



Analysis of U.S. freight-train derailment severity using zero-truncated negative binomial regression and quantile regression



Xiang Liu^{a,*}, M. Rapik Saat^{a,1}, Xiao Qin^{b,2}, Christopher P.L. Barkan^{a,3}

^a Rail Transportation & Engineering Center (RailTEC), University of Illinois at Urbana-Champaign, Newmark Civil Engineering Laboratory, 205 North Mathews Avenue, Urbana, IL 61801, United States

^b Department of Civil and Environmental Engineering, South Dakota State University, CEH 148, Box 2219, Brookings, SD 57007, United States

ARTICLE INFO

Article history:

Received 22 January 2013

Received in revised form 24 April 2013

Accepted 29 April 2013

Keywords:

Rail

Derailment severity

Zero-truncated negative binomial

Quantile regression

ABSTRACT

Derailments are the most common type of freight-train accidents in the United States. Derailments cause damage to infrastructure and rolling stock, disrupt services, and may cause casualties and harm the environment. Accordingly, derailment analysis and prevention has long been a high priority in the rail industry and government. Despite the low probability of a train derailment, the potential for severe consequences justify the need to better understand the factors influencing train derailment severity. In this paper, a zero-truncated negative binomial (ZTNB) regression model is developed to estimate the conditional mean of train derailment severity. Recognizing that the mean is not the only statistic describing data distribution, a quantile regression (QR) model is also developed to estimate derailment severity at different quantiles. The two regression models together provide a better understanding of train derailment severity distribution. Results of this work can be used to estimate train derailment severity under various operational conditions and by different accident causes. This research is intended to provide insights regarding development of cost-efficient train safety policies.

© 2013 Elsevier Ltd. All rights reserved.

1. Introduction

Railways are vital for U.S. national economy. While the society derives significant benefits from rail transportation, there are certain safety risks that must be managed and minimized to a feasible extent. Train accidents cause damage to infrastructure and rolling stock, disrupt services, and may cause casualties and harm the environment. Derailments accounted for 72% of freight-train accidents in the United States from 2001 to 2010 (Liu et al., 2012). Correspondingly, the analysis and prevention of train derailments is a high priority in the rail industry and government.

The probability of a train derailment has been studied by previous researchers (e.g., Nayak et al., 1983; Treichel and Barkan, 1993; Dennis, 2002; Anderson and Barkan, 2004; Kawprasert, 2010; Liu et al., 2011, 2012). In addition to analyzing the likelihood of a derailment, understanding the magnitude and variability of derailment severity is equally important. In this paper, derailment severity is measured by number of cars derailed after a train derailment

occurs. The generic use of “cars” refers to all vehicles (including locomotives, railcars and cabooses), unless specifically stated otherwise. Quantifying the relationship between train derailment severity and associated affecting factors could aid the rail industry and government to develop, evaluate, prioritize and implement cost-effective safety improvement strategies.

2. Literature review

Simulation and statistical analysis are the two basic approaches used in previous studies to model train derailment severity. Simulation models predict the response of railroad vehicles to specific track and environmental conditions. These models are typically based on detailed nonlinear wheel-rail interaction models. For example, Yang et al. (1972, 1973) developed a simulation model to determine the effect of ground friction, mating coupler moment, and brake retarding force on the number of cars derailed. They found that the position of the first car involved in the derailment (called point-of-derailment, or POD) and derailment speed could affect the number of cars derailed (Yang et al., 1972, 1973). In the late 1980s, Yang et al.'s model was extended by considering coupler failure and independent car motion (Coppens et al., 1988; Birk et al., 1990). The precision of simulation models is subject to the accuracy of modeling train derailment dynamics.

* Corresponding author. Tel.: +1 217 244 6063.

E-mail addresses: liu94@illinois.edu (X. Liu), mohdsaat@illinois.edu (M.R. Saat), xiao.qin@sdstate.edu (X. Qin), cbarkan@illinois.edu (C.P.L. Barkan).

¹ Tel.: +1 217 333 6974.

² Tel.: +1 605 688 6355.

³ Tel.: +1 217 244 6338.

In addition to simulation models, train derailment severity can also be estimated based on historical data. Saccomanno and El-Hage (1989, 1991) developed a truncated geometric model to estimate the mean number of cars derailed as a function of derailment speed, residual train length and accident cause. The model was modified by Anderson (2005) and Bagheri (2009), respectively. There is interest to consider new factors that may affect derailment severity. Last but not least, all previous derailment severity models focused on analyzing the mean number of cars derailed. Depending on factors discussed later in this paper, other distributional statistics may also need to be understood, such as quantiles.

3. Train derailment severity

The number of cars derailed in freight-train derailments on U.S. Class I⁴ railroad mainlines, from 2001 to 2010, is plotted in Fig. 1. On average, a freight-train derailment resulted in approximately 10 cars derailed, and the median severity is 6 cars derailed.

