PARAMETER
JURNAL MATEMATIKA, STATISTIKA DAN TERAPANNYA

e-ISSN. 2829-6303
Vol. 04 No. 1, April 2025
Page 63-72

BINARY LOGISTIC REGRESSION ANALYSIS OF TRAFFIC ACCIDENT RISK FACTORS IN INDONESIA

Seila Amalia[1], Amelia Putri[2], Michael Dolly Sianturi[3] , Risca Octaviani Hutapea[4] , Albert Servant Ndruru[5] , Arnita[6]

[1,2,3,4,5,6]Program Studi Statistika, Universitas Negeri Medan
J. William Iskandar Ps. V St, Deli Serdang, 20221, Sumatera Utara, Indonesia

E-mail Correspondence Author: *seilaamalia@gmail.com*

Abstract

This study analyzes the factors influencing the severity of traffic accidents in Indonesia using binary logistic regression. Data from Kaggle includes variables such as age, gender, driving experience, lighting conditions, and weather conditions. The results indicate that poor lighting and adverse weather significantly increase the likelihood of fatal accidents by 88.8% and 96.5%, respectively. The logistic regression model achieves 76% accuracy with a good data fit (Hosmer-Lemeshow p-value = 0.144). These findings provide valuable insights for traffic safety policies and infrastructure development.

Keywords: *Binary Logistic Regression, Risk Factors, Statistical Analysis, Traffic Accidents*

: https://doi.org/10.30598/parameterv4i1pp63-72

1. INTRODUCTION

Traffic accidents are among the leading causes of preventable deaths, both in developed and developing countries, including Indonesia [1]. The high incidence of traffic accidents in Indonesia is influenced by various factors such as driver behavior, road conditions, and insufficient safety facilities. Research by Lestari and Anjarsari revealed that in Banjarbaru City, certain segments of Ahmad Yani Street are accident-prone areas (blackspots), mainly due to low awareness of road safety and a lack of supporting infrastructure [2].

In a different location, a study by Juwita and Maharani on the Negeri Sakti-Bernung road segment in Lampung identified that accidents frequently occur on damaged roads and sharp curves, with the highest accident rates recorded during peak hours [3]. Another study by Syahriza also highlighted that human factors are the dominant cause of traffic accidents in Indonesia, significantly impacting public health and the economy. Collaborative efforts encompassing road safety policies, public education, and infrastructure improvements are essential to support the Sustainable Development Goals (SDGs) target of reducing traffic accidents by 50% by 2030 [1].

Traffic accidents remain a significant issue in various regions of Indonesia, including Sorong Regency, Padangsidimpuan City, and Buleleng Regency. In Sorong, the primary causes include human error, such as driving under the influence, and poor road infrastructure, such as potholes and inadequate lighting. Preventive and repressive measures have been implemented to address this issue. In Sorong, public education and law enforcement against traffic violators are the main strategies to reduce accidents [4]. In Padangsidimpuan, both technical and non-technical approaches have been applied to mitigate accident-prone areas through infrastructure improvements and traffic regulation [5].

According to WHO data, traffic accidents are among the leading causes of death worldwide, with high incidence rates, especially in developing countries. Binary logistic regression analysis is a commonly used method to identify risk factors influencing accident status. For instance, in the context of students in Jabodetabek, factors such as age, gender, mode of transportation, and activity location significantly determine accident status [6]. Similarly, in Riau Province, the logit model has been employed to reveal the influence of the victim's vehicle, opposing vehicle, and accident location on victim severity levels.

Statistical approaches like binary logistic regression have been widely used to analyze risk factors for traffic accidents. This method effectively identifies relationships between predictor variables and binary response variables, such as accident status (occurred or not occurred). For example, research [7] demonstrated that binary logistic regression can identify significant factors influencing divorce phenomena, such as age and education. Similarly, the logit model has been applied to evaluate the impact of the victim's vehicle, opposing vehicle, and accident location on victim severity. This approach was also utilized by [8] to model factors affecting students' academic achievements. Through logit transformation, binary logistic regression enables the identification of significant patterns, making it a crucial tool in supporting evidence-based interventions to reduce accident rates.

