Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Application of Quantile Mixed Model for modeling Traffic Barrier Crash Cost

Mahdi Rezapour[a,*], Khaled Ksaibati[b], Milhan Moomen[a]

[a] Research Associate Wyoming Technology Transfer Center, 1000 E University Ave, Dept. 3295, Laramie, WY, 82071, United States
[b] Wyoming Technology Transfer Center University of Wyoming, 1000 E. University Avenue Department 3295, Laramie, WY, 82071, United States

ARTICLE INFO

Keywords:
crash cost
quantile mixed model
sparse data
non-parametric modeling
hierarchical model

ABSTRACT

Run-off the road crashes account for a significant proportion of severe injuries to vehicle occupants. Traffic barriers have been installed with an objective to keep vehicles on the roadway, and prevent them from hitting natural obstacles like trees or boulders. However, still injuries and fatalities of barrier crashes account for high proportion of fatalities on roadway. Due to challenging geometrics characteristics of Wyoming's roadway, a high mileage of barriers has been installed in the state. The high mileages of barriers result in a high number of barrier crashes in terms of crash frequency and severity due to high exposure. Previous studies mainly focused on crash frequency or individual crash severity. However, it has been recognized the importance of accounting for both aspects of crash severity, and crash frequency. So, in this study, crashes are aggregated across different barriers, and those crashes were converted into costs by considering the impacts of both crash severity and frequency. However, one of the main challenges of this type of dataset is highly skewness of crash data due to its sparseness nature. An improper use of model distribution of crash cost would result in biased estimations of the covariates, and erroneous results. Thus, in order to address this issue, a semi-parametric method of quantile regression technique was implemented to account for the skewness of the response by relaxing model distribution parameters. Also, to account for the heterogeneity in the dataset due to barriers' types, a random intercept model accounting for the structure of the data was implemented. In addition, interaction terms between significant predictors were considered. Understanding what factors with which magnitude contribute to the barrier crash costs is crucial for the future barriers' optimization process. Thus, contributory factors to barriers crash cost with high, medium, and low values, corresponding to 95th, 70th, and 60th percentiles were considered, and a comparison was made across these models. It was found, for instance, that although factors such as rollover, driving under the influence, and presence of heavy truck all have contributory impacts on the cost of crashes, their impacts are greater on higher quantiles, or higher barriers' costs. These models were compared from various perspectives such as intra class correlation (ICC), and standard error of coefficients. This study highlights the changes in coefficient estimates while modeling crash costs.

1. Introduction

More than a million people die every year due to traffic crashes worldwide, with more than 50 million are severely injured on roadways (OECD Publishing, 2017). The cost of these crashes in the U.S. alone approximate around $242 billion dollars, including various factors such as productivity lost, medical cost, property damage, and congestions costs (Blincoe et al., 2015).

Around 8,000 people died in the U.S. in 2016 alone due to collisions with fixed objects which is higher than fixed objects crashes recorded in 2015 (FARS & National Highway Traffic Safety Administration, 2016). Roadway departure crashes tend to be hazardous. For instance based on the literature review, while run-off the road crashes accounted for 16% of all crashes, they resulted in 31% of fatal crashes (Nason, 2005). Based on the federal highway administration (FHWA), a roadway departure crash (RwD) is defined as a crash in which a vehicle crosses an edge line, or a centerline and leave the traveled way (FHWA, 2019). It should be noted that the severity of these crashes is especially worse in rural and mountainous areas like Wyoming (Rezapour et al., 2018).

Different countermeasures can be implemented by policy makers to

[*] Corresponding author.
E-mail addresses: mrezapou@uwyo.edu (M. Rezapour), khaled@uwyo.edu (K. Ksaibati), mmoomen@uwyo.edu (M. Moomen).

https://doi.org/10.1016/j.aap.2020.105795
Received 23 January 2020; Received in revised form 25 May 2020; Accepted 15 September 2020
Available online 8 October 2020

address this high crash costs such as roadside enhancement, education, and enforcement. However, the first step for taking any countermeasure is through identification of the contributory factors to crashes. For identification of those factors Practitioners need to implement a sound methodological approach to account for the structure of datasets to minimize the biased results, and use a right response based on the needs of policy makers. This holds true especially when the objective is optimizing traffic barrier or roadside design.

For this study as our final objective is to conduct cost-benefit analysis, and as the dataset included various barrier types, it was important to deal with the dataset correctly. There are different ways of dealing with multi-level datasets (Raudenbush & Bryk, 2002). First technique is to disaggregate the dataset based on the levels of dataset, but this will violate the independence and homogeneity hypothesis (Chen & Jou, 2019). The second option is to ignore the variations across different groups and conduct the analysis on the dataset as a whole which would ignore variation across different levels (Raudenbush & Bryk, 2002). The third option is to take advantage of both aggregated and disaggregated technique which could be reflected in the mixed intercept model that accounts for the hierarchy of the dataset.

Another issue of analyzing traditional datasets such as crash frequency data is that it is assumed that all types of crash severity are equivalent. To address this issue, it is more practical to have equivalent property damage only (EPDO) as a response in the dataset, but these types of dataset are notorious for not following any distribution. The situation would be worsened if the cost is used instead of EPDO for response. It is also worth investigating the impacts of various predictors across various quantiles to see how the impacts of those predictors' coefficients vary across those quantiles.

Do the impacts of predictors vary across barriers with extreme costs compared with barriers that experience only a single crash? How the model goodness of fit varies across different hypothesis, various quantiles? These are just some of the questions that this study seek to answer.

This study is conducted to model traffic barriers' crash costs based on different contributory factors. Also, this study accounted for the hierarchy of the dataset by incorporating a random intercept model for grouping variables. Due to variation across different roadway systems, this study only focused on two-lane highway system. It should be noted the basic idea of the selected models is that included predictors would exert various impacts on crash costs with different coefficients at different quantiles (Koenker and Bassett, 1978). So, it would be of interest to evaluate those impact on traffic barrier crash costs.

