Accident Analysis and Prevention 185 (2023) 107019

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Mitigating the impact of outliers in traffic crash analysis: A robust Bayesian regression approach with application to tunnel crash data

Zhenning Li [a,*,1], Haicheng Liao [b,1], Ruru Tang [a], Guofa Li [c], Yunjian Li [d], Chengzhong Xu [b,*]

[a] State Key Laboratory of Internet of Things for Smart City and Department of Civil and Environmental Engineering, University of Macau, Macau SAR 999078, China
[b] State Key Laboratory of Internet of Things for Smart City and Department of Computer and Information Science, University of Macau, Macau SAR 999078, China
[c] College of Mechanical and Vehicle Engineering, Chongqing University, Chongqing 400044, China
[d] Institute of Applied Physics and Materials Engineering, University of Macau, Macao SAR 999078, China

ARTICLE INFO

Keywords:
Robit model
Bayesian inference
Robust regression
Traffic safety modeling
Tunnel crash

ABSTRACT

Traffic crash datasets are often marred by the presence of anomalous data points, commonly referred to as outliers. These outliers can have a profound impact on the results obtained through the application of traditional methods such as logit and probit models, commonly used in the domain of traffic safety analysis, resulting in biased and unreliable estimates. To mitigate this issue, this study introduces a robust Bayesian regression approach, the robit model, which utilizes a heavy-tailed Student's t distribution to replace the link function of these thin-tailed distributions, effectively reducing the influence of outliers on the analysis. Furthermore, a sandwich algorithm based on data augmentation is proposed to enhance the estimation efficiency of posteriors. The proposed model is rigorously tested using a dataset of tunnel crashes, and the results demonstrate its efficiency, robustness, and superior performance compared to traditional methods. The study also reveals that several factors such as night and speeding have a significant impact on the injury severity of tunnel crashes. This research provides a comprehensive understanding of the outliers treatment methods in traffic safety studies and offers valuable recommendations for the development of appropriate countermeasures to effectively prevent severe injuries in tunnel crashes.

1. Introduction

Recent years have witnessed considerable applications of advanced econometric models in the field of traffic safety analysis to analyze the effects of different contributing factors on injury severity or crash frequency. Among these models, logit and probit regression models are the most commonly used in practice. However, they have recently been criticized because their estimators are not robust when the dataset contains outliers (Krueger et al., 2023; Liu, 2004; Newson and Falcaro, 2022). Outliers are usually considered as extreme or unexpected observations in the statistical domain, i.e., in some way deviate from the general pattern of the majority or "bulk" of the data (van Dyk and Meng, 2001). In traffic accident datasets, outliers are widely present. Since the dataset was mainly collected and reported by police officers after crashes, there is no doubt that this processing may introduce some errors and mistakes into the dataset. Although some outliers can be easily identified by statistical measures and removed from the dataset, many outliers (especially the ones in the small dataset) should not be removed directly because they actually reflect the real scenarios of the crash and are not incorrect data (El-Basyouny and Sayed, 2010). For example, some uncommon causes (e.g., fire, falling from a bridge) may result in mass fatalities even if other factors are considered to be favorable (e.g., sunny weather, appropriate speed, going straight, etc.). On the other hand, some crashes may only result in property damage only (PDO) injuries, even though all contributing factors are considered to be dangerous (e.g., drunk driving, speeding, limited visibility, etc.) (Li et al., 2018). The presence of a small proportion of outliers in the real data can have a significant distorting effect on the mean and variance. Therefore, it violates the basic assumption of the logit and probit models for data errors that they are normally distributed or at least can be approximately estimated with central limit theorem. Consequently, when the dataset contains outliers, the classical approach may provide biased estimates with poor performance in terms of breakdown points and influence functions (Markatou and He, 1994). However, how to deal

* Corresponding author.
E-mail address: li2016@hawaii.edu (Z. Li).
[1] Equal contribution.

https://doi.org/10.1016/j.aap.2023.107019
Received 28 November 2022; Received in revised form 5 February 2023; Accepted 2 March 2023
Available online 11 March 2023

with such outliers has not been discussed in depth in the field of traffic safety analysis, even though it is deeply entrenched in the dataset.

As classified by El-Basyouny and Sayed (2010), there are majorly three different types of methods for analyzing traffic crash data with possible outliers, including the "outlier ignoring" methods, the "outlier rejecting" methods, and the "outlier accommodating" methods. They concluded that, of all three methods, the outlier accommodating method, which uses a robust method to reduce the weight of potential outliers, provides the best estimates when all the data are taken into consideration. A severe problem with the other two methods is that the outlier editing procedure (i.e., deleting and truncating atypical data) may lead to inferences failing to account for the uncertainty in the exclusion process. In particular, the standard errors of these distributions tend to be too small.

When it comes to the outlier accommodating methods, a common way is to use the hierarchical-mixture models with different priors (e.g., Tukey-Huber model, contaminated normal distribution model), which assumes that the majority of the observations come from a primary distribution and the few remaining outliers come from an alternative distribution with large variance. However, this type of mixture, which requires expert settings, does not adequately reflect the nature of the data itself, and therefore also leaves the model estimates with biases beyond negligence (Gelman and Hill, 2006). A more comprehensive discussion about the shortcomings of these approaches can be found in the article of Rahimian and Mehrotra (2019).

Another alternative idea in the outlier accommodation approach is to accommodate the outliers and observations together with a robust distribution and use statistical indicators to determine the parameters of the distribution. Robust distributions are heavy-tailed or fat-tailed, with density tails that tilt slower toward zero than normal density tails. Thus, the parameter estimates are able to fit most of the data well when the data contain outliers, and also when the data have no outliers. The most frequently used heavy-tailed distributions for fully parametric approaches to robust modeling and inference include Weibull distribution, Log-normal distribution, Student's t distribution, and so on. Considering the common settings in traffic analysis, a two-tailed distribution is more suitable for analyzing the impacts of contributing factors. With all these in mind, the heavy-tailed Student's t distribution becomes the ideal distribution for our problem. It is symmetric and bell-shaped, in the same way with the normal distribution. However, the Student's t distribution has heavier tails, indicating that it is not that sensitive to the outliers and is useful for understanding the statistical behavior of certain types of ratios of random quantities (Barron, 2019). In addition, the tail weights of the t-distribution vary with the degrees of freedom ν, thus, the t-distribution has a better ability to accommodate different types of outliers in the dataset. Moreover, with different degrees of freedom, it can also approximate other widespread distributions, such as the Cauchy distribution (when $\nu = 1$), the normal distribution (when $\nu \to +\infty$), and so on (Roy, 2012).

Our focus in this study is on a dataset of tunnel accidents that poses unique challenges in terms of data analysis. With a relatively low frequency of incidents and a considerable variance among the contributing factors, it becomes evident that some records in the dataset are noticeably different from the majority of the data, resulting in the presence of outliers. In order to address these complexities, we introduce a cutting-edge robust Bayesian model with binary response, named the robit model, that leverages the heavy-tailed Student's t distribution as the link function to analyze the impact of various factors on tunnel accidents. Additionally, we propose a sandwich algorithm built on the data augmentation method to enhance the estimation of the posteriors, along with a novel approach to estimating degrees of freedom based on the data, rather than empirical estimates. The results of our model estimation demonstrate the robustness and efficiency of our approach, with better model performance compared to traditional probit and logit models for this particular dataset. As far as we are aware, this is the pioneering study to employ such a model in the domain of traffic safety analysis.

The remainder of this paper is organized as follows. The following section presents the tunnel dataset. Section 3 details the specification of the proposed robit model. Section 4 illustrates the estimation results of the model. The paper is finally concluded in Section 5.

