = \arg\min_{\beta \in R^k} \left[\sum_{i=1}^{n} \rho_p \left(T(Z; p) - \mathbf{X}\boldsymbol{\beta}(p) \right) \right] \quad (7)$$

For any quantile *p* between zero and one, $\hat{\boldsymbol{\beta}}(p)$ is called the *p*th 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.

Like other count models, the expected value of the conditional quantile regression is specified as a linear function of the covariates, as shown in Equation 6. The estimated coefficients βs are partial- or semielasticities because Equation 6 is a semi-log function where $\partial \ln y/\partial x_k = \beta_k$. In other words, β_k means the percentage change in *Y* associated with a one-unit change in X_k. Comparing βs from different conditional quantile functions can be problematic because *Y* is different. *Y* corresponds to different quantiles of the response variable that are conditional upon the covariate values. To facilitate the comparison of covariate effects across different quantiles, the marginal effect that measures a one-unit change in *X* on the dependent variable *Y* needs to be calculated for each coefficient. The marginal effect is calculated as the partial derivative $\partial[Q_Z(p|\mathbf{X}) - p]/\partial x$, the product of $\beta(p)$ and $[Q_Z(p\,|\,\bar{\mathbf{X}}) - p]$ (*13*). For a continuous variable, the marginal effect is $\beta(p)[Q_Z(p\,|\,\bar{\mathbf{X}}) - p]$, where $\beta(p)$ is the estimated coefficient at the *p*-quantile. For a dummy variable, the marginal effect is $Q_Z(p\,|\,\bar{\mathbf{X}}, x_k = 1) - Q_Z(p|\bar{\mathbf{X}}, x_k = 0) = [\exp(\gamma(p)) - 1][Q_Z(p|\bar{\mathbf{X}}) - p)]$, where $\gamma(p)$ is the estimated coefficient for dummy variable x_k at the *p*-quantile.

Machado and Santos Silva's jittering algorithm QCOUNT was implemented in Stata by Miranda (*23*). QCOUNT not only calculates the QR estimates for coefficients, as well as the coefficients' statistical inferences, but also computes the marginal effects of these coefficients, which is a more effective measure of the impact of covariates on the response variable across different areas of the distribution. Because generating a uniform variable is a random process, QCOUNT allows users to specify the number of repeats and then averages the estimates of all of the iterations, as suggested by Machado and Santos Silva (*13*). According to several empirical studies, 1,500 iterations is considered sufficient to achieve a stable coefficient estimate (*13, 26, 27*). Another specification is the value of ζ, which is a small positive number. In QCOUNT, ζ is set to be 10^{-4}, which is consistent with other studies. The coefficients for the covariates are estimated by the linear programming process that minimizes the weighted residual function, as formulated in Equation 3.

TABLE 1 Summary Statistics for Dependent Variable Crash (Total Number of Crashes, 2004–2008)

Percentile	Value	Percentile	Value
10	0	90	5
25	0	95	9
50	1	99	14
75	2		

NOTE: Mean = 1.81; standard deviation (SD) = 3.61; variance = 13.00; skewness = 6.35; kurtosis = 83.33.

EMPIRICAL STUDY OF CRASH FREQUENCY OF HIGHWAY SEGMENT

Data Description

Crash information and roadway inventory data were requested, respectively, from the South Dakota Department of Public Safety and South Dakota Department of Transportation. The data acquired to generate crash prediction models included a five-year crash data set from 2004 to 2008 and the corresponding highway functional classification, annual average daily traffic, and geometric characteristics. The majority of South Dakota highways are in rural areas. The rural minor arterial highway type is chosen for this study because it is one of the principal highway types in the state. A total of 1,231 roadway segments with the necessary data elements are available in this class, ranging from 0.1 mi to 26.2 mi in length, with a mean of 2.7 mi. The total number of crashes occurring during the 5-year time period on these segments ranges from zero to 64, with an average of 1.81 crashes and a standard deviation of 3.61 crashes. Tables 1–3 list the data summary statistics. Detailed information on data collection and processing is available in a previous report (*28*).