2. Literature

Most of the past studies about traffic safety either focused on analyzing the frequency of crashes (Anastasopoulos & Mannering, 2009; Lord & Mannering, 2010; Qin & Reyes, 2011), or the severity of crashes (Liu et al., 2013; Qin & Reyes, 2011). On the other hand, some studies used equivalent property damage only (EPDO) crashes to account for both crash frequency and severity. However, the challenge of modeling EPDO compared with crash frequency is that the data would no longer follows the distribution of count data which are often modeled by gamma, negative binomial, or Poisson distributions. For EPDO, different crashes based on their severities would be converted to property damage only (PDO) to make a fair comparison across different locations. Various techniques have been used for modeling sparse EPDO crashes which this study would go over few of them.

A standard negative binomial regression model was adopted in the literature review to model EPDO (Oh et al., 2010). It was discussed that EPDO introduces a great deal of dispersion into the data. The study also noted that while the underlying assumptions of discrete distributions such as the negative binomial distribution make identification of contributory factors easier, it brings the assumptions under question.

A hurdle framework was proposed for modeling EPDO crashes (Ma et al., 2016). Hurdle model is a two-component models with a component for modeling positive counts, and a hurdle component which models the zero counts. For the positive part of the EPDO crash, three distributions including lognormal, gamma and normal were used and then these three hurdle models were compared against the Tobit model and the random parameter Tobit model. The results indicated that hurdle model outperformed the other models.

On the other hand, the quantile regression model is another popular method for modeling EPDO as the dependent variable. The following paragraph will go over few of the studies that used quantile regression approach to model EPDO as their dependent variable.

Quantile regression model was used to have flexibility in estimating trends at different quantiles, and accounting for heterogeneity in the dataset (Qin et al., 2010). The method was used for identification of risk-prone intersections. A comparison was also made between quantile regression at 95% quantile and generalized linear model (GLM) based on Poisson-gamma distribution. The results indicated that the two models are identical in signs but different slightly in magnitude.

A study was conducted to identify crash blackspots using quantile regression model (Washington et al., 2014). The proposed methodology identified covariate effects on various quantiles instead of the population mean. The study used Korean road segments as its dataset and the results were presented for 90%, 95% and 97% quantiles.

Although some studies used standard quantile regression, no traffic study has implemented the quantile mixed linear model. Thus, the following sections will discuss a study that used this method in areas other than transportation.

A study conducted to predict the constraint effect of environmental on density and diversity of macroinvertebrate (Fornaroli et al., 2015). In this study a linear model for quantile regression with a subject-specific random intercept was used to account for within-subject correlation. The effects included the presence or absence of waterfall which was accounted for by the random effect intercept.

Based on the literature review and to the best knowledge of the authors of this study, no study about traffic safety has been conducted using the linear mixed quantile model. Although a few studies have been undertaken using the linear quantile model, it is important also to account for the heterogeneity due to the structure of the dataset by incorporating a random intercept. Otherwise, the results would be biased. Thus, in this study we accounted for the structure of dataset, guardrail versus box beam barrier, by incorporating a random intercept which accounts for a grouping variable.

3. Method

Contrary to standard ordinary least square in linear regression, quantile regression does not assume a parametric assumption for the response, no constant variance (Rodriguez and Yao, 2017). This is one of the challenges for the dataset used in this study. The idea behind the linear quantile mixed model (LQMM) is that the covariates exert different impacts at different quantiles of the outcome distribution, and the degree of unobserved heterogeneity may be defined by variance parameters (Geraci and Bottai, 2014). The LQMM is based on the asymmetric Laplace (AL) distribution. This distribution forms a subset of geometric distribution. Compared with normal distribution, the AL distribution has steeper peaks and heavier tails, making it suitable for cost datasets (Kozubowski & Podgórski, 2001). A continuous random variable follows AL density with parameters (μ, σ, τ), the density follows the equation below (Yu & Zhang, 2005a)

$$P(\omega\mu,\ \sigma,\ \tau) = \frac{\tau(1-\tau)}{\sigma} exp\left\{ -\frac{1}{\sigma}\rho_{\tau}(\omega-\mu)\right\}, \tag{1}$$

Where μ is the location parameter, σ is the scale parameter, τ is the skewness parameter, and $\rho_{\tau}(.)$ is the loss function. It should be noted that a standard quantile regression has a similar linear format as general linear model as follows:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i, \ \mathrm{i} = 1, \ldots, \mathrm{n} \tag{2}$$

Where y_i is the response or crash cost, and for the ith observation, β_1 *to* β_p are different coefficients estimates and x_{ip} are predictors.

In the above equation, β_j*'s* in a simple linear regression which would be estimated by least square minimization as follows:

$$\frac{min}{\beta_0, \ldots, \beta_p}\left(\sum_{i=1}^{n}(y_i + \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 \tag{3}$$

On the other hand, quantile regression, with level/quantile τ of the response could be written as follows:

$$Q_\tau(y_i) = \beta_0(\tau) + \beta_1(\tau)x_{i1} + \ldots + \beta_p(\tau)x_{ip}, \ \mathrm{i} = 1, \ldots, \mathrm{n} \tag{4}$$

