Civil Engineering Research Journal
ISSN: 2575-8950

Juniper Publishers
key to the Researchers

Research Article
Volume 3 Issue 5 - March 2018
DOI: 10.19080/CERJ.2018.03.555622

Civil Eng Res J

Statistical Analysis of Vehicle Crashes on Mississippi Coastal Two-Lane Highways

Zhao Ma, Ningning Wang and Feng Wang*

Dept. of Civil & Environmental Engineering, Jackson State University, Mississippi

Submission: December 27, 2017; **Published:** March 05, 2018

***Corresponding author:** Feng Wang, Dept. of Civil & Environmental Engineering, Jackson State University, Jackson, Mississippi 39217, Tel: 601-979-1094; Fax: 601-979-3238; Email: feng.wang@jsums.edu

Abstract

The traffic fatality per capita of Mississippi has been at about twice the national average level in the last five years. Although tremendous efforts have been made to develop the State Highway Safety Plan, limited attention has been paid to good understanding of the characteristics of the crashes in Mississippi. Due to the relatively high percentage of heavy trucks in and out of the seaports, the traffic crashes on two-lane county roads of the coastal area are suspected to be more severe. A binary logistic regression model is developed to expose the factors that contribute to the crash severity on the two-lane county roads of the Mississippi coastal area. The study conducts statistical analyses using the crash data of the past three years to find possible relationships between crash severities on these two-lane highways and the factors or combinations of factors of time of day, environment, roadway, vehicle, driver, and driver behavior. The analysis results indicate that those crashes that involved with vehicles, dark lighted and dusk or dawn light conditions, drivers with no driver license or suspended driver license, high speed, speeding, and none restraint usage or helmet usage, tend to increase the probability to be more severe crashes. On the other hand, it is also presented that male drivers decrease the likelihood of fatal or injury crashes. The predictive power of the model is tested under 10-fold cross validation. The results show that the model has significantly higher predictive power than a non-information guess.

Keywords: Two-lane highway; County roads; Binary logistic regression; Crash severity; Cross validation

Introduction

Significance of study

Vehicle crashes are considered among top 10 leading causes of deaths in the United States. According to the data from the National Highway Traffic Safety Administration (NHTSA), more than 30,000 people died from vehicle crashes every year since 1949. The numbers of people that died from vehicle crashes are 32,479, 33,782, and 32,719 during the years 2011, 2012, and 2013, respectively, while the numbers of people who were injured in these three years are 2,217 2,362 and 2,313 thousands, respectively [1]. Vehicle crashes, which take a major weight of traffic safety, have been a nationwide focus in the United States. The current traffic safety situation in Mississippi has been of great concerns. From Table 1, it is indicated that vehicle crashes caused around 600 fatalities in Mississippi each year in the past three years. But the fatality rate per capita, calculated at around 20 fatalities per 100,000 population which is almost twice as high as the US average level, is actually among the highest in the country. The fatality rate assessed at over 1.5 fatalities per million vehicle miles travelled (VMT), is also much higher than that of the nationwide average. The NHTSA's Traffic Safety Facts: State Traffic Crash Data have shown that the fatality rate is decreasing over the past decades. The fatality rate per million VMT also decreased by 58% from 1975 to 2013 in Mississippi. However, considering the original high level and the nationwide decreasing trends (The fatality rate per VMT in the US decreased by 67% in the same time period), Mississippi is still among the top states in traffic crash rate. All these facts clearly indicate that the traffic safety situation in Mississippi is still severe. Three Mississippi coastal counties, Harrison, Hancock, and Jackson, which have a total population around 400,000 out of the approximately 3 million population in 82 Mississippi counties in the state, are areas with high values in both population and vehicle crash fatalities. Gulfport is located at the center of the Mississippi coast and the second largest city in Mississippi after the state capital Jackson. High freight traffic is generated due to the transshipping of freight from cargo vessels to trucks and intermodal operations at Port of Gulfport, which is No. 19 in the US in terms of containership and among the top 50 US ports by port calls and vessel type [2]. Table 2 was derived from

vehicle crash records collected by the Mississippi Department of Transportation (MDOT) in year 2011-2013, which would explain the motivation of this study. From Table 2, it can be shown that county road is the road class with the highest fatality rate and injury rate. Among the 4025 crash records on Mississippi coastal county roads, 3684 (91.53%) were on 2-lane highways, which indicates that 2-lane county roads are hazardous spots that entail more attention. Mining out factors that contribute to more frequent and severe crashes would be the first step to improve the undesirable situation on 2-lane highways of the Mississippi coastal area.

