Journal of Engineering Research and Reports

13(4): 1-9, 2020; Article no.JERR.57708
ISSN: 2582-2926

Analysis of Traffic Accident Severity at Intersection Using Logistic Regression Model

Yao Tzu Hsu[1], Shun Chi Chang[2*] and Tzu Hsin Hsu[1]

[1]Department of Transportation and Logistics, Feng Chia University, No. 100, Wenhua Rd., Xitun Dist., Taichung City 407, Taiwan.
[2]Ph.D. Program for Civil Engineering, Water Resources Engineering and Infrastructure Planning, Feng Chia University, No. 100, Wenhua Rd., Xitun Dist., Taichung City 407, Taiwan.

Authors' contributions

This work was carried out in collaboration among all authors. Author YTH was an advisor. Authors SCC and THH proposed the research framework of the study, managed the literature searches, carried out the logistic regression analysis, managed the analyses of the study, wrote the protocol and wrote the first draft of the manuscript. Author YTH viewed the manuscript. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/JERR/2020/v13i417106
Editor(s):
(1) Dr. Grzegorz Sierpiński, Silesian University of Technology, Poland.
Reviewers:
(1) K. T. Geetha, Avinashilingam University, India.
(2) Shuoben Bi, Nanjing University of Information Science and Technology, China.
Complete Peer review History: http://www.sdiarticle4.com/review-history/57708

Original Research Article

Received 02 April 2020
Accepted 09 June 2020
Published 22 June 2020

ABSTRACT

Accident severity analysis is an important issue in the field of traffic safety study, and intersections are also locations of relatively high accident rates in the roadway network. Therefore, the main purpose of this study is to establish a prediction model of intersection severity based on the binary logistic regression model of data mining technology. The data source of intersection accident is obtained from the Taichung City Police Department in Taiwan in 2018 and there are 27461 valid samples. The dependent variable is the severity of intersection accident. The independent variables include 9 variables such as month, time of accident, weather condition, light conditions, road type, road surface condition, traffic control type, accident type and vehicle type, and are analyzed by the forward selection (Wald). The research results show that time of accident, road surface condition, accident type and vehicle type have significant effects. The confusion matrix is used to verify the reliability of the model, and the results can be used as the references for reducing the degree of accident injury at the intersection in the future.

**Corresponding author: Email: charlie4045@gmail.com;*

Keywords: Road traffic accident; accident severity; binary logistic regression (BLR); at-grade intersection.

1. INTRODUCTION

Road traffic accidents have long been one of the urgent issues in the transportation field. In the road network system, due to the possible conflicts of various vehicles at the intersection, the movement of the vehicle is much more complicated than other locations, resulting in a higher accident rates at the intersection. According to the statistics of road traffic accidents published by the National Police Agency, Ministry of the Interior of Taiwan [1], in recent years, more than 50% of A1 and A2 accidents occurred at intersections, as shown in Table 1. It can be seen from Table 1 that the accident rate at intersections showed a slight upward trend from 2011 to 2015. Although it decreased slightly from 2016 to 2018, there were still 58.68% of accidents at intersections in 2018. In addition, according to the research report of the National Highway Traffic Safety Administration [2], 36% of traffic accidents in the United States are related to intersections. Tay and Rifaat [3] also pointed out that traffic accidents at intersections in Singapore accounted for about 35% of the total number of accidents, and serious injuries and deaths related to traffic accidents accounted for about 32%. It shows that intersection accidents in many regions of the world have become a serious social problem and deserve further investigation.

Khalili and Pakgohar [4] pointed out that in statistical methods, especially the regression models play an important role in identifying the key influencing factors of the severity of road traffic accidents, such as the logistic regression model. The general linear regression model is used to predict continuous variables. When the dependent variable is a nominal variable, the logistic regression model is used for analysis. In the field of transportation, the severity of accidents is usually divided into categories such as death, injury, and uninjured. Therefore, when conducting related research on the analysis of the severity of road traffic accidents, the logistic regression model is a very suitable analysis method.

