Scientific Research Publishing

Journal of Transportation Technologies, 2017, 7, 190-205
http://www.scirp.org/journal/jtts
ISSN Online: 2160-0481
ISSN Print: 2160-0473

Road Crash Prediction Models: Different Statistical Modeling Approaches

Azad Abdulhafedh

University of Missouri-Columbia, MO, USA
Email: asa8cd@mail.missouri.edu

How to cite this paper: Abdulhafedh, A. (2017) Road Crash Prediction Models: Different Statistical Modeling Approaches. *Journal of Transportation Technologies*, 7, 190-205.
https://doi.org/10.4236/jtts.2017.72014

Received: December 20, 2016
Accepted: April 27, 2017
Published: April 30, 2017

Abstract

Road crash prediction models are very useful tools in highway safety, given their potential for determining both the crash frequency occurrence and the degree severity of crashes. Crash frequency refers to the prediction of the number of crashes that would occur on a specific road segment or intersection in a time period, while crash severity models generally explore the relationship between crash severity injury and the contributing factors such as driver behavior, vehicle characteristics, roadway geometry, and road-environment conditions. Effective interventions to reduce crash toll include design of safer infrastructure and incorporation of road safety features into land-use and transportation planning; improvement of vehicle safety features; improvement of post-crash care for victims of road crashes; and improvement of driver behavior, such as setting and enforcing laws relating to key risk factors, and raising public awareness. Despite the great efforts that transportation agencies put into preventive measures, the annual number of traffic crashes has not yet significantly decreased. For instance, 35,092 traffic fatalities were recorded in the US in 2015, an increase of 7.2% as compared to the previous year. With such a trend, this paper presents an overview of road crash prediction models used by transportation agencies and researchers to gain a better understanding of the techniques used in predicting road accidents and the risk factors that contribute to crash occurrence.

Keywords

Crash Prediction Models, Poisson, Negative Binomial, Zero-Inflated, Logit and Probit, Neural Networks

1. Introduction

Road traffic accidents are the world's leading cause of death for individuals be-

*PhD in Civil Engineering.

tween the ages of one and twenty-nine [1]. Throughout the world, cars, buses, trucks, motorcycles, pedestrians, animals, taxis and other categories of travelers, share the roadways, contributing to economic and social development in many countries. Yet each year, many vehicles are involved in crashes that are responsible for millions of deaths and injuries. Globally, every year, about 1.25 million people are killed in motor vehicle crashes and approximately 50 million more are injured. Following current trends, about two million people could be expected to be killed in motor vehicle crashes each year by 2030 [1]. Currently, road crashes are ranked as the ninth most serious cause of death in the world, and without new initiatives to improve road safety, fatal crashes will likely rise to the third place by the year 2020 [1]. In developed countries, road traffic death rates have decreased since the 1960s because of successful interventions such as seat belt safety laws, enforcement of speed limits, warnings about the dangers of mixing alcohol consumption with driving, and safer design and use of roads and vehicles. For example, road traffic fatalities have declined by about 25.0 percent in the United States from 2005 to 2014 and the number of people injured has decreased 13.0 percent from 2005 to 2014 [2]. In Canada, the number of road traffic fatalities has declined by about 62.0 percent from 1990 to 2014, and the number of injuries has declined by about 68.0 percent during the same period [3]. However, traffic fatalities have increased in developing countries from 1990 to 2014 (*i.e.* 44.0 percent in Malaysia and about 243.0 percent in China) [1]. Developing countries bear a large share of the burden, accounting for 85.0 percent of annual deaths and 90.0 percent of the disability-adjusted life years. More than one-half of all road traffic deaths globally involve people ages 15 to 44, during their most productive earning years. Moreover, the disability burden for this age group accounts for about 60.0 percent of all disability-adjusted life years. The costs and consequences of these losses are significant. Three-quarters of all poor families who lost a member in a traffic crash reported a decrease in their standard of living, and about 61.0 percent reported having to borrow money to cover expenses following their loss [4]. The World Bank estimates that road traffic injuries cost 2.0 percent to 3.0 percent of the Gross National Product of developing countries, or twice the total amount of development aid received worldwide by developing countries [5]. Although transportation agencies often try to identify the most hazardous road sites, and put great efforts into preventive measures, such as illumination and policy enforcement, the annual number of traffic crashes has not yet significantly decreased. For instance, 35,092 traffic fatalities were recorded in the US during 2015, an increase of 7.2% as compared to the previous year [6]. The fatality rate per 100 million vehicle miles traveled (VMT) increased 3.7% between 2014-2015. Thirty-five States had more motor vehicle fatalities in 2015 than in 2014. Every month except November saw increases in fatalities from 2014 to 2015, and the highest increases occurred in July and September [6]. Given this trend, it is imperative to gain a better understanding of the risk factors that may be associated with traffic crashes. This paper aims at presenting an overview of road crash prediction models used by transportation agencies and researchers to help understanding the techniques used in predicting road accidents and the risk factors that contribute to crash occurrence.