The literature has investigated the effect of accident cause on train derailment severity (e.g., Saccomanno and El-Hage, 1989, 1991; Barkan et al., 2003; Anderson, 2005; Bagheri, 2009; Bagheri et al., 2011; Liu et al., 2011, 2012). Broken rails are the most common cause of freight-train derailments on U.S. Class I mainlines (Barkan et al., 2003; Liu et al., 2012). On average, a broken-rail-caused freight-train derailment caused 14 cars derailed, compared with 7 cars derailed in a bearing-failure-caused derailment. Although a bearing failure has the potential to cause a severe train derailment, 50% of them caused a single-car derailment. Because broken rails likely to pose greater risk than other causes due to its high frequency and severity, this paper focuses on modeling train derailment severity for this cause. However, the methodology can be adapted to other accident causes.

4. Data

4.1. Data source

Train derailment data were from the Rail Equipment Accident (REA) database maintained by the Federal Railroad Administration (FRA) of U.S. Department of Transportation (U.S. DOT). Railroads in the U.S. are required to submit detailed accident reports on all accidents that exceeded a specified monetary threshold of damage costs to on-track equipment, signals, track, track structures and roadbed. The reporting threshold is periodically adjusted, and has increased from \$5700 in 1990 to \$9400 in 2011 (FRA, 2011). The REA database contains detailed train accident information such as total damage costs, number of cars derailed, track type, train length, derailment speed and others. In some previous studies, monetary damage has been used to assess the severity of train derailments. However, the financial cost of a derailment is subject to many variables, such as the cost difference between locomotives and railcars, or the difference in repairing regular track versus special trackwork. Instead, number of cars derailed may better represent train derailment severity under certain circumstances (Barkan et al., 2003).

4.2. Explanatory variables

Several factors may affect train derailment severity, including residual train length, derailment speed, train power distribution

and proportion of loaded railcars in the train. The explanation to each variable is presented below.

4.2.1. Residual train length

Residual train length is defined as the number of railcars following the point-of-derailment (POD), where POD is the position of the first car derailed. Residual train length describes the maximum number of cars potentially subject to derailment (Saccomanno and El-Hage, 1989, 1991). Previous studies have found that a greater residual train length is expected to result in more cars derailed, given all else being equal (e.g., Saccomanno and El-Hage, 1989, 1991; Anderson, 2005; Bagheri, 2009; Bagheri et al., 2011).

4.2.2. Derailment speed

Nayak et al. (1983), Treichel and Barkan (1993), Saccomanno and El-Hage (1991), Saccomanno and El-Hage (1989), Anderson (2005), Bagheri (2009), Bagheri et al. (2011) and Liu et al. (2011) all showed a positive correlation between the mean number of cars derailed and derailment speed.

4.2.3. Distribution of train power

No previous study analyzed whether distributed train power could affect train derailment severity. In this study, freight-trains are classified by two types: (1) non-distributed-power trains with only head locomotives and (2) distributed-power trains with head-end locomotives and additional locomotives in other positions (typically in the middle and/or in the rear). A binary variable (1 represents a distributed-power train, 0 otherwise) is created to examine the hypothesis that the two types of trains do not have statistically different derailment severities.

4.2.4. Proportion of loaded cars

This is another new factor considered in this paper. The proportion of loaded cars in the train is defined as the ratio of number of loaded cars normalized by total number of cars (both empty and loaded) in the train. The null hypothesis is that a train carrying a larger proportion of loaded cars may derail more cars. A larger proportion of loaded cars in the train may also indicate greater kinetic energy in the derailment, thereby causing more cars to derail, given all else being equal.

Table 1 presents some descriptive statistics of the studied explanatory variables. The Spearman correlation coefficients⁵ are presented in Table 2. The significant correlation is between train power distribution and the proportion of loaded cars in the train, at 5% significance level. It indicates that a derailed train having a higher proportion of loaded cars is more likely to be equipped with distributed power.

5. Zero-truncated negative binomial (ZTNB) model

5.1. Model development

The number of cars derailed represents non-negative count data, whose mean value can be estimated using regression techniques. Poisson regression and negative binomial (NB) regression are among the most popular count data regression methods used in accident analysis (e.g., Maccullagh and Nelder, 1989; Miaou, 1994; Hauer, 2001; Wood, 2002, 2005; Lord et al., 2005; Lord and Mannering, 2010). The Poisson model is suitable for data whose mean is equal to its variance, whereas the NB model assumes that the Poisson mean follows a gamma distribution. The NB model has been used for analyzing over-dispersed data (the variance is

⁴ A group of the largest railroads accounting for 97% of traffic (ton-miles) and 94% of total freight rail revenue in the U.S. (AAR, 2011).

⁵ Pearson correlation coefficients were also computed and yielded similar results.