The main purpose of this study is to analyze the severity of accidents at intersections through a binary logistic regression model and explore related influence factors. The remnant of the paper is then organized as follows: The literature review on issues such as intersection accident analysis, traffic accident analysis applying logistic regression model is discussed in section 2; section 3 is to describes the concept of binary logistic regression model, method of model performance evaluation, and the collecting of data; section 4 presents the results of binary logistic regression analysis, and the suggestions and conclusions of the study are presented in section 5.

2. LITERATURE REVIEW

Because intersections have a higher accident rate than other locations in the road network, in the field of road traffic safety, many studies have focused on issues related to intersection analysis to explore important factors that affect intersection accidents. Ertunc et al. [5] pointed out that to effectively reduce the occurrence of traffic accidents, it must rely on accurate data for research. In this study, Arcmap-10 software was used to create a database that used the data of fatal traffic accident at the intersection of Antalya city center from 2009 to 2010, and the accidents were statistically evaluated visually and graphically. Ahmed et al. [6] explored the factors that include road width, land use, lane markings, and traffic control affecting the safety of unsignalized intersections, and pointed out that for effective accident analysis, the attributes of the data set must be classified certainly. The research results showed that the most prone to accidents are minor road intersections with single lane markings and no signal control in non-urban areas. Fan [7] used the fault tree analysis to analyze the traffic accident database of Shanghai University of Engineering and Technology from 2008 to 2017, and discussed the causes of intersection accidents in depth. In the study, results of the fault tree analysis are used to find the relationship between the occurrence of traffic accidents at intersections and influence factors such as people, vehicle, road and environment. From these studies, we knew that traffic accidents are an important issue in the field of transportation, especially those related to intersections. The researchers tried to find out the relationship between the causes of traffic accidents and the impact factors.

Accident severity analysis is also a very important research topic in the field of road traffic safety, and related research results will help reduce accident injuries. Tay and Rifaat [3] discussed the impact of various accident-related factors on the severity of accidents at intersections by the ordered probit model. The research results showed that vehicle type, road type, collision type, driver characteristics and time are important factors that affect the severity of intersection accidents. Zhang et al. [8] uses the accident database of the US DOT-Fatality Analysis Reporting System (FARS) to analyze fatal accidents at intersections, and uses the ordered probit model to find out the factors that affect the severity of accidents. The results showed that the driver's age, vehicle type, and accident type are all significant influence factors. Asgarzadeh et al. [9] discussed the influence of intersection and street design variables on severity of bicycle-motor vehicle crashes. The results showed that pavement conditions, road characteristics, vehicle types, etc. are all important influence factors. George et al. [10] studied road traffic accident severity for various vehicles and discussed the impact of accident type and weather conditions on the severity of the accident. These studies revealed that time, vehicle type, and accident type are the key factors that affect the severity of road traffic accidents. Ditcharoen [11] mentioned that the deaths and injuries caused by road traffic accidents have attracted much attention around the world. The research reviewed the important factors that affect the severity of road traffic accidents, as well as the analysis methods commonly used in related research in the past, such as logistic regression, and summarized the results of past studies to found that speed, type of vehicle, whether drinking alcohol, and driving fatigue are all significant factors affecting the severity of an accident. Logistic regression has applicability in establishing a classification prediction model. Therefore, Ma et al. [12] have referred that this model is often used as an analysis method in transportation-related research in recent years. Xi et al. [13] used logistic regression to establish a prediction model for the severity of traffic accidents on curve. The model incorporates three aspects of driver characteristics, driving environment characteristics and road environment characteristics, with a total of 15 traffic accident attributes as independent variables. The research results showed that weather, roadside protection facilities and sidewalk structures are the most important factors for the severity of traffic accidents on the curve. Khalili and Pakgohar [4] pointed out that road defects are considered to be one of the important factors leading to accidents. In the study, logistic regression was used to explore the impact of road defects on the severity of accidents. The research results showed that insufficient road width and shoulder height difference are the most important factors. The results of these studies have shown that the logistic regression has good applicability in the analysis of road traffic accidents, but the accident characteristics of intersections have not been discussed.

