Analysis of the Harmfulness of Abnormal Riding Behaviors of Electric Bicycles Based on Improved Multiclass Logistic Regression Model

Yuzhe Qiu[1,a], Yingshun Liu[2,b]

[1]Department of Traffic and Transportation Engineering, Nanjing University of Science and Technology, Nanjing,Jiangsu,China
[2]Department of Traffic and Transportation Engineering, Nanjing University of Science and Technology, Nanjing,Jiangsu,China

Abstract: To analyze the harmfulness of abnormal riding behaviors of electric bicycles in-depth, the research focuses on the 2022 electric bicycle accident data in a specific city in China. Based on an improved multiclass logistic regression model, the relationship between different abnormal riding behaviors and the severity of electric bicycle traffic accidents is explored. Firstly, the severity of accidents is categorized into three levels as the dependent variable, while driver attributes and various hazardous driving behaviors serve as independent variables to construct the multiclass logistic regression model. Secondly, the model is optimized by eliminating irrelevant independent variables and improving the link function. Finally, the harmfulness of abnormal riding behaviors of electric bicycles is analyzed based on the results of the regression model. The results indicate that eight factors significantly influence the dependent variable, with three factors, including driving under the influence of alcohol, being more likely to lead to fatal accidents, requiring focused attention for intervention and regulation.

1. Introduction

1.1 Research background and literature review

Electric bicycles have positive implications for alleviating urban traffic pressure and facilitating citizen mobility. However, despite the steady decrease in road traffic accidents in our country, the number of traffic accidents involving electric bicycles is showing a contrary upward trend.[1] Therefore, strengthening the monitoring of abnormal riding behaviors of electric bicycles holds significant importance.

In the field of research in this domain, Ke [2] constructed a binary logistic regression model and, based on the results of the regression model, selected significant independent variables and influencing factors. They then carried out single-evidence variable and multiple-evidence variable coupled inference analysis using the TAN model, quantifying the magnitude of their impact and analyzing heterogeneity within the data. Wang [3] applied a binary logistic regression model to identify 12 significant factors from four aspects: electric bicycle rider information, motor vehicle information, road segment information, and accident information. Chai [4] utilized a multiclass logistic regression model and identified eight significant factors, including elderly riders. Li [5] employed a random forest model and found that accident type, injured body parts, and Physical barrier category were the three most important factors. Ting, Bryan [6] proposed a fast method for multivariate probit estimation using a two-stage composite likelihood approach, laying the foundation for extending beyond the pure binary setting. Zhao [7] proposed an abnormal driving behavior recognition algorithm based on spatio-temporal convolutional networks and soft thresholding. Experimental results showed that the model outperformed the state-of-the-art best baseline model by 2.24%. Castignani [8] introduced a mechanism for distracted driving detection using non-intrusive vehicle sensor data. The results demonstrated that a larger decision window led to higher performance.

The aforementioned studies have explored the factors influencing the severity of electric bicycle traffic accidents using statistical regression methods and data mining techniques. However, the current research has the following limitations: (1) The categorization of independent variables such as driver age and accident severity are not detailed enough. For example, driver age is only classified into two categories based on a threshold of 60 years, without considering age as a continuous variable. (2) The non-significant independent variables that do not contribute to explaining the dependent variable are not eliminated based on the goodness of fit and the significance p-values from likelihood ratio tests.

[a]122110011241@njust.edu.cn
[b]yingshun@njust.edu.cn

Therefore, this study will utilize an improved multiclass logistic regression model, taking the urban road electric bicycle traffic accident data from a specific city in 2022 as an example to analyze the harmfulness of different abnormal riding behaviors of electric bicycles. By employing a multiclass approach, the categorization of independent variables will be more detailed. Furthermore, by improving the model based on the significance of independent variables, the accuracy of the analysis results will be enhanced, aiming to provide theoretical support for improving the level of shared electric bicycle traffic safety management.

1.2 Data collection and preprocessing

This study collected traffic accident data from urban roads in a specific city in 2022, involving casualties in incidents related to electric bicycles. As this study focuses solely on analyzing traffic accidents caused by abnormal riding behaviors of electric bicycles, only traffic accidents resulting from abnormal riding behaviors of electric bicycles were included. Among these accidents, minor incidents accounted for 38.5% of the total, accidents causing injuries accounted for 43.9%, and fatal accidents accounted for 17.6%.

First, the data needs to be divided into dependent variables and independent variables. The dependent variable is set as the accident severity, with the values defined as follows: Y_1 for minor accidents, Y_2 for accidents causing injuries, and Y_3 for fatal accidents. Based on the data characteristics and referring to relevant previous studies [4], 10 fields are selected from the attributes of electric bicycle riders and characteristics of dangerous driving behaviors as potential independent variables (X) that may have a significant impact on the dependent variable. These independent variables are further categorized into continuous and discrete variables, and their descriptive statistics are presented in Table 1.