Where τ is the quantile level, and the β_j*'s* would be estimated by solving the following equation:

$$\frac{min}{\beta_0(\tau), \ldots, \beta_p(\tau)} \sum_{i=1}^{n} \rho_\tau(y_i + \beta_0(\tau) - \sum_{j=1}^{p} x_{ij}\beta_j(\tau)) \tag{5}$$

where ρ_τ is a check loss and could be written as:

$$\rho_\tau() = \tau\max(,0) + (1-\tau)\max(.,0) \tag{6}$$

Location of the parameter μ is the τ_{th} quantile of ω so for a fixed τ the τ_{th} regression quantile could be written as (Geraci and Bottai, 2014):

$$y_i = \mu^{(\tau)} + \varepsilon^{(\tau)} \tag{7}$$

Where $\varepsilon^{(\tau)} \sim AL(0, \sigma, \tau)$

From the Linear quantile mixed model (LQMM), and based on the above equation, the joint density of (y,u) based on M clusters in the τ_{th} quantile could be written as (Geraci and Bottai, 2014)

$$p\left(y, u\theta_x^{(\tau)}, \sigma^{(\tau)}, \varphi^{(\tau)}\right) = p\left(y\theta_x^{(\tau)}, \sigma^{(\tau)}, u\right)p\left(u\varphi^{(\tau)}\right) = \prod\nolimits_{i=1}^{M} p\left(y_i|\theta_x^{(\tau)}, \sigma^{(\tau)}, u_i\right)p\left(u_i|\varphi^{(\tau)}\right) \tag{8}$$

Where u is a random effect, $\theta_x^{(\tau)} \in R^p$ isavectorofunknownfixedeffect, $\sigma^{(\tau)}$ is scale of a joint AL model, $\varphi^{(\tau)}$isaq× qcovariancematrix, qisnumberofobservations, and M is number of clusters.

Fitting a group as a fixed effect assumes that the groups' means are independent of one another, sharing a common residual variance. However, fitting a group with a random intercept assumes that the groups' means are only subsets drawn from a global set of population (Harrison et al., 2018). The random intercept model accounts for subject-specific random intercept which could account for within-group correlation (Geraci & Bottai, 2006). For instance, median regression with random effects would be written as (Geraci & Bottai, 2006).

$$G_{y_i|u_i}(0.5|X_i, u_i) = X_i\beta + u_i, i = 1, \ldots., i \tag{9}$$

Where $\mathrm{u_i}$ is a random intercept.

The statistical analysis was conducted in this study in R software (Geraci and Bottai, 2014). This method uses gradient-search method to minimize the negative integrated log-likelihood in equation 10, which use the Clarke's derivative to find the path of steepest descent.

$$\iota_{app}(\theta, \sigma|y) = \sum_{i=a}^{M} log\{\sum_{k_1=1}^{K} \ldots \sum_{k_1=1}^{K} p(y_i|\theta_x, \sigma, (\Psi^T)^{1/2}\upsilon_{k_1 \cdots k_q}) \times \prod\nolimits_{i=1}^{q}\omega_{ki} \tag{10}$$

Where ι_{app} is the approximated AL-based log-likelihood, $\theta = (\theta_x^T, \ \theta_z^T)^T$ is parameter of interest, Ψ is q ×q covariance matrix, $\upsilon_{k_1 \cdots k_q}$ are nodes and $\omega_{ki} = 1$,…,q are weights.

It should be noted that the assumption of Asymmetric Laplace distribution is merely ancillary and is used just to cast estimation of M-quantile regression models into a maximum likelihood context (Marino et al., 2018).

In this study, intra class correlation (ICC) is used to show the total variability in the cost as a response that is not explained by the population average. ICC can be written as:

$$ICC = \frac{\widehat{\varphi}_u^2}{\widehat{\varphi}_u^2 + \widehat{\varphi}_\varepsilon^2} \tag{11}$$

Where $\widehat{\varphi}_u^2$ is an estimated variance of an unobserved random effect, and $\widehat{\varphi}_\varepsilon^2$ is a measure of residual variance or variance of unobserved noise. $\widehat{\varphi}_\varepsilon^2$ is based on residual scale parameter ($\widehat{\sigma}$) and quantile value, tau. $\widehat{\sigma}$ is the maximum likelihood estimation of scale parameter of an AL distribution (Yu & Zhang, 2005b). ICC provides information about the proportion of variance explained by clustering. It should be noted allowing ICC to be close to zero suggests that the variables are not correlated with one another and a grouping is unnecessary.

Various quantiles values were used for coefficient estimation to highlight and evaluate the changes in the estimated coefficients based on various quantile. In this study three quantile were considered, 95%,70% and 60%. Besides making a comparison across various quantiles to see the changes across the coefficient, considering the three quantiles would help for optimization process to use various equations for different barriers belonging to different quantiles. For instance, if a barrier severity level, based on cost, belongs to a high quantile of 0.95, the point estimates of that level should be used for optimization process of that specific barrier.

It is also expected that a higher quantile would result in a lower fit due to the presence of extreme values as the model would consider very high values in fitting the model. For a quantile, the estimated coefficient would be calculated from a conditional distribution based on that specific quantile.

4. Data

Different datasets from various resources were used in this study. Traffic barrier and roadside information were obtained from Wyoming department of transportation (WYDOT). The crashes were recorded for the years 2007-2017. A crash was extracted and incorporated into the final dataset if the first harmful event columns indicated collision resulted from hitting a traffic barrier. A traffic barrier crash was excluded from the dataset if more than one vehicle was involved in a crash due to complexity of the situation, and an existence of confounding factors. Due to differences across highway and interstate systems in Wyoming, data was also filtered to include only traffic barrier crashes in the highway system.

Roadway geometrics and traffic counts were obtained from an inventory dataset maintained by WYDOT. Traffic barrier height, length and offset and roadside characteristics of over 1.3 million feet of barrier were collected by WYDOT. The process of combining the datasets involved three stages. First, traffic, geometric, crash data, and traffic barrier geometric information were combined to the crash data based on road mileposts, and highway system IDs.

After this step, the dataset was collated based on traffic barriers grouping ID. Thus, there were indications for the sum of crash frequency based on severity level and average of predictors such as driver and traffic characteristics that the barrier experienced. For instance, if a barrier with an ID of 1 experiences 2 PDO crashes, with one driver under normal conditions and another driver under the influence, the response would be 2×$\mathrm{cost_{PDO}}$ and driver condition indicator would be the average of the driver condition as 0.5.