In summary, there have been many issues related to the analysis of intersection accidents in the past research, such as the analysis of the cause of intersection accidents. However, there are little researches on the use of logistic regression to carry out analysis of intersection accidents. Stoltzfus [14] pointed out that logistic regression model is a powerful prediction tool, and more and more scholars apply it to the study of accident prediction. In the field of data mining, the logistic regression model is often used in the study of category prediction. Its main concept is to predict the possible events and probability based on some data attributes. It can be seen from the above discussion that some scholars have applied this model to the analysis of traffic accidents, and have achieved good results. Therefore, this study will attempt to analyze the severity of intersection accidents through binary logistic regression.

3. METHODOLOGY AND DATA COLLECTION

The main purpose of this study is to use the logistic regression model in data mining technology to investigate the factors that influence the severity of accidents at intersections, and to explore the effects of various factors to find out the more critical factors. The research framework was provided in Fig. 1. This section is to explain the concept of the binary logistic regression model and evaluation method of model reliability, and also describe the collection and screening of intersection accident data and the selection of independent variables.

3.1 Binary Logistic Regression

The logistic regression model is suitable for predicting the probability of binary variables. In the analysis of the severity of traffic accidents, the dependent variable is usually divided into some categories such as injured and uninjured or fatal and non-fatal, and the independent variables can include continuous variables and nominal variables in the model. Logistic regression is a special form of log-linear models [15]. Logistic distribution is the most widely used distribution function to analyze binary variable, and the logistic regression is applicability in the analysis of probability models. The binary logistic regression model is that set a binary variable as a dependent variable in the log-linear model, and then include a set of independent variables that affect the dependent variable according to the research goal, as shown in equation (1). The parameter estimation in the model is based on the method of maximum likelihood, which is different from the general traditional linear regression.

Table 1. Total road accidents and intersection accident numbers (A1+A2)

Year	2011	2012	2013	2014	2015	2016	2017	2018
Intersection Accidents (A1+A2)	140,319	149,070	166,872	185,789	184,073	183,574	175,849	187,968
Total Road Accidents (A1+A2)	235,776	249,465	278,388	308,742	305,413	305,556	296,826	320,315
Ratio of Intersection accidents	59.51%	59.76%	59.94%	60.18%	60.27%	60.08%	59.24%	58.68%

Notes:
1. A1 refers to a traffic accident that caused the death of a person on the spot or within 24 hours.
2. A2 refers to a traffic accident that caused injuries or killed more than 24 hours.

Data source: [1]

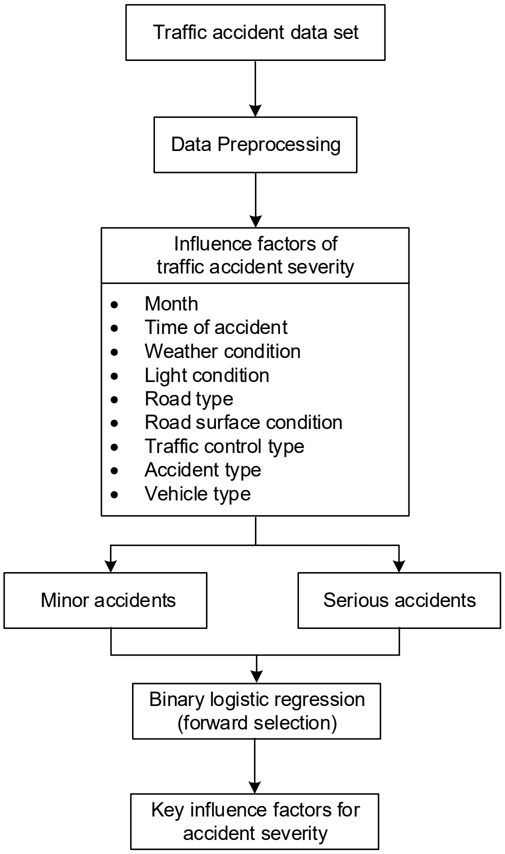

Fig. 1. Proposed framework for analysis

$$F = \log(P/(1-P)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \qquad (1)$$