2. The Importance of Traffic Accidents Prediction Models

Traffic accidents prediction models are very useful tools in highway safety, given their potential for determining both the frequency of accident occurrence and the contributing factors that could then be addressed by transportation policies. Vehicular crash data can be used to model both the frequency of crash occurrence and the degree of crash severity. Crash frequency refers to the prediction of the number of crashes that would occur on a specific road segment or intersection in a time period [7]. Crash severity methods generally explore the relationship between crash severity injury categories and contributing factors such as driver behavior, vehicle characteristics, roadway geometry, and road-environment conditions. Traffic accident related-fatalities and injuries can be prevented or at least minimized by a joint involvement from multiple sectors (*i.e.* transportation agencies, police, health departments, education institutions) that oversee road safety, vehicles, and the drivers themselves. Effective interventions include design of safer infrastructure and incorporation of road safety features into land-use and transport planning; improvement of vehicle safety features; improvement of post-crash care for victims of road crashes, and improvement of driver behavior, such as setting and enforcing laws relating to key risk factors, and raising public awareness [8]. Transportation agencies and research institutions often seek to identify the most dangerous road sites, and this will require modeling road crash data to determine both crash frequency and crash severity degree. In addition, traffic accidents prediction models can also assist with the development of generalized theories concerning road safety. A range of basic laws have been put forth to help explain the relationship between the occurrence of road crashes and potential risk factors, such as: the universal law of learning, which implies that the crash rate tends to decline as the number of kilometers travelled increases; the law of rare events, which states that rare events, such as environmental hazards, would have more effect on crash rates than regular events; and the law of complexity, which implies that the more complex the traffic situation road users encounter, the higher the probability of crash occurrence [9].

3. Factors Affecting Road Traffic Accidents

A traffic accident may have many contributing factors, such as those related to driver behavior, road geometry, traffic volumes, vehicle, and environment. The influence of such variables on crash occurrence could significantly vary on a case-by-case basis, but in general, both behavioral factors related to the driver's errors, and non-behavioral factors related to road geometry, traffic flow conditions, vehicle, and environment are thought to significantly affect traffic crashes [10]. Research has revealed that there are generally six major groups of risk factors affecting traffic crash occurrence [11] [12] [13] [14] [15]:

1) Driver behavior: alcohol and drug use, reckless operation of vehicle, failure to properly use occupant protection devices, the use of cell phones or texting, and fatigue.

2) Vehicle factors: vehicle type, and the engineering and the safety design standards for vehicle performance. For example, the design of windshield glass and the location and durability of gas tanks can increase safety. Passenger protection systems in vehicles (*i.e.* air bags, safety belts), if used, can eliminate injuries or reduce their severity.

3) Roadway characteristics: road geometries and road side conditions, such as well-designed curves and grades, wide lanes, adequate sight distance, clearly visible striping, flared guardrails, good quality shoulders, roadsides free of obstacles, well-located crash attenuation devices, and well-planned use of traffic signals.

4) Traffic volumes: average annual daily traffic (AADT) or the vehicle miles travelled (VMT). AADT is the average number of vehicles passing a point along a particular road section each day. Thus, AADT represents the vehicle flow over a road section on an average day of the year. VMT refers to the distance travelled by vehicles on roads. It is often used as an indicator of traffic demand and is commonly applied to evaluate mobility patterns and travel trends.

5) Environmental factors: weather conditions, and light conditions.

6) Time factors: the season of the year, the month of the year, weekdays, and the hour of crash occurrence.

4. The Costs of Road Traffic Accidents

The highest cost of traffic crashes is in the loss of human lives; however, society also bears the consequences of many costs associated with motor vehicle crashes. Highway crashes currently cost the USA about $1078.0 billion a year, approximately 5.0 percent higher than 2000. Total costs include both economic costs and societal harm [16]. In the year 2010, 3.9 million people were injured and 32,999 killed in 13.6 million motor vehicle crashes in the US [2]. The economic costs of these crashes totaled $242.0 billion including lost productivity, medical costs, legal and court costs, emergency service costs, insurance administration costs, congestion costs, property damage, and workplace losses. The $242.0 billion cost of motor vehicle crashes represents the equivalent of nearly $784.0 for each person living in the United States, and 1.6 percent of the $14.96 trillion U.S. Gross Domestic Product for 2010 [16]. When quality of life valuation is considered, the total value of societal harm from motor vehicle crashes in 2010 was $836.0 billion, roughly three and a half times the value measured by economic impacts alone. Lost market and household productivity accounted for $77.0 billion of the total $242.0 billion economic costs, while property damage accounted for $76.0 billion. Medical expenses totaled $23.0 billion. Congestion caused by crashes, including travel delay, excess fuel consumption, greenhouse gases and criteria pollutants accounted for $28.0 billion. Each fatality resulted in an average discounted lifetime cost of $1.4 million. Each critically injured survivor cost an average of $1.0 million [16].