Table 1: Traffic Crash Data for MS vs. US (Source: NHTSA)

Data Item	2011		2012		2013	
	US	MS	US	MS	US	MS
Fatalities per 100,000 Drivers	15.28	32.7	15.84	29.72	15.42	31.13
Registered Vehicles (thousands)	2,57,512	2,037	2,65,647	2,052	2,69,294	2,074
Fatalities per 100,000 Vehicles	12.57	30.94	12.63	28.36	12.15	29.56
Population (thousands)	3,11,592	2,979	3,13,914	2,985	3,16,129	2,991
Fatalities per 100,000 Population	10.39	21.15	10.69	19.5	10.35	20.49
Vehicle Miles Travelled (millions)	2,946,131	38,851	2,968,815	38,667	NA	NA
Fatalities per 100 million VMT	1.1	1.62	1.13	1.51	NA	NA
Total Killed	32,367	630	33,561	582	32,719	613

Table 2: Crashes in Mississippi Coastal Counties in 2011-2013 Assorted by Road System.

Data Item	Fatality		Injury		Property Damage Only	
Road System	Frequency	Percentage	Frequency	Percentage	Frequency	Percentage
Interstate	25	0.61	1262	30.93	2793	68.46
US Highway	42	0.31	4429	32.46	9173	67.23
State Highway	29	0.41	2159	30.63	4861	68.96
City Street	19	0.14	4241	30.61	9597	69.26
County Road	29	0.72	1650	40.99	2346	58.29
Parking Lot or Private Drive	2	0.33	134	21.97	474	77.7
Off Road	0	0	3	11.54	23	88.46
State Park	0	0.61	0	30.93	6	68.46

Statistical models

In the past twenty years, numerous studies have applied statistical models on crash analysis. Traditionally, negative binomial models have been applied to assess highway safety based on crash counts and crash rates [3-8]. Crash frequency models were developed for each collision type [3]. They analyzed the individual collision types by comparing the aggregate model results. The results indicated that annual daily traffic (AADT), lane number, and the presence of turning lanes were positively related with all collision types, while median width and light condition were negatively related with different collision types. However, Pande and Abdel-Aty [9] pointed out the limitations of the negative binomial models. Lengths of the segments selected to aggregate the crash data were hard to be determined. There was not an agreeable way of optimizing the selection of the segment lengths for crash studies.

Logistic regression models have been widely used in analyses of crash severity as a response variable. Dissanayake and Lu [10] modeled crash severity for single-vehicle fixed object crashes involving young drivers. By using sequential binary logistic regression, they modeled the crash severity wit five categories, which were no injury, possible injury, non-capacitating injury, incapacitating injury and fatal. Factors such as alcohol or drug influence, ejection in the crash, point of impact, rural locations, curved or graded crash location and speed of vehicle significantly increased the probability of more severe crashes. Restraint device usage and male drivers were considered to reduce the crash severity level. It was also found that factors such as weather, residence location, and physical condition did not have significant influences on crash severity using this model. Binary logistic regressions have been used when the response variable is binary. Lui et al. [11] presented significant findings modeling

How to cite this article: Zhao M, Ningning W, Feng W. Statistical Analysis of Vehicle Crashes on Mississippi Coastal Two-Lane Highways.Civil Eng Res J. 2018; 3(5): 555622. DOI: 10.19080/CERJ.2018.03.555622

crash severity with a multivariate approach. Shanker et al. [3] developed a predictive model of crash severity with a nested logit model. Kusano & Garbler [12] tested the predictive power of logistic regression and machine learning, and concluded that logistic regression slightly outperformed machine learning. The authors also mentioned that the improvement of prediction accuracy is very meaningful to reduce the odds of death by guiding the trauma team to take seriously injured occupants into a trauma center to receive necessary treatments.

The objectives of this paper are to identify the factors that are likely to lead to more severe crashes, and to build up a predictive model which can significantly increase the prediction accuracy of the crash severity. The organization of the paper is as follows: after the Introduction section, the Methodology section presents the multivariate logistic regression model and model validation, followed by description of the data used in the analysis. In the Data Analysis and Results section, the variable selection in the data analysis process and regression analysis results are presented followed by careful checks and discussions on the prediction power of the regression model used for the study. Finally findings of this study are summarized in Conclusions.