Results and Discussion

Previous crash prediction models have focused on estimating the conditional mean crash count. However, the resulting estimates of effects were not necessarily indicative of the nature of these effects on the low tails (low number of crashes) or high tails (high number of crashes) of the crash distribution. The characteristics of the sites that

TABLE 2 Summary Statistics for Continuous Independent Variable

Continuous Independent Variable	Description	Mean	SD	Range
Length	Segment length (mi)	2.75	3.60	[0.1, 26.2]
ADT	Average daily traffic	1,180	1,009	[110, 6845]
SURF_WI	Surface width (ft)	28.82	8.43	[18, 91]
SHLDR_WI	Shoulder width (ft)	4	2.49	[0, 12]
SPD_LIM	Speed limit (mph)	54.89	13.43	[20, 65]
V_DEN	Vertical curve density (curves/mi)	4.71	4.23	[0, 39.96]

TABLE 3 Summary Statistics for Categorical Independent Variable

Categorical Independent Variable	Description	Value	Frequency	Proportion (%)
ACCESS	Access control	None	1,192	96.8
		Part	39	3.2
SRS	Shoulder rumble strip	No	1,144	92.9
		Yes	87	7.1
SHLDR_TY	Shoulder surface type	Asphalt	750	60.9
		Blotter	20	1.6
		Concrete	142	11.5
		Gravel	191	15.5
		None	115	9.3
		Recycled	13	1.1

experience low or high numbers of crashes are of particular interest and can be described by a family of conditional quantile functions, as specified in Equation 8:

$$\ln\left(Q_Z\left(p\,|\mathbf{X}\right)-p\right)=\beta_0+\beta_1\text{lnVMT}+\beta_2\text{SURF\_WI}+\beta_3\text{SHLDR\_WI}+\beta_4\text{SPD\_LIM}+\beta_5\text{V\_DEN}+\gamma_1\left(\text{SRS}=\text{Y}\right)\quad(8)$$

where

VMT = vehicle miles traveled (millions),
SURF_WI = surface width (ft),
SHLDR_WI = shoulder width (ft),
SPD_LIM = speed limit (mph),
V_DEN = vertical curve density (curves/mi), and
SRS = Y = presence of shoulder rumble strip.

VMT is calculated as the product of the annual average daily traffic, the segment length, and the number of days between 2004 and 2008, divided by 1 million.

The covariates included in the QR models were chosen by a conventional Poisson gamma model, which is the benchmark statistical distribution for modeling crash counts. All of the covariates are statistically significant at the 5% significance level in the Poisson gamma model. The details for developing the Poisson gamma model and selecting the variables are referred to in an earlier research report and are omitted here for brevity (*28*). The reason for keeping the same covariates in the QR model as in the mean regression model is to illustrate the difference in the amount of information presented. Because 48% of the crash counts in the data set are zeros, the upper tail of the crash distribution is certainly more interesting. The coefficients of the covariates were estimated for the 25-, 50-, 75-, 85-, and 95-quantiles. In the lower tail, a variation in the conditional quantiles of $Q_Z(p|\mathbf{X})$ may be due to the random error. Table 4 lists the QR estimates for these quantiles with the standard errors in parentheses. Poisson gamma estimates for the covariate coefficients are provided in the last column of the table.

Similarities and differences can be clearly identified by comparing each quantile level to the others. The sign of each statistically significant coefficient is consistent across all quantiles tested, whereas the value of the coefficient varies considerably from one quantile to another. Figure 1 illustrates the quantile regression results of the coefficients. The solid line represents the estimates of the coefficients for the 25, 50, 75, 85, and 95 percentiles, which are enveloped by two dashed lines representing a 95% confidence interval. The horizontal dotted line presents the estimates of the coefficients from the Poisson gamma model. A positive relationship between crash frequency and lnVMT is presented at all quantile levels, but the trend of effect decreases from 1.164 at the low crash count brackets to 0.727 at the high crash count brackets. The rural minor arterial crash data suggest that the effect of lnVMT is substantially lower at the high tail of the crash frequency distribution than at the low tail of the distribution. This effect implies that for segments with a high number of crashes, travel demand management that tries to reduce traffic exposure (VMT) may lead to a smaller reduction in crash percentage than the areas with a relatively low numbers of crashes. Arguably, the sites with high crash frequencies may have a greater VMT because of a positive causal relationship. From a different perspective, this may suggest that any significant increase in traffic volume at locations with a historically low number of crashes may trigger a considerable surge in crashes compared with the sites already experiencing a high number of crashes.