Among them, P is the probability of occurrence of the event, β_0 is the intercept, $\beta_1\cdots\ \beta_n$ is the coefficient of each independent variable, and $x_1\cdots x_n$ is the independent variable. However, the key point to be discussed in the logistic regression analysis is the probability value P of the event occurrence. Therefore, the result must convert the odds value of the event into P, as shown in equation (2):

$$P = e^F/(1+e^F) \qquad (2)$$

3.2 Method of Model Performance Evaluation

In the classification method of machine learning, in order to test the quality of the estimation results of the classification model and the prediction effectiveness, the confusion matrix or classification error matrix is used to verify. The concept is to calculate the ratio of the predicted value to the actual value [16,17], to evaluate the predictive ability of the model, as shown in Table 2, the matrix is divided into four parts, C1, C2, C3, and C4. C1 means that the classification model correctly judges the intersection accident as a minor accident; C2 indicates that the intersection accident is a serious accident, but the classification model misjudged it as a minor accident; C3 indicates that the intersection accident is a minor accident, but the classification model misjudged it as a serious accident; C4 indicates that the classification model correctly judges the intersection accident as a serious accident.

In this study, after the establishment and analysis of the logistic regression model, the effectiveness of the model prediction results will be examined through the confusion matrix. The confusion matrix is to compare the predicted result with the actual data. All samples are divided into two categories: minor accidents and serious accidents. In the prediction of the logistic regression model, 0.5 is used as the cutting point. When the prediction probability is greater than 0.5, the sample is summarized as "serious accident"; when the predicted probability is less than 0.5, the sample is classified as "minor accident". The analysis of confusion matrix is to evaluate whether the prediction result is good by calculating three indicators. The three indicators are:

1. The accuracy of minor accidents predicted =C1/(C1+C2) (3)
2. The accuracy of minor accidents predicted=C4/(C3+C4) (4)
3. The overall accuracy = (C1+C4)/(C1+C2+C3+C4) (5)

The above three calculation indicators have a value range between 0% and 100%. The bigger the value is the better, which means that the model prediction is more accurate.

3.3 Data Collection and Variables Classification

The road traffic accident data used in this research is taken from Taichung City Government Open Data Platform [18]. The traffic accident data on the platform each year is compiled from the police agency 's accident investigation report. The accident data set records various types of road traffic accidents and related information about accidents, such as time, weather, and the severity of accident. This study screened out the intersection accident data in 2018 from the database for analysis, and then deleted the missing values and irrelevant road accident attributes. After pre-processing the data, a total of 38,660 valid samples will be analyzed. The dependent variable used in the study is the severity of the accident, and the independent variables include 9 accident-related attributes such as month, time period, weather, light conditions, road type, pavement status, sign type, accident type, and vehicle type. Because the dependent variable is a binary variable, the severity of the accident is divided into two categories: minor accidents and serious accidents. In this study, minor accidents were defined as no person or only one person was injured, serious accidents were defined as someone are killed or more than 2 people are injured. The classification content of each variable is shown in Table 3.

Table 2. Confusion matrix

		Actual class	
		Minor accident	**Serious accident**
Predicted class	Minor accident	C1	C2
	Serious accident	C3	C4