Methodology

Logistic Regression Model

Logistic regression is one of the popular regression methods which describe the relationship between explanatory variables and a discrete response variable. The explanatory variables can be either categorical or numerical, or a mixture of both. The model is generally used to handle categorical variables. A binary logistic regression is good to use when the dependent variable is a bivariate. In this study, the dependent variable can only take on two values: y = 1 for fatality or injury, and y = 0 for property damage only. The probability that a fatality or injury takes place is modeled as logistic distribution by Equation (1):

$$P\left(y_i = 1/X\right) = \pi\left(X_i\right) = \frac{e^{\beta X_i}}{1+e^{\beta X_i}} \quad (1)$$

And the logit of the binary logistic regression model is presented in Equation (2):

$$g\left(X_i\right) = \ln\left[\frac{\pi\left(X_i\right)}{1-\pi\left(X_i\right)}\right] = \beta X_i \quad (2)$$

Where $P(.)$ stands for the probability of a severe accident $(y_i = 1)$; X_i is the vector of independent variables for the i^{th} observation; (X_i) is the conditional probability of a fatality or injury that occurs when an accident is present; is the coefficient vector, which directly determine the odds ratio involved in the fatality or injury; $g\left(X_i\right)$ is the link function.

The odds ratio for the j^{th} independent variable that is equal to $e^{\beta j}$ represents the relative value by which the odds of the fatality or injury increase or decrease when the value of the j^{th} predictor is increased by 1.0 units.

The estimation of the coefficient vector is processed by the maximum likelihood method [13].The likelihood function is given by Equation (3):

$$L\left(\beta/y, X\right) = \Pi_{i=1}^{n} P\left(y_i/X_i\right) \quad (3)$$

Where $L(.)$ stands for the maximum likelihood a crash severity given the observed data. Combine Equations (2) and (3), noting that $P\left(y_i = 0/X_i\right) = 1 - P\left(y_i = 1/X_i\right)$:

$$L\left(\beta/y, X\right) = \Pi_{i=1}^{n} P\left(y_i/X_i\right) = \Pi_{i=1}^{n} P\left(y_i = 1/X_i\right)^{y_i} P\left(y_i = 0/X_i\right)^{1-y_i} \quad (4)$$

Taking logs, the log-likelihood function can be written as follows:

$$L\left(\beta/y, X\right) = \sum_{i=1}^{n} y_i \ln P\left(y_i = 1/X_i\right) + (1-y_i) \ln P\left(y_i = 0/X_i\right) \quad (5)$$

Iterations were applied to maximize the log-likelihood function and achieve the estimation of the coefficient vector. Due to the complex computation, the open source statistical analysis software R program was adopted to conduct the estimation of the coefficient vector in this study.

Different plausible models were built and tested for goodness of fit using the Wald chi-square measures, and the best fitted model was selected as the final model. A set of null and alternative hypotheses were assumed to construct different models and tested in the Analysis of Variances (ANOVA). Specifically, under the null hypothesis, the reduced model is the adequate model while under the alternative hypothesis the full model is the adequate model. Iteratively the full model is reduced based on each of the hypothesis tests, while the Wald chi-square test was applied to determine whetherb or not to statistically reject the null hypothesi based on the pre-selected p-value for the level of significance.

Model validation

The 10-fold cross validation process was applied in this study to test the predictive power of the selected model. If the predictors are trained to a dataset and then the same data are used to test the model's accuracy, and the model that over-fits the dataset is generally considered to have the best performance. However, this classifier may perform poorly comparing to a more flexible classifier with new data [12]. The 10-fold cross validation method was adopted in order to eliminate the overly optimistic estimates of model performance.

The procedure of 10-fold cross validation includes:

1) divide the dataset into 10 even subsets;

2) use 9 subsets to train the model and 1 subset to test the model; and

3) repeat the steps in 1) and 2) for10 times until all subsets are tested as the testing dataset.