TABLE 4 Estimated Coefficients for $Q_Z(p|\mathrm{X})$

	QR					Poisson-Gamma (Means Model)
	$Q_Z(0.25\,|\,\mathbf{X})$	$Q_Z(0.5\,|\,\mathbf{X})$	$Q_Z(0.75\,|\,\mathbf{X})$	$Q_Z(0.85\,|\,\mathbf{X})$	$Q_Z(0.95\,|\,\mathbf{X})$	
LnVMT	1.164[a] (0.055)	1.100[a] (0.048)	1.023[a] (0.051)	0.941[a] (0.065)	0.727[a] (0.058)	0.934[a] (0.033)
SURF_WI	–0.027 (0.014)	–0.018 (0.012)	–0.010 (0.012)	–0.018 (0.011)	–0.021[a] (0.009)	–0.025[a] (0.006)
SHLDR_WI	–0.059[a] (0.026)	–0.092[a] (0.023)	–0.079[a] (0.023)	–0.064[a] (0.029)	–0.071[a] (0.018)	–0.085[a] (0.015)
SPD_LIM	0.004 (0.008)	0.000 (0.007)	–0.008 (0.008)	–0.018[a] (0.009)	–0.023[a] (0.005)	–0.019[a] (0.003)
V_DEN	–0.030 (0.027)	–0.049[a] (0.021)	–0.082[a] (0.027)	–0.093[a] (0.037)	–0.068[a] (0.023)	–0.061[a] (0.01)
SRS	–0.466 (0.289)	–0.369 (0.193)	–0.461 (0.254)	–0.488[a] (0.194)	–0.352 (0.232)	–0.368[a] (0.142)
Constant	–1.266 (0.773)	–0.251 (0.624)	0.749 (0.783)	2.024[a] (0.812)	3.116[a] (0.441)	1.285[a] (0.306)

[a] Statistically significant at 5% level of significance.