5. Literature Review

Early crash analysis models were generally based on simple multiple linear regression methods assuming normally distributed errors. However, researchers soon discovered that crash occurrence could be better fitted with a Poisson distribution. Hence, a Poisson regression model based upon a generalized linear framework was soon adopted over conventional multiple linear regression techniques. Several such Poisson regression approaches for exploring the relationship between the risk factors and crash frequency have been proposed [15] [16] [17] [18] [19] [20]. However, it has been found that Poisson regression approaches have one important constraint that the mean must be equal to the variance which if violated, the standard errors estimated by the maximum likelihood method, will be biased, and the test statistics derived from the model will be incorrect. Recent studies have shown that crash data are usually over-dispersed, when the variance exceeds the mean, therefore, incorrect estimation of the likelihood of crash occurrence could result in applications of the Poisson regression model [7]. In efforts to overcome the problem of over-dispersion, researchers began to employ the Negative Binomial (NB) distribution (also called the Poisson-Gamma) instead of the Poisson distribution, which relaxes the mean equals to variance constraint, and hence can accommodate over-dispersion in crash data counts [7]. NB models have been widely used in crash frequency modeling [14] [15] [19] [21] [22] [23]. However, NB models have some limitations such as the inability to handle under-dispersion of crash counts when the mean of the crash counts is higher than the variance. Although rare, this phenomenon can arise when the sample size is very small, leading to erroneous parameter estimates [24] [25]. To address the limitations of NB models, Poisson-lognormal models have been proposed, in which the error term is Poisson-lognormal rather than gamma-distributed to better handle the under-dispersed crash counts [21] [26] [27]. Another widely used type of crash prediction model is the zero-inflated Poisson and zero-inflated negative binomial models, which have been introduced mainly to deal with the over-dispersion problem caused by excessive zeroes (*i.e.* locations where no crashes can be observed) in traffic data counts. The zero-inflated models have shown great flexibility, although their applicability in crash prediction has been criticized because of the long term mean equals zero in the safe state that could produce some biased estimates [7] [22]. Generalized additive modeling approaches have also been proposed which provide smoothing functions for the explanatory variables. However, these models typically include more parameters than the traditional count models, and therefore their applicability to the crash prediction has been very limited [28] [29]. Random- parameters models have been applied to take the effect of the unobserved heterogeneity from one roadway site to another, however, their application in practice has been very limited [30] [31] [32]. The finding that road crashes are poorly explained by linear functions of independent variables, has encouraged the exploration of non-linear approximators such as fuzzy logic and neural networks. For example, a fuzzy logic approach was used for prediction of urban highway crash occurrence and it was found that the use of fuzzy sets in crash prediction is indeed a viable approach [33]. Neural networks have been applied to highway safety applications as predictive tools, such as in driver behavior analysis, pavement maintenance, vehicle detections, traffic signal control, and vehicle emissions, however, their application to crash analysis has been limited [28] [34] [35]. For instance, an artificial neural network was utilized to analyze the freeway crash frequency in Taiwan, and the results indicated that an artificial neural network can provide a consistent alternative method for analyzing crash frequency [36]. Also, a group of artificial neural networks was applied to model the non-linear relationships between the injury severity levels and crash-related factors. The findings indicated that artificial neural network models can predict crashes more effectively than the traditional statistical methods [37]. In crash severity models, a wide variety of statistical approaches such as the binary and the multinomial logit models, nested logit models, mixed logit models and ordered probit models have been investigated. For example, the ordered probit model was applied to predict crash severity on roadway sections, signalized intersections and toll plazas in Florida [38]. A mixed logit model was applied that used the injury outcome of the crash using limited crash data to investigate the proportion of crashes of each severity level on a specific roadway segment over a specified time period. Then, the number of crashes by severity level was determined without the need for detailed crash-specific data [39]. Also, a multinomial logistic regression was applied to model the severity injury of different vehicle collision patterns in urban highways in Arkansas, and the researchers recommended the use of the MNL over other models [40].

6. A Review on the Statistical Approaches of Road Crash Prediction Models

There are different statistical approaches for modeling traffic crashes. The following approaches present some of the mostly used methods.

6.1. Multiple Linear Regression

Early models of traffic accident models were based on the simple multiple linear regression approach assuming normally distributed errors. The general form of the linear crash prediction model can be expressed as follows:

$$Y \mid \theta \sim Dist\left(\theta\right) \text{with} \theta = f\left(X, \beta, \varepsilon\right) \quad (1)$$

where,

Y: the dependent variable (*i.e.* crash frequency),

θ: the crash dataset,

Dist(θ): the model distribution,

X: a vector representing different independent variables (*i.e.* risk factors),

β: a vector of regression coefficients,

f(.): link function that relates X and Y together, *ε*: the disturbance or error terms of the model.

6.2. Poisson Regression

Although multiple linear regression models have been widely applied, it has been found that crash occurrence can often be better fitted with a Poisson distribution. One frequent pitfall is to model crash data as continuous data by applying an ordinary least square regression [41]. This approach is inappropriate because regression models can produce predicted values that are non-integers and can also predict values that are negative, both of which are inconsistent with continuous data modeling. In addition, many distributions of crash data are positively skewed with many observations in the data set having a value of 0.0. The high number of zeros in the data set prevents the transformation of a skewed distribution into a normal one, which is a requirement of normal distribution. An alternative is to use a Poisson distribution or one of its variants. Poisson distributions have a number of advantages over an ordinary normal distribution, including a skew, discrete distribution, and the restriction of predicted values to non-negative numbers [41]. Hence, generalized linear modeling variates of the Poisson regression model have been proposed to explore the relationship between the risk factors and traffic accident modeling [15] [17] [18] [19]. Poisson regression has been applied to a wide range of transportation count data, including crash frequency. A Poisson regression model is similar to an ordinary linear regression, with two exceptions. First, it assumes that the errors follow a Poisson (not normal) distribution. Second, rather than modeling the response variable *Y* as a linear function of the regression coefficients, it models the natural log of the response variable, *In(Y)*, as a linear function of the coefficients [7]. The Poisson model can be expressed as follows:

$$P\left(n_{\mathrm{i}}\right)=\frac{\lambda iEXP\left(-\lambda i\right)}{n!} \tag{2}$$

where,

$P(n_{\mathrm{i}})$: the probability of n crashes occurring on a highway segment *i*,

n_i: the number of observations per time period (such as a year),

λ_{i}: the expected crash frequency on road segment *i* per time period (*i.e.* the mean of distribution) which can be estimated as follows:

$$\lambda_i = EXP\left(\beta X_i\right) \tag{3}$$

where

X_{i}: a vector of the independent variables (*i.e.* risk factors),

β: a vector of the estimates (coefficients) of the independent variables X_i.