The total crash cost as a dependent variable was obtained from multiplying the sum of crash frequency and cost of each crash type from Table 1. If a barrier ID did not receive any crash, it was removed from the dataset. There was a total of 1,350 barrier IDs on the highway system with 515 of these barriers not experiencing any crash. It should be noted that ID is a number assign to each of the barriers in the state for

Table 1
Cost of different crashes based on figures from the Wyoming department of transportation (WYDOT).

Type crash	Cost
Fatality	9,604,727
Suspected serious injury	464,837
Unknown	149,551
Suspected minor injury	132,181
Possible injury	75,331
PDO	34612

identification, which is used in this study as barriers ID. Each barrier, barrier ID, is unique with different characteristics.

All these barriers accounted for 246 miles of barrier length. It should be noted from Table 1 that the cost of fatality is more than 277 times PDO crashes, highlighting the resultant distribution of barrier costs. The difference between severe crash and PDO is due to societal crash cost including various factors such as insurance administration, workplace cost, legal costs, and household productivity.

In Fig. 1, theoretical quantile versus sample quantile (QQ) plot of the log transferred of response, on the left, and the distribution of the response, on the right, are presented. Although a log transformation was used for the left part of Fig. 1, the figure still shows a severe departure from the Gaussian assumption, and presence of outlying observations. It should be noted for QQ plot that similar predictors as the included predictors in LQMM were used. The right part of Fig. 1 shows a distribution of response, crash cost. As can be seen from this figure, the crash cost is very sparse. For instance while majority of crashes were only one PDO crash with the cost of 34,612, see the highest pick on the left of the figure, there is only one crash with highest cost of 19,244,060, which is not clear due to only having one observation.

Based on Fig. 1, it is clear that a traditional distribution like the linear or gamma distribution cannot handle the severe over-dispersion and sparse nature of the cost response, heavy tail. M-quantile regression specification is appropriate for the data considering the need for a robust approach. Also, the dependence between observations recorded from the same cluster, barrier type, should be taken into consideration to avoid bias about regression coefficient inference. This can be done by inclusion of a random intercept accounting for barrier type in the model. For the semi-parametric model, the model distribution would be left unspecified and estimated directly from the data (Alfò et al., 2017).

Table 2 presents summary statistics of significant predictors in the statistical analysis. Due to very low frequency of cable, and concrete barriers, those observations were removed from the dataset. The dataset therefore included box beam and guardrail barriers only. Average annual daily traffic (AADT), length of a curve and length of barrier were incorporated in the model as a means of exposure, normalizing the dataset. These variables were found to be important and were therefore retained in the model. Again, the observations were aggregated from crash dataset into traffic barrier dataset. In the original crash dataset, most of the predictors were categorical, which are highlighted in Table 2. For instance, the average of alcohol involvement is 0.124, which means that most of the drivers hitting barriers were under normal condition, normal condition as 0 versus others as 1.

Table 2
Descriptive analysis of the significant variables

Variable	Mean	Max	Min	Sd.dev
Type (box-beam = 1[a], W-beam = 2)	1.5	2	1	0.498
Barrier height (continuous) ft	2.44	3.4	1.5	0.235
AADT (continuous)	3,17	27670	95	3405.69
Alcohol involvement (No alcohol involvement[a] versus others)	0.12	1	0	0.286
Barrier length (continuous)	874	35,471	14.396	1816.786
Heavy truck involvement (Non heavy truck involvement * versus others)	0.045	1	0	0.182
Side slope flat (reference)	206	———	———	———
Side slope(Filled)	540	———	———	———
Side slope(cut)	90	———	———	———
Shoulder width (continuous)	4.59	58	0	3.330
Restrain status (Unrestrained drivers[a] versus others)	0.66	1	0	0.427
Road condition (Non dry road condition[a]versus others)	0.57	1	0	0.446
Rollover (Non rollover crash[a] versus others)	0.034	1	0	0.170
Length of a curve (continuous) ft	934	3700	100	
Citation record (Drivers did have a citation record[a] versus others)	0.439	1	0	0.441

[a] reference category.

5. Results

The barrier crash cost was modeled by a QLMM because the crash cost has a sparse distribution, and distribution of the residual does not look like any parametric distribution. To account for a poor model distribution due to the presence of outliers, and severe skewness of the response a semi-parametric model was chosen over parametric ones. Also, random intercept for barrier types was employed to estimate a conditional quantile, adjusting for covariates. Group was set as a barrier type in the syntax of the model to incorporate barrier-specific random effects to account for within barriers' clustering. The benefit of this analysis is that the crash cost model captures a holistic view of crashes by considering both crash frequency and crash severity.

Barriers crash cost with low, medium, and high values for he included covariate were considered. In this study. the high, medium and

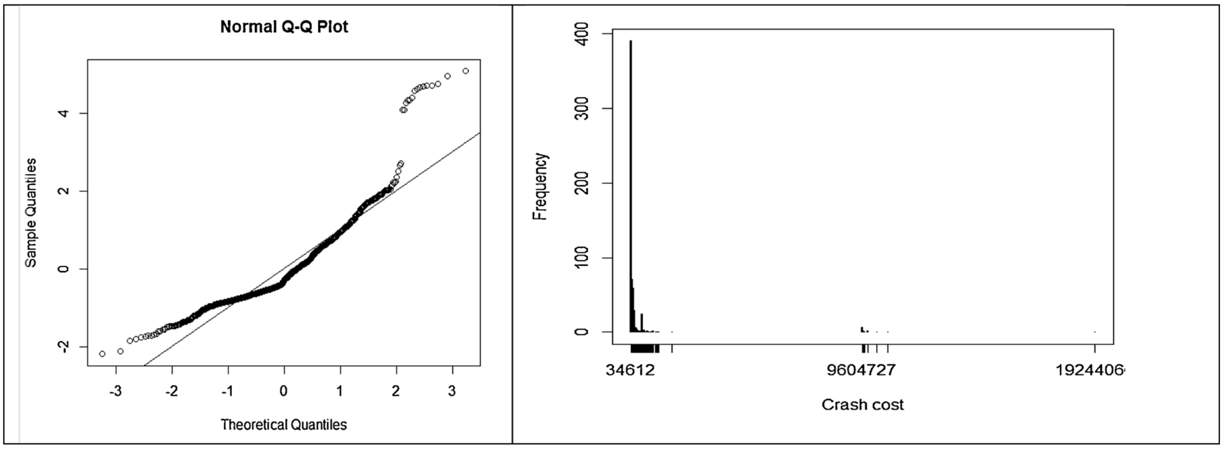

Fig. 1. The quantile-quantile plot of log of crash cost (left) and frequency of response (right)