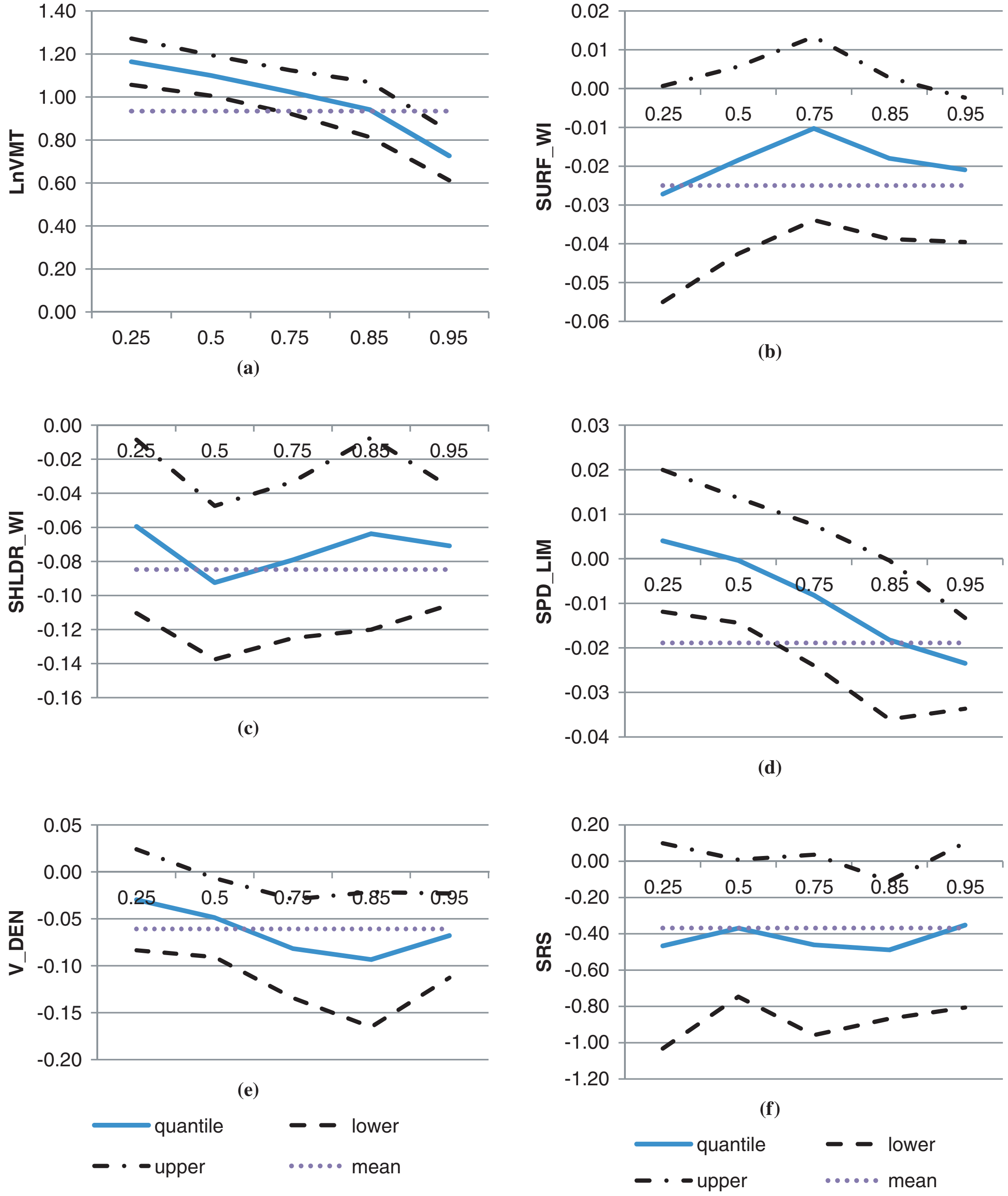

FIGURE 1 Quantile plots for variable coefficients.

Given the limited right-of-way, widening the surface width, the shoulder width, or a combination of both becomes a critical decision. Gross et al. compared the safety cost-effectiveness of several lane–shoulder width configurations for fixed total paved widths as a countermeasure for roadway departure crashes and concluded that certain lane–shoulder configurations have the potential to cost-effectively reduce crashes on rural, two-lane, undivided roads (*29*). In this study, surface width was defined as the edge of travel lane to travel lane and did not include shoulder width. A negative relationship can be found between crash counts and both surface width and shoulder width, indicating that wider surfaces and shoulders may contribute to fewer crashes. A close comparison between surface width and shoulder width reveals that the coefficients of surface width are not statistically significant at the 5% level of significance at all estimated quantile levels. Moreover, the values of coefficient estimates suggest that widening the shoulders is more effective in reducing crashes than widening the surfaces. Therefore, empirical QR results show that although widening the surfaces and widening the shoulders may both reduce crashes, the safety performance of shoulder widening is more stable than surface widening because the coefficient is statistically significant at the 5% significance level at all quantile levels tested.

For the other variables in the model—posted speed limit, vertical curve density, and shoulder rumble strip—the value and statistical significance of the estimated coefficients vary at different quantiles. Vertical curve density (the number of vertical curves per mile) is negatively related to the number of crashes at all test quantile levels, except for the 25 percentile. On the contrary, installing shoulder rumble strips is only statistically significant at the 85-percentile, even though the coefficient has a positive impact on highway safety. For the statistically significant coefficients of posted speed limit, the relationship with the crash frequency is negative, suggesting a positive safety impact; the effect gradually changes from –0.008 to –0.023.

As expected, the Poisson gamma coefficient estimates are all within the boundary of the quantiles of interest. The signs of the Poisson gamma coefficients are consistent with their QR counterparts; however, the sizes vary at different quantile levels compared with the QR estimates. Some are close to the high tail and others are close to the median. Obviously, mean regression only yields one set of coefficient estimates, which are not able to reflect the subtle to substantial covariate effects in different areas of the crash distribution.