1.1. Data

The current study utilizes a dataset collected by the Guangdong Provincial Department of Traffic Police in China, with a specific focus on the G94 Jiangmen-Zhaoqing Expressway. This major transportation corridor extends over 107.7 km and boasts advanced engineering design, including a six-lane configuration in both directions, with three lanes in each direction, and a design speed limit of 100 km/h.

To ensure the reliability and pertinence of the data, a rigorous filtration and cleaning procedure was performed on the dataset. The final sample consisted of 485 crashes that occurred within three separate tunnels along the expressway: the Dawangding Tunnel (measuring 2.2 km in length), the Jiangjunshan Tunnel Group (comprised of five subtunnels with a combined length of 3.7 km), and the Maozhanling Tunnel (measuring a total of 9.5 km and being the longest tunnel in Guangdong Province). Only crashes that took place in the access, entrance, transition, exit zones, outside the tunnel exit, and mid-zone were retained in the sample (Pervez et al., 2022b).The study period covered the years 2014 to 2019, offering a comprehensive examination of road safety and crash trends over a five-year span.

While the sample size of the dataset may be considered modest, certain outliers in the descriptive factors of some accidents were observed, which deviate significantly from the overall observations. Record #137, for instance, is representative of a serious injury accident involving an experienced bus driver on a sunny summer morning, resulting from spontaneous vehicle combustion. Record #262 highlights a young male driver who collided with the roadway shoulder, resulting in a superficial injury following excessive speed and a vehicle rollover. An additional example of a record presenting outliers is a crash involving a driver who suffered a heart attack while behind the wheel (Record #312). The vehicle veered off the road and collided with the tunnel wall, resulting in serious injury to the driver but minimal damage to the vehicle. However, after careful examination, we found that these records were not caused by erroneous recording, but in fact they reflect the real crash situation. Therefore, they are kept in the dataset although then may introduce potential outlying issues in the dataset. The detailed information of these records is provided in Table 1.

In the data sample, the injuries sustained in the crashes were divided into two categories: severe injury and non-severe injury. Severe injury involved death or disability resulting from the crash within seven days, while non-severe injury included non-disability injury and property damage only. Out of the 485 crashes in the final dataset, 61 were severe and 424 were non-severe. The variables in the dataset were grouped into four categories based on their relevance to the study: environmental characteristics (such as season, day of week, time of day, and weather), roadway geometric characteristics (such as road grade and curve), vehicle and crash characteristics (such as vehicle type and collision type) and driver characteristics (including age, gender, and driver experience).

2. Methodology

In this study, the dependent variable is dichotomous including two responses, i.e., severe injury and non-severe injury, thus a binary regression model is suitable for analyzing the impacts of corresponding factors on the injury severity. Suppose y_i is the injury severity for the i-th crash where y_i is either 0 or 1. A binary regression model using Bernoulli distribution as the distribution function in a Generalized Linear Models fashion could be written as

Table 1
Statistical description of the variables in the tunnel accident dataset.

Variable	Description	Mean
Injury severity	Non-severe* = 1, otherwise = 0	0.874
Environmental characteristics		
Season	Spring = 1, otherwise = 0	0.237
	Summer = 1, otherwise = 0	0.272
	Autumn* = 1, otherwise = 0	0.210
	Winter = 1, otherwise = 0	0.280
Day of week	Weekday* = 1, otherwise = 0	0.687
Time of day	Daytime* = 1, otherwise = 0	0.726
Weather	Clear* = 1, otherwise = 0	0.431
	Cloudy = 1, otherwise = 0	0.274
	Rain = 1, otherwise = 0	0.295
Roadway geometric characteristics		
Road grade	Downgrade = 1, otherwise = 0	0.130
	Flat = 1, otherwise = 0	0.588
	Upgrade* = 1, otherwise = 0	0.282
Curve	Straight* = 1, otherwise = 0	0.538
Vehicle and crash characteristics		
Vehicle type	Car* = 1, otherwise = 0	0.773
	Bus and van = 1, otherwise = 0	0.054
	Truck = 1, otherwise = 0	0.173
Collision type	Rear-end = 1, otherwise = 0	0.652
	Hitting fixed object = 1, otherwise = 0	0.163
	Rollover* = 1, otherwise = 0	0.095
	Other = 1, otherwise = 0	0.091
Driver characteristics		
Driver behavior	Speeding = 1, otherwise = 0	0.016
	Fatigue driving = 1, otherwise = 0	0.037
	Violation of distance keeping = 1, otherwise = 0	0.676
	No violation* = 1, otherwise = 0	0.270
Driver age	20–29 = 1, otherwise = 0	0.190
	30–39* = 1, otherwise = 0	0.365
	40–49 = 1, otherwise = 0	0.359
	>50 = 1, otherwise = 0	0.087
Driver experience	0–4 = 1, otherwise = 0	0.379
	5–9* = 1, otherwise = 0	0.313
	10–14 = 1, otherwise = 0	0.163
	>15 = 1, otherwise = 0	0.144
Driver gender	Male* = 1, otherwise = 0	0.934

Note: * selected as the reference category.

$$Y_i \sim \text{Bernoulli}(p_i) \tag{1}$$

where y_i equals 1 with probability $p_i = \text{pr}(y_i = 1|x_i, \beta)$ and 0 with the probability $1 - p_i$. Assume $F^{-1}(p_i) = x_i^T\beta$ is the link function, and $x_i, (i = 1, 2\cdots, n)$ is a $p \times 1$ vector of corresponding factors, and β is a $p \times 1$ vector of coefficients. If the link function $F(\cdot)$ is the cumulative distribution function of the normal or logistic distribution, the model becomes a probit or logistic one. However, as widely evidenced by previous studies, the estimates of coefficients for both models are not robust to the outliers (Little and Rubin, 2019; Shafieezadeh Abadeh et al., 2015). Considering the nature of the data set, this study proposes a more robust model, namely robit regression model, which replaces the link function of above two models with the Student's t link function (Liu, 2004).

As further suggested by Gelman and Hill (2006), the robit model, different from the probit and logistic models, allows for a better model fit by effectively reducing the weights of discordant data points in the presence of outliers. In addition, both probit and logistic models can be well approximated by the robit links with different degrees of freedom (Kim et al., 2008). Previous studies showed that the probit link could be approximated with high accuracy by a robit model with large degrees of freedom, while the logistic link could be well approximated with seven degrees of freedom. Thus, in certain cases, the robit model can be considered as a more general binary regression model because it is able to approach other widely used models, for instance, probit and logistic, by replacing links with robit links with appropriate degrees of freedom and can provide a robust estimate when outliers exist.

More formally, the robit regression model for the data is

$$\text{pr}(y_i = 1|x_i, \beta) = 1 - \text{pr}(y_i = 0|x_i, \beta) = F_\nu\left(x_i^T\beta\right) \tag{2}$$

where $F_\nu(\cdot)$ is the cumulative distribution function of the univariate Student's t distribution and ν is the corresponding degrees of freedom which is known and fixed. The tail of the Student's t-density is heavier than the normal and logistic distribution, and it provides a robust alternative to these distributions in cases where outliers are suspected in the data (Kang and Schafer, 2007). The density of this link function is given by

$$f_\nu(x) = \frac{\Gamma\left(\frac{(\nu+1)}{2}\right)}{(\pi\nu)^{\frac{1}{2}}\Gamma\left(\frac{\nu}{2}\right)\left(1+\frac{x^2}{\nu}\right)^{\frac{(\nu+1)}{2}}} (x \in (-\infty, \infty)) \tag{3}$$

It can be readily observed that this robit(ν) model becomes the probit regression model when $\nu \to \infty$.