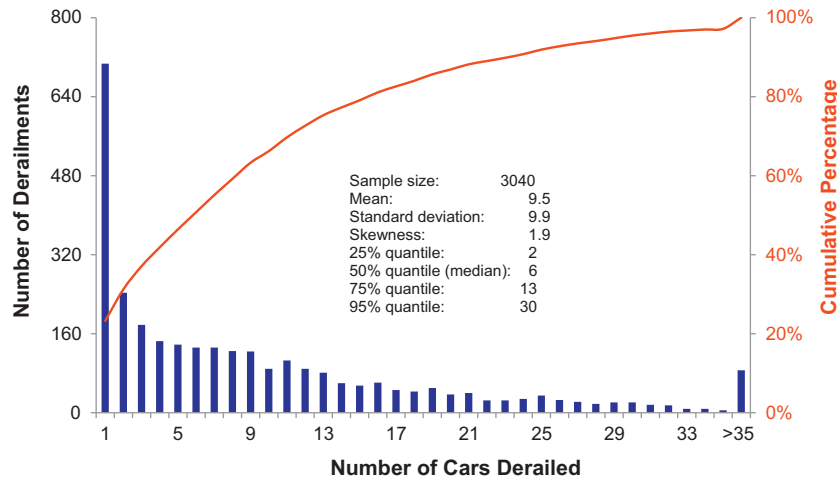


Fig. 1. Distribution of number of cars derailed per FRA-reportable freight-train derailment, all accident causes combined, U.S. Class I freight railroad mainlines, 2001–2010.

Table 1
Descriptive statistics of explanatory variables for broken-rail-caused freight-train derailments, U.S. Class I mainlines, 2001–2010.

Variables	Mean	Standard deviation	Minimum	Maximum	Type
Residual train length	51.1	35.2	1	152	Count
Derailement speed (mph)	28.8	14.1	4	70	Continuous
Proportion of loaded cars in the train	0.8	0.3	0	1	Continuous
Train power distribution	0.1	0.3	0	1	Binary

Table 2
Spearman correlation coefficient between explanatory variables for broken-rail-caused freight-train derailments, U.S. Class I mainlines, 2001–2010.

	Derailement speed	Proportion of loading cars	Train power distribution
Residual train length	−0.04	−0.03	0.11
Derailement speed		−0.09	0.05
Proportion of loading cars			0.29*

* Significant as 5% level.

greater than mean) by some previous researchers (e.g., Maccullagh and Nelder, 1989; Hauer, 2001; Long, 1997; Lord et al., 2005; Hilbe, 2007).

Both the Poisson and NB distributions include zeros, so they cannot be directly used to analyze data excluding zero counts, such as the train derailment severity data in this study. The smallest number of cars derailed in a derailment is 1. The Poisson or NB probability functions and their respective log-likelihood functions need to be modified to account for the exclusion of zeros. Gurmu (1991) and Grogger and Carson (1991) discussed methodologies to analyze zero-truncated count data. Compared to the traditional count data models (Poisson or negative binomial), the zero-truncated models calculate the probability of response variable based on positive count data using Bayes’s Theorem (Gurmu, 1991; Grogger and Carson, 1991; Long, 1997; Hilbe, 2007). Below shows the probability mass function (Eq. (1)), mean (Eq. (2)), variance (Eq. (3)), likelihood function (Eq. (4)) and response surface (Eq. (5)) of a zero-truncated negative binomial (ZTNB) model. The comparison of a ZTNB and NB model shows that the ZTNB model accounts for the exclusion of zeros, thus may have a greater probability and mean value of the response variable, given all else being equal. A detailed discussion of the ZTNB model can be found in Gurmu (1991) and Grogger and Carson (1991):

$$Pr(y_i|y_i > 0) = \frac{\Gamma(y_i + \alpha^{-1})/y_i! \Gamma(\alpha^{-1})(\alpha^{-1}/(\alpha^{-1} + \mu_i))^{\alpha^{-1}} (\mu_i/(\alpha^{-1} + \mu_i))^{y_i}}{1 - (1 + \alpha\mu_i)^{-\alpha^{-1}}} \quad (1)$$

$$E(y_i|y_i > 0) = \frac{\mu_i}{Pr(y_i > 0)} = \frac{\mu_i}{1 - (1 + \alpha\mu_i)^{-\alpha^{-1}}} \quad (2)$$

$$Var(y_i|y_i > 0) = \frac{E(y_i|y_i > 0)}{Pr(y_i > 0)^\alpha} [1 - Pr(y_i = 0)^{1+\alpha} E(y_i|y_i > 0)] \quad (3)$$

$$L = \prod_{i=1}^N Pr(y_i|y_i > 0) = \prod_{i=1}^N \frac{\Gamma(y_i + \alpha^{-1})/y_i! \Gamma(\alpha^{-1})(\alpha^{-1}/(\alpha^{-1} + \mu_i))^{\alpha^{-1}} (\mu_i/(\alpha^{-1} + \mu_i))^{y_i}}{1 - (1 + \alpha\mu_i)^{-\alpha^{-1}}} \quad (4)$$

$$\log(\mu_i) = \beta_0 + \beta_1 X_{1i} + \dots + \beta_k X_{ki} \quad (5)$$

where $Pr(y_i|y_i > 0)$ is the probability mass function of zero-truncated negative binomial distribution, $E(y_i|y_i > 0)$ is the expectation of zero-truncated negative binomial distribution, $Var(y_i|y_i > 0)$ is the variance of zero-truncated negative binomial distribution, α is the over-dispersion parameter, L is the likelihood function, μ_i is the estimated derailment severity for the i th observation, y_i is the observed derailment severity for the i th observation, β_k is the parameter coefficient of the k th predictor variable ($k=0$ for intercept), X_{ki} is the value of the k th predictor variable for the i th observation.

