

Passenger Hailing Safety PASW Modeler and Big Data Statistical Analysis Study

S.H. Meng¹(✉), A.C. Huang², T.J. Huang³, J. Chen¹, and J.S. Pan¹

¹ College of Information Science and Engineering,
Fujian University of Technology, Fuzhou, Fujian 350118, China
menghui@fjut.edu.cn

² Department of Electrical Engineering,
National Sun Yat-Sen University, Kaohsiung, Taiwan

³ Guo-Guang Laboratory School,
National Sun Yat-Sen University, Kaohsiung, Taiwan

Abstract. This paper presents a study based on passenger hailing safety big data collection and PASW statistical analysis. A regression analysis model was used on data collected to study whether two or more variables were correlated. Changes in the direction and strength of correlation, and a regression analysis of arguments given estimates for the conditional expectation of the dependent variables, fully revealed the complex dependence. In addition, the RSS, which reflects the influence of random errors on dependent variables, measured the influence of the variance of factors other than passengers' hailing safety, data collection, and statistics analysis. In the linear regression analysis model, to improve traffic safety prediction and control, R² represents the contribution rate of analytic variables to a forecast change.

Keywords: Big data · Regression analysis model · Residuals Sum of Squares (RSS or SSE) · Statistical product and · Service Solutions (PASW or SPSS)

1 Introduction

Traffic safety for drivers, passengers, and pedestrians has caused widespread concern. According to a big data analysis in a traffic report released by the Traffic Police in Nanjing, China, more than 40 % of the traffic accidents resulted from drivers failing to focus or concentrate on driving. This resulted in casualties, most of whom were youth. According to relevant statistics, the annual number of deaths in traffic accidents worldwide reached approximately 600,000, and up to 12,000,000 people were injured in car accidents. Therefore, casualties and financial losses caused by car accidents exceed that of fire, flood, and other disasters combined [1]. Car accidents are known as 'the no. 1 public hazard of the civilized world'.

According to the statistics, most car accidents occurred when drivers could not stay constantly focused during driving, were fatigued, or were hailed by passengers. Thus, the drivers could not concentrate on their driving, lane departures caused by careless small movements, and did not notice nearby vehicles. All of these caused safety implications for drivers, passengers, and pedestrians, and especially for public

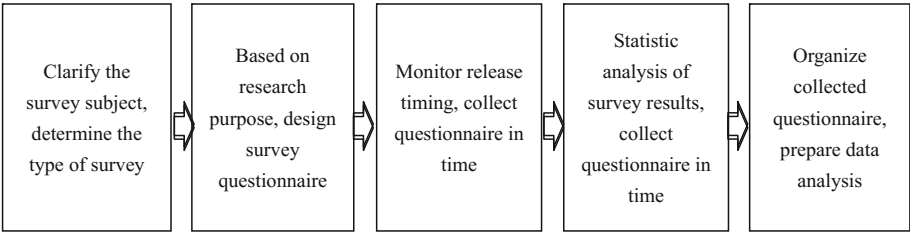
transportation drivers [2]. On the other hand, the study noted that many people hail cars anywhere along the road. Some even did so under dangerous conditions, such as in the driveways [3].

Based on a quantitative model analysis of passenger hailing safety, the safety of drivers, passengers, and pedestrians was studied by doing the following: taking safety as a dependent variable; using driving sight distance, hailing sight distance, and hailing behaviour as arguments; and building a linear regression model between dependent variables and arguments. To understand the influence of safety factors among them, the scope of research emphasized inductive analysis and surveys, intelligent data analysis, safety requirements for passengers and drivers, and core value mining. These data were obtained and converted into a professional value detection system [4].

2 Statistical Methods

To support a comprehensive passenger ride-hailing prompt system, we performed a large sample survey, data collection, storage, and data analysis, and curated the relevant data, is shown in Table 1 data statistics workflow.

Table 1. Data statistics workflow.



The first part deals with the basic information from the respondents, including gender, age range, and occupation. This part has multiple-choice questions to be answered by respondents.