The methodology of estimation in statistical modeling holds a critical significance in determining the validity of the inferences drawn from the data. In instances where the sample size is ample and the outliers are uncommon, maximum likelihood estimation (MLE) has been widely utilized as a conventional approach for determining the parameters of probit and logistic models. MLE operates under the principle of maximizing the likelihood of observing the data, given the model, to estimate the values of the parameters. However, when the sample size is limited, MLE may prove to be inadequate in fully capturing the intricacies of the underlying data structure, thereby leading to biased estimates for the probabilities of unseen samples. This can have a deleterious impact on the estimation process and the conclusions drawn from the data.

In response to this issue, Bayesian estimation offers a more appropriate solution. Bayesian estimation constitutes a framework that accounts for uncertainty and incorporates prior knowledge into the estimation process. Instead of determining the maximum likelihood estimates, Bayesian estimation calculates the posterior distribution of the parameters, given the data and the prior distribution. This integration of prior knowledge and modeling of uncertainty leads to more robust inferences even in the face of limited sample size. Therefore, in cases where the sample size is restricted and the occurrence of outliers is substantial, Bayesian estimation can be deemed a more suitable approach compared to MLE. This study opted for a Bayesian approach in analyzing the data, as it was deemed the most appropriate method for accurately estimating the probabilities of unseen samples, given the limited sample size and the substantial presence of outliers.

For the Bayesian robit, a prior distribution for the vector of regression parameters β needs to be assumed first. With Eq. (3), a multivariate Student's t prior on β can be given as

$$\pi(\beta) = \frac{\Gamma\left(\frac{(\nu_0+p)}{2}\right)}{(\pi\nu_0)^{\frac{p}{2}}\Gamma\left(\frac{\nu}{2}\right)}|\Sigma_0|^{\frac{1}{2}}\left[1+\nu_0^{-1}\beta^T\Sigma_0\beta\right]^{-\frac{p+\nu_0}{2}} \tag{4}$$

where the prior for the parameter β is 0-centered $p \times 1$ multivariate Student's t distribution $t_p\left(0, \Sigma_0^{-1}, \nu_0\right)$, where Σ_0 is a $p \times p$ positive definite scatter matrix, and ν_0 is the known degrees of freedom. $\pi(\beta)$ is the marginal prior for β under Zellner's g-prior (Zellner, 1986) when $\Sigma_0 = aX^TX$, where X is a $n \times p$ matrix whose element in the ith row is x_i^T, and a is a constant. Correspondingly, the posterior density $\pi_\nu(\beta|y)$ is given as

$$\pi_\nu(\beta|y) = \frac{1}{m_\nu(y)} l_\nu(\beta|y) \times \pi(\beta) \tag{5}$$

where $m_\nu(y)$ is the normalizing constant, and is given by

$$m_\nu(y) := \int_{\mathbb{R}^p} l_\nu(\beta|y) \times \pi(\beta) d\beta \tag{6}$$

and $l_\nu(\beta|y)$ is the likelihood function and given by

$$l_\nu(\beta|y) = \Pi_{i=1}^{n} \left(F_\nu\left(x_i^T\beta\right)\right)^{y_i} \left(1 - F_\nu\left(x_i^T\beta\right)\right)^{1-y_i} \tag{7}$$

Previous studies have shown that inference based on Eq. (7) can often be reduced to calculating the posterior expectations (Liang et al., 2011), that is

$$E_\pi h := \int_{\mathbb{R}^p} h(\beta)\pi_\nu(\beta|y) d\beta \tag{8}$$

However, $E_\pi h$ is a ratio of two intractable integrals that cannot be easily obtained in closed form. Moreover, in most of the traffic safety dataset, the number of contributing factors, p, is large, thus vanilla Monto Carlo methods become problematic because they need solid requirements of independent and identically distributed samples. Recent studies have suggested that data augmentation (DA) algorithms have the potential to be used to explore the posterior density $\pi_\nu(\beta|y)$ (Liang et al., 2011; Roy, 2012). Nevertheless, DA algorithms tend to suffer from slow convergence like their deterministic counterpart, the EM algorithm (van Dyk and Meng, 2001). Therefore, in this study, following Hobert and Marchev (2008), we use an alternative MCMC algorithm to estimate parameters, namely the *sandwich algorithm (SW)*, which is equivalent to the DA algorithm mathematically, with faster and more efficient convergence to the stationary distribution. In the following, we will first introduce the DA algorithm and then the SW algorithm.

Suppose $t_\nu(\mu, 1)$ denotes the univariate Student's t distribution whose location is μ, scale is 1 and degrees of freedom is ν. We conduct the DA algorithm with the fact that t distribution can be represented as a scale mixture of normal distributions. Then, let $z = (z_1, z_2, \cdots, z_n)^T$ be n independent variables with $z_i \; t_\nu\left(x_i^T\beta, 1\right)$, $z_i|\lambda_i \; N\left(x_i^T\beta, \frac{1}{\lambda_i}\right)$, and $\lambda = (\lambda_1, \lambda_2, \cdots, \lambda_n)^T$ with $\lambda_i \; \text{Gamma}\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$, then z_i can be regarded as the latent variable of y_i. Similarly, the multivariate Student's t prior $\pi(\beta)$ for β can also be expressed using a scale mixture of multivariate normal distributions, that is

$$\tau_0 \; \text{Gamma}\left(\frac{\nu_0}{2}, \frac{\nu_0}{2}\right) \tag{9a}$$

and

$$\beta|\tau_0 \; N_p\left(0, \frac{\Sigma_0^{-1}}{\tau_0}\right) \tag{9b}$$

Let (z, λ, τ_0) be the *augmented data* to realize $\pi(\beta, (z, \lambda, \tau_0)|y)$, which is the joint posterior density of (β, λ, z) given y:

$$\int_{\mathbb{R}} \int_{\mathbb{R}^+} \int_{\mathbb{R}^+} \pi(\beta, (z, \lambda, \tau_0)|y) \mathrm{d}z\mathrm{d}\lambda\mathrm{d}\tau_0 = \pi_\nu(\beta|y) \tag{11}$$

Therefore, the marginal density of $\pi(\beta, (z, \lambda, \tau_0)|y)$ is the target posterior density $\pi_\nu(\beta|y)$. Thus, we can obtain the posterior density by making draws from the conditional densities on β and the augmented data, i.e., $\pi(\beta|z, \lambda, \tau_0, y)$ and $\pi(z, \lambda, \tau_0|\beta, y)$. The current state β to the next state β' can be obtained by the following DA algorithm:

DA Algorithm:

Step 1	Draw $\{(\lambda_i, z_i), i = 1, 2, \cdots, n\}$ by first drawing the truncated t distribution $z_i \; Tt_\nu(x_i^T\beta, y_i)$, then draw $\lambda_i \; \text{Gamma}\left(\frac{\nu+1}{2}, \frac{\nu + (z_i - x_i^T\beta)^2}{2}\right)$ and independently draw $\tau_0 \; \text{Gamma}\left(\frac{\nu_0 + p}{2}, \frac{\nu_0 + \beta^T\Sigma_0\beta}{2}\right)$
Step 2	Then draw $\beta' \tilde{\text{N}}_p\left(\widehat{\beta}, \left(X^T\Lambda X + \tau_0\Sigma_0\right)^{-1}\right)$

As further proved by Brooks et al. (2011), the DA algorithm is Harris ergodic and can finally converge to the target density in Eq. (5). The interested reader is referred to the book of Brooks et al. (2011), and references cited therein. However, as mentioned earlier, the DA algorithm has recently been often criticized for its low convergence efficiency.

