Proceedings of the 2nd International Conference on Mathematical Physics and Computational Simulation
DOI: 10.54254/2753-8818/38/20240546

Research on the prediction of traffic accident by linear regression

Yuqing Wang

Shenzhen College of International Education, Shenzhen, 518000, China

s23767.wang@stu.scie.com.cn

Abstract. Traffic accident is getting increasingly serious. Although previous researchers use a variety of methods to predict the traffic accident, there are numerous demerits that need to be improved. This article demonstrates 12 variables that impact the traffic accident with 679 samples of accidents in UK from 2012 to 2014. This paper first analyses the relevance between dependent and independent variables, and also two independent variables to show the correlation between each factor. By using the multiple linear regression, it is concluded that although some independent variables do not have relationship with the dependent variable 'urban or rural area', Accident Severity, Number of Casualties, Road Type, Speed limit, Junction Control show significant relationship with the dependent variable. The paper also considers the 95% confidence interval in order to compare the effective density of data. Overall, the prediction of traffic accident is based on a number of factors and a sizable sample of accidents to summarize the impact that traffic accidents bring.

Keywords: Traffic accident, multiple linear regression, confidence interval.

1. Introduction

Traffic accident causes serious problems, leading to harm both humans' physical and mental well-being, and spending a substantial quantity of money on treating wounds [1]. According World Health Organization, the number of traffic accident deaths are 1.19 million in 2023, illustrating traffic accident is still a major cause of death [2]. In addition, traffic accident creates congestion, causing the transportation system to malfunction. Then the emission of waste gas intensifies, which will further ruin the air [3]. In general, whether it causes human casualties or environmental health, it will cause an inefficient use of social resources. In this way, the entire planet is working hard to contribute to decrease traffic accident. Therefore, this paper aims to help people recognize the methods of predicting the traffic accident in order to minimize the risk to humans and nature.

Analyzing the variables that influence the traffic accident can not only support medical personnel to plan and respond to the accidents, but also can provide detailed information to road decision maker in order to prevent future traffic accident [4]. The fork design of the road, driving habits, weather conditions and numerous other factors affect the traffic accident [5]. Early researchers study various model of prediction. For instance, Diderot et al. considered that intelligent transportation system (ITS) is a model that combine humans, vehicles and roads to improve the transportation efficiency, like reducing the amount of traffic congestion and traffic accidents by applying several sources of traffic data and communication system. However, the costs of related technologies are substantial, leading to

DOI: 10.54254/2753-8818/38/20240546

a few developed countries use this system [6]. Alqatawna et al. illustrated that Artificial Neural Networks (ANN) is a model of data manipulation, which is flexible to handle big data by using nonlinear relationships between variables. The ANN can effectively evaluate the traffic accident and predict its risks, but the model requires numerous input parameters and the output is challenging to interpret, leading to low degree of credibility [7]. Ye et al. claimed that a traffic accident prediction model called extended belief rule-based system (EBRBS) has the advantage of using professional understanding to enhance input and output data analysis. However, this model has low efficiency and rarely used in transportation industry engineering [8]. Additionally, Alkheder et al. showed three models of prediction, including decision tree, Bayesian Network and linear Support Vector Machine. Those models present the relationship between each factor that affect the traffic accident, while the outcome of Bayesian Network is more precise compared to others [9]. Last but not the least, Yang et al. stated that Geographical Information System (GIS) is used to analyze the spatial distribution of accidents by collecting the information of longitude, latitude and so on, but the data of maps usually update slowly [10]. Overall, the prediction models have their own merits and limitations.

In summary, the research on the models of traffic accident prediction is important for humans as it can decrease the injury rate by accidents. This article will analyze a prediction methodology and make recommendations for future research.

2. Methodology

2.1. Data source

The dataset used in this paper is fetched from the Kaggle website (1.6 million UK traffic accidents), which was collected from the UK government from 2012 to 2014. All the accident data comes from police report, which does not include minor incidents.

2.2. Variable selection

The data used in this paper include 679 accidents and contains 12 variables (Accident Severity, Number of Vehicles, Number of Casualties, Day of Week, Road Type, Speed limit, Junction Control, Pedestrian Crossing-Physical Facilities, Light Conditions, Weather Conditions, Road Surface Conditions, Urban or Rural Area), where the variable ‘urban and rural area’ is the dependent factor.