low correspond to the 95th, 70th, and 60th percentiles of barriers' crash costs. Table 3 presents the model estimation results for those quantiles with block-bootstrap with 50 as the number of replications. In general, the interpretation of the model parameter is specific to the quantile under consideration. The estimates of the parameters at q-quantile could be interpreted as an impact of one unit change in the covariate on a given q-quantile of the response (Marino et al., 2018).

It should be noted a relatively high estimated coefficient is due to having crash costs in dollars in the response. So, a unit change of predictor on response is for every single dollar. In order to make a comparison between various quantile models, models with similar predictors were chosen across the three models. The Akaike information criterion (AIC), and ICC were used for the model comparisons.

Although in most parts the results show similar significant predictors across the models for different quantiles, the values of the coefficients differ across the models. For instance, although the heavy truck variable for the 95% quantile has a contributory impact on the crash cost, this value was found to be insignificant for the other two quantiles, with various magnitudes. On the other hand, while a reduction impact of non-dry road conditions on crash costs was found to be significant for the 95% and 60% quantile, this effect was found not to be important for the 70% quantile.

In terms of side slope, it was found that although cut side slope and fill side slop have close coefficient estimates, cut side slope coefficient estimates were found to be statistically insignificant for all the quantiles. Fill side slope quantile, on the other hand, was found to contribute to the cost of barriers crashes. This results is in accordance with a previous study, finding vehicles that the roads are less likely to involve in severe crashes if the side slope is flat (Marquis & Weaver, 1976).

On the other hand, although length of curve has a preventive significant impact on crash costs on the 95% quantile and 60%, it was insignificant for the other quantiles. The negative impact of this predictor is expected and related to a positive correlation of this variable, length of a curve and radius.

In terms of driver behavior, while alcohol involvement and driver restrain were both significant for all the quantiles, their impacts were more predominant for the 95% quantile. This highlights the higher impacts of these predictors in causing a higher crash cost, in a higher

Table 3
Linear mixed quantile models for different quantile values.

variable	Value	Std. error	95% confidence intervals		*p*-value
			2.5%	97.5%	
95% Bootstrapped quantile regression, AIC = 26,913, BIC = 26993 DF = 17, $\widehat{\varphi}_u^2 = 4.93\text{e-}32$, $\widehat{\varphi}_\varepsilon^2 = 335,4817$,ICC ~ 0					
Intercept	7.34e+05	1.07E+06	−1.42E+06	2.90E+06	0.494
Barrier height	−1.4456e+05	1.14E+05	−3.73E+05	8.36E+04	0.208
AADT	1.3519e+01	3.71E+00	6.06E+00	2.10E+01	0.001
Alcohol involvement	6.5675e+05	1.89E+05	2.77E+05	1.04E+06	0.001
Barrier length	2.5677e+02	1.05E+02	4.61E+01	4.67E+02	0.018
Heavy truck involvement	6.3172e+05	3.28E+05	−2.65E+04	1.29E+06	0.050
Side slope (filled)	8.8727e+04	3.57E+04	1.71E+04	1.60E+05	0.016
Side slope (cut)	6.3066e+04	3.46E+04	−6.52E+03	1.33E+05	0.074
Shoulder width	−2.0428e+05	7.79E+04	−3.61E+05	−4.76E+04	0.011
Citation record	−9.8197e+04	5.71E+04	−2.13E+05	1.65E+04	0.091
Restrain status	8.3456e+06	3.83E+06	6.57E+05	1.60E+07	0.033
Road condition	−2.4523e+05	1.19E+04	−2.69E+05	−2.21E+05	< 0.001
Roll over	4.9964e+05	7.59E+04	3.47E+05	6.52E+05	< 0.001
Length of a curve	−1.0133e+02	2.96E+01	−1.61E+02	−4.18E+01	0.0011
Shoulder width: barrier height	8.8041e+04	3.43E+04	1.91E+04	1.57E+05	0.013
70% Bootstrapped quantile regression, AIC = 24,615, BIC = 24695, DF = 17, $\widehat{\varphi}_u^2 = 3.641\text{e+}11$, $\widehat{\varphi}_\varepsilon^2 = 820,135$, ICC = .44					
Intercept	1.08E+06	3.12E+05	4.51E+05	1.70E+06	0.001
Barrier height	−1.27E+05	9.90E+04	−3.26E+05	7.23E+04	0.206
AADT	8.67E+00	4.18E+00	2.56E-01	1.71E+01	0.043
Alcohol involvement	1.03E+05	1.23E+03	1.00E+05	1.05E+05	< 0.001
Barrier length	4.81E+01	1.20E+01	2.40E+01	7.22E+01	< 0.001
Heavy truck involvement	1.06E+04	1.39E+04	−1.74E+04	3.85E+04	0.451
Side slope(filled)	6.46E+04	1.37E+04	3.71E+04	9.21E+04	< 0.001
Side slope(cut)	4.97E+03	7.88E+03	−1.09E+04	2.08E+04	0.530
Shoulder width	−9.76E+04	3.70E+04	−1.72E+05	−2.31E+04	0.011
Citation record	−2.62E+04	1.55E+04	−5.74E+04	5.04E+03	0.098
Restrain status	1.08E+05	4.23E+04	2.30E+04	1.93E+05	0.013
Road condition	−3.97E+04	2.10E+04	−8.18E+04	2.44E+03	0.064
Roll over	1.32E+05	4.13E+04	4.91E+04	2.15E+05	0.002
Length of a curve	−5.68E+00	6.67E+00	−1.91E+01	7.72E+00	0.398
Shoulder width: barrier height	4.06E+04	1.47E+04	1.11E+04	7.01E+04	0.007
60% Bootstrapped quantile regression, AIC = 24,197,BIC = 24277, DF = 17, $\widehat{\varphi}_u^2 = 1.341\text{e+}11$, $\widehat{\varphi}_\varepsilon^2 = 496,980$, ICC = .35					
Intercept	1.03E+06	7.63E+04	8.73E+05	1.18E+06	< 0.001
Barrier height	−5.04E+04	3.81E+04	−1.27E+05	2.61E+04	0.191
AADT	6.13E+00	1.68E+00	2.76E+00	9.51E+00	0.001
Alcohol involvement	7.75E+04	4.63E+03	6.82E+04	8.68E+04	< 0.001
Barrier length	3.60E+01	9.56E+00	1.67E+01	5.52E+01	0.0004
Heavy truck involvement	−9.01E+03	8.43E+03	−2.59E+04	7.94E+03	0.290
Side slope(filled)	3.94E+04	9.35E+03	2.06E+04	5.82E+04	0.0001
Side slope (cut)	3.05E+03	4.29E+03	−5.57E+03	1.17E+04	0.4807
Shoulder width	−4.94E+04	2.41E+04	−9.79E+04	−9.10E+02	0.046
Citation record	−8.55E+03	7.44E+03	−2.35E+04	6.41E+03	0.256
Restrain status	6.64E+04	2.83E+04	9.59E+03	1.23E+05	0.022
Road condition	−2.10E+04	6.95E+03	−3.50E+04	−7.02E+03	0.004
Roll over	6.98E+04	1.01E+04	4.95E+04	9.01E+04	< 0.001
Length of a curve	3.87E+00	1.01E+00	1.84E+00	5.90E+00	0.0003
Shoulder width: barrier height	2.07E+04	9.67E+03	1.31E+03	4.02E+04	0.034