This model is estimable by standard maximum likelihood methods, with the log likelihood (*LL*) function given as:

$$LL\left(\beta\right)=\sum_{1}^{n}\left[-EXP\left(\beta X_i\right)+n\left(\beta Xi\right)-Ln\left(n!\right)\right] \tag{4}$$

One assumption of Poisson Models is that the mean and the variance are equal, an assumption that is sometimes violated [7]. This can be dealt with by using a dispersion parameter if the difference is small, or by using a negative binomial regression model if the difference is large [42].

6.3. Negative Binomial Regression Model (NB)

In order to overcoming the problem of over-dispersion, the Negative Binomial (NB) distribution (also called the Poisson-Gamma) has been investigated as an alternative to the Poisson distribution given that it relaxes the condition of mean equals to variance, and hence can take into account over-dispersion in the crash data counts [7]. As a result, NB models have been widely applied in crash frequency modeling [14] [15] [19] [21] [22] [23].

The NB uses a Gamma probability distribution and can relax the assumption of the mean equals the variance and, hence, the NB can accommodate over-dispersion that may exist in the crash data counts [43]. A primary source of over-dispersion is the clustering of data, and the possible omission of relevant independent variables influencing the Poisson rate across observations [44]. In order to obtain the NB model, the Poisson regression can be rewritten by adding an error term to its expected number of crashes, and becomes [7]:

$$\lambda i = EXP\left(\beta Xi + \varepsilon_i\right) \tag{5}$$

where $EXP\,(\varepsilon_i)$ is a gamma-distributed error with mean equals one and variance equals α. The addition of this term allows the variance $VAR\,(n_i)$ to differ from the mean $E\,(n_i)$ as shown in Eq. 6:

$$VAR\left(n_i\right) = E\left(n_i\right)\left(1 + \alpha E\left(n_i\right)\right) \tag{6}$$

This error term is called the over-dispersion parameter, and both α and β can be estimated from the maximum likelihood function. When α is zero, the model becomes Poisson regression, and if α is found to be significantly different from zero, then the NB regression can be used instead of the Poisson regression model to handle the over-dispersion in crash data. However, the NB model also has some limitations such as its inability to handle the case of under-dis- persion of the data count, when the mean of the crash counts is higher than the variance [25] [44].

6.4. Poisson-Lognormal Regression Model

To address the limitations of the NB models, the Poisson-lognormal model was introduced, in which the error term is Poisson-lognormal rather than gamma-distributed so as to better handle under-dispersed data counts [21] [26] [27]. The Poisson-lognormal model is similar to the negative binomial model, however, the $EXP\,(\varepsilon_i)$ term used in the model is lognormal-rather than gamma-distributed. The Poisson-lognormal model provides more flexibility than the negative binomial model, but it does have some limitations, such as, its complex estimation of parameters due to the fact that the Poisson-lognormal distribution does not have a closed form [26].

6.5. Zero Inflated Poisson and Negative Binomial Regression Models

Another widely used crash frequency modeling approach is the zero-inflated

Poisson and zero-inflated negative binomial models, which have been introduced primarily to deal with the over-dispersion problem caused by excessive zeroes (*i.e.* locations where no crashes can be observed) in traffic data counts. The zero-altered procedure allows modeling the crash frequencies in two states, namely; the zero-crash state, and the non-zero crash state (where crash frequencies follow Poisson or negative binomial distribution), and the probability of a section being in zero or non-zero states can be found by a binary logit or probit model. In crash data, large numbers of zero observations are commonly present largely due to under reporting of minor crashes at these sites, the presence of dangerous crash sites (*i.e.* non-zero crash sites) in close proximity to the neighboring zero crash sites rendering the zero-crash sites to the safe mode, and given that some of zero crash sites may be free from only certain type of crashes, not all types of crashes [45]. Zero-inflated models attempt to account for such excess zeros. A dual state crash system may be assumed, in which one state is the zero crash state that can be regarded as virtually safe during the observation period, while the other state is the non-zero crash state. For example, consider vehicle crash occurring per year on 1-kilometer sections of highway. For straight sections of roadway with wide lanes, low traffic volumes, and no roadside objects, the likelihood of a vehicle crash occurring may be extremely small, but still present because an extreme human error could randomly cause an accident. These sections are considered to be in a zero-crash state that refer to situations where the likelihood of an event occurring is extremely rare in comparison to the non-zero state where crash occurrence is inevitable and follows some count distribution [46]. To address the zero-inflated modeling processes, the zero-inflated Poisson (ZIP) and the zero-inflated negative binomial (ZINB) regression models have been developed. The probabilities of the two possible zero- and non-zero states are: p_i for the zero crash state, and (1-p_i) for the non-zero crash state, and the overall probability of crashes is the sum of the probabilities from each state. The probability of crash frequency in the zero state can be modeled as:

$$Pr\left(n_i = 0\right) = p_i + \left(1 - p_i\right) R_i\left(0\right) \tag{7}$$

where $R_i(0)$ is the probability of zero crashes that occurs in the zero state. The probability of crash frequency in the non-zero state can be modeled as:

$$Pr\left(n_i > 0\right) = \left(1 - p_i\right) R_i\left(n_i\right) \tag{8}$$

where $R_i\ (n_i)$ is the probability of non-zero crashes in the non-zero state. Maximum likelihood estimates can be used to estimate the parameters of both ZIP and ZINB regression models and confidence intervals are constructed by likelihood ratio tests. In zero-inflated models, the two state process is assumed to follow a logit (logistic) or probit (normal) probability process [45]. Zero-inflated models have shown great flexibility in both states, although their applicability to crash prediction has been criticized because of the long term mean equals to zero in the safe state, and hence, biased estimates may result [7].

6.6. Conway-Maxwell Poisson Regression Models

The Conway-Maxwell Poisson model has been recently investigated with respect to highway safety issues, but its application in crash frequency modeling has been rather limited [7]. Generalized additive models have been explored given that they can provide smoothing functions for the explanatory variables. The Conway-Maxwell Poisson distribution is a generalization of the Poisson distribution that can handle both under-dispersed and over-dispersed crash data. The main advantage of this model is to handle the under-dispersion in crash data that cannot be modeled by the Poisson model or the Negative Binomial model. However, the low sample-mean, and small sample size of the under-dispersed crash data can influence the estimated parameters, and therefore, it has been limited in the application of crash frequency [24]. However, in practice, the estimation of these models can become very difficult as they require more parameters, a problem that has likely impeded their application to crash frequency prediction [29] [47].

6.7. Random-Parameter Models

Random-parameters models have also been investigated to take the effect of the unobserved heterogeneity from one roadway site to another [31] [32].

The motivation for random-parameter models is to account for unobserved heterogeneity across observations. Random-parameter models can be derived by assuming that the estimated parameters vary across observations according to some distribution. Estimated parameters can be modeled as [48]:

$$\beta_n = \beta + \omega_n \tag{9}$$

where

β_n: a vector of estimated parameters of the n observations,

ω_n: a randomly distributed term.

With this equation, the Poisson, and the Negative Binomial parameters become:

$$\lambda i \left| \omega_n \right. = EXP\left(\beta_n X_n\right) \tag{10}$$

$$\lambda i \left| \omega_n \right. = EXP\left(\beta n \; Xn + \varepsilon_n\right) \tag{11}$$

6.8. Artificial Neural Networks and Fuzzy Logic models

Given that a linear function may not sufficiently explain the relationship between the dependent variables and the associated independent variables in crash modeling, non-linear approximators such as fuzzy logic and neural networks have also been explored. Artificial Neural Networks (ANNs) are a class of computational intelligence tools that can be used for prediction and classification problems. ANNs can model very complex non-linear functions to high accuracy levels using a process of learning that is similar to the learning procedure of the cognitive system in the human brain. The network body is composed of input layers, hidden layers, and output layers. These models can be trained to approximate any nonlinear function to a required degree of accuracy using a learning algorithm (such as back propagation) that would give the desired output, in a supervised learning process. ANNs have some advantages over the statistical models. For instance, regression models need a pre-defined relationship or functional form between the dependent variable (crash frequency) and the independent explanatory variables that can be estimated by some statistical approaches, whereas the ANNs do not require the establishment of these functional forms, and can be easily applied in the analysis. On the other hand, the ANNs differ from the statistical models in that they behave as black-boxes and do not provide interpretation for the parameter estimates [15] [18] [35] [36]. Fuzzy logic applications have increasingly been proven to have a significant crash-predicting capability in recent years [49]. Fuzzy logic system is defined as the nonlinear mapping of an input data set to a scalar output data, and the first step of the process (known as fuzzification) consists of gathering a crisp set of input data that will be converted to a fuzzy set using fuzzy linguistic variables, fuzzy linguistic terms, and membership functions. After that, an inference is made based on a set of fuzzy rules, and then, the resulting fuzzy output is mapped to a crisp output using the membership functions, in the defuzzification step [33].

6.9. Logit and Probit Models

Logit and Probit models can be applied to study crash severity modeling. The data used in modeling crash severity is often attributed with many details relating to the crash occurrence (*i.e.* such as the number of vehicles involved, age of victims, weather conditions, types of vehicles involved, and crash type) which can be integrated in statistical models. Since the dependent variable (*i.e.* crash severity) usually has two or more outcome categories (*i.e.* fatal, injury, property-damage-only), logit and probit models are often used to model the severity of crash data. Discriminant analysis could also be used to model crash severity, but given its rigid assumptions, logit and probit models have been viewed as preferable [32] [50]. Binary models consider two response outcomes (*i.e.* fatal vs. non-fatal or injury vs. property-damage-only), and multinomial models consider three or more response outcomes. Traffic accident severity models can be generally classified as either nominal or ordinal. Although there is no consensus on which model is the best, as the selection of the model is often governed by the characteristics of the data, some researchers have opted for nominal models over ordinal models. The rationale for this choice is likely due to the influence that independent variables in ordinal models could exert on the ordered discrete outcome probabilities. That is, in closely related categories (*i.e.* no injury and possible injury) there may be some shared unobserved effects among adjacent injury categories. Failing to account for such correlation could generate incorrect inferences [32] [51]. Others still prefer ordinal models due to their simplicity and overall performance, especially when less detailed data are available. Binary models consider two outcomes, and multinomial models consider three or more outcomes. In binomial logit or probit models, the dependent variable, Y, can take one of two values 0.0 or 1.0. For example, injury or non-injury, fatal or non-fatal. The general shape of binomial logit model is (assuming $\pi_i = Pr(Y_i = 1)$):