Table 1 Definition and Descriptive Statistics of Explanatory Variables

Factors	Variable Name	Variable Explanations and Values	Average Values
Drivers' Information	Gender	1 for male, 0 for female	0.645
	Age	Continuous variable ranging from 16 to 75 years	36.351
Dangerous Driving Behaviors	Speeding	1 if Speeding, 0 otherwise	0.244
	Not driving in designated lane	1 if not driving in designated lane, 0 otherwise	0.316
	Not maintaining safe distance	1 if not maintaining safe distance, 0 otherwise	0.079
	Running red lights	1 if running red lights, 0 otherwise	0.101
	Drunk driving	1 if drunk driving, 0 otherwise	0.028
	Fatigue driving	1 if fatigue driving, 0 otherwise	0.017
	Overloading	1 if overloading, 0 otherwise	0.088
	Driving in the opposite direction	1 if driving in the opposite direction, 0 otherwise	0.354

1.3 Construction of a multiclass logistic regression model

First, variable categorization needs to be performed. Unlike the traditional approach of dichotomizing age, this study treats age as a covariate and treats gender and accident factors as factors, which are then entered into the multinomial logistic regression model separately.

Next, an analysis of the significance p-values is conducted. In the context of model fitting, if a p-value is less than 0.05, it indicates that the model has at least one significant predictor variable. Conversely, when the model proves to be ineffective, it becomes necessary to eliminate independent variables that exhibit weak correlations with the dependent variable. Furthermore, in the parallel lines test, the significance of the conventional model is compared to a threshold of 0.05. If the significance is greater than 0.05, it indicates that the parameter estimates are reliable.

Finally, the appropriate link function is chosen based on the dependent variable. For a multinomial logistic regression model, the default link function is the logit function:

$$\mathrm{l}(\gamma) = \log(\gamma/1-\gamma) \tag{1}$$

In the given scenario where the probabilities of different severity levels of traffic accidents are not evenly distributed, the logit function may not be the most suitable link function. In order to improve the accuracy of the regression model analysis, the article will use the Negative log-log model and Probit model for simulation calculations. The Negative log-log model and Probit model are alternative link functions that can better accommodate the uneven probabilities of occurrence for different severity levels in this context.

Among them, the Negative log-log model is often applied in scenarios where the probabilities of lower-level categories are higher. Its expression is as follows:

$$l(\gamma) = -\log\big(-\log(1-\gamma)\big) \tag{2}$$

The Probit model is often applied in scenarios where the data follows a normal distribution. Its expression is as follows:

$$\phi^{-1}(\gamma) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{\gamma^2}{2}} \mathrm{d}\gamma \tag{3}$$

The analysis process of the improved multinomial logistic regression model is illustrated in the Figure 1.

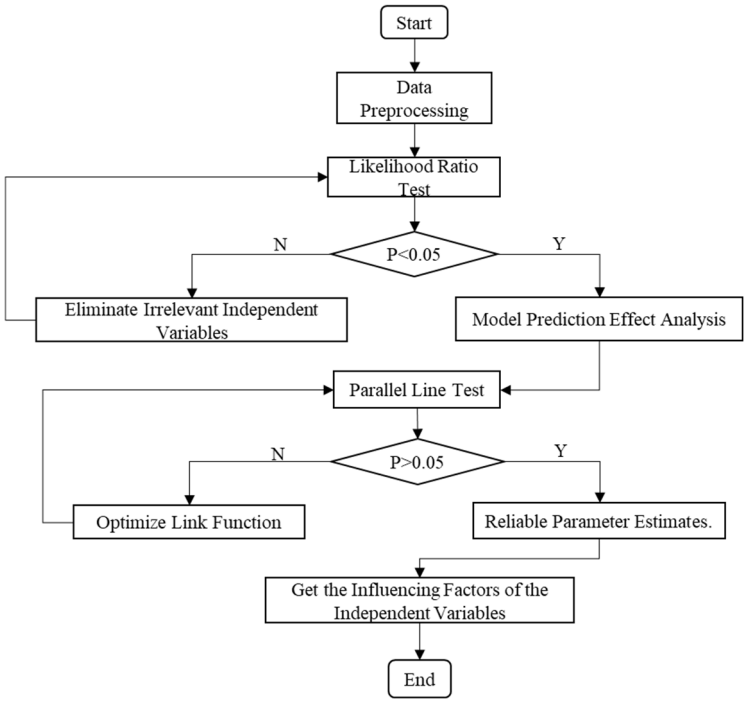

Figure 1 Flow chart of the improved multi-class logistic regression model

2.Conclusions

First, likelihood ratio test and parallel lines test were conducted on the model. The results of the likelihood ratio test are shown in Table 2