In terms of train derailment severity analysis, the response variable is the number of cars derailed. A ZTNB model is developed based on a set of 458 broken-rail-caused freight-train derailments on Class I mainlines from 2001 to 2010. The model accounts for main effect, higher-order component and their interaction terms of explanatory variables. A logarithmic transformation of residual

Table 3
ZTNB modeling results for broken-rail-caused freight-train derailments, Class I mainlines, 2001–2010.

Variable	Coefficient	Standard error	P-Value
Intercept	1.38	0.71	0.05
RL	−0.03	0.23	0.91
DS	−0.26	0.17	0.13
LO	−0.57	0.34	0.09
(RL) ²	−0.06	0.02	0.01
RL × DS	0.24	0.04	0.00
RL × LO	0.21	0.08	0.01

Notes: RL, logarithmic residual train length; DS, logarithmic derailment speed; LO, loading factor (proportion of loaded cars in a train).

train length and derailment speed was reported to have a better fit of train derailment severity data (Saccomanno and El-Hage, 1989, 1991). The same finding was found based on the data used in this paper. Table 3 shows the “final” model using hierarchical model selection technique based on the deviance (Agresti, 2007). The detailed model selection process is presented in Appendix A.

Correspondingly, the train derailment severity model is as below:

$$Z = \exp[1.38 - 0.03 \text{RL} - 0.26 \text{DS} - 0.57 \text{LO} - 0.06(\text{RL})^2 + 0.24 \text{RL} \times \text{DS} + 0.21 \text{RL} \times \text{LO}] \quad (6)$$

where Z is the estimated number of cars derailed per broken-rail-caused freight-train derailment on Class I mainlines. Using Eq. (6), the first-order partial-derivative of derailment severity with respect to each predictor variable is:

$$\frac{\partial Z}{\partial \text{RL}} = Z(-0.03 - 2 \times 0.06 \times \text{RL} + 0.24 \text{DS} + 0.21 \text{LO}) \quad (7)$$

$$\frac{\partial Z}{\partial \text{DS}} = Z(-0.26 + 0.24 \text{RL}) \quad (8)$$

$$\frac{\partial Z}{\partial \text{LO}} = Z(-0.57 + 0.21 \text{RL}) \quad (9)$$

There are several observations based on Eqs. (7)–(9):

- (1) The sensitivity of derailment severity to residual train length is also dependent on derailment speed and loading factor. If $\text{RL} < 2\text{DS} + 1.75 \text{LO} - 0.25$, then $(\partial Z/\partial \text{RL}) > 0$ (RL is the logarithmic residual train length). It means that derailment severity increases by residual train length when the residual train length is below a threshold determined by derailment speed and loading factor. The greater the speed and/or loading factor, the more likely that derailment severity has a positive correlation with residual train length, all else being equal. Whereas, given certain values of low speed or low loading factor, it is possible that derailment severity decreases when residual train length increases. However, the latter scenario is rare in the historical accident data (4 out of 458, or 0.08%).
- (2) The sensitivity of derailment severity with respect to derailment speed is dependent on residual train length. If $\text{RL} > 1.1$, then $(\partial Z/\partial \text{DS}) > 0$. It means that a greater derailment speed is associated with more cars derailed, all else being equal. This applies to 99% of observations in the historical data (453 out of 458).
- (3) The relationship between derailment severity and loading factor is dependent on residual train length as well. When $\text{RL} > 2.7$, then $(\partial Z/\partial \text{LO}) > 0$. It means that derailment severity increases if there is a larger proportion of loaded cars in the train. This applies to the majority of observations (394 out of 458, or 86%). Note that the remaining 14% accidents had either shorter train length or the derailment initiated from the rear end. Note that

loading factor herein does not provide information regarding the actual position of each loaded car, which is also an important factor affecting train derailment dynamics. Unfortunately, car position information is generally not publicly available, thus it may introduce uncertainty when applying the model to predict train derailment severity under these circumstances.

To summarize, there is a significant interaction effect between residual train length, derailment speed and loading factor. This interaction may be related to the energy accumulation in train accidents. Because of the interaction effects between explanatory variables, caution should be made when interpreting the effect of an individual factor on train derailment severity. Future research is needed to better understand train derailment process based on more factors and data.