How to cite this article: Zhao M, Ningning W, Feng W. Statistical Analysis of Vehicle Crashes on Mississippi Coastal Two-Lane Highways.Civil Eng Res J. 2018; 3(5): 555622. DOI: 10.19080/CERJ.2018.03.555622

In this study, the receiver operator characteristic (ROC) curve and confusion matrix are used to show the predictive power of a logistic regression model. ROC curve, which consists of the true positive rate and false positive rate, has the advantage of showing the predictive power in a stable shape, while the confusion matrix exposes the predictive power numerically and in a straight forward manner.

Dataset

The data for this study were provided by the Mississippi Department of Transportation (MDOT). The original dataset consists of vehicle crash records in Mississippi for years 2011, 2012, and 2013. The dataset with 3684 usable observations was achieved by screening and cleansing the data for crashes on Mississippi coastal county roads.

Seven types of crash information were used that included: 1) crash data, 2) temporal data, 3) environment data, 4) road data, 5) vehicle data, 6) driver data, and driver behavior data. The crash data includes severity and number of vehicles involved. Temporal data is equivalent to day of week and time of day. Environment data contains light condition, weather, and road surface condition. Road data consists of whether or not at an intersection and pavement surface material. Vehicle data shows vehicle type. Driver data provides information of driver's age, gender, race, and driver licensure status. Driver behavior data refers to estimated speed, speeding, and restraint usage. Table 3 shows the frequency and percentage of explanatory variables. A total of 3864 records were retrieved for the county roads in the Mississippi coast, with 1549 fatal or injury records and 2135 property damage only records.

Table 3: Summary of Crash Statistics in Mississippi Coastal Counties.

Type of variable	Variable	Category	Frequency	Percentage
Crash	Severity	Fatality or Injury	1549	42.05
		Property Damage Only	2135	57.95
	Vehicle Involved	1	1337	36.29
		2	2159	58.6
		>2	188	5.1
Temporal	Day of Week	Weekday	2699	73.26
		Weekend	985	26.74
Environment	Light Condition	Daylight	2480	67.32
		Dark Lighted	332	9.01
		Dark Unlighted	775	21.04
		Dusk or Dawn	97	2.63
	Weather	Clear or Cloudy	3183	86.4
		Rain	451	12.24
		Fog/Smog/Smoke	50	1.36
	Surface Condition	Dry	2967	80.54
		Wet	717	19.46
Road	Intersection	Yes	1994	54.13
		No	1690	45.87
	Surface Material	Asphalt	3556	96.53
		Concrete	93	2.52
		Other	35	0.95
Vehicle	Vehicle Type	Passenger Car	1689	45.85
		Light Truck	1055	28.64
		SUV	626	16.99
		Van	143	3.88
		Motorcycle	90	2.44
		Other	81	2.2

How to cite this article: Zhao M, Ningning W, Feng W. Statistical Analysis of Vehicle Crashes on Mississippi Coastal Two-Lane Highways.Civil Eng Res J. 2018; 3(5): 555622. DOI: 10.19080/CERJ.2018.03.555622

Driver	Age	<25	1216	33.01
		25-34	804	21.82
		35-44	553	15.01
		45-54	516	14.01
		55-64	318	8.63
		65-74	176	4.78
		>75	101	2.74
	Race	White	3180	86.32
		Black	372	10.1
		Hispanic	61	1.66
		Other	71	1.93
	Gender	Male	2114	57.38
		Female	1570	42.62
	Driver License	Valid	3326	90.28
		No License	108	2.93
		Suspended	180	4.89
		Suspended-DUI	43	1.17
		Expired	16	0.43
		Other	8	0.22
Driver Behavior	Speed	<25mph	1438	39.03
		25-40mph	1128	30.62
		40-55mph	921	25
		55-70mph	151	4.1
		>70mph	46	1.25
	Speeding	Yes	1852	50.27
	Restraint Usage	Shoulder Lap Belt	3382	91.8
		Helmet	72	1.95
		None	230	6.24

The missing values and unreasonable data were removed during initial data processing. The removed data represented a small proportion of the dataset, which means removing them would not lead to bias in data analysis and unreliable analysis results.