2. RESEARCH METHOD

2.1. Types of Research

This study is a quantitative research employing a descriptive and inferential approach. The data used in this research is sourced from a publicly available dataset on the Kaggle platform, titled "Road Traffic Accidents," provided at https://www.kaggle.com/datasets/saurabhshahane/road-traffic-accidents. This dataset contains information related to traffic accidents involving 50 respondents, which is relevant for analysis to identify patterns, trends, or factors influencing accident incidents.

The utilization of this secondary dataset aims to provide insights based on structured data, thereby supporting statistical analysis and drawing conclusions based on empirical evidence. The research process involves data cleaning, data exploration, statistical analysis, and visualization. The sample used in this study consists of 50 respondents.

2.2. Research Variables

The following variables are used in this study.

Table 1. Variable Description

Variables	Description	Category
Y	Accident Severity	1 = Fatal 2 = Serious Injury 3 = Minor Injury
X1	Driver Age	1 = <18 Years 2 = 18-30 Years 3 = 31-50 Years 4 = Unknown
X2	Driver Gender	1 = Male 2 = Female 3 = Unknown
X3	Driving Experience	1 = <1 Year 2 = 1-2 Years 3 = 2-5 Years 4 = 5-10 Years 5 = >10 Years 6 = Unknown 7 = No license
X4	Lighting Condition	1 = Daylight 2 = Darkness -lights 3 = Darkness - no 4 = Darkness - no lights on
X5	Weather Conditions	1 = Normal 2 = Rain

2.3. Research Procedure

The procedures for data analysis in this study are as follows:

1. Understanding the characteristics of traffic accident data in Indonesia.
2. Examining the independence between predictor and response variables.
3. Utilizing multinomial logistic regression through the following steps:
   a. Estimating parameters to assess the influence of predictor variables.
   b. Forming the logit function for each risk level category.
   c. Conducting a simultaneous test to evaluate the overall effect of predictor variables.
   d. Performing partial tests to determine the specific influence of each predictor.
4. Identifying the strength of influence for each predictor.
5. Evaluating the model with actual data and predictions:
   a. Goodness-of-Fit Test: This test ensures that the regression model fits the actual data. Tests include the Hosmer-Lemeshow test to examine the alignment between predicted results and actual data.
   b. Wald Test: This test is used to examine the significance of each individual predictor variable. Each variable is tested to determine whether its coefficient significantly influences the response variable.
   c. Likelihood Ratio Test (LRT): This test evaluates the overall significance of the model to determine whether all predictor variables collectively influence the response variable.
6. Measuring the prediction accuracy of the model:
   a. True Positive (TP): Cases where the model correctly classifies the risk level according to actual data.
   b. True Negative (TN): Cases where the model correctly identifies negative cases.
   c. False Positive (FP) and False Negative (FN): Misclassifications measured to identify areas for model improvement.

3. RESULTS AND DISCUSSION

3.1 Fit Test Model (Hosmer and Lemeshow Test)

In the study by [9], it is explained that the Hosmer-Lemeshow Test is used to assess the fit of a logistic regression model by dividing the data into several groups based on predicted probabilities. Each group is then compared with the observed and expected occurrences. The validity of this test is determined by the resulting p-value; if the p-value is greater than 0.05, there is insufficient evidence to reject the null hypothesis, indicating that the model is considered to fit the data. The study also highlights the importance of selecting the appropriate number of groups, as the number of groups can influence the test results and the interpretation of model fit.