5.2. Limitation of ZTNB model

The ZTNB model is suitable for analyzing the data excluding zeros. However, like other classical count data models, it has some limitations:

- (1) The model analyzes the mean response variable. However, the mean may not fully represent the data distribution. For example, for bearing-failure-caused accident causes (Fig. 2b), the mean derailment severity is 7, while the median is 1. Solely using the conditional mean may over-estimate derailment severity for a large proportion of observations. It raises a statistical question: how can we analyze the other distributional statistics (such as quantiles) in the regression?
- (2) The ZTNB or NB model is based on (homogeneous) Bernoulli assumption. Whether this assumption is true in reality may be subject to uncertainty. The ZTNB model requires that each car has (approximately) equal derailment probability following the point-of-derailment (POD). While this assumption is difficult to verify in the field, some previous models adopted this assumption and reported reasonable goodness-of-fit (Saccomanno and El-Hage, 1989, 1991; Anderson, 2005; Bagheri, 2009; Bagheri et al., 2011). Therefore, it is of interest to understand whether certain (restrictive) assumptions regarding data distribution can be relaxed. This motivates the introduction of quantile regression, which is the focus of the remaining sections.

6. Quantile regression

In addition to the mean, additional distributional statistics, specifically quantiles, are analyzed in this section. Quantiles are taken from the cumulative distribution function (CDF) of a random variable. Let p be a number between 0 and 1, and the p quantile of the distribution of a random variable Y is denoted by Q(p).

$$p = F(Q(p)) = \int_{-\infty}^{Q(p)} f(y) dy \quad (10)$$

$$Q(p) = F^{-1}(p) = \inf\{y : F(y) \geq p\}, \quad 0 \leq p \leq 1 \quad (11)$$

where $F^{-1}(p)$ is the inverse function of the cumulative distribution function (CDF), Inf is the smallest y that satisfies $F(y) \geq p$.

Compared to the ZTNB model that estimates the conditional mean, quantile regression (QR) estimates the conditional quantiles (such as the median). QR was originally developed by Koenker and Bassett (1978), and it has been applied in various research fields (e.g., Taylor, 1999; Arias et al., 2001; Machado and Mata, 2001; Nielson and Rosholm, 2001). The limited application of QR in transportation safety research includes Hewson (2008), Qin et al. (2010), Qin and Reyes (2011), and Qin (2012). The authors are not aware of

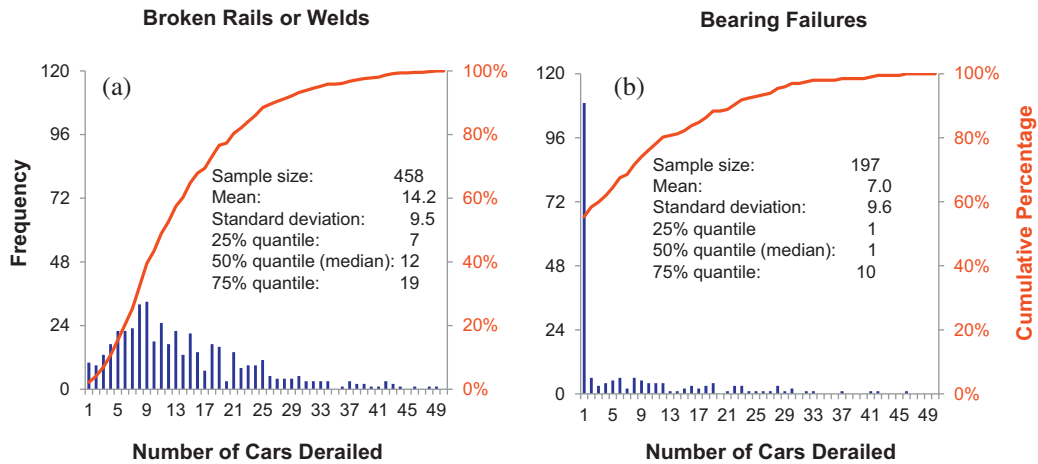


Fig. 2. Freight-train derailment severity distribution by major accident cause, U.S. Class I mainlines, 2001–2010.

any previous research that applied QR to rail safety research. Compared to the ZTNB model, quantile regression provides additional understanding of the distribution of derailment severity. We use the following linear programming proposed by [Koenker and Bassett \(1978\)](#) to estimate parameter coefficients in quantile regression:

$$\min \left[\sum_{y_i \geq Q_Y(p)} p|y_i - Q_Y(p|X)| + \sum_{y_i < Q_Y(p)} (1-p)|y_i - Q_Y(p|X)| \right] \quad (12)$$

where

$$\log(Q_Y(p|X)) = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k$$

To illustrate the concept of quantile regression, we consider the following example. For example, we aim to analyze the conditional quantile of derailment severity (number of cars derailed), denoted by $Q_Y(p|X)$, based on explanatory variables. It is assumed that the logarithmic transformation of $Q_Y(p|X)$ is a linear combination of predictor variables, that is $\log(Q_Y(p|X)) = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k$, where X is the main effect, higher-order component or the interaction of predictor variables. In order to determine the “optimal” set of parameter coefficients, we define an objective function. According to [Koenker and Bassett \(1978\)](#), the objective function is defined in the following way. First, all observations (train derailments) are classified into two groups based on derailment severity relative to a quantile value determined by specific values of explanatory variables and parameter coefficients. In the first group, all train derailments have derailment severity greater than the specified quantile. The other group has derailment severity below the quantile. Within each group, we can calculate the total absolute deviation between the observation-specific derailment severity and specified quantile value. The objective function is defined as the weighted average of total deviation within each group. The