Data Analysis and Results

Variable selection

The R program was applied to process and analyze the data in the following 4 steps. First, all variables were included in the binary logistic regression model to test the significance of every variable. Then, variables with values larger than >0.05 were removed iteratively to fit the reduced models. The third step is to conduct the likelihood ratio test to determine if we need to reject the null hypothesis that the reduced model is true. Repeat the above steps until the reduced model was rejected and the process of selecting variables for the regression model is then terminated. Table 4 shows the analysis results for variable selection, which are described in the following paragraphs. Table 5 presents the ANOVA results obtained from the R program. As shown in the table, Model No. 3 is the final model in our study. It includes variables with at least one significant category. However, age was an exception due to the relatively small proportion (2.74%) of the significant category for drivers aged 75 and above. Removing insignificant variables can prevent unnecessary disturbance without losing the predictive power of the model.

Table 4: Variable Selection.

Regression Results			Full Model		Reduced Model	
Variable	**Reference**	**Category**	**Estimate**	**P-value**	**Estimate**	**P-value**
(Intercept)			-1.389	<0.0001	-1.2674	<0.0001

How to cite this article: Zhao M, Ningning W, Feng W. Statistical Analysis of Vehicle Crashes on Mississippi Coastal Two-Lane Highways.Civil Eng Res J. 2018; 3(5): 555622. DOI: 10.19080/CERJ.2018.03.555622

Vehicle Involved	1	2	0.3203	0.0008	0.3157	0.0005
		>2	0.9579	<0.0001	0.9164	<0.0001
Day of Week	Weekday	Weekend	0.0699	0.3945		
Light Condition	Daylight	Dark Lighted	0.369	0.0038	0.3435	0.0061
		Dark Unlighted	0.1434	0.1324	0.1352	0.1488
		Dusk or Dawn	0.4988	0.025	0.5275	0.0169
Weather	Clear or Cloudy	Rain	-0.2156	0.1986		
		Fog/Smog	0.0086	0.9778		
Surface Condition	Dry	Wet	0.0805	0.5616		
Intersection	Yes	No	0.0083	0.9169		
Surface	Asphalt	Concrete	-0.3241	0.1723		
		Other	0.3708	0.3089		
Vehicle Type	Passenger Car	Light Truck	-0.0345	0.7029		
		SUV	0.1332	0.1835		
		Van	0.1919	0.3059		
		Motorcycle	1.5957	0.129		
		Other	0.1292	0.613		
Age	<25	25-34	0.1003	0.312		
		35-44	-0.0.05	0.791		
		45-54	0.1279	0..2741		
		55-64	0.142	0.301		
		65-74	-0.1606	0.3779		
		>75	0.467	0.0334		
Race	White	Black	0.1797	0.134		
		Hispanic	0.0617	0.827		
		Other	-0.4226	0.1205		
Gender	Female	Male	-0.1945	0.0119	-0.2037	0.0051
Driver License	Valid	No License	0.5356	0.0175	0.5783	0.0008
		Suspended	0.4532	0.0081	0.4881	0.0038
		Suspended-DUI	0.6595	0.0558	0.6598	0.0525
		Expired	0.3981	0.4515	0.3484	0.5091
		Other	-0.019	0.9801	0.0124	0.9869
Speed	<25mph	25-40mph	0.4859	<0.0001	0.4702	<0.0001
		40-55mph	0.9077	<0.0001	0.8729	<0.0001
		55-70mph	0.8485	0.0001	0.8031	0.0003
		>70mph	1.8951	<0.0001	1.904	<0.0001
Speeding	No	Yes	0.2628	0.0236	0.3657	0.021
Restraint Usage	Shoulder Lap	Helmet	1.7128	0.1367	3.2904	<0.0001
	Belt	None	1.5356	<0.0001	1.6159	<0.0001
ANOVA Test Results:						
	Residual Df	Residual Dev	Degree of Freedom		Deviance	Pr(>Chi)
Full Model	3644	4537.7				
Reduced Model	3665	4564.8	-21		-27.079	0.1683

How to cite this article: Zhao M, Ningning W, Feng W. Statistical Analysis of Vehicle Crashes on Mississippi Coastal Two-Lane Highways.Civil Eng Res J. 2018; 3(5): 555622. DOI: 10.19080/CERJ.2018.03.555622

Table 5: ANOVA Test Results.