The second part is a survey of the basic hailing behaviours of passengers, including hailing habits, attitudes towards hailing, and the demand for a smart driving assistance system. This part is to be answered by respondents based on their own personal riding experiences.

The third part is the basic driver behaviour survey, which includes driving habits, an understanding of the hailing methods of passengers, and their demands on a smart driving assistance system.

The fourth part takes the data collected through the design platform as impact factors, aggregates them into an information summary table, and establishes multiple variable linear regression analysis models using actual survey statistics [5–7]. Then, the survey results and complete data are analysed to obtain survey results with higher authenticity and reliability.

3 Data Entry: List of Samples Included in Data

- (1) Respondent's gender: males account for 61.94 % of the total survey, females account for 38.06 %.
- (2) Age distribution of respondents: in April 2014, the China Nanjing Traffic Police released a big-data analysis of traffic reports, of which the casualties were mainly youth. Based on relevant statistics, this study collected extensive samples in the range of 21–25 years old, 26–30 years old, and 31–40 years old as the study intervals.
- (3) The driver has certain range of sight distance while driving. This has a certain impact on traffic safety. The driver not only has to constantly pay attention to the road conditions ahead, but also pay attention to whether there is a pedestrian ahead or if a passenger hails a ride. In addition, there are obstacles. Thus, the driving sight distance becomes very important. This is important for safety among drivers, passengers, and pedestrians. Therefore, statistics for driving sight distance are a very important consideration in traffic safety studies. According to statistics, 46.27 % of the crowd has a driving sight distance in sunny weather of up to 100 m, 47.76 % of the crowd has a sight distance up to 50 m, and only 5.9 % of people have a sight distance of approximately 20 m.
- (4) Low visibility on rainy days, in addition to windshields covered with rain, are equivalent to the driver wearing a pair of dark sunglasses. Therefore, the statistics for driving sight distance on rainy days is very important to traffic safety studies. According to statistics, only 5.22 % of the crowd has a driving sight distance of up to 100 m on rainy days, 45.52 % of the crowd has a sight distance up to 50 m, and 49.25 % has a sight distance of approximately 20 m. For drivers who have to pay attention to traffic conditions and traffic safety, but are also distracted by looking for passengers hailing from the street, this increases the safety implications for drivers.

Passengers certainly expect to get a ride as soon as possible, and hope the drivers can drive them in shortest possible time. This can guarantee a car's loading rate, and meet the needs of prospective passengers. According to the statistical analysis, approximately 41.04 % of the crowd can notice people waving in the front, 13.43 % of the crowd does not notice people waving in the front, and the rest of the crowd (45.52 %) is to be determined depending on specific distance.

- (5) On the other hand, in the statistics for passengers' hailing behaviours (by a survey of whether passengers hailed a taxi in the middle of the road, or at other unsafe locations), a total of 14.93 % of people often hailed a taxi at unsafe locations, 28.36 % would do it sometimes, while 19.04 % depended on the situation. Only 37.31 % of the people stated they would not have behaved in this manner under normal conditions.

Based on the statistics, the study considered hailing vehicles at a safe place as safe hailing behaviour. Depending on situations with unsafe hailing behaviour, we performed a probability estimation of China Taiwan, Mainland China, Japan, and the USA. This estimation uses the specific circumstances of population proportions, after

being converted by probability. We obtained the information shown in the figure below. Statistics show that based on unsafe hailing behaviour estimated from population risk statistics, Mainland China's unsafe population exceeds nine billion. Thus, Mainland China is ranked first among the four countries listed.

4 Model Building and Analyses

Through the sample analysis, the study classified collected data into two categories (sight distances on sunny days and rainy days). Sample data (such as hailing vehicles at some unsafe locations, and so on) were included in a safety analysis of a passenger hailing prompt system, while sample data collected from hailing methods and hailing failures were classified in a passenger study using a multivariate linear regression mathematics model [5–7]. First, the study quantified the collected data and defined the variables. The study described the quantification of the variables with an example of the problems in the questionnaire. For example: Will you hail a taxi in the middle of road or at other unsafe places? A: Sometimes, B: Often, C: Never, D: Depends.