Table 3. Data categories and variables

Variable symbol	Variable name	Variable content
Dependent variable		
y	Accident severity	0= minor accidents 1= serious accidents
Independent variable		
x_1	Month	1= winter (12~2), 2=spring (3~5), 3= summer (6~8), 4= fall (9~11)
x_2	Time of accident	1=0~2, 2=2~4, 3=4~6, 4=6~8, 5=8~10 6=10~12, 7=12~14, 8=14~16, 9=16~18 10=18~20, 11=20~22, 12=22~24
x_3	Weather condition	1=sunny day, 2=cloudy day, 3=bad
x_4	Light condition	1=day light, 2=road light, 3=no light
x_5	Road type	1=three-way, 2=four-way, 3=multi-way
x_6	Road surface condition	1=dry, 2=bad
x_7	Traffic control type	1=traffic control signals (without pedestrian signals) 2=traffic control signals (with pedestrian signals) 3=flashing light signals 4=no signals
x_8	Accident type	1=vehicle-pedestrian, 2=vehicle-vehicle, 3=single vehicle
x_9	Vehicle Type	1=big-sized vehicle, 2=small-sized vehicle, 3= motorcyclist or cyclist, 4=others

4. RESULTS AND DISCUSSION

In this study, the binary logistic regression was used to establish a model for predicting the severity of intersection accidents. The dependent variable is the binary variable of the severity of the accident. The serious accident is set to 1 and the minor accident is set to 0; The model also includes 9 influence factors such as month, time of accident, weather condition, light condition, road type, road surface condition, traffic control type, accident type, and vehicle type as independent variables, and the variables are screened by the forward selection method. If the p-value of the variable is less than 0.05, and the variable is selected in the prediction model.

The F-test of the binary logistic regression model reaches a significant level, indicating that at least one of the independent variable coefficients in the model is not 0, and the regression model has predictive power. In the model, the stepwise forward regression method (Wald) is used to analyze the independent variables, and the variables that having a significant impact on the severity of the accident are included in the model one by one. There are 4 steps in total that select vehicle type (x_9) in step 1, select accident type (x_8) in step 2, select time of accident (x_2) in step 3, and select road surface condition (x_6) in step 4. Therefore, in the end, the model for predicting the severity of accidents at the intersection contains a total of four influence factors. The results are shown in Table 4.

When the accident occurred between 22:00 and 24:00, the probability of serious accidents is that 1.232 times of 06: 00 ~ 08: 00, 1.42 times of 08: 00 ~ 10: 00, 1.603 times of 10: 00 ~ 12: 00, 1.387 times of 12: 00 ~ 14: 00, 1.458 times of 14: 00 ~ 16: 00,1.348 times of 16: 00 ~ 18: 00, and 1.266 times of 18: 00 ~ 20: 00. The probability of serious accidents due to good road surface conditions is 1.46 times that of bad road surface conditions. It indicates that serious accidents are more likely to occur in the middle of the night and at the good road surface conditions. The possible reason is that the traffic flow is low in the middle of the night, and the driver's running speed is usually higher in the case of good road surface conditions. Traffic enforcement authorities should enhance police patrol and ban in the middle of the night. The probability of accident type of single vehicle being a serious accident is 0.077 times that of vehicle-pedestrian collision, and 0.122 times that of vehicle-vehicle collision. It shows that vehicle-pedestrian collision is more likely to cause serious accidents. Since pedestrians are the most vulnerable people on the road, the road authority should plan a perfect pedestrian environment, for example to ensure the sidewalks clear. The probability of serious accidents in vehicle type of others is 4.975 times that of big-sized vehicle and 5.525 times that of small-sized vehicle.

After the analysis of the logistic regression model, in order to understand whether the effectiveness of the constructed road traffic accident severity prediction model is good, a confusion matrix is used to verify the effectiveness of the prediction. The results of analysis are shown in Table 5, which is a cross