$$\text{Logit}\left(\pi_i\right)=\log\left[\frac{\pi_i}{1-\pi_i}\right]==X_i\beta \qquad (12)$$

where,

XX_i: a vector of explanatory variables (*i.e.* risk factors),

β: a vector of regression coefficients.

As π approaches zero, *logit* (π) tends toward $-\infty$; and as π approaches 1.0, *logit* (π) tends toward $+\infty$ [52]. The binomial probit model is an alternative to the binomial logit model, in which the *probit* (π_i) is the standard cumulative normal distribution function (θ^{-1}) that can be expressed as:

$$\text{Probit}\left(\pi_i\right)=\theta^{-1}\left(\pi_i\right)=Xi\beta \qquad (13)$$

There are many types of the multinomial models that can be used in modeling crash severity, such as, the multinomial logistic regression (MNL), the nested logistic regression, the mixed logistic regression, and the multinomial probit models. For example, The MNL tries to find the best fitted model to describe the relationship between the polytomous dependent variable with more than two categories and a set of independent variables. The logistic regression model is a non-linear transformation of the linear regression model, as it consists of an S-shaped distribution function [53]. The logit distribution constrains the estimated probabilities that lie between 0.0 and 1.0, as shown in Figure 1.0. The logistic regression function is bounded by 0.0 and 1.0, whereas the linear regression function may predict values above 1.0 and below 0.0.The logistic (logit) function can be expressed as:

$$\text{Logit}\left(p\right)=b_0+b_1X_1+b_2X_2+\cdots+b_kX_k \qquad (14)$$

where,

p: the probability of presence of an outcome of interest,

X_k: the vector of k independent variables,

b_0: the regression coefficient on the constant term (intercept),

b_k: the vector of regression coefficients on the independent variables X_k,

The odd ratio is the probability of the event divided by the probability of the nonevent, and is defined as follows [50] [53]:

$$\text{odd ratios} = p/(1-p) \qquad (15)$$

When $p = 0$, then odd (p) = 0, when $p = 0.5$, then odd (p) = 1.0, and when $p = 1.0$, then odd (p) = ∞. The logit transformation is defined as the logged odds:

$$\text{Logit}\left(p\right)=\ln\left[p/(1-p)\right] \qquad (16)$$

The transformation from odds to log of odds is the log transformation, and this is a monotonic transformation. That is, the greater the odds, the greater the log of odds and vice versa. Logit (p) can be back-transformed to p by the following formula:

$$p=\frac{1}{1+e^{-logit(p)}} \qquad (17)$$

The transformation from probability to odds is a monotonic transformation as well, meaning the odds increase as the probability increases or vice versa.

Probability ranges from 0.0 and 1.0. Odds range from 0.0 and positive infinity [53] [54].

7. Conclusion

Traffic crash prediction models are very useful tools in road safety programs used by transportation agencies, police, health departments, education institutions that oversee road safety, vehicles, and the driver's education. They can be used to predict both the frequency of crash occurrence and the contributing factors that could then be addressed by transportation policies. According to the world health organization (WHO), road crashes are ranked as the ninth most serious cause of death in the world, and present the world's leading cause of death for individuals between the ages of one and twenty-nine. Each year, traffic accidents are responsible for killing about 1.25 million people and injuring approximately 50 million more. Following current trends, about two million people could be expected to be killed in motor vehicle crashes each year by 2030. The World Bank estimates that road traffic injuries cost 2.0 percent to 3.0 percent of the Gross National Product of developing countries. Given a such trend, this paper presented different types of traffic crash prediction models to gain a better understanding of the techniques used to predict road accidents and their contributing risk factors. A wide range of statistical approaches were presented including, Poisson regression, Negative Binomial regression, Zero-Inflated models, logit and probit models, and machine learning methods.

References

[1] World Health Organization (2015) Global Status Report on Road Safety 2015. http://www.who.int/violence_injury_prevention/road_safety_status/2015/en/

[2] NHTSA—National Center for Statistics and Analysis (NCSA) (2016) NHTSA Studies Vehicle Safety and Driving Behavior to Reduce Vehicle Crashes. http://www.nhtsa.gov/NCSA

[3] Transport Canada (2016) Road Safety in Canada. http://www.tc.gc.ca/eng/motorvehiclesafety/tp-tp15145-1201.htm

[4] Beirness, D.J. and Beasley, E. (2011) A Comparison of Drug- and Alcohol-Involved Motor Vehicle Driver Fatalities. Canadian Centre on Substance Abuse, Ottawa.