“optimal” set of parameter coefficients is determined by optimizing the objective function. This optimization can be performed by various methods such as simplex algorithm, interior point method or smooth algorithm ([Chen, 2005](#)). QR exhibits several advantages over traditional count data regression models according to [Qin et al. \(2010\)](#), [Qin and Reyes \(2011\)](#), and [Qin \(2012\)](#):

- First, QR does not require the data to follow a specific probability distribution, such as normal or Poisson distribution. Therefore, it may be suitable for analyzing the data exhibiting strong skewness.
- Second, QR is more robust against outliers. Compared to mean, the median and other quantile values may be less sensitive to outliers and multi-modality.
- QR estimates multiple rates of change on different parts of the response variable distribution, and provides a complete view about the effects of predictors.

To date, QR is provided in several statistical software packages, such as QREG in Stata and QUANTREG in SAS. However, none of these standard statistical procedures were designed for discrete variables. In terms of count data, the objective function in the optimization is not differentiable, making it difficult to model quantiles directly as a continuous function of predictor variables ([Machado and Silva, 2005](#)). Smoothing approaches are needed to apply QR to count data ([Machado and Silva, 2005](#)). In this study, we use the jittering method developed by [Machado and Silva \(2005\)](#). Machado and Silva’s jittering algorithm was implemented in a procedure called QCOUNT in Stata by [Miranda \(2007\)](#). Using the QCOUNT in Stata, we develop a quantile regression model based on the freight-train derailment severity data. [Table 4](#) presents the QR estimates for selected quantiles and the associated standard errors. For comparison, we use the same parameters in developing ZTNB model shown

Table 4
Estimated coefficients in quantile regression versus ZTNB model, broken-rail-caused Class I mainline freight-train derailments, 2001–2010.

Variables	0.2 quantile		0.4 quantile		0.6 quantile		Mean	
	Coefficient	P-Value	Coefficient	P-Value	Coefficient	P-Value	Coefficient	P-Value
Intercept	1.90	0.01	1.47	0.00	0.97	0.06	1.38	0.05
RL	-0.43	0.11	-0.26	0.04	0.03	0.83	-0.03	0.91
DS	-0.25	0.16	-0.25	0.01	-0.21	0.06	-0.26	0.13
LO	-0.95	0.01	-0.50	0.10	-0.22	0.45	-0.57	0.09
(RL) ²	-0.02	0.43	-0.04	0.00	-0.07	0.00	-0.06	0.01
RL × DS	0.24	0.00	0.26	0.00	0.24	0.00	0.24	0.00
RL × LO	0.32	0.00	0.21	0.01	0.15	0.05	0.21	0.01

Notes: RL, logarithmic residual train length; DS, logarithmic derailment speed; LO, loading factor.

in Table 3. The analysis shows that the relationship between derailment severity and affecting factors may differ at different quantiles. There are two principal observations:

- First, the sign of coefficients for significant parameters is the same at the studied quantiles and the mean. It indicates that the effect of the same variable, although may be different in magnitude, is consistent in terms of direction. So change of a factor may consistently affect train derailments across all severities.
- The parameter coefficient in modeling the mean derailment severity using ZTNB model is within the range of coefficients using quantile regression between 20 and 60 percentiles. This is not surprising because the mean is also a special quantile value.

One advantage of quantile regression is that it can provide additional information regarding other quantiles. This may be supplementary to classical regression models on the conditional mean and may be particularly suitable when certain restrictive assumptions (e.g., Poisson distribution) do not hold under certain circumstances.

7. Comparison between conditional-mean and conditional-quantile models

In the last three decades, several statistical models have been developed to analyze count data in transportation safety research, of which the majority concentrate on estimating the conditional mean given covariates. However, the mean may be insufficient to describe highly skewed or multi-modal data distributions. For instance, estimating the conditional mean may over-estimate train derailment severity for a large proportion of derailments if the data are skewed toward low-severity derailments. As an alternative approach, quantile regression provides estimates at different quantiles, and provides additional views in analyzing the entire derailment severity distribution. Another advantage of quantile regression is its robustness against outliers in the dataset. A few outliers may significantly affect the mean and variance, but may not affect certain quantiles. Consequently, quantile regression may be useful in analyzing datasets that may have input errors.