Table 4. Model estimation and odds ratio for independent variables

Parameter	Coefficient (B)	Wald	Exp(B)	p-value
Intercept	-2.519	133.068	.081	<0.001
Time of accident		92.333		<0.001
Time of accident (1)	.054	.179	1.056	.672
Time of accident (2)	.187	.975	1.206	.324
Time of accident (3)	.094	.345	1.098	.557
Time of accident (4)	-.208	7.365	.812	.007
Time of accident (5)	-.351	22.337	.704	<0.001
Time of accident (6)	-.471	37.182	.624	<0.001
Time of accident (7)	-.327	18.001	.721	<0.001
Time of accident (8)	-.376	23.395	.686	<0.001
Time of accident (9)	-.299	16.759	.742	<0.001
Time of accident (10)	-.235	9.582	.790	.002
Time of accident (11)	-.050	.381	.951	.537
Road surface condition(1)	.388	47.215	1.474	<0.001
Accident type		396.201		<0.001
Accident type(1)	2.559	331.318	12.922	<0.001
Accident type(2)	2.108	366.753	8.229	<0.001
Vehicle type		2822.003		<0.001
Vehicle type(1)	-1.604	45.768	.201	<0.001
Vehicle type(2)	-1.707	94.684	.181	<0.001
Vehicle type(3)	.039	.050	1.040	.823

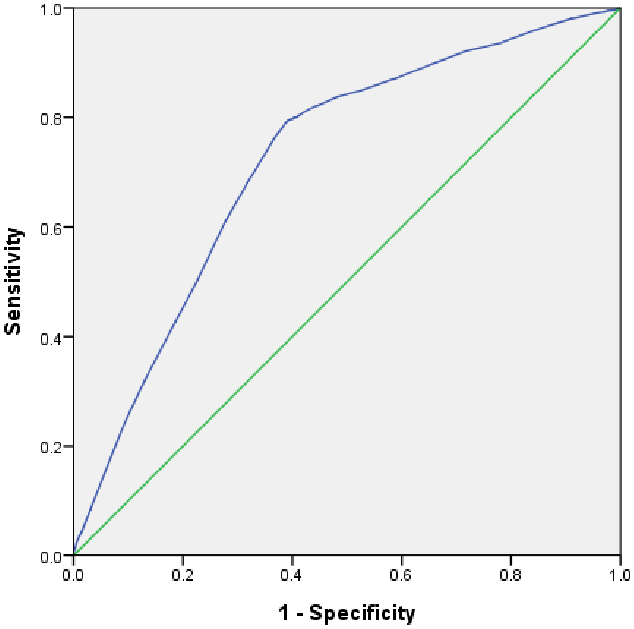

Fig. 2. ROC curve

Table 5. Results of confusion matrix analysis

		Actual class		Accuracy
		Minor accidents	**Serious accidents**	
Predicted class	Minor accidents	19333	6896	73.71%
	Serious accidents	608	624	50.65%
		Overall Accuracy		72.67%

Table 6. Area under the curve

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% confidence interval	
			Lower bound	**Upper bound**
.723	.003	<.001	.716	.729

table of 2 × 2 classifications of observations and predictions, showing that the accuracy of minor accidents predicted is 73.71%, the accuracy of serious accidents are correct predicted is 50.65%, and the overall accuracy is 72.67%. The examination results of the ROC curve are shown in Table 6 and Fig. 2. The area under the curve is 0.723, which is significantly greater than 0.5, indicating that the prediction effect is good. From the results of confusion matrix and ROC curve, it can see that the factors such as time of accident, road surface condition, accident type and vehicle type selected in this study are good for predicting the severity of accidents at intersections.

5. CONCLUSION AND SUGGESTION

When the dependent variable is a nominal variable, if researchers want to build a predictive model, the logistic regression model can play its role. In the past, many studies have applied this model to issues related to traffic accident analysis, and obtained fruitful results, which shows the suitability of the logistic regression model in the field of traffic safety. This study analyzes the intersection accident data of Taichung City in 2018 published by the Taichung City Police Department in Taiwan to discuss the factors that affect the severity of the intersection accident.

In this study, the severity of accidents is divided into two categories: Minor accidents and serious accidents, and the possible influence factors are considered that include month, time period, weather, light conditions, road type, pavement status, sign type, accident type, and vehicle type. This paper attempts to use the binary logistic regression to establish a model for predicting the severity of intersection accidents, and verifies the suitability of the model through the confusion matrix. The results show that time of accident, road surface condition, accident type and vehicle type have a significant impact on the severity of the accident, and the confusion matrix analysis results show that the prediction results of the model are good.