In general, the DA algorithm has two steps, i.e., $\beta \to (z, \lambda, \tau_0) \to \beta'$. While in the SW algorithm, an additional step is sandwiched between Step 1 and Step 2 in the DA, and the other steps in the DA algorithm are kept. Therefore, the SW algorithm can be viewed as $\beta \to (z, \lambda, \tau_0) \to \left(z', \lambda', \tau_0'\right) \to \beta'$. Let $\pi(z, \lambda, \tau_0|y)$ be the marginal density of the augmented data (z, λ, τ_0) from Eq. (10), then the middle step of the SW algorithm, i.e., $(z, \lambda, \tau_0) \to \left(z', \lambda', \tau_0'\right)$ can be obtained by making a draw in accordance with a Markov transition function $R((z, \lambda, \tau_0), \cdot)$, which is defined on the basis of the marginal density $\pi(z, \lambda, \tau_0|y)$ and is invertible with respect to $\pi(z, \lambda, \tau_0|y)$ (Hobert and Marchev, 2008).

Following the group action method of Hobert and Marchev (2008), Eq. (10) can be calculated as

$$\begin{aligned}\pi(z, \lambda, \tau_0|y) \propto & \frac{\exp\left\{-\frac{1}{2}\left[z^T\Lambda^{\frac{1}{2}}(I - W)\Lambda^{\frac{1}{2}}z\right]\right\}}{\left|X^T\Lambda X + \tau_0\Sigma_0\right|^{\frac{1}{2}}} |\Lambda|^{\frac{\nu-1}{2}} e^{-\frac{1}{2}\Sigma\lambda_i} \tau_0^{\frac{p+\nu_0}{2}-1} \\ & \times e^{-\frac{\tau_0\nu_0}{2}} \prod_{i=1}^{n} \left[I_{\mathbb{R}^-}(z_i) I_{\{0\}}(y_i) + I_{\mathbb{R}^+}(z_i) I_{\{1\}}(y_i)\right]\end{aligned} \tag{12}$$

where $W = \Lambda^{\frac{1}{2}} X \left(X^T\Lambda X + \tau_0\Sigma_0\right)^{-1} X^T\Lambda^{\frac{1}{2}}$. Suppose H is the multiplicative topological group $\mathbb{R}^+$ which is unimodular with Haar measure, and $\varrho(\mathrm{d}h) = \frac{\mathrm{d}h}{h}$, where $\mathrm{d}h$ denotes Lebesgue measure on $\mathbb{R}^+$. Let $\mathscr{Z} = \{z_i\} \subset \mathbb{R}^n$ be an n-fold Cartesian product of positive and negative halves, and the ith element is positive if $y_i = 1$ and negative otherwise. We then consider a transformation $\left(z', \lambda', \tau_0'\right) = L_h(z, \lambda, \tau_0) = (hz, h\lambda, h\tau_0)$, suppose $h \in H$ and an integrable function $g : \mathscr{Z} \times \mathbb{R}^{n,+} \times \mathbb{R}^+ \to \mathbb{R}$ then

$$\int_{\mathscr{Z}} \int_{\mathbb{R}^{n,+}} \int_{\mathbb{R}^+} g(z, \lambda, \tau_0) \mathrm{d}z\mathrm{d}\lambda\mathrm{d}\tau_0 = \psi(h) \int_{\mathscr{Z}} \int_{\mathbb{R}^{n,+}} \int_{\mathbb{R}^+} g(L_h(z, \lambda, \tau_0)) \mathrm{d}z\mathrm{d}\lambda\mathrm{d}\tau_0 \tag{13}$$

$$\begin{aligned}\pi(\beta, (z, \lambda, \tau_0)|y) = & \frac{1}{m_\nu} \left[\prod_{i=1}^{n} \left\{I_{\mathbb{R}^-}(z_i) I_{\{0\}}(y_i) + I_{\mathbb{R}^+}(z_i) I_{\{1\}}(y_i)\right\}\right] \phi\left(z_i; x_i^T\beta, \frac{1}{\lambda_i}\right) q\left(\lambda_i; \frac{\nu}{2}, \frac{\nu}{2}\right) \\ & \times \phi_p\left(\beta; 0, (\tau_0\Sigma_0)^{-1}\right) q\left(\tau_0; \frac{\nu_0}{2}, \frac{\nu_0}{2}\right); \lambda_i, \tau_0 \in \mathbb{R}^+, z_i \in R, \beta \in \mathbb{R}^p\end{aligned} \tag{10}$$

where $I(\cdot)$ is the indicator function, $\phi(\cdot)$ is the density of univariate normal distribution, $\mathbb{R}^- = (-\infty, 0)$, $\mathbb{R}^+ = (0, \infty)$, and $q(x; a, b)$ is the gamma density at x whose shape parameter is a and scale parameter is b, i.e., $q(x; a, b) = b^a x^{a-1} e^{-bw}/\Gamma(a))$. Correspondingly, $\pi_\nu(\beta|y)$ can be given as

where $\psi(h) = h^{2n+1}$, and $\psi\left(h^{-1}\right) = \frac{1}{\psi(h)}$ and $\psi(h_1 h_2) = \psi(h_1)\psi(h_2)$ for all $h, h_1, h_2 \in H$. Following Hobert and Marchev (2008), suppose that the density function $\varphi(h)$ is defined on H and depends on the augmented data (z, λ, τ_0), where

$$\begin{aligned}&\varphi(h)dh\propto\pi(hz, h\lambda, h\tau_0|y)\psi(h)\varrho(\mathrm{d}h)\\ &\propto h^{\frac{3n+n\nu+\nu_0}{2}-1}\exp\left\{-\frac{h}{2}\left[\left(\nu\Sigma_{i=1}^{n}\lambda_i+\tau_0\nu_0\right)+h^2z^T\Lambda^{\frac{1}{2}}(I-W)\Lambda^{\frac{1}{2}}z\right]\right\}dh\end{aligned} \tag{14}$$

Since Σ_0 is positive semidefinite, thus $\left(X^T\Lambda X\right)^{-1}-\left(X^T\Lambda X+\tau_0\Sigma_0\right)^{-1}$ is positive semidefinite (Hobert and Marchev, 2008), and $I-\Lambda^{\frac{1}{2}}X\left(X^T\Lambda X\right)^{-1}X^T\Lambda^{\frac{1}{2}}$ is an idempotent matrix. So

$$\begin{aligned}z^T\Lambda^{\frac{1}{2}}(I-W)\Lambda^{\frac{1}{2}}z\& &= z^T\Lambda z-z^T\Lambda X\left(X^T\Lambda X+\tau_0\Sigma_0\right)^{-1}X^T\Lambda z\\ &\geqslant z^T\Lambda z-z^T\Lambda X\left(X^T\Lambda X\right)^{-1}X^T\Lambda z\\ &= z^T\Lambda^{\frac{1}{2}}\left(I-\Lambda^{\frac{1}{2}}X\left(X^T\Lambda X\right)^{-1}X^T\Lambda^{\frac{1}{2}}\right)\Lambda^{\frac{1}{2}}z\\ &\geqslant 0\end{aligned} \tag{15}$$

Since $z^T\Lambda^{\frac{1}{2}}(I-W)\Lambda^{\frac{1}{2}}z \geq 0$, $\varphi(h)$ is a valid density. With above settings, the SW algorithm is given as follows:

Sandwich Algorithm:

Step 1	First draw $\{(\lambda_i, z_i), i = 1, 2, \cdots, n\}$ by drawing the truncated t distribution z_i $Tt_\nu\left(x_i^T\beta, y_i\right)$, then draw λ_i Gamma$\left(\frac{\nu+1}{2}, \frac{\nu+\left(z_i-x_i^T\beta\right)^2}{2}\right)$ and separately draw τ_0 Gamma$\left(\frac{\nu_0+p}{2}, \frac{\nu_0+\beta^T\Sigma_0\beta}{2}\right)$
Step 2	Draw h $\varphi(h)$ where $\varphi(h)$ is given in Eq. (14)
Step 3	Then draw $\beta\,\bar{\mathrm{N}}_p\left(h\widehat{\beta}, \frac{1}{h}\left(X^T\Lambda X+\tau_0\Sigma_0\right)^{-1}\right)$

In addition, the value of degrees of freedom can be determined by the data, rather than just using a pre-specified fixed value as in most previous studies. We believe that this processing leads to better model fitting results. In this Bayesian framework, the value of ν is determined by providing maximum the marginal likelihood of the data $m_\nu(y)$ in Eq. (6). Suppose a family of robit models with different degrees of freedom $\nu \in \mathcal{N}$, and $\nu_* \in \mathcal{N}$ is an appropriately chosen fixed value. Let $\eta_{\nu,\nu_*} = \frac{m_\nu(y)}{m_{\nu_*}(y)}$, for picking up models that are superior to the others for all $\nu \in \mathcal{N}$, we can calculate and thereafter compare the values of η_{ν,ν_*}, and ν that could provide the largest value of η_{ν,ν_*} will be the chosen one. η_{ν,ν_*} can be estimated as shown below.

Let $\left\{\beta^{(i)}\right\}_{i=1}^{N}$ be the Markov chain with distribution $\pi_{\nu_*}(\beta|y)$ which is produced by the SW algorithm. According to the ergodic theorem, the starting value $\beta^{(1)}$ has no influence on the value of η_{ν,ν_*} when N is large enough, since

$$\begin{aligned}&\frac{1}{N}\Sigma_{i=1}^{N}\frac{l_\nu\left(\beta^{(i)}|y\right)}{l_{\nu^*}\left(\beta^{(i)}|y\right)}\approx\int\frac{l_\nu(\beta|y)}{l_{\nu^*}(\beta|y)}\pi_{\nu_*}(\beta|y)d\beta\\ &=\frac{m_\nu(y)}{m_{\nu_*}(y)}\int\frac{\frac{l_\nu(\beta|y)\pi(\beta)}{m_\nu(y)}}{\frac{l_{\nu^*}(\beta|y)\pi(\beta)}{m_{\nu_*}(y)}}\pi_{\nu_*}(\beta|y)d\beta\\ &=\frac{m_\nu(y)}{m_{\nu_*}(y)}\int\frac{\pi_\nu(\beta|y)}{\pi_{\nu_*}(\beta|y)}\pi_{\nu_*}(\beta|y)d\beta\\ &=\frac{m_\nu(y)}{m_{\nu_*}(y)}\#\end{aligned} \tag{16}$$

Therefore, for $\nu \in \mathcal{N}$, η_{ν,ν_*} can be continuously estimated by sampling the posterior $\pi_{\nu_*}(\beta|y)$ Eq. (16). In order to avoid the potential unstable estimation brought by the possible significant difference between $l_\nu(\beta|y)$ and $l_{\nu^*}(\beta|y)$, an alternative way is to choose a finite number of points $\nu_1, \nu_2, \cdots, \nu_q \in \mathcal{N}$ to replace $l_{\nu^*}(\beta|y)$ by a linear combination $\Sigma_{i=1}^{q}\alpha_i l_{\nu_i}(\beta|y)\frac{m_{\nu_*}(y)}{m_{\nu_i}(y)}$ for $i = 2, 3, \cdots, q$, $\Sigma_{i=1}^{q}\alpha_i = 1$, and $m_{\nu 1} = m_{\nu^*}$.

Suppose that we have a sample $\left\{\beta^{(l)}\right\}_{l=1}^{N}$ from $\pi_{\nu_d}(\beta|y) = \Sigma_{i=1}^{q}\alpha_i\pi_{\nu_i}(\beta|y)$, then

$$\begin{aligned}\frac{1}{N}\Sigma_{i=1}^{N}\frac{l_\nu\left(\beta^{(l)}|y\right)}{\Sigma_{i=1}^{q}\alpha_i l_{\nu_i}\left(\beta^{(l)}|y\right)\frac{m_{\nu_*}(y)}{m_{\nu_i}(y)}}&\approx\int\frac{l_\nu(\beta|y)}{\Sigma_{i=1}^{q}\alpha_i l_{\nu_i}(\beta|y)\frac{m_{\nu_*}(y)}{m_{\nu_i}(y)}}\pi_{\nu_d}(\beta|y)d\beta\\ &=\eta_{\nu,\nu_*}\#\end{aligned} \tag{17}$$

Similarly, let $\left\{\beta_j^{(l)}\right\}_{l=1}^{N_j}$ be the sample from posterior densities $\pi_{\nu_j}(\beta|y)$ for $j = 1, 2, \cdots, q$. Assume $\alpha_i = N_i/N$ where $= \Sigma_{i=1}^{q}$, based on the ergodic theorem (Doss, 2010), the estimator of η_{ν,ν_*} can be given as

$$\begin{aligned}\widehat{\eta}_{\nu,\nu_*}&=\frac{1}{N}\Sigma_{j=1}^{q}\Sigma_{l=1}^{N_j}\frac{l_\nu\left(\beta_j^{(l)}|y\right)}{\Sigma_{i=1}^{q}\alpha_i l_{\nu_i}\left(\beta_j^{(l)}|y\right)\frac{m_{\nu_*}(y)}{m_{\nu_i}(y)}}\\ &=\frac{1}{N}\frac{1}{m_{\nu_*}}\Sigma_{j=1}^{q}\Sigma_{l=1}^{N_j}\frac{l_\nu\left(\beta_j^{(l)}|y\right)\pi(\beta)}{\Sigma_{i=1}^{q}\alpha_i l_{\nu_i}\left(\beta_j^{(l)}|y\right)\Big/m_{\nu_i}(y)}\\ &\approx\frac{m_\nu(y)}{m_{\nu_*}(y)}\Sigma_{j=1}^{q}\int\frac{\alpha_i l_\nu(\beta|y)}{\Sigma_{i=1}^{q}\alpha_i l_{\nu_i}(\beta|y)}\pi_{\nu_d}(\beta|y)\mathrm{d}\beta\\ &=\frac{m_\nu(y)}{m_{\nu_*}(y)}\\ &=\eta_{\nu,\nu_*}\#\end{aligned} \tag{18}$$

Therefore, the estimator $\widehat{\eta}_{\nu,\nu_*}$ is consistent with η_{ν,ν_*}, and can be used to find the optimal degrees of freedom ν based on the data.

The selection of an appropriate statistical or econometric model is a crucial step in data analysis. To assist in this process, a trio of commonly used model selection criteria have been developed: the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), and the Hannan-Quinn Information Criterion (HQIC). These criteria are designed to strike a balance between the fit of the model to the data and the parsimony of its structure. Mathematically, these criteria are expressed as follows:

$$\mathrm{AIC} = -2\times\mathrm{LL}+2k \tag{19a}$$

$$\mathrm{BIC} = -2\times\mathrm{LL}+2k\ln(n) \tag{19b}$$

$$\mathrm{HQIC} = -2\times\mathrm{LL}+2k\ln(\ln(n)) \tag{19c}$$

where LL represents the log-likelihood of the final model, k is the number of significant parameters, and n is the number of observations. The best model is selected by comparing the AIC, BIC, and HQIC values obtained from fitting multiple models to the data. The model that exhibits the lowest value for any of these criteria is considered to be the optimal choice, providing a suitable balance between fitting the data and simplicity of structure.