Table 1. Variables introduction

Variables	Logogram	Range	Mean	Variance
Accident Severity	x_1	1-3	2.89	0.10
Number of Vehicles	x_2	1-5	1.82	0.32
Number of Casualties	x_3	1-10	1.22	0.45
Day of Week	x_4	1-7	4.08	3.71
Road Type	x_5	1-4	2.03	0.19
Speed Limit	x_6	20-70	33.96	97.70
Junction Control	x_7	1-3	2.47	0.69
Pedestrian Crossing-Physical Facilities	x_8	1-3	1.35	0.38
Light Conditions	x_9	1-3	1.28	0.23
Weather Conditions	x_{10}	1-4	1.20	0.27
Road Surface Conditions	x_{11}	1-4	1.25	0.23
Urban or Rural Area	Y	1-2	1.22	0.17

*Accident Severity, Road Type, Junction Control, Pedestrian Crossing-Physical Facilities, Light Conditions, Weather Conditions, Road Surface Conditions, Urban or Rural Area: The higher the number is (from 1 to 4), the more likely the condition is to cause a traffic accident.

DOI: 10.54254/2753-8818/38/20240546

From Table 1, it demonstrates a simple overview of data. For instance, for the factor of light conditions, number 1 means the situation is in daylight, number 2 represents that street lights present and lit under darkness, and number 3 means that street lights do not present and lit under darkness, so the number 3 shows this light condition is more likely to cause traffic accident. Among the table, the data of accident severity has the smallest degree of dispersion due to the smallest value of variance 0.10, indicating the data of accident severity is stable. Additionally, the data of speed limit has the largest degree of dispersion, which means these data vary greatly.

2.3. Method introduction

The paper uses a linear regression model, which is a statistical technique for modeling relationship between the dependent variable and the independent variables, and the correlation between two independent variables. In simple dimension, linear regression can be visualized as a line. In higher dimension, linear regression can be visualized as hyperplane. Linear regression in the paper is used to analyze the correlation between variables and predict traffic accidents. In addition, model results are shown to demonstrate the usefulness of linear regression in analyzing the factor that impact traffic accident. The general equation for multiple linear regression is:

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{11} x_{11} \tag{1}$$

3. Results and discussion

3.1. Data processing

The paper shows how eleven variable that influence the traffic accident. As shown in Figure 1:

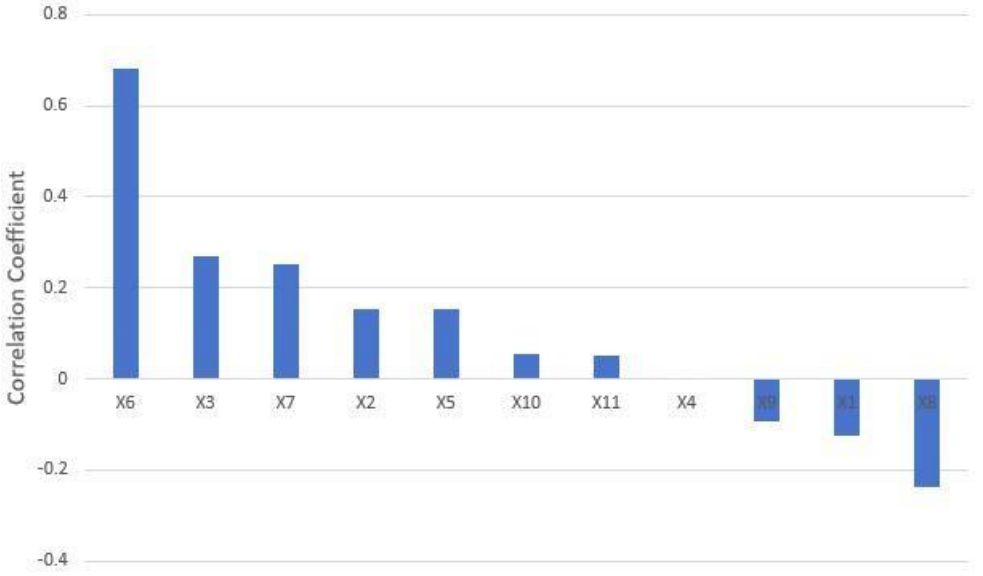

Figure 1. Correlation Between Dependent and Independent Variables.

From Figure 1, it shows that speed limit has the most positive relation to the factor of driving in urban or rural area, which means the driving speed can influence the happening of traffic accident especially in rural area. Besides, variable of Pedestrian Crossing-Physical Facilities demonstrates the most negative correlation with the dependent variable, so the crossing facilities are more important in urban area.