Then the study set the variables as follows, using 1, 2, 3, 4 instead of A, B, C, D as answers: 'Sometimes' as 1, 'Often' as 2, 'Never' as 3, and 'Depends' as 4. For the variables in multiple choice problems, the study used a multiple dichotomy method. The fundamental idea is to set each option in a problem as a variable, and separate each option into two options (select the option, or not select the option). For example, a multiple-choice question has three options A, B, and C. Select is 1, and unselect is 0. Using the quantification table information, the study could perform a Goodness of Fit Analysis, regression equation significance test, and a regression coefficient significance test to obtain the corresponding parameters. These were followed by a data analysis, summary, and conclusion.

5 PASW Safety Model Analysis

The dependent variable is safety. The variables are driving sight distance, hailing sight distance, and hailing behavior, is shown in Table 2.

Table 2. Variables^a included/excluded.

Model	Variables included	Method
1	Hailing behavior, Driving sight distance, Hailing sight distance ^b	Input

a. Dependent variables: (safety)

b. All required variables included

Using Explained Sum of Squares define coefficient of multiple determination (CMD) in total square's ratio:

$$R^2 = \frac{U}{SST} \tag{1}$$

$R = \sqrt{R^2}$, it was called coefficient of multiple determination (CMD). The related relationships are more closely like Y and Independent variable x^1, \dots, x^m , if R is more bigger. Ordinarily, it was considered to be related to the establishment of the relationship when $R > 0.8$.

Linear regression: the goodness of fit test is to determine the coefficient R2. The greater the value, the better the fit. By observing the adjusted coefficient of 0.840, the goodness of fit is high, is shown in Table 3. (The closer to 1, the more accurate the regression equation coefficients and parameters, and the better the fit for the regression. A goodness of fit above 0.8 is generally considered high.)

Table 3. Goodness of fit analysis.^b

Model	R	R ²	Adjusted R ²	Estimated Standard Error
1	0.919 ^a	0.844	0.840	1.542159864858501

a. Estimated variables: (constant)

b. Dependent variable: safety

The ‘regression sum of squares’ shows the explanatory parts of the variance of response variables by arguments contained in the regression model. The ‘residual sum of squares’ represents the variance of response variables that was not explained by the variables contained in the regression model. These two values are associated with the sample size and the number of arguments in the model. The larger the sample size, the greater the corresponding variance. df is the degree of freedom, which is the number of free value variables. F indicates the F test statistics, which are used to test the significance of the regression equation. Since the P value of the significance test of the regression equation is 0.000 (smaller than the significance level of 0.05), the linear relationship is pretty good, is shown in Tables 4 and 5. The significance test is passed. This tested the overall hypothesis and indicated that the sample data deduction of an actual population, and the overall null hypothesis, were significant and reasonable.

The points in the P-P diagram surround a line, which shows the residuals approximately obey a normal distribution. At this point, the regression model passed various tests and achieved a good fit. Therefore, a regression model (with safety as a dependent variable and driving sight distance, haling sight distance, and hailing behaviour as variables) to improve the safety of drivers and passengers was established

Table 4. Regression equation significance test.^a

Model	Sum of Squares	dF	Mean Squares	F	Significance
Regression	1672.360	3	557.453	234.396	0.000
Residuals	309.173	130	2.378		
Sum	1981.534	133			

Table 5. Regression coefficients significance test.^a

Model	Nonstandardized coefficients		Standard coefficient	t	Sig.
	B	Standard error			
(Constants)	1.333	0.495		2.691	0.008
Driving sight distance	0.110	0.005	0.785	22.060	0.000
Hailing sight distance	0.094	0.011	0.312	8.717	0.000
Hailing behavior	0.268	0.125	0.077	2.151	0.033