The change in the dependent variable corresponding to the change in each predictor can usually be expressed as the value of coefficient β. For instance, β_k means the percentage change in Y associated with a one-unit change in X_k. It does not matter whether the change is relative or absolute for mean regression estimates because there is only one set of coefficients. For quantile regression, the change matters because quantile estimates are with respect to different quantiles of the dependent variable. Comparing low quantiles with high quantiles in terms of percentage change in Y may not be appropriate. A better comparison can be made with the marginal effects of covariates that take the magnitude of Y into consideration. In other words, marginal effects suggest the change in the response variable—in this study, the number of crashes—as the change of one unit in the regressor. The marginal effects for the conditional quantiles of the jittered data Z are calculated by setting all continuous variables to their means and all dummy variables to their modes (*23*). Then the changes on the conditional quantile of interest are a function of QR coefficient estimates at that quantile level. Table 5 shows the results of the marginal effects for the covariates. All of the marginal effects show a greater impact of covariates at the high tail than at the median or low tail in this empirical data set, which may not be true for other data. For example, in Miranda's research on women's preferences toward number of children, some variables fluctuated from the low quantile (fewer children) to the high quantile (more children) (*27*).

A nuisance caused by the jittering algorithm is the production of a slightly biased estimate for quantiles. As shown in Equation 9, Q_Y can be retrieved from Q_Z:

$$Q_Y(p|\mathbf{X}) = \lceil Q_Z(p|\mathbf{X}) - 1 \rceil \tag{9}$$

where $\lceil\ \rceil$ represents the ceiling function (i.e., the next largest integer). A change in a covariate may or may not be sufficient to change the p-quantile of the dependent variable Y because that covariate's marginal effect may be rounded to the nearest integer without increasing or decreasing the value of $Q_Y(p|\mathbf{X})$, even though the covariate is statistically significant to $Q_Z(p|\mathbf{X})$. In other words, different quantiles of Z may correspond to a single Y because of the relationship formulated between Y and Z. Again, this is one of the fundamental differences between quantiles and the mean for count data. In count data, quantiles have to be integers, but the mean can be a continuous variable. Given the resolution level of integers and the data characteristics (48% are zeros), it is more meaningful to estimate the covariate effects than to estimate actual quantiles.

CLOSING REMARKS

Crash data are heterogeneous in nature because they are collected from different sources at different locations at different times. This data heterogeneity creates challenges for researchers and practitioners who are dedicated to improving highway safety. In the past decade, several promising modeling strategies have been proposed to handle overdispersed crash data. The prevailing research focuses on estimating the conditional mean crash count. However, the conditional mean method has inherent limitations and is less efficient at describing highly skewed or multimodal distributions.

Poisson gamma, the benchmark statistical model for crash counts, is employed to estimate the covariate coefficients. The mean regression provides some information about the effect on the response of a unit change in a covariate. The limitation is that mean regression only accounts for the effect of this change at the mean of the dependent variable. Because the effect can vary over the range of covariate values, it is plausible that the use of typical values leads to a distorted (incomplete) view of highly skewed data sets. Nevertheless, the study confirms a high consistency between the mean estimates and the QR estimates. The value of the mean estimate is encapsulated in the range of QR estimates, and both have the same sign. Hence, QR can be a useful alternative to mean-based crash modeling.

Furthermore, this paper illustrates that focusing on means may cause important aspects contained in the data to be missed. A more detailed analysis using a QR model for crash count data confirms that crash predictors have varying effects on crash distribution. Empirical results for South Dakota rural minor arterials suggest that the impact of traffic exposure is substantially lower in the high tail of the crash distribution than in the low tail. These results imply that for roadway segments with high crash frequencies, the same reduction in travel demand (VMT) may not be equally effective for areas with a relatively low numbers of crashes. This finding might also suggest that any significant increase in traffic volume at locations with a historically low number of crashes may trigger a more considerable surge in crashes than at the sites that already experience a high number of crashes. The analysis also reveals that widening the shoulder width is more effective at the high tail of the crash distribution than widening the surface width. The marginal effects of covariates, as alternative