DOI: 10.54254/2753-8818/38/20240546

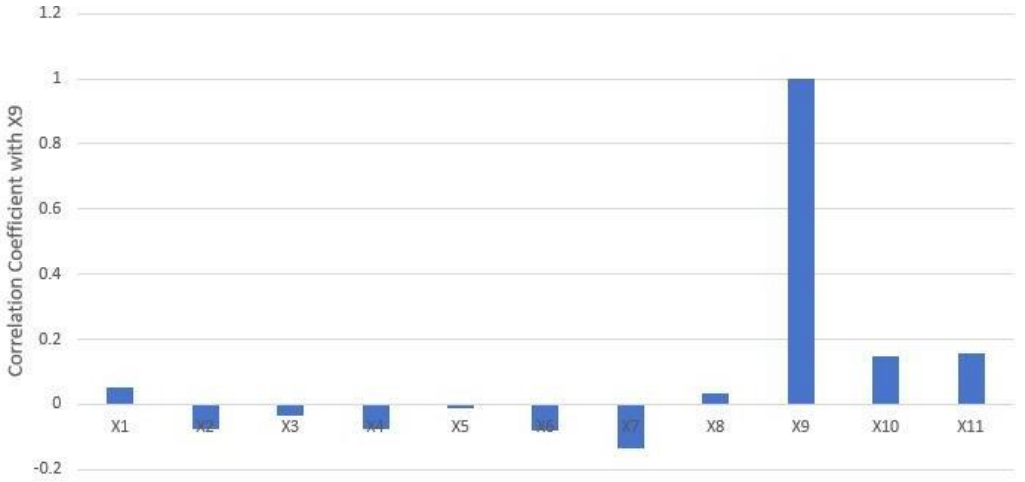

Figure 2. Correlation Between Independent Variables and x_9.

From Figure 2, it can be seen the correlation between independent variables and one variable of light conditions. The data shows that level of accident severity, pedestrian crossing-physical facilities, weather conditions and road surface conditions are positively correlate to the level of light conditions, while others show negative correlation with the level of light, which means traffic accident is more likely to happen in the condition of dim light with wet weather and on the street with the crowd. Overall, what affect the traffic accident are comprehensive since accidents occur in various conditions.

3.2. Model results

This paper analyses the linear regression of some factors. Let the variable 'urban or rural area' be the dependent variable, others factors be the independent variables (as shown in table 1). Through the analyzation of the statistic, the coefficient of determination R^2, which is used to measure the proportion of the variation in the dependent variable that can use the independent variables to judge the explanatory power of the model, has the value of 0.503, which means the model has a 50.3% fit, so those independent factors account for 50.3% of the variation in urban or rural area. Moreover, some significant data can be selected through the linear regression coefficient table:

Table 2. Linear Regression Coefficient Table.

	Beta	VIF	Tolerance	p
Constant	0.503	-	-	0.001
x_1	-0.145	1.018	0.983	0.000
x_2	0.017	1.093	0.915	0.407
x_3	0.039	1.152	0.868	0.033
x_4	0.006	1.013	0.987	0.295
x_5	0.076	1.083	0.923	0.005
x_6	0.026	1.225	0.816	0.000
x_7	0.033	1.292	0.774	0.032
x_8	-0.031	1.231	0.813	0.126
x_9	-0.022	1.067	0.973	0.366
x_{10}	-0.002	2.055	0.487	0.950
x_{11}	0.023	2.047	0.489	0.485

DOI: 10.54254/2753-8818/38/20240546

From Table 2 above, the values of beta are shown. All the values of variance inflation factor (VIF) are less than five, and all the values of tolerance are greater than 0.2. Besides, the Durbin-Watson statistic is near the number 1.5, which means there is no correlation in the model, so the data does not have collinearity problem, showing that the model has a good fit with the nature of stable and reliable. The p value for five independent variables x_1, x_3, x_5, x_6, x_7 is less than 0.05, so those factors have significance. In conclusion, variable x_1 has a significant negative impact on Y due to the negative value of beta, while variables x_3, x_5, x_6, x_7 have significant positive impact on Y due to the positive value of beta. Among them, the p value for x_1 and x_6 is 0.000, which are the smallest value, so the data of those two variables are the most significant. Considering the statistics above, the multiple linear regression equation can be written as:

$$E(Y) = 0.503 - 0.145x_1 + 0.017x_2 + \cdots + 0.023x_{11} \quad (2)$$

3.3. Model evaluation

Estimates of population parameters can be evaluated by using 95% confidence interval. The execution of the confidence interval is the degree to which the true value of the parameter has a certain probability surrounding the measurement result, which gives the degree of confidence in measured value of the measured parameter. The 95% confidence interval can be calculated in the formula:

$$\sum x/n \pm Z_{a/2}S/\sqrt{n} \quad (3)$$

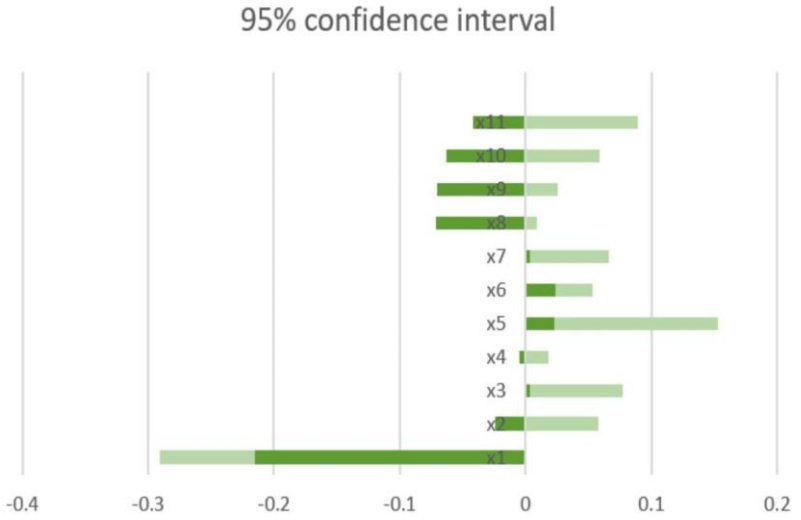

Figure 3. 95% Confidence Interval.

From Figure 3, factor Accident Severity has the largest confidence interval, indicating that it has the highest accuracy. Factor Speed limit has the smallest confidence interval, showing that it has the lowest accuracy.

4. Conclusion

The paper is based on 679 samples of traffic accidents from 2012 to 2014. The method of linear regression is employed to analyze the factors that impact traffic accident by showing the correlation coefficient of 12 variables.

During the use of linear regression, all the statistic are effective with no collinearity issues because of the suitable value for VIF and tolerance. It can be seen that four factors 'Number of Casualties, Road Type, Speed Limit, Junction Control' have significant positive influence on the variable of urban and rural area, but only one factor 'Accident Severity' have significant negative impact on the

DOI: 10.54254/2753-8818/38/20240546

dependent factor. What's more, the 95% confidence interval of the statistic illustrates the accuracy, which can determine the practicality and the significance of statistics.

By means of this research, traffic accidents can be prevented by keeping an eye on such variables when driving, resulting in fewer the injuries and deaths. Relevant departments can also big data and multiple linear regression to predict which location and conditions have the highest probability of traffic accidents. However, there are certain shortcomings, such as the amount of data is small and is not sufficiently recent, so some abnormal data will appear, affecting data fitting and analysis. In this way, further researchers can improve this and reinforce the method of the prediction of traffic accident.

References

[1] Alhaek F, Liang W, Rajeh T M, et al. 2024 Learning spatial patterns and temporal dependencies for traffic accident severity prediction: A deep learning approach. Knowledge-Based Systems, 286, 111406.

[2] World Health Organization. 2023 Global status report on road safety 2023. Retried from: https://www.who.int/publications/i/item/9789240086517.

[3] Chen J, Tao W, Jing Z, et al. 2024 Traffic accident duration prediction using multi-mode data and ensemble deep learning. Heliyon.

[4] Alhaek F, Liang W, Rajeh T M, et al. 2024 Learning spatial patterns and temporal dependencies for traffic accident severity prediction: A deep learning approach. Knowledge-Based Systems, 286, 111406.

[5] Mao X, Yuan C, Gan J, et al. 2019 Risk factors affecting traffic accidents at urban weaving sections: Evidence from China. International journal of environmental research and public health, 16(9), 1542.

[6] Diderot C D, Bernice N W A, Tchappi I, et al. 2023 Intelligent transportation systems in developing countries: Challenges and prospects. Procedia Computer Science, 224, 215-222.

[7] Alqatawna A, Álvarez A M R and García-Moreno S S C 2021 Comparison of multivariate regression models and artificial neural networks for prediction highway traffic accidents in spain: A case study. Transportation research procedia, 58, 277-284.

[8] Ye F F, Yang L H, Wang Y M, et al. 2023 A data-driven rule-based system for China's traffic accident prediction by considering the improvement of safety efficiency. Computers & Industrial Engineering, 176.

[9] Alkheder S, AlRukaibi F and Aiash A 2020 Risk analysis of traffic accidents' severities: An application of three data mining models. ISA transactions, 106, 213-220.

[10] Yang Y, Shao Z, Hu Y, et al. 2022 Geographical spatial analysis and risk prediction based on machine learning for maritime traffic accidents: A case study of Fujian sea area. Ocean Engineering, 266, 113106.