quantile. The results highlighted the impact of alcohol crashes accounting for more than $50 billion or 22% of all crash economic costs every year (Blincoe et al., 2015). The impact of barrier length on crash cost is expected: the higher the length of a traffic barrier, the more cost is associated with that barrier. It is interesting to note that again this impact is more highlighted for higher quantiles. It is worthy to mention that intercept values in the models highlight the impact of barrier types. A positive value indicates that the box-beam as a reference category is safer than W-beam. Drivers' citation record was incorporated in the model due to the importance of this predictor and even though the p-values are slightly higher than 0.05. The results indicated that drivers with citation record have a higher associated cost compared with drivers with no citation record (citation record as 0).

Interaction terms have important implications in interpretation of the results. They aid in a better understanding of the combined impacts of traffic barrier and roadway characteristics. Another benefit of inclusion of interaction terms is that they could capture heterogeneity across different models. The significance of the two-term interaction between barrier height and shoulder width means that the impact of those variables cannot be separated. The interaction was considered based on a previous study (Rezapour et al., 2019). It should be noted that a random intercept is incorporated in the model to account for the correlation of repeated observations within barriers' types.

5.1. Models' goodness of fit and ICC comparison

The coefficients across various quantiles ae calculated based on conditional distribution of the response, allowing the distribution shape to be adjusted based on predictors. Beside comparison of the models based on their coefficients' estimates, it is worth comparing models from goodness of fit perspective. Besides the results of three quantiles, Table 3 presents the standard error, which highlights the spread of the data, confidence interval (CI), and p-value. In addition, on top of each model section, AIC, degree of freedom (DF), $\widehat{\varphi}_u^2$: the estimated variance of an unobserved random effect, or intercept of covariance matrix of the random effect, $\widehat{\varphi}_\varepsilon^2$: the variance of error term ε, and ICC values are presented.

The estimated variance parameters of the random effects, $\widehat{\varphi}_u^{\,2}$, is very small at $\tau = 0.95$ but larger at the 75% and 60% quantiles. This is an indication that there is very little heterogeneity due to barriers; types across crashes with various barriers' types for the 95% quantile compared to the other two quantiles. This finding is supported by the very low ICC value of 4.93e-32 for this quantile indicating a lack of correlation among crashes across the different barrier types. In other words, very little variability in the response is explained by the barriers' grouping for 95% quantile model. This might be due to the fact that other factors are at play for barriers with very height crash costs, and the costs cannot be accounted for by barrier types.

In terms of ICC, for instance, for the third model the random effect has an estimated variance of $\widehat{\varphi}_u^2 = 1.341\mathrm{e}+11$, resulting in an ICC of .35. This indicates that 35% of the variability of the crash cost is due to unobserved heterogeneity across traffic barrier types. Very low ICC for a higher quantile is in accordance with a previous study that ICC follows an inverted U curve (Borgoni et al., 2018)

Another implication of the ICC values is that while no variance related to the grouping levels (barrier types) was observed (ICC = 0) for the first model, multilevel analyses are relevant for the other two analyses. In other words, the results of ICC favor retaining random intercepts for the last two models and discarding that for the first model.

Across the three presented models, the optimal solution could be chosen based on a minimum AIC value. As can be seen, a decrease in a quantile resulting in a lower AIC. In other words, 95% quantile resulted in a worst fit while the 60% quantile outperform the other two models. This a possibly due to having less extreme observations in the dataset. Also, there is no correlation between ICC and AIC values. While the second model grouping accounted for a highest ICC value, not necessarily it resulted in a lowest AIC. In summary, for highest quantile, there are so much unseen factors that barriers' type as grouping could account for only 0% of the whole heterogeneity.