TABLE 5 Marginal Effects for $Q_Z(p|X)$

	$Q_Z(0.25\|\mathbf{X})$	$Q_Z(0.5\|\mathbf{X})$	$Q_Z(0.75\|\mathbf{X})$	$Q_Z(0.85\|\mathbf{X})$	$Q_Z(0.95\|\mathbf{X})$
lnVMT	0.257[a] (0.02)	0.524[a] (0.033)	0.933[a] (0.055)	1.355[a] (0.114)	2.077[a] (0.169)
SURF_WI	–0.006 (0.003)	–0.009 (0.006)	–0.009 (0.011)	–0.026 (0.016)	–0.060[a] (0.027)
SHLDR_WI	–0.013[a] (0.006)	–0.044[a] (0.011)	–0.072[a] (0.022)	–0.092[a] (0.041)	–0.203[a] (0.055)
SPD_LIM	0.001 (0.002)	0.000 (0.003)	–0.007 (0.007)	–0.026[a] (0.014)	–0.067[a] (0.015)
V_DEN	–0.007 (0.006)	–0.023[a] (0.010)	–0.075[a] (0.025)	–0.134[a] (0.051)	–0.194[a] (0.065)
SRS	–0.085 (0.043)	–0.151 (0.067)	–0.348 (0.157)	–0.576[a] (0.194)	–0.869 (0.501)

[a] Significant at 5% level of significance.

measures to QR coefficient estimates, provide a direct observation of changes in the quantity, not percentage, of the crash frequency in response to a one-unit change in covariates.

In summary, QR is a statistical method that can be used to effectively describe the data heterogeneity through different regression equations at different quantiles. Data heterogeneity implies more than one relationship between a response variable and explanatory variables measured on a subset of these factors. QR estimates multiple relationships from the low quantile to the high quantile of the response, yields rich information on different parts of the response distribution, and provides a complete understanding about the sophisticated associations between variables.

ACKNOWLEDGMENT

The author thanks Gerald Wang in the Department of Economics at South Dakota State University for his assistance with QCOUNT.