Model No.	Explanatory Variables	Residual DF	Residual Dev	DF	Deviance	Pr (>Chi)
1	Intercept, Light Condition, Vehicle Involved, Restraint Usage, Speed, Speeding, Gender, Driver License, Day of Week, Vehicle Type, Race, Age, Weather, Intersection, Surface Condition, Surface Material	3644	4537.7	--	--	--
2	Intercept, Light Condition, Vehicle Involved, Restraint Usage, Speed, Speeding, Gender Driver License, Day of Week, Vehicle Type, Race, Age	3650	4542.9	6	5.2559	0.5114
3	Intercept, Light Condition, Vehicle Involved, Restraint Usage, Speed, Speeding, Gender, Driver License	3665	4564.8	15	21.823	0.1125
4	Intercept, Light Condition, Vehicle Involved, Restraint Usage, Speed, Gender, Driver License	3666	4570.1	1	5.3471	0.0208

Binary logistic regression results

The odds ratios (OR) for the final predictive model are shown in Table 6, which stand for the ratios of log probabilities of involving in a fatal or injuring crash of a category over the reference category of a selected independent variable. Table 6 also lists the estimated coefficients (B) of the selected variables and the categories of the variables to predict the log probabilities of a severe crash. The coefficients are actually the log probabilities of a severe crash occurring due to the independent variables, and each coefficient stands for the change in the log probability of a severe crash associated with one-unit change in the independent variable. The seven variables retained in the final model include the number of vehicles involved, lighting condition, gender, driver license, estimated speed, speeding, and restraint usage. The effects of the seven independent variables in the model are discussed in the following paragraphs.

Table 6: Binary Logistic Regression: Parameter Estimates and Odds Ratios.

Variable	Category	Estimate (B)	Odds Ratio (OR)	95% Confidence Limits		P-value
(Intercept)		-1.2674				<0.0001
Vehicle Involved	1					

How to cite this article: Zhao M, Ningning W, Feng W. Statistical Analysis of Vehicle Crashes on Mississippi Coastal Two-Lane Highways.Civil Eng Res J. 2018; 3(5): 555622. DOI: 10.19080/CERJ.2018.03.555622

	2	0.3157	1.3712	1.1493	1.6379	0.0005
	>2	0.9164	2.5002	1.7833	3.5123	<0.0001
Light Condition	Daylight					
	Dark Lighted	0.3435	1.4099	1.1022	1.8014	0.0061
	Dark Unlighted	0.1352	1.1448	0.9525	1.3751	0.1488
	Dusk or Dawn	0.5275	1.6947	1.0993	2.6182	0.0169
Gender	Female					
	Male	-0.2037	0.8157	0.7073	0.9406	0.0051
Driver License	Valid					
	No License	0.5783	1.783	1.1652	2.752	0.0008
	Suspended	0.4881	1.6292	1.1712	2.2713	0.0038
	Suspended-DUI	0.6598	1.9344	1.0005	3.8317	0.0525
	Expired	0.3484	1.4169	0.4831	3.9678	0.5091
	Other	0.0124	1.0125	0.2011	4.3002	0.9869
Speed	<25mph					
	25-40mph	0.4702	1.6002	1.2716	2.0123	<0.0001
	40-55mph	0.8729	2.3938	1.8103	3.1664	<0.0001
	55-70mph	0.8031	2.2324	1.4487	3.4444	0.0003
	>70mph	1.904	6.7127	2.9167	17.5069	<0.0001
Over speed	No					
	Yes	0.3657	1.3043	1.0413	1.6354	0.021
Restrained	Shoulder Lap Belt					
	Helmet	3.2904	26.8537	11.7693	77.4934	<0.0001
	None	1.6159	5.0322	3.5996	7.159	<0.0001

Vehicles involved: All things being equal, the OR results indicate that the log probability ofhaving a severe (fatal or injuring) crash due to a crash involved with two vehicles is 1.3712 times as much as the log odd of crashes due to only single vehicles. Similarly, the crashes involving with more than two vehicles have a log probability of a fatal or injure severity 2.500 times the log odd of the crashes involving with only single vehicles.

Light Condition: A light condition has four categories. The OR value indicates that the log probability of having a fatal or injury accident due to a dark light condition is 1.4099 times that due to a daylight condition. Similarly, the log probabilities of having a fatal or injury crash due to dark unlighted condition and dusk or dawn condition are 1.14 and 1.69 times that due to a daylight condition respectively. The reason why the dark lighted and dusk or dawn conditions increase the probability of a fatal or injury crash might be because of the misestimate of vision loss under these conditions. Being aware of vision loss under a dark unlighted condition, the drivers would tend to pay closer attention to the road condition and drive more carefully, therefore could reduce the severity level of a crash.