3. Results and discussion

In accordance with the methodology espoused by Doss (2010), 10 candidates for the degrees of freedom ν were selected from the set $\{0.05, 0.25, 0.5, 0.8, 1.25, 2.5, 4, 5, 7, 12\}$ through empirical means. The first stage of the procedure involved the separate execution of 10 chains, each utilizing one of the aforementioned ν candidates, over 10,000 iterations. These programs were run using R and the RStan library (Stan Development Team, 2020) on a desktop system featuring a 12th Gen Intel(R) Core (TM) i7-12700 2.10 GHz processor and an impressive 32.0 GB of RAM. A reference value of $\nu_* = 1.25$ was established through

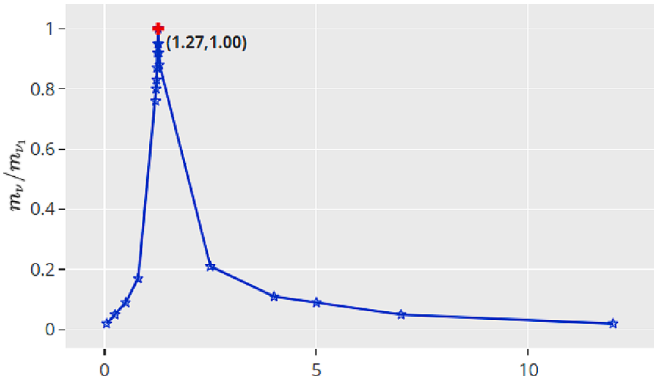

Fig. 1. Comparison of model performance with different degrees of freedom.

preliminary MCMC simulations using the SW algorithm, as this value was found to yield the maximum marginal likelihood $m_\nu(y)$. This first step took around 4 min with the proposed SW algorithm and about 27 min with the classic DA algorithm. In the subsequent stage, 10 new chains with the degrees of freedom set $\{1.21, 1.22, 1.23, 1.24, 1.25, 1.26, 1.27, 1.28, 1.29, 1.30\}$, corresponding to the ν values in step 1, are separately run for another 10,000 iterations. In order to acquire variability in the estimates, the entire procedure was repeated for 20 times with different start points, and the maximum standard error of these groups of $\eta_{\nu,\nu*}$ is less than 0.01. The results indicate that our estimates are stable. As depicted in Fig. 1, the final ν value of 1.27 was determined to be the optimal choice, as it resulted in the largest $\eta_{\nu,\nu*}$. This outcome led to the selection of a robit model with 1.27 degrees of freedom for use in the present study, effectively rendering the probit and logit models as unsuitable for the data at hand.

Table 2 presents a comprehensive evaluation of the proposed robit model in relation to the conventional probit and Logit models, through the use of multiple statistical indices that gauge both the model's complexity and its conformity to the underlying data. Of particular significance is the log-likelihood of the final model, which represents a critical metric of model fit. The results indicate a remarkable superiority of the robit model, with a log-likelihood value of −166.01, significantly lower than that of the probit (−193.75) and logit (−182.02) models. The robit model's superiority is further substantiated by its largest number of parameters, as reflected in its lowest Akaike Information Criterion (AIC) value of 352.02, compared to the probit (403.50) and Logit (380.04) models. The Bayesian Information Criterion (BIC) and Hannan-Quinn Information Criterion (HQIC) provide additional support for the robit model's superiority, as evidenced by its lowest BIC and HQIC values, in comparison to the probit and logit models.

Furthermore, the results presented in Fig. 2 reveal that the convergence of the robit model is significantly faster when utilizing the proposed SW algorithm, in comparison to the data augmentation algorithm. This highlights the superior efficiency of the proposed model. Given the cumulative weight of evidence provided by the various comparison results, it can be conclusively stated that the proposed robit model with $\nu = 1.27$ is the most appropriate model among all candidates for the studied dataset. As such, the discussions and conclusions presented in the subsequent sections of this paper are based solely on this model.

Table 2
Model comparison results.

Model	Robit	Probit	Logit
Log-likelihood of intercept-only model	−337.04		
Log-likelihood of final model	−166.01	−193.75	−182.02
Number of parameters	10	8	8
AIC	352.02	403.50	380.04
Observation	485	485	485
BIC	393.86	436.97	413.51
HQIC	368.46	416.65	393.19

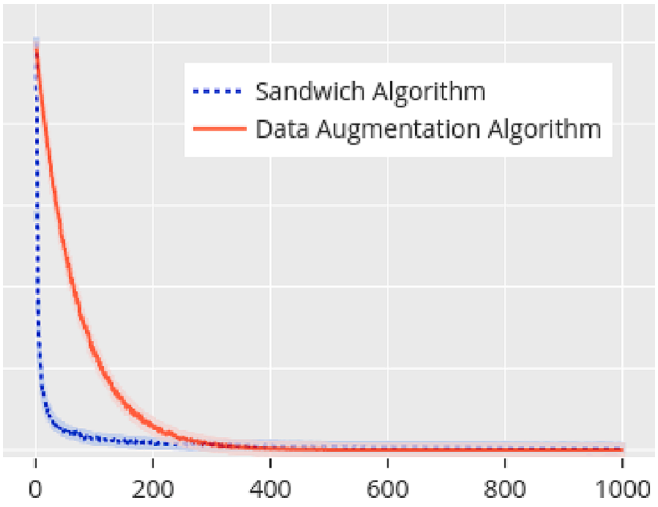

Fig. 2. Comparison of convergence speed of different algorithms (Shading indicates the error range).

Table 3
Estimation results of robit, probit and logit models.

Variable	Robit Coef.* (S.D**)	Probit Coef. (S.D)	Logit Coef. (S.D)
Constant	−4.06 (1.17)	−3.26 (1.75)	−4.23 (1.54)
Environmental characteristics			
Summer (relative to Autumn)	–***	0.24 (0.07)	–
Weekend (relative to weekday)	0.63 (0.12)	–	0.54 (0.22)
Night (relative to daytime)	−0.45 (0.22)	−0.53 (0.18)	−0.41 (0.19)
Rain (relative to clear)	−0.79 (0.34)	–	−0.64 (0.40)
Cloudy (relative to clear)	–	–	0.17 (0.06)
Roadway geometric characteristics			
Downgrade (relative to upgrade)	–	0.15 (0.02)	–
Curve (relative to straight)	0.27 (0.09)	–	–
Vehicle and crash characteristics			
Bus and van (relative to car)	–	−0.18 (0.02)	–
Truck (relative to car)	0.96 (0.16)	–	0.74 (0.22)
Rear-end (relative to rollover)	1.09 (0.34)	–	0.97 (0.47)
Hitting fixed object (relative to rollover)	1.22 (0.24)	0.72 (0.16)	–
Driver characteristics			
Speeding (relative to no violation)	2.67 (0.33)	–	1.78 (0.52)
Fatigue driving (relative to no violation)	1.62 (0.28)	1.76 (0.29)	–
>50 (relative to 30–39)	–	0.61 (0.17)	–

Notes:
Coef.: Coefficient.
S.D.: Standard deviation.
–: not significant for that particular model.

Table 3 presents the coefficient estimates and standard deviations (in parentheses) for the robit, probit, and logit models. The data in the table highlights the impact of various environmental, vehicle, crash and driver-related characteristics on the dependent variable. Of these variables, nine are found to significantly influence injury severity in the robit model (at a 0.05 level of significance). The marginal effects of these

Table 4
Marginal effects of significant variables in robit model.