[5] World Bank (2015) The World Bank-Transport for Development. http://blogs.worldbank.org/transport/why-vehicle-safety-matters-crash-related-deaths?cid=EXT_WBBlogSocialShare_D_EXT

[6] NHTSA—National Center for Statistics and Analysis (NCSA) (2016) NHTSA. https://www.nhtsa.gov/press-releases/us-dot-announces-steep-increase-roadway-deaths-based-2015-early-estimates

[7] Lord, D., and Mannering, F. (2010) The Statistical Analysis of Crash Frequency Data: A Review and Assessment of Methodological Alternatives. *Accident Analysis and Prevention*, **44**, 291-305.

[8] Mohan, D. (2002) Road Safety in Less-Motorized Environments: Future Concerns. *International Journal of Epidemiology*, **31**, 527-532. https://doi.org/10.1093/ije/31.3.527

[9] Elvik, R. (2006) Laws of Accident Causation. *Accident Analysis and Prevention*, **38**, 742-747.

[10] Caliendo, C., Guida, M. and Parisi, A. (2007) A Crash-Prediction Model for Multilane Roads. *Accident Analysis and Prevention*, **39**, 657-670.

[11] Greibe, P. (2003) Accident Prediction Models for Urban Roads. *Accident Analysis and Prevention*, **35**, 273-285.

[12] Delen, D., Sharada, R. and Bessonov, M. (2006) Identifying Significant Predictors of Injury Severity in Traffic Accidents Using a Series of Artificial Neural Networks. *Accident Analysis and Prevention*, **38**, 434-444.

[13] Gelman, A. and Hill, J. (2007) Data Analysis Using Regression and Multilevel Hierarchical Models. Cambridge University Press, London.

[14] Kim, D.G., Lee, Y., Washington, S. and Choi, K. (2007) Modeling Crash Outcome Probabilities at Rural Intersections: Application of Hierarchical Binomial Logistic Models. *Accident Analysis and Prevention*, **39**, 125-134.

[15] Abdulhafedh, A. (2016) Crash Frequency Analysis. *Journal of Transportation Technologies*, **6**, 169-180.

[16] Blincoe, J., Miller, R., Zaloshnja, E. and Lawrence, A. (2015) The Economic and Societal Impact of Motor Vehicle Crashes, 2010. National Highway Traffic, Washington DC.

[17] Park, S. and Lord, D. (2007) Multivariate Poisson-Lognormal Models for Jointly Modeling Crash Frequency by Severity. *Transportation Research Record*, **2019**, 1-6. https://doi.org/10.3141/2019-01

[18] Ma, J., Kockelman, K.M. and Damien, P. (2008) A Multivariate Poisson-Lognormal Regression Model for Prediction of Crash Counts by Severity, Using Bayesian Methods. *Accident Analysis and Prevention*, **40**, 964-975.

[19] El-Basyouny, K. and Sayed, T. (2009) Collision Prediction Models Using Multivariate Poisson-Lognormal Regression. *Accident Analysis and Prevention*, **41**, 820-828.

[20] Lord, D. and Bonneson, A. (2007) Development of Accident Modification Factors for Rural Frontage Road Segments in Texas. *Transportation Research Record*, **2023**, 20-27.

[21] Daniels, S., Brijs, T., Nuyts, E. and Wets, G. (2010) Explaining Variation in Safety Performance of Roundabouts. *Accident Analysis and Prevention*, **42**, 292-402.

[22] Malyshkina, N. and Mannering, F. (2010) Markov Switching Multinomial Logit Model: An Application to Accident-Injury Severities. A*ccident Analysis and Prevention*, **41**, 829-838.

[23] Geedipally, R., Lord, D. and Dhavala, S. (2012) The Negative-Binomial Lindley Generalized Linear Model: Characteristics and Application Using Crash Data. *Accident Analysis and Prevention*, **45**, 258-265.

[24] Lord, D. (2006) Modeling Motor Vehicle Crashes Using Poisson-Gamma Models: Examining the Effects of Low Sample Mean Values and Small Sample Size on the Estimation of the Fixed Dispersion Parameter. *Accident Analysis and Prevention*, **46**, 751-766.

[25] Oh, J., Washington, S. and Nam, D. (2006) Accident Prediction Model for Railway-Highway Interfaces. *Accident Analysis and Prevention*, **38**, 346-356.

[26] Lord, D. and Miranda-Moreno, F. (2008) Effects of Low Sample Mean Values and Small Sample Size on the Estimation of the Fixed Dispersion Parameter of Poisson-Gamma Models for Modeling Motor Vehicle Crashes: A Bayesian Perspective. *Accident Analysis and Prevention*, **46**, 751-770.

[27] Aguero-Valverde, J. and Jovanis, P. (2008) Analysis of Road Crash Frequency with

Spatial Models. *Transportation Research Record*, **2061**, 55-63. https://doi.org/10.3141/2061-07

[28] Xie, Y. and Zhang, Y. (2008) Crash Frequency Analysis with Generalized Additive Models. *Transportation Research Record*, **2061**, 39-45. https://doi.org/10.3141/2061-05

[29] Li, X., Lord, D., Zhang, Y. and Xie, Y. (2009) Predicting Motor Vehicle Crashes Using Support Vector Machine Models. *Accident Analysis and Prevention*, **40**, 1611-1618.