REFERENCES

1. Shankar, V., J. Milton, and F. Mannering. Modeling Accident Frequencies as Zero-Altered Probability Processes: An Empirical Inquiry. *Accident Analysis and Prevention,* Vol. 29, No. 6, 1997, pp. 829–837.
2. Qin, X., J. N. Ivan, and N. Ravishanker. Selecting Exposure Measures in Crash Rate Prediction for Two-Lane Highway Segments. *Accident Analysis and Prevention,* Vol. 36, No. 2, 2004, pp. 183–191.
3. Lord, D., S. Washington, and J. N. Ivan. Further Notes on the Application of Zero-Inflated Models in Highway Safety. *Accident Analysis and Prevention,* Vol. 39, No. 1, 2007, pp. 53–57.
4. Park, B. J., and D. Lord. Application of Finite Mixture Models for Vehicle Crash Data Analysis. *Accident Analysis and Prevention,* Vol. 41, No. 4, 2009, pp. 683–691.
5. Mitra, S., and S. Washington. On the Nature of Over-Dispersion in Motor Vehicle Crash Prediction Models. *Accident Analysis and Prevention,* Vol. 39, No. 3, 2007, pp. 459–468.
6. Miaou, S.-P., and D. Lord. Modeling Traffic Crash-Flow Relationships for Intersections: Dispersion Parameter, Functional Form, and Bayes Versus Empirical Bayes Methods. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1840,* Transportation Research Board of the National Academies, Washington, D.C., 2003, pp. 31–40.
7. Miranda-Moreno, L. F., F. Fu, F. F. Saccomanno, and A. Labbe. Alternative Risk Models for Ranking Locations for Safety Improvement. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1908,* Transportation Research Board of the National Academies, Washington, D.C., 2005, pp. 1–8.
8. Geedipally, S. R., and D. Lord. Effects of Varying Dispersion Parameter of Poisson-Gamma Models on Estimation of Confidence Interval of Crash Prediction Models. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2061,* Transportation Research Board of the National Academies, Washington, D.C., 2008, pp. 46–54.
9. Anastasopoulos, P., and F. Mannering. A Note on Modeling Vehicle-Accident Frequencies with Random-Parameters Count Models. *Accident Analysis and Prevention,* Vol. 41, No. 1, 2009, pp. 153–159.
10. El-Basyouny, K., and T. Sayed. Accident Prediction Models with Random Corridor Parameters. *Accident Analysis and Prevention,* Vol. 41, No. 5, 2009, pp. 1118–1123.
11. Barrodale, I., and F. Roberts. Solution of an Overdetermined System of Equations in the L1 Norm [F4] (Algorithm 478). *Communications of the ACM,* Vol. 17, No. 6, 1974, pp. 319–320.
12. Koenker, R., and G. Bassett. Regression Quantiles. *Econometrica,* Vol. 46, No. 1, 1978, pp. 33–50.
13. Machado, J. A. F., and J. M. C. Santos Silva. Quantiles for Counts. *Journal of the American Statistical Association,* Vol. 100, No. 472, 2005, pp. 1226–1237.
14. Koenker, R., and K. F. Hallock. Quantile Regression. *Journal of Economic Perspectives,* Vol. 15, No. 4, 2001, pp. 143–156.
15. Nielson, H. S., and M. Rosholm. The Public–Private Sector Wage Gap in Zambia in the 1990s: A Quantile Regression Approach. *Empirical Economics,* Vol. 26, No. 1, 2001, pp. 169–182.
16. Machado, J. A. F., and J. Mata. Earning Functions in Portugal 1982–1994: Evidence From Quantile Regressions. *Empirical Economics,* Vol. 26, No. 1, 2001, pp. 115–134.
17. Buchinsky, M. Quantile Regression with Sample Selection: Estimating Women's Return to Education in the U.S. *Empirical Economics,* Vol. 26, No. 1, 2001, pp. 87–113.
18. Wehby, G. L., J. C. Murray, E. E. Castilla, J. S. Lopez-Camelo, and R. L. Ohsfeld. Quantile Effects of Prenatal Care Utilization on Birth Weight In Argentina. *Health Economics,* Vol. 18, No. 11, 2009, pp. 1307–1321.
19. Hewson, P. Quantile Regression Provides a Fuller Analysis of Speed Data. *Accident Analysis and Prevention,* Vol. 40, No. 2, 2008, pp. 502–510.
20. Qin, X., M. Ng, and P. Reyes. Identifying Crash-Prone Locations with Quantile Regression. *Accident Analysis and Prevention,* Vol. 42, No. 6, 2010, pp. 1531–1537.
21. Qin, X., and P. Reyes. Conditional Quantile Analysis for Crash Count Data. *Journal of Transportation Engineering,* Vol. 137, No. 9, 2011, pp. 601–607.
22. Stata: Data Analysis and Statistical Software. http://www.stata.com/help.cgi?qreg. Accessed June 2011.
23. Miranda, A. QCOUNT: Stata Program to Fit Quantile Regression Models for Count Data. http://ideas.repec.org/c/boc/bocode/s456714.html. Accessed June 2011.
24. Chen, C. An Introduction to Quantile Regression and the QUANTREG Procedure. *Proc., 30th Annual SAS Users Group International Conference,* Cary, N.C., SAS Institute, Inc., Cary, N.C., 2005, pp. 1–24.
25. The Comprehensive R Archive Network. http://cran.r-project.org/web/packages/quantreg/index.html. Accessed June 2011.
26. Winkelmann, R. Reforming Health Care: Evidence from Quantile Regressions for Counts. *Journal of Health Economics,* Vol. 25, No. 1, 2006, pp. 131–145.
27. Miranda, A. Planned Fertility and Family Background: A Quantile Regression for Counts Analysis. *Journal of Population Economics,* Vol. 21, No. 1, 2008, pp. 67–81.
28. Qin, X., and A. Wellner. *Development of Safety Screening Tool for High Risk Rural Roads in South Dakota.* Research report MPC-11-231. Upper Great Plains Transportation Institute, North Dakota State University, Fargo, 2011.
29. Gross, F., P. P. Jovanis, and K. A. Eccles. Safety Effectiveness of Lane and Shoulder Width Combinations on Rural, Two-Lane, Undivided Roads. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2103,* Transportation Research Board of the National Academies, Washington, D.C., 2009, pp. 42–49.

The Statistical Methods Committee peer-reviewed this paper.