Variable	Marginal Effect
Weekend (relative to weekday)	0.016
Night (relative to daytime)	−0.013
Rain (relative to clear)	−0.019
Curve (relative to straight)	0.005
Truck (relative to car)	0.024
Rear-end (relative to rollover)	0.032
Hitting fixed object (relative to rollover)	0.037
Speeding (relative to no violation)	0.072
Fatigue driving (relative to no violation)	0.045

nine variables are presented in the Table 4. The impact of each of these variables is further explored in the discussion that follows.

The robit model results, as shown in Tables 3 and 4, reveal that weekend tunnel crashes (compared to weekday crashes) have a notably adverse impact on injury severity, with a marginal effect of 0.016. This suggests that the probability of injury severity being elevated during tunnel crashes on weekends is higher compared to those on weekdays. One potential explanation for this could be the decreased traffic volume on the Jiangmen-Zhaoqing expressway over the weekend, which is approximately 26% less, resulting in a higher probability of high-speed driving and even excessive speeding (with an average speed increase of 12%). The constricted driving space and inadequate lighting conditions within tunnels may also contribute to exacerbating the seriousness of injuries during high-speed accidents (Huang et al., 2018).

Table 4 demonstrates that compared to daytime driving, driving at night presents a lower probability of severe injuries, as evidenced by the marginal effect of −0.013. This reduction in injury severity is likely attributed to the stark contrast in luminance levels between the tunnels and the surrounding highways during daylight hours (Pervez et al., 2022a). The phenomenon of 'black-hole' and 'white-hole' conditions in tunnels has been shown to have a significant impact on driver performance and can lead to severe crashes (Li et al., 2019a). Additionally, it is worth mentioning that these types of accidents tend to occur with higher frequency in the entrance and exit zones of tunnels.

The estimation results reveal that when it comes to the impact of weather conditions on the likelihood of severe injuries incurred during driving, adverse weather conditions (rain) appear to present a significantly lower risk when compared to clear weather conditions. This observation can be attributed to the fact that drivers tend to exhibit greater caution and vigilance when navigating roads during inclement weather, as opposed to during sunny weather. These findings are consistent with previous studies conducted on this topic, further strengthening the argument that driving in tunnels at adverse weather conditions presents a lower risk of severe injury (Feng et al., 2016; Li et al., 2019b).

The results in Tables 3 and 4 indicate a significant relationship between the presence of curved tunnels and an elevated probability of severe crashes. This finding is consistent with previous studies that have explored the impact of tunnel curvature on road safety (Pervez et al., 2022b). The limited vision distance offered by curved tunnels presents a challenge for drivers, as it makes it more difficult to maintain proper lane position while navigating these roadways. This, in turn, increases the risk of severe crashes.

Estimation results show that the presence of trucks is a significant contributor to the heightened likelihood of severe injuries. This result is unsurprising, given the treacherous driving conditions that the mountainous terrain presents. The steep slopes in the area pose significant braking challenges for trucks, rendering it increasingly difficult for them to navigate the roads with safety and stability. Furthermore, the constricted road space within the tunnels only exacerbates these difficulties, elevating the probability of accidents (Yu et al., 2019).

The results of the study reveal a significant relationship between crash type and the severity of injury outcomes. It has been observed that rear-end collisions exhibit a higher propensity for resulting in more serious injuries. This may be due to the driver's need to simultaneously manage the lateral position within the confines of the tunnel, while also ensuring the longitudinal headway is within a safe range. The resulting distractions may make the driver overlook critical safety considerations, leading to an increased likelihood of severe injuries. These findings are consistent with prior research that has investigated the relationship between crash type and injury severity (Pervez et al., 2022b). Moreover, the analysis revealed that collisions with fixed objects also exhibit a higher probability of resulting in severe injury outcomes. This is likely a result of the high collision energy generated by such incidents, which can result in catastrophic consequences.

It is not surprising that speeding can significantly increase the likelihood of severe injuries. This may be due to the fact that speeding reduces the time a driver has to react to avoid a collision in a dangerous situation, increases the stopping distance of the vehicle, and reduces the ability of road safety structures to protect the occupants of the vehicle in a collision (Matsuo et al., 2020). Fatigue driving also contributes to increased injury severity. This may be due to the slower reaction time of a fatigued driver with reduced attention, awareness, and control of the vehicle. This finding is also in line with previous studies (Se et al., 2021).

4. Conclusions and limitations

This study proposes an efficient and robust Bayesian robit model which adopts the Student's t distribution as the link function to deal with the outliers problem in the traffic safety dataset. A sandwich algorithm which is built on the data augmentation algorithm is proposed to increase the estimation speed of posteriors. The robit model is applied to a tunnel crash dataset and compared with the baselines including probit and logit models. The model estimation results show that the robit model has superior goodness of fit than the other models, indicating that the robit model is potentially more suitable for traffic accident analysis, especially in the face of outliers and the presence of classification imbalances in the small data sets.

The study's findings shed light on the need for certain safety measures to be implemented in order to improve the traffic safety of tunnels. To begin with, it is imperative that the lighting levels be thoroughly evaluated and adjusted as required when entering the tunnel, as well as in the transition zone. This will help to mitigate the likelihood of abrupt lighting changes, which could prove hazardous to drivers. Furthermore, the implementation of variable message signs (VMS) is highly recommended, especially in long tunnels. These signs can provide drivers with vital information regarding accidents and driving requirements, thereby enhancing road safety. Additionally, the installation of speed indicator devices in all road tunnels should be considered, as this would help drivers to maintain a safe speed, and thus reduce the risk of collisions. Moreover, the type and location of signage should be reassessed, as information overload can be a major challenge for drivers approaching tunnel entrances. The goal should be to simplify driving tasks and reduce the likelihood of accidents, thus ensuring that all drivers are able to traverse the tunnel safely. By implementing these countermeasures, we can make significant strides in improving the traffic safety of tunnels and protecting the well-being of all road users.

The present study is not devoid of limitations, albeit, these limitations do not undermine the overall import of the findings. These limitations could be considered in future endeavors within the domain of traffic safety analysis. The following limitations are noteworthy:

Unobserved heterogeneity: Despite the implementation of a robust Bayesian regression approach in the form of the robit model, the findings of this study may still be influenced by the presence of unobserved heterogeneity in the crash data. Unobserved heterogeneity pertains to individual factors that can impact the likelihood of a crash, yet are not easily measurable or observable. These factors may encompass driver-specific characteristics such as skill level,

experience, and fatigue, as well as vehicle-specific factors such as age, make, and model, and maintenance history (Anastasopoulos and Mannering, 2011). If these factors are not considered in the analysis, the results may be biased, and the relationships between other variables, such as road design and crash frequency, may not accurately reflect the true relationships. To address unobserved heterogeneity in traffic crash research, researchers can use techniques such as mixed effects models or propensity score matching to control for these factors and improve the accuracy of their results (Alnawmasi and Mannering, 2022; Lord and Mannering, 2010). While the robit model effectively reduces the impact of outliers, it does not address the issue of unobserved heterogeneity. This is a limitation that future research endeavors in the domain of traffic safety analysis should aim to address.

Limitations in large datasets: The robit model was evaluated on a relatively small dataset of tunnel crashes, and while it demonstrated improved robustness and efficiency compared to traditional methods, it remains uncertain as to whether the improvement achieved by the robit model would be equally significant in the context of larger datasets. Further research is necessary to ascertain the scalability and performance of the robit model in such scenarios.