It is sensible to have a comparison across the models with other possible models. Similat to count models, there are not as many options for comparison due to the continuous response type of the data. Hierarchical linear model could be discarded due to the nature of the sparse nature of cost. Hierarchical Gamma or log-normal, being very close in terms of the performance, could be considered. However, as Fig. 1 depicted even the log transformation of the response is skewed and cannot be considered as normal. In addition, the aforementioned models are all parametric and could not account for various quantiles.

Thus, we considered standard quantile linear model, with no grouping as a comparison. Although AIC or BIC penalize for the incorporated number of variables for a fair comparison, the barrier type which was considered as grouping in LQMM was considered as an explanatory variable in LQM. The results of AIC for LQM for 95%, 70% and 60% are 27,117, 25,052, and 24,818 respectively, which are higher than the correspondents LQMM, which highlights a better fit of LQMM compared with LQM.

6. Conclusion

Although the majority of studies in the literature only focused on crash frequency or crash severity, it is important to account for both characteristics by incorporating factors such as crash cost. Also, while some studies analyzed EPDO as a response category, only few studies considered analyzing crash cost. This is due to sparse nature of crash cost especially for an area like Wyoming with a low crash number, and high crash fatality rates. This increase complexity of cost dataset requiring a robust and versatile method for building a model.

Non-parametric or semi parametric analyses could be employed to address the nature of the crash cost data. Besides the necessarily assumptions that need to be met, a regression model would only answer questions about the impact of different predictors on the mean of crash cost. However, it cannot answer the important questions about how different predictors' impact on crash cost differ across low versus high cost locations. This study was conducted to capture a more comprehensive picture of the impacts of different predictors on barrier crash costs.

The residual distribution of crash cost data does not fit any known distribution such as normal, or gamma,. A non-parametric method such as quantile regression model could be used to account for the distributional limitation of this type of data. Another benefit of regression quantile model is that it is expected that different predictors would exert various impacts with different coefficients at different quantile, which is worth investigating. Also, to account for the heterogeneity of the model, random intercept was introduced to account for the source of unobserved heterogeneity across different barrier types.

The results of this study indicated that although the random intercept model, accounting for different barrier levels, were warranted for the 70% and 60% quantiles, no such variability was observed for the 95% quantile. This is due to a possible reason that for barriers with very high crash costs, there are much confounding factors contributing to these crashes that the impact of barrier types is minimal.

It was interesting to see that the magnitude of the coefficients for driver actions, driver under influence and other driver conditions increase with an increase in a quantile, highlighting the impact of these predictors on higher barriers crash costs. All the interaction terms across different important factors, especially roadway and barrier characteristics were evaluated and incorporated in the model. It was found that the two-term interaction between should width, and barrier height are important, and the impact of any of these predictors should be considered as an interaction term. In this paper we put all the effort such as inclusion of hierarchy and interaction terms to account for the heterogeneity in the dataset.

In this study we incorporated traffic, and length of barrier as means of exposure. For future studies, an optimization would be conducted to minimize the cost of barrier crash cost based on available resources, considering exposure parameters in the model. The result of the future study would answer if changes in the roadway characteristics would be cost effective based on associated cost and future cost reduction.

6.1. *Concluding remarks*

In summary the following point have been reached:

1 A higher quantile not only impact the magnitude of the coefficients but also impact the significance of various incorporated coefficients, which might be due to unobserved confounding factors, especially for high barriers crash costs.
2 Higher ICC is not necessarily correlated with a lower quantile. Although ICC was 0 for 95% quantile, the value is maximum for 70% quantile as ICC = 0.44. Compared with ICC of 0.35 for quantile = 60%. Also, ICC = 0 for highest quantile might be due to much confounding factors that would ignore the impact of barrier types as grouping.
3 For all the coefficients, higher standard error can be observed for a higher quantile. This is expected to be due to the presence of outliers.
4 A higher confidence interval could be observed for higher quantiles due to higher uncertainty. Although a higher magnitude in coefficients could be due to the impact of extreme values in higher quantile, this could be also due to uncertainty in coefficient estimated which need more investigation.

For this study we just considered barriers with crashes as barriers with no crashes did not have drivers' characteristic such as alcohol involvement, or citation record as a crash has not been occurred. Inclusion of barriers with no crashes is important in the state as much of the barriers are not based on recommended designs, and much of them have not experienced any crash. To achieve the aforementioned criteria the following analysis could be considered in the future studies to incorporate barriers with no crashes as follows:

1 It is possible to only consider variables that are similar across barriers with crashes and with no crashes. Those included predictors such as barriers' types, geometric characteristics, traffic, and barrier length. That model could be implemented on both barriers with and without crashes.
2 Instead of using cost as response, EPDO could be used. For this type of response various model such as negative binomial could be conducted on both barriers with and without crashes.
3 As negative binomial might not perform optimally for excess number of zeroes, two component models, hurdle or zero-inflated models, are expected to perform better. Those two-component models would have two layers: one model for barriers with zero count crashes and one model for barriers with crashes. In order to account for grouping factor that we considered in this study; hierarchical model could be a closest model to the implemented model in those studies.

As discussed, after identification of factors to barriers 'crashes, the final objective is to conduct cost-benefit analysis. This would be implemented through quantile machine learning technique. The algorithm would be trained over the original dataset. Then, variables especially barriers geometric characteristics, such as barriers' height would be optimized to their optimal values. The trained model would be implemented again over a new dataset and cost-benefit output would be estimated.

For instance, barriers' optimum height is 27 inches for box-beam. In many places, the barriers' height is less than that value. Thus, first the cost would be predicted based on the barrier current height. Then, the barrier's height would be changed to 27 inches. The trained algorithm would be conducted again, and cost would be predicted. The difference would be the cost/benefit output.