Table 2 Results of the likelihood ratio test

Parameters	Initial Likelihood Ratio Test			Optimized Likelihood Ratio Test		
	-2 Log-Likelihood	Chi-square	Significance	-2 Log-Likelihood	Chi-square	Significance
Final	1584.308	19.45	0.093	343.422	15.505	0.002
Age	1584.408	1.819	0.103	345.004	1.582	0.053
Gender	1586.127	0.101	0.951	——	——	——
Overloading	1585.309	1.001	0.106	344.209	0.787	0.075
Speeding	1585.153	0.845	0.055	344.772	1.35	0.009
Driving in the wrong direction	1586.372	2.064	0.056	345.964	2.541	0.028
Fatigue driving	1588.181	3.873	0.644	——	——	——
Drunk driving	1589.053	4.745	0.093	347.996	4.574	0.001
Failure to drive in designated lanes	1587.962	3.654	0.161	347.057	3.635	0.016
Running red lights	1584.844	0.536	0.065	344.034	0.612	0.074
Failure to maintain safe distance	1584.576	0.268	0.075	343.738	0.315	0.085

In Table 2, the initial likelihood ratio test shows a significance level (P) greater than 0.05, indicating a weak correlation between at least one predictor variable and the dependent variable. Further comparison of the significance of different predictor variables based on Table 2 reveals that gender and fatigue driving are not significant factors affecting the severity of accidents. After removing these factors, a subsequent likelihood ratio test and parallel lines test were conducted, and their results are shown in the right half of Table 2. In the likelihood ratio test, the significance level (P) is less than 0.05, indicating that at least one partial regression coefficient of a predictor variable is non-zero. This suggests that the model, which includes the aforementioned predictor variables, has a better goodness of fit compared to a model that only includes the intercept term. This test confirms that the final model is superior to a model containing only the intercept term.

The results of the parallel lines test are shown in Table 3. The null hypothesis of the parallel lines test assumes that the slope coefficients are equal in each respective category. By comparing the p-values of the parallel lines test, we can further assess the suitability of the logit, negative log-log, and probit link functions for this model. The results indicate that only the probit link function with a p-value of 0.065>0.05 passes the parallel lines test. This

suggests that the multinomial logistic regression model with the probit link function can accept the null hypothesis, indicating that the regression equations are parallel for each category.

Table 3 Results of the parallel lines test

Link function	-2 Log-Likelihood	Chi-square	Significance
logit	343.311	6.104	0.036
Negative log-log	1338.992	4.585	0.041
Probit	1351.518	5.919	0.065

From the above discussion, it can be concluded that the data in this study, after undergoing selection, is suitable for a multinomial logistic regression model. The chosen link function is the Probit function. Based on this, Table 4 presents the parameter estimation results of the regression model.

Table 4 Parameter estimation results based on the multinomial logistic regression model

Dependent variables [a]	Independent variables	B	Significance	Exp(B)	95% Confidence Interval for Exp(B)	
					Lower Limit	Upper Limit
Minor Accidents	Intercept	0.792	0.031			
	Age	-0.001	0.866	0.999	0.982	1.015
	Speeding	0.153	0.500	1.165	0.747	1.819
	Failure to maintain safe distance	0.125	0.001	1.134	0.603	2.131
	Drunk driving	-0.916	0.000	0.400	0.165	0.968
	Running red lights	-0.195	0.835	0.823	0.132	5.130
	Overloading	0.161	0.002	1.175	0.613	2.250
	Driving in the wrong direction	-0.253	0.179	0.776	0.537	1.123
	Failure to drive in designated lanes	0.371	0.002	1.450	0.977	2.152
Injury-causing accidents	Intercept	0.963	0.007			
	Age	-0.739	0.002	0.478	0.073	3.146
	Speeding	0.002	0.013	1.002	0.645	1.557
	Failure to maintain safe distance	0.001	0.998	1.001	0.535	1.874
	Drunk driving	-0.778	0.003	0.460	0.200	1.054
	Running red lights	-0.807	0.001	0.446	0.061	3.266
	Overloading	0.269	0.403	1.308	0.697	2.455
	Driving in the wrong direction	-0.293	0.002	0.746	0.520	1.071
	Failure to drive in designated lanes	0.207	0.299	1.230	0.832	1.817

"a": represents the reference category for severity, with "Fatal Accidents " as the reference category.

This study, through the optimized classification of accident severity and age, concludes the following: In contrast to fatal accidents, age is not a significant variable ($p>0.05$) affecting the propensity between minor accidents and fatal accidents. However, when comparing accidents causing injury to fatal accidents, the older age group is less inclined to be involved in accidents causing injury. Holding other factors constant, the probability of an injury-causing accident is 0.478 times that of a fatal accident.