In the past, most of the accident severity analysis was based on probit regression or traditional statistical methods. This study verified the applicability of logistic regression in accident severity analysis, and explained the research results from the perspective of the odds ratio. In addition, most of the studies on the severity of accidents at intersections did not analyze the variables of intersection control types, such as signalizated intersection, unsignalizated intersection, and flash intersection. Although according to the results of this study, this variable has not significant, it is also the difference between this study and previous studies. The results can be used as the references for the development of improvement strategies to reduce intersection accidents. At the same time, future relevant research can also be extended based on this study. There are many influence factors related to the severity of road traffic accidents. How to choose more appropriate factors as model parameters and how to classify the nominal variables appropriately are topics that can be discussed in depth in future research.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Statistical Data of National Police Agency, Ministry of the Interior (NPAMI), Taiwan; 2019.
   (Accessed November 2019)
2. National Highway Traffic Safety Administration, NHTSA. Crash factors in intersection-related crashes: An on-scene perspective. Washington, DC; 2010.

3. Tay R, Rifaat SM. Factors contributing to the severity of intersection crashes. Journal of Advanced Transportation. 2007;41:245-265.
4. Khalili M, Pakgohar A. Logistic regression approach in road defects impact on accident severity. Journal of Emerging Technologies in Web Intelligence. 2013;5(2):132-135.
5. Ertunc E, Cay TS, Ömer M. Intersection road accident analysis using geographical information systems: Antalya (Turkey) example. 7th International Conference on Application of Information and Communication Technologies (AICT); 2013.
6. Ahmed A, Sadullah AFM, Yahya AS. Accident analysis using count data for unsignalized intersections in Malaysia. Procedia Engineering. 2014;77:45-52.
7. Fan F. Study on the cause of car accidents at intersections. Open Access Library Journal. 2018;5:1-11.
8. Zhang Y, Fu C, Cheng S. Exploring driver injury severity at intersection: An ordered probit analysis. Advances in Mechanical Engineering. 2014;1-11.
9. Asgarzadeh M, Verma S, Mekary RA, Courtney TK, Christiani DC. The role of intersection and street design on severity of bicycle-motor vehicle crashes. Injury Prevention. 2017;23:179–185.
10. George Y, Athanasios T, George P. Investigation of road accident severity per vehicle type. Transportation Research Procedia. 2017;25:2076–2083.
11. Ditcharoen A, Chhour B, Traikunwaranon T, Aphivongpanya N, Maneerat K, Ammarapala V. Road traffic accidents severity factors: A review paper. 5th International Conference on Business and Industrial Research (ICBIR). 2018;339-343.
12. Ma Z, Shao C, Yue H, Ma S. Analysis of the logistic model for accident severity on urban road environment. IEEE Intelligent Vehicles Symposium. 2009;983-987.
13. Xi JF, Liu HZ, Cheng W, Zhao ZH, Ding TQ. The model of severity prediction of traffic crash on the curve. Mathematical Problems in Engineering. 2014;2014:1-5.
14. Stoltzfus JC. Logistic regression: A brief primer. Academic Emergency Medicine. 2011;18:1099-1104.
15. Chan HC, Chang CC, Hung YJ. Establishment of predicting landslide susceptibility for Alisan forestry railway by logistic regression model. Journal of Soil and Water Conservation. 2012;44(4):421-436. Chinese
16. Hair JF, Anderson RE, Tatham RL, Black WC. Multivariate data analysis. 7th Ed., Macmillan, New York; 2016.
17. Han TC, Wang CM, Hung IH. Forecasting probability of tanker accidents using logistic regression model. Maritime Quarterly. 2017;26(4):103-119. Chinese
Available:https://www.npa.gov.tw/NPAGip/wSite/np?ctNode=12552&mp=1
18. Taichung City Government Open Data Platform, Taiwan; 2019.
(Accessed October 2019)
Available:https://opendata.taichung.gov.tw/

© 2020 Hsu et al.; This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Peer-review history:
The peer review history for this paper can be accessed here:
http://www.sdiarticle4.com/review-history/57708