Gender: Compared with female drivers, male drivers are statistically less likely to be involved in a fatal or injury crash, with an odd ratio of 0.8157.

Driver license: The results show that a driver with an expired driver license is not statistically significant for a severe crash. However, a driver without a driver license or with a driver license suspended is statistically significant for a severe crash. Compared with a validly licensed driver, the odd ratios of the log probability of involving in a fatal or injuring crash, for a driver without a driver license and a driver with a driver license suspended are 1.78 and 1.63 respectively.

Speed: Compared with the reference speed category of less than 25 mph, all other 4 categories are statistically significant in involvement with a severe crash. Compared with the reference speed category, the odds ratios of the log probability of involving in a fatal or injuring crash, for driving speeds at 25-40 mph, 40-55 mph, 55-70 mph, and over 70 mph are 1.6002, 2.3938, 2.2324, and 6.7127 respectively. Obviously this analysis result shows that the higher the driving speed, the larger the probability of involving in a fatal or injury crash.

Speeding: The odds ratio of 1.3 in involvement of a fatal or injury crash of a speeding behavior over the non-speeding driving with respect to the speed limit well indicates that speeding in driving is statistically significant in causing a severe crash.

Restraint usage: The usage of shoulder lap belt is set to be the reference category. The odds ratio of helmet usage at 26.85

How to cite this article: Zhao M, Ningning W, Feng W. Statistical Analysis of Vehicle Crashes on Mississippi Coastal Two-Lane Highways.Civil Eng Res J. 2018; 3(5): 555622. DOI: 10.19080/CERJ.2018.03.555622

actually indicates that a motorcycle driver wearing a helmet is 26.85 times more likely to be in a fatal or injury crash, compared with the log probability of a driver utilizing shoulder lap belt involving a severe crash. The odds ratio of a driver with no restraint usage at 5.03 well indicates the great importance of restraint usage in avoiding fatality and injury in driving.

Prediction power

As shown in Table 7, the confusion matrix under the 10-fold cross validation method has measured the predictive power of the final model developed in the binary logistic regression. The accuracy of the prediction at 0.6509 is significantly higher than the non-information guess at a value of 0.5795, which indicates that the logistic regression has a significantly higher predictive power than a non-information guess. The receiver operator characteristic (ROC) curve was also plotted to present the predictive power. The advantage of ROC curve is its stability in shape. The area under curve (AUC) is the index to evaluate ROC curve. A larger AUC means better predictive power. Figure 1 presents the AUC of the ROC curve of this logistic regression model which was assessed at 0.6805.

Table 7: Confusion Matrix.

	Reference		
Prediction	**Property Damage Only**		**Fatality or Injury**
Property Damage Only	1846		997
Fatality or Injury	289		552
Accuracy: 0.6509 95% CI: (0.6353,0.6663)		No Information Rate: 0.5795	P-value [Acc>NIR]: <0.0001

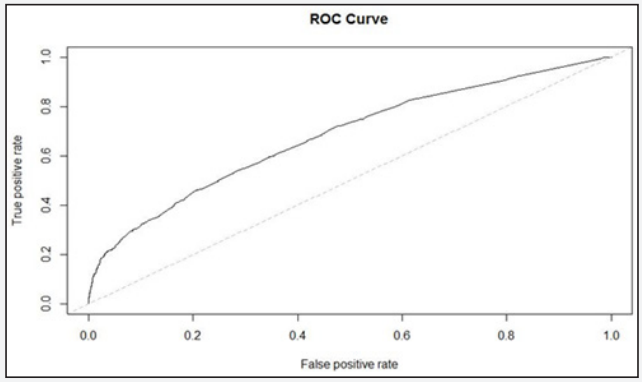

Figure 1: Receiver operator characteristic curve under 10-fold cross validation.

Conclusion

This study uses a binary logistic regression model to identify the factors that possibly contribute to the crash severities on two-lane county roads in the Mississippi coastal area. Vehicle crash severity is considered as dependent variable. Seven types of data, which include crash data, temporal data, environment data, roadway data, vehicle data, driver data, and driver behavior data are the explanatory variables used in predicting crash severity. The analysis is conducted using the R program. The variables with at least one category statistical significance were retained in the final model.