As an important element of accident risk modeling, an appropriate measurement of derailment severity allows for the evaluation of the effectiveness of an accident prevention strategy. Quantile regression can be used to better understand the effects of explanatory variables on the response variable over a range of values. Specifically, if an explanatory variable has stable parameter

Model	Predictors	Log-likelihood	Number of parameters (including intercept)	Model comparison	Deviance difference
1	RL+DS+LO+PO	-1367.4	5		
2	RL+DS+LO	-1367.4	4	(2)-(1)	0.002 (df=1)
3	RL+DS+LO+(RL) ² +(DS) ² +(LO) ²	-1361.8	7	(3)-(2)	11.187 (df=3)
4	RL+DS+LO+(RL) ²	-1363.1	5	(4)-(3)	2.61 (df=2)
5	RL+DS+LO+(RL) ² +RL×DS+RL×LO+DS*LO	-1348.2	8	(5)-(4)	29.9 (df=3)
6	RL+DS+LO+(RL) ² +RL×DS+RL×LO	-1348.2	7	(6)-(5)	0.006 (df=1)
7	RL+DS+LO+(RL) ² +RL×DS+RL×LO+(RL) ² ×DS+(RL) ² ×LO	-1348.1	9	(7)-(6)	0.20 (df=2)

estimates at different quantiles, an accident prevention strategy affecting this variable may be effective for all derailments regardless of their severities. However, if the parameter estimates vary widely at different quantiles, it indicates that the accident prevention strategy may affect low-severity derailments differently than high-severity derailments. A better understanding of the effectiveness of an accident prevention strategy can result in an efficient allocation of resources to maximize the safety. To conclude,

quantile regression provides a tool to model the entire distribution of a response variable. Depending on the questions to address, the two approaches can supplement one another to provide assistance in train safety policy making.

However, it should be noted that quantile regression has its own limitations. For instance, development of quantile regression does not rely on the maximum likelihood method, thereby certain model selection criteria (e.g., AIC, BIC) may not directly be used for identifying the “best” model. The methodologies for evaluating the goodness-of-fit of a quantile regression model require further research. In addition, quantile regression originates for continuous variable and has only recently been adapted to count data (Machado and Silva, 2005). More work is needed to advance the theory and practicability of quantile regression in order to supplement the existing count data models.

8. Conclusions

Zero-truncated negative binomial (ZTNB) and quantile regression (QR) models are developed to understand the distribution of freight-train derailment severity and the effect of influencing factors. A ZTNB model is used to estimate the mean severity, accounting for the data excluding zeros. A QR model is used to estimate a specific quantile value and it is suitable for analyzing the entire data distribution accounting for possible data skewness or outliers or when certain distributional assumptions are not met. Combining the two models facilitates a better understanding of train derailment severity distribution. These methods can be adapted to various other related research problems.

Acknowledgements

The first author was supported in part by research grants from the Association of American Railroads (AAR), BNSF Railway, NEXTRANS University Transportation Center and National University Rail (NURail) Center. The authors are very grateful to the constructive comments from the anonymous reviewers. These comments aid substantial improvement of the draft manuscript. We also thank Dr. Yanfeng Ouyang from the University of Illinois at Urbana-Champaign (UIUC) for his useful comments on the manuscript. However, the authors are responsible for all analysis presented in this paper.

Appendix A. Model selection of zero-truncated negative binomial model

References

Agresti, A., 2007. *An Introduction to Categorical Data Analysis*. John Wiley & Sons, Inc., Hoboken, New Jersey.

Anderson, R.T., Barkan, C.P.L., 2004. Railroad accident rates for use in transportation risk analysis. *Transportation Research Record* 1863, 88–98.

Anderson, R.T., 2005. *Quantitative Analysis of Factors Affecting Railroad Accident Probability and Severity*. University of Illinois at Urbana-Champaign, Urbana (M.S. Thesis).