Authors' statements

Mahdi Rezapour: Methodology. **Khaled Ksaibati:** Funding acquisition, and Supervision. **Milhan Moomen:** Data curation and Investigation.

Declaration of Competing Interest:

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper

Acknowledgments

The authors would like to acknowledge that this work is part of project #RS03218 funded by the Wyoming Department of Transportation (WYDOT). The subject matter, all figures, tables, and equations not previously copyrighted by outside sources are copyrighted by WYDOT, State of Wyoming and the University of Wyoming. All rights reserved copy righted in 2016.

References

Alfò, M., Salvati, N., Ranallli, M.G., 2017. Finite mixtures of quantile and M-quantile regression models. Statistics and Computing 27 (2), 547–570.

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

Blincoe, L., Miller, T.R., Zaloshnja, E., Lawrence, B.A., 2015. The Economic and Societal Impact of Motor Vehicle Crashes, 2010 (Revised).

Borgoni, R., Del Bianco, P., Salvati, N., Schmid, T., Tzavidis, N., 2018. Modelling the distribution of health-related quality of life of advanced melanoma patients in a longitudinal multi-centre clinical trial using M-quantile random effects regression. Statistical Methods in Medical Research 27 (2), 549–563.

Chen, T., Jou, R., 2019. Using HLM to investigate the relationship between traffic accident risk of private vehicles and public transportation. Transportation Research Part A: Policy and Practice 119, 148–161.

FARS, N., National Highway Traffic Safety Administration, 2016. Fatality analysis reporting system. On-Line: Http://Www-Fars.Nhtsa.Dot.Gov/Main/Index.Aspx,.

Fornaroli, R., Cabrini, R., Sartori, L., Marazzi, F., Vracevic, D., Mezzanotte, V., Canobbio, S., 2015. Predicting the constraint effect of environmental characteristics on macroinvertebrate density and diversity using quantile regression mixed model. Hydrobiologia 742 (1), 153–167.

Geraci, M., Bottai, M., 2006. Quantile regression for longitudinal data using the asymmetric laplace distribution. Biostatistics 8 (1), 140–154.

Geraci, M., Bottai, M., 2014. Linear quantile mixed models. Statistics and Computing 24 (3), 461–479.

Harrison, X.A., Donaldson, L., Correa-Cano, M.E., Evans, J., Fisher, D.N., Goodwin, C.E., Inger, R., 2018. A brief introduction to mixed effects modelling and multi-model inference in ecology. PeerJ 6, e4794.

Koenker, R., Bassett Jr, G., 1978. Regression quantiles. Econometrica: Journal of the Econometric Society 33–50.

Kozubowski, T.J., Podgórski, K., 2001. Asymmetric laplace laws and modeling financial data. Mathematical and Computer Modelling 34 (9-11), 1003–1021.

Liu, X., Saat, M.R., Qin, X., Barkan, C.P., 2013. Analysis of US freight-train derailment severity using zero-truncated negative binomial regression and quantile regression. Accident Analysis & Prevention 59, 87–93.

Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. Transportation Research Part A: Policy and Practice 44 (5), 291–305.

Ma, L., Yan, X., Wei, C., Wang, J., 2016. Modeling the equivalent property damage only crash rate for road segments using the hurdle regression framework. Analytic Methods in Accident Research 11, 48–61.

Marino, M.F., Tzavidis, N., Alfò, M., 2018. Mixed hidden markov quantile regression models for longitudinal data with possibly incomplete sequences. Statistical Methods in Medical Research 27 (7), 2231–2246.

Marquis, E.L., Weaver, G.D., 1976. Roadside slope design for safety. Transportation Engineering Journal of the American Society of Civil Engineers 102 (1), 61–74.

Nason, N., 2005. TRAFFIC SAFETY FACTS 2005-A compilation of motor vehicle crash data from the fatality analysis reporting system and the general estimates system, national highway traffic safety administration. National Center for Statistics and Analysis, US Department of Transportation, Washington, DC, 20590.

OECD Publishing, 2017. Road safety annual report 2017 OECD Publishing.

Oh, J., Washington, S., Lee, D., 2010. Property damage crash equivalency factors to solve crash Frequency–Severity dilemma: Case study on south korean rural roads. Transportation Research Record 2148 (1), 83–92.

Qin, X., Ng, M., Reyes, P.E., 2010. Identifying crash-prone locations with quantile regression. Accident Analysis & Prevention 42 (6), 1531–1537.

Qin, X., Reyes, P.E., 2011. Conditional quantile analysis for crash count data. Journal of Transportation Engineering 137 (9), 601–607.

Raudenbush, S.W., Bryk, A.S., 2002. Hierarchical linear models: Applications and data analysis methods Sage.

Rezapour, M., Wulff, S.S., Ksaibati, K., 2018. Effectiveness of enforcement resources in the highway patrol in reducing fatality rates. IATSS Research.

Rezapour, M., Wulff, S.S., Ksaibati, K., 2019. Examination of the severity of two-lane highway traffic barrier crashes using the mixed logit model. Journal of Safety Research 70, 223–232.

Rodriguez, R.N., Yao, Y., 2017. April). Five things you should know about quantile regression. Proceedings of the SAS global forum 2017 conference, Orlando 2–5.

Washington, S., Haque, M.M., Oh, J., Lee, D., 2014. Applying quantile regression for modeling equivalent property damage only crashes to identify accident blackspots. Accident Analysis & Prevention 66, 136–146.

Yu, K., Zhang, J., 2005a. A three-parameter asymmetric laplace distribution and its extension. Communications in Statistics—Theory and Methods 34 (9-10), 1867–1879.

Yu, K., Zhang, J., 2005b. A three-parameter asymmetric laplace distribution and its extension. Communications in Statistics—Theory and Methods 34 (9-10), 1867–1879.