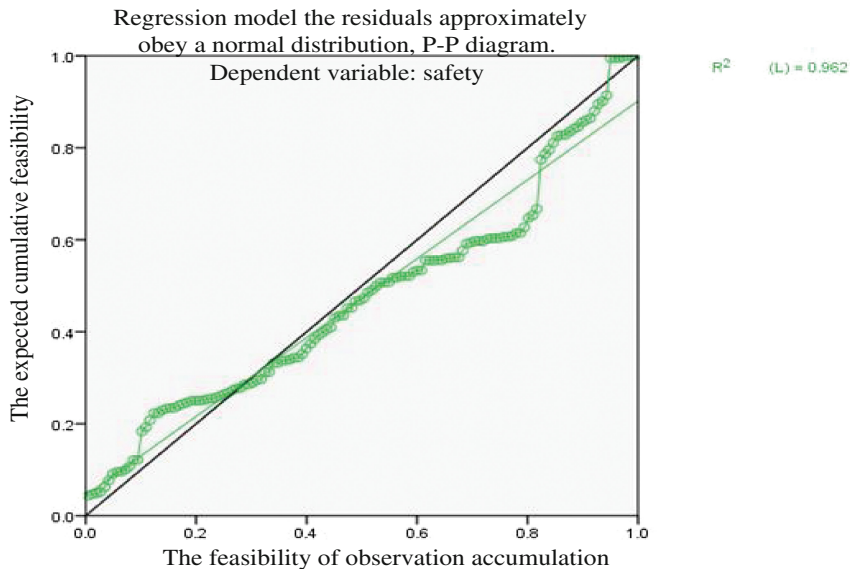


Fig. 1. Safety as dependent variable.

after analysis with $R^2 = 0.962$ shown in Fig. 1. This approaches 1 and with a diagnosis of residual normality.

Most of the points in the diagram gather along the line with a slope of 1 (ideal normal distribution line), which shows that the regression model passed the residual normal distribution test. This implies that the greater the overlap between the residual distribution curve and the normal distribution line, the higher the consistency between the two distributions. The established regression model has a positive significance.

6 Conclusions

Currently, no taxi has a prompt system that detects whether a passenger is hailing from the front. Instead, drivers are relied on to autonomously notice the presence of passengers who want a ride. Thus, drivers while driving must divide their attention to

predict whether someone wants to hail the vehicle. This makes the drivers prone to traffic accidents, and can make drivers miss prospective passengers, thus reducing the loading rate.

This paper began with respondents' basic information, and used it to investigate people's driving and hailing behaviours, safety awareness, and demand analysis. Using statistical data from a population survey, a data analysis was used as basis to design hailing data. This was done for the safety of drivers and passengers, to obtain statistics that sensed distances, and to better ensure people's safety.

References

1. Yang, Z.: A Study of Measurement model for Economic Losses in Traffic Accidents. Shandong University of Science and Technology (2005)
2. Han, Y.: Pedestrian Detection System of the Advanced Driving Assistance Systems. Xidian University (2014)
3. Xu, H.: A Pedestrian Recognition Algorithm of a Visual Sensor Type in Vehicle Collision Prevention System. Guangdong University of Technology (2012)
4. Meng, S.H., Huang, A.C., Huang, T.J.: Passengers Hailing Reminding Method and Apparatus. Taiwan Patent (2016)
5. Tian, B.: Multiple linear regression analysis and its practical application. Yinshan Acad. J. (Natural Science Edition) (2011)
6. Cao, F.: Multivariate linear model parameter estimation. Wuhan University of Science and Technology (2006)
7. Wang, H., Meng, J.: Multiple linear regression analysis. J. Beijing Univ. Aeronaut. Astronaut. (2007)

Intelligent Data Analysis and Applications

Proceedings of the Third Euro-China Conference on

Intelligent Data Analysis and Applications, ECC 2016

Pan, J.-S.; Snášel, V.; Sung, T.-W.; Wang, X.D. (Eds.)

2017, XIX, 288 p. 157 illus., Softcover

ISBN: 978-3-319-48498-3