- 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.
- Association of American Railroads (AAR), 2011. Class I Railroad Statistics.
- Bagheri, M., 2009. Risk-Based Model for Effective Marshalling of Dangerous Goods Railway Cars. Department of Civil and Environmental Engineering, University of Waterloo, Ontario (Ph.D. Dissertation).
- Bagheri, M., Saccomanno, F., Chenouri, S., Fu, L.P., 2011. Reducing the threat of in-transit derailments involving dangerous goods through effective placement along the train consist. *Accident Analysis and Prevention* 43, 613–620.
- Barkan, C.P.L., Dick, C.T., Anderson, R.T., 2003. Analysis of railroad derailment factors affecting hazardous materials transportation risk. *Transportation Research Record* 1825, 64–74.
- Birk, A.M., Anderson, R.J., Coopens, A.J., 1990. A computer simulation of a derailment accident. Parts I & II. *Journal of Hazardous Materials* 25, 121–147, 149–165.
- Chen, C., 2005. An Introduction to Quantile Regression and the Quantreg Procedure. SAS Institute, Inc., Cary, NC.
- Coppens, A.J., Wong, J.D.E., Bibby, A., Birk, A.M., Anderson, R.J., 1988. Development of a Derailment Accident Computer Simulation Model. Transport Canada Report No. TP 9254E. Prepared for the Transportation Development Centre and Transport of Dangerous Goods, Ottawa, Canada.
- Dennis, S.M., 2002. Changes in railroad track accident rates. *Transportation Quarterly* 56 (4), 161–174.
- Federal Railroad Administration (FRA), 2011. Railroad Equipment Accident/Incident Reporting Threshold 1957–2011. FRA, U.S. Department of Transportation, Washington, DC.
- Grogger, J.T., Carson, R.T., 1991. Models for truncated counts. *Journal of Applied Econometrics* 6, 225–238.
- Gurmu, S., 1991. Tests for detecting over-dispersion in the positive Poisson regression model. *Journal of Business and Economic Statistics* 9, 215–222.
- Hauer, E., 2001. Overdispersion in modeling accidents on road sections and in empirical Bayes estimation. *Accident Analysis and Prevention* 33 (6), 799–808.
- Hewson, P., 2008. Quantile regression provides a fuller analysis of speed data. *Accident Analysis and Prevention* 40, 502–510.
- Hilbe, J.M., 2007. *Negative Binomial Regression*. Cambridge University Press, Cambridge, UK.
- Kawprasert, A., 2010. Quantitative Analysis of Options to Reduce Risk of Hazardous Materials Transportation by Railroad. Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, Urbana (Ph.D. Dissertation).
- Koenker, R., Bassett, G., 1978. Regression quantiles. *Econometrica* 46, 33–50.
- Liu, X., Barkan, C.P.L., Saat, M.R., 2011. Analysis of derailments by accident cause: evaluating railroad track upgrades to reduce transportation risk. *Transportation Research Record* 2261, 178–185.
- Liu, X., Saat, M.R., Barkan, C.P.L., 2012. Analysis of causes of major train derailment and their effect on accident rates. *Transportation Research Record* 2289, 154–163.
- Long, J.S., 1997. *Regression Models for Categorical and Limited Dependent Variables*. SAGE Publications, Inc., Thousand Oaks, CA.
- Lord, D., Washington, S.P., Ivan, J.N., 2005. Poisson-gamma and zero inflated regression models of motor vehicle crashes: balancing statistical fit and theory. *Accident Analysis and Prevention* 37 (1), 35–46.
- Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodologies alternatives. *Transportation Research Part A* 44, 291–305.
- Maccullagh, P., Nelder, J.A., 1989. *Generalized linear models*, 2nd ed. Chapman & Hall, London.
- 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 count. *Journal of American Statistical Association* 100 (472), 1226–1237.
- Miaou, S.P., 1994. The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. *Accident Analysis and Prevention* 26 (4), 471–482.
- Miranda, A., 2007. QCOUNT: Stata Program to Fit Quantile Regression Models for Count Data. <http://ideas.repec.org/c/boc/bocode/s456714.html> (accessed in February 2012).
- Nayak, P.R., Rosenfield, D.B., Hagopian, J.H., 1983. Event Probabilities and Impact Zones for Hazardous Materials Accidents on Railroads. Report DOT/FRA/ORD-83/20. FRA, U.S. Department of Transportation.
- 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.
- Qin, X., Ng, M., Reyes, P., 2010. Identifying crash-prone locations with quantile regression. *Accident Analysis and Prevention* 42 (6), 1531–1537.
- Qin, X., Reyes, P., 2011. A conditional quantile analysis for crash count data. *Journal of Transportation Engineering* 137 (9), 601–607.
- Qin, X., 2012. Quantile effects of causal factors on crash distributions. *Transportation Research Record* 2279, 40–46.
- Saccomanno, F.F., El-Hage, S., 1989. Minimizing derailments of railcars carrying dangerous commodities through effective marshaling strategies. *Transportation Research Record* 1245, 34–51.
- Saccomanno, F.F., El-Hage, S.M., 1991. Establishing derailment profile by position for corridor shipments of dangerous goods. *Canadian Journal of Civil Engineering* 18, 67–75.
- Taylor, J., 1999. A quantile regression approach to estimating the distribution of multiperiod returns. *Journal of Derivatives* 24, 64–78.
- Treichel, T.T., Barkan, C.P.L., 1993. Working Paper on Mainline Freight Train Accident Rates. Research and Test Department, Association of American Railroads, Washington, DC.
- Wood, G.R., 2002. Generalised linear accident models and goodness of fit testing. *Accident Analysis and Prevention* 34 (4), 417–427.
- Wood, G.R., 2005. Confidence and prediction intervals for generalised linear accident models. *Accident Analysis and Prevention* 37 (2), 267–273.
- Yang, T.H., Manos, W.P., Johnstone, B., 1972. A Study Continuation of Derailment Behavior Final Report (Phase 08 Report on Computer Derailment Study). RPI/AAR Report RA-08-1-12 (R-135). Railroad Tank Car Safety Research and Test Project, Association of American Railroads, Washington, DC.
- Yang, T.H., Manos, W.P., Johnstone, B., 1973. Dynamic analysis of train derailments. 72-WA/RT-6. In: Presented at the 1972 Winter Annual Meeting of ASME. In: *Rail Transportation Proceedings*, The American Society of Mechanical Engineers, New York, p. 8.