[30] Milton, J.C., Shankar, V. and Mannering, F. (2008) Highway Accident Severities and the Mixed Logit Model: An Exploratory Empirical Analysis. *Accident Analysis and Prevention*, **40**, 260-266.

[31] Anastasopoulos, P.C. and Mannering, F. (2009) A Note on Modeling Vehicle Accident Frequencies with Random-Parameters Count Models. *Accident Analysis and Prevention*, **41**, 153-159.

[32] Washington, P., Karlaftis, G. and Mannering, F. (2010) Statistical and Econometric Methods for Transportation Data Analysis. 2nd Edition, Chapman Hall, Control and Reporting Center, Boca Raton.

[33] Meng, H., Zheng, L. and Qing, M. (2009) Traffic Accidents Prediction and Prominent Influencing Factors Analysis Based on Fuzzy Logic. *Accident Analysis and Prevention*, **9**, 87-92.

[34] Abdelwahab, H.T. and Abdel-Aty, M.A. (2002) Artificial Neural Networks and Logit Models for Traffic Safety Analysis of Toll Plazas. *Accident Analysis and Prevention*, **1784**, 115-125. https://doi.org/10.3141/1784-15

[35] Riviere, C., Lauret, P., Ramsamy, M. and Page, Y. (2006) A Bayesian Neural Network Approach to Estimating the Energy Equivalent Speed. *Accident Analysis and Prevention*, **38**, 248-259.

[36] Chang, L.Y. (2005) Analysis of Freeway Accident Frequencies: Negative Binomial Regression versus Artificial Neural Network. *Accident Analysis and Prevention*, **43**, 541-557.

[37] Cameron, A.C., and Trivedi, P.K. (1998) Regression Analysis of Count Data. Cambridge University Press, Cambridge, UK. https://doi.org/10.1017/CBO9780511814365

[38] Abdel-Aty, M. (2003) Analysis of Driver Injury Severity Levels at Multiple Locations Using Ordered Probit Models. *Journal of Safety Research*, **34**, 597-603.

[39] Chang, L.Y., and Wang, H. (2006) Analysis of Traffic Injury Severity: An Application of Nonparametric Classification Tree Techniques. *Accident Analysis and Prevention*, **38**, 1019-1027. https://doi.org/10.1016/j.aap.2006.04.009

[40] Bham, G., Javvadi, B. and Manepalli, U. (2012) Multinomial Logistic Regression Model for Single-Vehicle and Multivehicle Collisions on Urban U.S. Highways in Arkansas. *Journal of Transportation Engineering*, **138**, 786-797 https://doi.org/10.1061/(ASCE)TE.1943-5436.0000370

[41] Glenberg, A. (1996) Learning from Data: An Introduction to Statistical Reasoning. 2nd Edition, Lawrence Erlbaum Associates, Mahwah.

[42] Hilbe, J. (2007) Negative Binomial Regression. Cambridge University Press, London. https://doi.org/10.1017/CBO9780511811852

[43] Hilbe, J. (2014) Modeling Count Data. Cambridge University Press, London. https://doi.org/10.1017/CBO9781139236065

[44] Amoros, E., Martin, J.L., & Laumon, B. (2003) Comparison of Road Crash Incidents

and Severity between Some French Counties. *Accident Analysis and Prevention*, **35**, 537-547. https://doi.org/10.1016/S0001-4575(02)00031-3

[45] Shankar, N., Milton, J.C. and Mannering, F. (1997) Modeling Accident Frequencies as Zero-Altered Probability Process: An Empirical Enquiry. *Accident Analysis and Prevention*, **29**, 829-837.

[46] Lambert, D. (1992) Zero-Inflated Poisson Regression with an Application to Defects in Manufacturing. *Technometrics*, **34**, 1-14. https://doi.org/10.2307/1269547

[47] Zhang, J.X., Chang, K.T. and Wu, J.Q. (2008) Effects of Dynamic Effect Model Resolution and Source on Soil Erosion Modeling: A Case Study Using the Water Erosion Prediction Project Model. *International Journal of Geographical Information Science*, **22**, 925-942. https://doi.org/10.1080/13658810701776817

[48] Greene, W. (2008) Econometric Analysis. 6th Edition, Prentice-Hall. Upper Saddle River.

[49] Wang, H., Zheng, L. and Meng, X.H. (2011) Traffic Accidents Prediction Model Based on Fuzzy Logic. *Communications in Computer and Information Science*, **201**, 101-108. https://doi.org/10.1007/978-3-642-22418-8_14

[50] Greene, W. (2012) Econometric Analysis. 7th Edition, Prentice Hall, Upper Saddle River.

[51] Savolainen, P., Mannering, F., Lord, D. and Quddus, M. (2011) The Statistical Analysis of Highway Crash-Injury Severities: A Review and Assessment of Methodological Alternatives. *Accident Analysis and Prevention*, **43**, 1666-1676.

[52] Mannering, F. and Grosdsky, L. (1995) Statistical Analysis of Motorcyclist: Perceived Accident Risk. *Accident Analysis and Prevention*, **27**, 21-31.

[53] Judge, G., Griffiths, W.E., Hill, R.C., Lutkepohl, H. and Lee, T.C. (1985) The Theory and Practice of Econometrics. 2nd Edition, Wiley, New York.

[54] Baltagi, B.H. (2011) Econometrics. 5th Edition, Springer, Berlin. https://doi.org/10.1007/978-3-642-20059-5