As discussed in the paper, the analysis results indicate that compared to single vehicle crashes, the crashes that involved with two or more than two vehicles have a higher probability to be fatal or injury crashes. It is shown that dark lighting and dusk or dawn, as compared to daylight, increase the probability of fatal or injury crashes. Males are found to be less likely to be involved in fatal or injury crashes. Drivers without driver license are more likely to be in more severe crashes. Drivers with driver license suspended, especially suspended due to DUI, have a larger chance to be in fatal or injury crashes. The probability to be in fatal or injury crashes is greater with a higher vehicle speed. Speeding also leads to more severe crashes. The restraint usage greatly impacts crash severity. No restraint usage apparently increases the probability of fatal or injury crashes. Compared with car driving, motorcycle driving even with helmets on are much more prone to severe crashes. The confusion matrix under the 10-fold cross validation method shows that the binary logistic regression model has a significantly high predictive power than a non-information guess.

Acknowledgement

The project received research funding support from the Institute for Multimodal Transportation (IMTrans) at Jackson State University. The IMTrans is member of the Maritime Transportation Research and Education Center (MarTREC) with the University of Arkansas (lead), Louisiana State University, and the University of New Orleans. MarTREC is one of the Tier I University Transportation Centers funded by the US DOT. Traffic engineers Christopher Kimbrell, Jim Willis, James Sullivan, and Wes Dean at the Mississippi Department of Transportation are thanked for providing data support to the study.

References

1. http://www-nrd.nhtsa.dot.gov/Pubs/812139.pdf. Accessed in June 2017.
2. US Department of Transportation (USDOT) (2017) Maritime Administration, US Waterborne Foreign Container Trade by US Custom Ports.
3. Shankar VN, FL Mannering, W Barfield (1995) Effect of roadway geometrics and environmental conditions on rural accident frequencies. Accid Anal Prev 27(3): 371-390.
4. Poch M, F Mannering (1996) Negative Binomial Analysis of Intersection-Accident Frequencies. Journal of Transportation Engineering 122(2): 105-113.

How to cite this article: Zhao M, Ningning W, Feng W. Statistical Analysis of Vehicle Crashes on Mississippi Coastal Two-Lane Highways.Civil Eng Res J. 2018; 3(5): 555622. DOI: 10.19080/CERJ.2018.03.555622

5. Wang Y, Ieda H, F Mannering (2003) Estimating Rear-End Accident Probabilities at Signalized Intersections: Occurrence-Mechanism Approach. Journal of Transportation Engineering l129(4): 377-384.

6. Savolainen PT, Tarko AP (2005) Safety impacts at intersections on curved segments. Transp Res Rec 1908: 130–140.

7. Shankar V, Milton J, Mannering F (1997) Modeling Accident Frequencies as Zero-Altered Probability Process: An Empirical Inquiry. Accident Analysis and Prevention 29(6): 829-837.

8. Shankar V, Mannering FL (1996) An exploratory multinomial logit analysis of single-vehicle motor cycle accident severity. J Saf Res 27(3): 183-194.

9. Pande A, M Abdel-Aty (2009) A novel approach for analyzing severe crash patterns on multilane highways. Accid Anal Prev 41(5): 985–994.

10. Dissanayake S, Lu J (2002) Analysis of severity of young driver crashes: sequential binary logistic regression modeling. Transportation Research Record 1784: 108-114.

11. Lui KJ, McGee D, Rhodes P, Pollock D (1988) An application of a conditional logistic regression to study the effects of safety belts, principal impact points, and car weights on drivers' fatalities. J Saf Res 19(4): 197–203.

12. Kusano K, Gabler HC (2014) Comparison and validation of injury risk classifiers for advanced automated crash notification systems. Traffic Injury prevention. pp 126-133.

13. Bishop YM, SE Fienberg, PW Holland (1975) Discrete multivariate analysis: Theory and practice, MIT Press, Cambridge, MA.

DOI: 10.19080/CERJ.2018.03.555622

How to cite this article: Zhao M, Ningning W, Feng W. Statistical Analysis of Vehicle Crashes on Mississippi Coastal Two-Lane Highways.Civil Eng Res J. 2018; 3(5): 555622. DOI: 10.19080/CERJ.2018.03.555622