Compared to the binary logistic regression model that dichotomizes age, the findings of this study are consistent with the conclusion that older individuals are more prone to fatal accidents [2]. Building upon this result, this method further analyzes the weak correlation between minor accidents and age. Among the predictor variables influencing the propensity towards accidents causing injury and fatal accidents, running red lights, driving under the influence, and driving in the wrong direction are three risky driving behaviors that are more likely to result in fatal accidents. Therefore, in the safety education for elderly electric bicycle drivers, these three factors should be emphasized as key points.

In summary, when age is treated as a continuous variable and incorporated into a multinomial logistic regression model, its influence may overlap or confound with other independent variables, leading to the failure of parallel lines test and a decrease in the model's explanatory power. However, by optimizing the link function, the model successfully passed the parallel lines test, circumventing this issue. Treating age as a continuous variable has the advantage of providing richer statistical information, allowing the model to capture the underlying trends and variations between age and

different levels of the dependent variable.

Based on the above analysis, the following conclusions can be drawn:

(1) Abnormal driving behaviors, including overloading and others, have a statistically significant impact on the severity of accidents. However, gender and fatigue driving do not show statistically meaningful effects. Drunk driving, running red lights, and driving in the wrong direction have a higher probability of causing fatal accidents compared to accidents causing injuries. This finding is consistent with the results of the Annual Report on Road Traffic Accidents [9]. Relevant authorities should prioritize regulatory measures targeting these dangerous driving behaviors.

(2) Compared to minor accidents and accidents causing injuries, accidents involving drunk driving are more likely to result in fatal accidents. The probability ratio of minor accidents, accidents causing injuries, and fatal accidents under the condition of drunk driving is 0.400:0.460:1. Therefore, it is necessary to strengthen awareness campaigns on the hazards of drunk driving for non-motorized vehicles, increase the frequency of alcohol testing for non-motorized vehicle users, and impose stricter penalties [10].

References

1. Huang, W., "Research on the Co-Governance Countermeasures of Shared Electric Bicycles," Urban Studies Journal, 2021, 42(6): 13-17. DOI: 10.3969/j.issn.2096-059X.2021.06.003.
2. Ke, X., Ding, L., Zhao, D., "Analysis of Factors Influencing the Severity of Electric Bicycle Traffic Accidents Based on Logistic-TAN," Journal of People's Public Security University of China (Natural Science Edition), 2023, 29(2): 47-54. DOI: 10.3969/j.issn.1007-1784.2023.02.007.
3. Wang, W., Shen, X., Wang, G., et al., "Analysis of Factors Influencing the Severity of Electric Bicycle Rider Accidents," China Safety Science Journal, 2019, 29(2): 20-25. DOI: 10.16265/j.cnki.issn1003-3033.2019.02.004.
4. Chai, H., Ma, B., Li, P., et al., "Multifactor Analysis of Severity in Motor Vehicle and Non-Motor Vehicle Accidents," Journal of Beijing University of Information Science and Technology (Natural Science Edition), 2022, 37(6): 38-45, 56. DOI: 10.16508/j.cnki.11-5866/n.2022.06.006.
5. Li, Y., Zhang, X., Wang, W., et al., "Analysis of Factors Influencing the Severity of Electric Bicycle Rider Accidents Based on Random Forest," Journal of Transportation Systems Engineering and Information Technology, 2021, 21(1): 196-200. DOI: 10.16097/j.cnki.1009-6744.2021.01.030.
6. Ting, Bryan Wright, Fred, ZHOU, Y.H., "Fast Multivariate Probit Estimation via a Two-Stage Composite Likelihood[J]. Statistics in Biosciences,2022,14(3):533-549. DOI:10.1007/s12561-022-09338-6.
7. Zhao, Y.Y., Jia, H.W., Luo, H.Y., et al., " An abnormal driving behavior recognition algorithm based on the temporal convolutional network and soft thresholding[J]. International journal of intelligent systems,2022,37(9):6244-6261. DOI:10.1002/int.22842.
8. Sasan J., German C., Thomas E., "Non-intrusive Distracted Driving Detection based on Driving Sensing Data[C]. //Proceedings of the 4th International Conference on Vehicle Technology and Intelligent Transport Systems: VEHITS 2018, Funchal, Madeira, Portugal, March 16-18, 2018.: Science and Technology Publications, Lda, 2018:178-186.
9. Wang, Z., Jiang, W., "Analysis of Illegal Behaviors of Electric Bicycles and Governance Strategies," Shandong Journal of Transportation Science and Technology, 2022(6): 8-12. DOI: 10.3969/j.issn.1673-8942.2022.06.003.
10. Wu, C., Zhang, J., Sun, W., "Research on Monitoring Technology for Traffic Violations of Electric Bicycles," Road Traffic Management, 2021(3): 25-27.