Generalizability: The results obtained from this study may not be applicable to other types of crashes or different geographic regions. The study focuses on tunnel crashes, and the findings may not be generalizable to other types of roadways or traffic conditions. Further research is necessary to determine the generalizability of the proposed robit model to other types of crashes.

To address these limitations and continue advancing our understanding, possible research directions include evaluating the generalizability of the robit model to other types of crashes and geographic regions, determining its efficiency and effectiveness in large datasets, incorporating additional factors into the analysis, evaluating the robustness and stability of the model, and exploring the integration of advanced technologies into traffic safety analysis.

CRediT authorship contribution statement

Zhenning Li: Data curation, Formal analysis, Methodology, Funding acquisition, Writing – original draft, Writing – review & editing. **Haicheng Liao:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing. **Ruru Tang:** Data curation, Formal analysis, Methodology, Writing – review & editing. **Guofa Li:** Conceptualization, Formal analysis, Methodology, Writing – review & editing. **Yunjian Li:** Data curation, Formal analysis, Methodology, Writing – original draft, Writing – review & editing. **Chengzhong Xu:** Conceptualization, Funding acquisition, Project administration, Supervision, Writing – review & editing.

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.

Data availability

Data will be made available on request.

Acknowledgements

This paper is supported the Science and Technology Development Fund of Macau SAR (File no. 0021/2022/ITP, 0081/2022/A2, SKL-IOTSC(UM)-2021-2023, 0123/2022/AFJ, and 0015/2019/AKP), GuangDong Basic and Applied Basic Research Foundation (No. 2020B515130004), and Key-Area Research and Development Program of Guangdong Province (No. 2020B010164003).

References

Alnawmasi, N., Mannering, F. 2022. A temporal assessment of distracted driving injury severities using alternate unobserved-heterogeneity modeling approaches. *Anal. Methods Accid. Res.*, *34*, 100216.

Anastasopoulos, P.C., Mannering, F.L., 2011. An empirical assessment of fixed and random parameter logit models using crash-and non-crash-specific injury data. Accid. Anal. Prev. 43 (3), 1140–1147.

Barron, J.T., 2019. A general and adaptive robust loss function. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4331–4339.

Brooks, S., Gelman, A., Jones, G., Meng, X.-L., 2011. Handbook of Markov Chain Monte Carlo. CRC Press.

Doss, H., 2010. Estimation of large families of Bayes factors from Markov chain output. Stat. Sin. 537–560.

El-Basyouny, K., Sayed, T., 2010. A method to account for outliers in the development of safety performance functions. Accid. Anal. Prev. 42 (4), 1266–1272.

Feng, S., Li, Z., Ci, Y., Zhang, G., 2016. Risk factors affecting fatal bus accident severity: Their impact on different types of bus drivers. Accid. Anal. Prev. 86, 29–39. https://doi.org/10.1016/j.aap.2015.09.025.

Gelman, A., Hill, J., 2006. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press.

Hobert, J.P., Marchev, D., 2008. A theoretical comparison of the data augmentation, marginal augmentation and PX-DA algorithms. Ann. Stat. 36 (2), 532–554.

Huang, H., Peng, Y., Wang, J., Luo, Q., Li, X., 2018. Interactive risk analysis on crash injury severity at a mountainous freeway with tunnel groups in China. Accid. Anal. Prev. 111, 56–62.

Kang, J.D.Y., Schafer, J.L. 2007. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data.

Kim, S., Chen, M.-H., Dey, D.K., 2008. Flexible generalized t-link models for binary response data. Biometrika 95 (1), 93–106.

Krueger, R., Bierlaire, M., Gasos, T., Bansal, P., 2023. Robust discrete choice models with t-distributed kernel errors. Stat. Comput. 33 (1), 2.

Li, Z., Ci, Y., Chen, C., Zhang, G., Wu, Q., Qian, Z.S., Prevedouros, P.D., Ma, D.T., 2019a. Investigation of driver injury severities in rural single-vehicle crashes under rain conditions using mixed logit and latent class models. Accid. Anal. Prev. 124, 219–229.

Li, Z., Chen, C., Wu, Q., Zhang, G., Liu, C., Prevedouros, P.D., Ma, D.T. 2018. Exploring driver injury severity patterns and causes in low visibility related single-vehicle crashes using a finite mixture random parameters model. *Anal. Methods Accid. Res.*, *20*, 1–14. https://doi.org/10.1016/j.amar.2018.08.001.

Li, Z., Wu, Q., Ci, Y., Chen, C., Chen, X., Zhang, G., 2019b. Using latent class analysis and mixed logit model to explore risk factors on driver injury severity in single-vehicle crashes. Accid. Anal. Prev. 129, 230–240.

Liang, F., Liu, C., Carroll, R., 2011. Advanced Markov Chain Monte Carlo Methods: Learning From Past Samples. John Wiley & Sons.

Little, R.J.A., Rubin, D.B., 2019. Statistical Analysis with Missing Data, Vol. 793. John Wiley & Sons.

Liu, C., 2004. Robit regression: a simple robust alternative to logistic and probit regression. In: Applied Bayesian Modeling and Casual Inference from Incomplete-Data Perspectives, pp. 227–238.

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

Markatou, M., He, X., 1994. Bounded influence and high breakdown point testing procedures in linear models. J. Am. Stat. Assoc. 89 (426), 543–549.

Matsuo, K., Sugihara, M., Yamazaki, M., Mimura, Y., Yang, J., Kanno, K., Sugiki, N. 2020. Hierarchical Bayesian modeling to evaluate the impacts of intelligent speed adaptation considering individuals' usual speeding tendencies: A correlated random parameters approach. *Anal. Methods Accid. Res.*, *27*, 100125.

Newson, R., Falcaro, M., 2022. Robit regression in Stata. Stata J.

Pervez, A., Lee, J., Huang, H. 2022. Exploring factors affecting the injury severity of freeway tunnel crashes: a random parameters approach with heterogeneity in means and variances. *Accid. Anal. Prev.*, *178*, 106835.

Pervez, A., Huang, H., Lee, J., Han, C., Li, Y., Zhai, X., 2022a. Factors affecting injury severity of crashes in freeway tunnel groups: A random parameter approach. J. Transp. Eng., Part A: Syst. 148 (4), 04022006.

Rahimian, H., Mehrotra, S., 2019. Distributionally robust optimization: A review. ArXiv Preprint. ArXiv:1908.05659.
Roy, V., 2012. Convergence rates for MCMC algorithms for a robust Bayesian binary regression model. Electron. J. Stat. 6, 2463–2485.
Se, C., Champahom, T., Jomnonkwao, S., Karoonsoontawong, A., Ratanavaraha, V. 2021. Temporal stability of factors influencing driver-injury severities in single-vehicle crashes: A correlated random parameters with heterogeneity in means and variances approach. *Anal. Methods Accid. Res.*, *32*, 100179.
Shafieezadeh Abadeh, S., Mohajerin Esfahani, P.M., Kuhn, D., 2015. Distributionally robust logistic regression. Adv. Neural Inf. Process. Syst. 28.
Stan Development Team. 2020. *RStan: the R interface to Stan.* http://mc-stan.org/.
van Dyk, D.A., Meng, X.-L., 2001. The art of data augmentation. J. Comput. Graph. Stat. 10 (1), 1–50.
Yu, H., Li, Z., Zhang, G., Liu, P. 2019. A latent class approach for driver injury severity analysis in highway single vehicle crash considering unobserved heterogeneity and temporal influence. *Anal. Methods Accid. Res.*, *24*, 100110.
Zellner, A., 1986. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. Bayesian Inference Decis. Techniq.