Table 2. Model Fit Test

Contingency Table for Hosmer and Lemeshow Test

		Accident severity = ,00		Accident severity = 1,00		
		Observed	Expected	Observed	Expected	Total
Step 1	1	10	10.497	1	.503	11
	2	9	10.153	2	.847	11
	3	12	12.707	2	1.293	14
	4	13	11.433	0	1.567	13
	5	9	8.459	1	1.541	10
	6	9	7.410	0	1.590	9
	7	11	10.436	2	2.564	13
	8	6	9.207	6	2.793	12
	9	9	7.796	2	3.204	11
	10	4	3.901	5	5.099	9

Table 3. Uji Hosmer and Lemeshow Test

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	12.175	8	.144

Based on the results of the Hosmer and Lemeshow Test, a chi-square value of 12.175 with a significance of 0.144 > 0.005 was obtained. Thus, the null hypothesis (Ho) is accepted (model FIT). This indicates that the binary logistic regression model is suitable for further analysis, as there is no significant difference between the predicted probabilities and the observed probabilities.

3.2 Parameter Significance Test

1) Coefficient of Determination Test
   In the study by **[10]**, it is explained that the Nagelkerke R Square is used to measure how well the independent variables can explain the dependent variable in a logistic regression model. The Nagelkerke R Square value ranges from 0 to 1, where a higher value indicates that the model has a better ability to explain data variation..

Table 4. Parameter Significance Test

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	95.746[a]	.107	.173

   This model has a moderate ability to explain data variation, with a Nagelkerke R Square of 0.173, indicating that the model explains approximately 17.3% of the variation in the dependent variable. This suggests that while the model can be used, there may be other factors not accounted for in the model that influence the outcome.
2) Uji F
   The F Test (Simultaneous Test), also known as the Overall Model Test or ANOVA Test, is used to assess the combined influence of all independent variables on the dependent variable. This test can be conducted by comparing the calculated F-value with the F-table value. If the calculated F-value > F-table, the hypothesis is accepted, meaning the model is significant. Conversely, if the calculated F-value < F-table, the hypothesis is rejected, indicating that the model is not significant **[11].**

Table 7. Omnibus Tests of Model Coefficients Table.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	12.765	5	.026
	Block	12.765	5	.026
	Model	12.765	5	.026

   With a p-value of 0.026 < 0.05, we reject the null hypothesis, which states that the model coefficients are equal to zero. This means that at least one predictor variable in the model significantly influences the dependent variable. The model used has statistically significant goodness-of-fit, making it suitable for further analysis.
3) Uji T
4) In the study by **[12]**, it is explained that the t-test is used to evaluate the significance of regression coefficients in a logistic regression model. This test aims to determine whether each independent variable significantly affects the dependent variable. The t-test results

are obtained by comparing the regression coefficient value with the standard error of that coefficient. If the p-value from the t-test is less than 0.05, it can be concluded that the independent variable significantly influences the dependent variable. The study also demonstrates that several tested independent variables have significant p-values, indicating that they statistically contribute to the constructed logistic regression model.

Table 6. T Test

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Age band of driver	-.606	.346	3.076	1	.079	.545
	Sex of driver	-.229	.486	.221	1	.638	.796
	Driving experience	-.234	.196	1.424	1	.233	.792
	Light conditions	.636	.346	3.370	1	.066	1.888
	weather conditions	.675	.960	.495	1	.482	1.965
	Constant	-.573	1.614	.126	1	.722	.564

The results of the test of the influence of independent variables on dependent variables are as follows:

a) The sig. value of the Age band of driver variable is 0.079 > 0.05, then Ho is accepted, which means that the Age band of driver variable does not have a significant effect on the severity of the accident.
b) The sig. value of the Sex of driver variable is 0.638 > 0.05, then Ho is accepted, which means that the Sex of driver variable does not have a significant effect on the severity of the accident.
c) The sig. value of the Driving experience variable is 0.233 > 0.05, then Ho is accepted, which means that the Driving experience variable does not have a significant effect on the severity of the accident.
d) The sig. value of the Physical condition of the treatment room is 0.893 > 0.05, then Ho is accepted, which means that the Physical condition of the treatment room does not have a significant effect on Patient Satisfaction.
e) The sig. value of the access variable in service is 0.138 > 0.05, then Ho is accepted, which means that the access variable in service does not have a significant effect on Patient Satisfaction.
f) The sig. value. The Light conditions variable is 0.066 > 0.05, so Ho is accepted, which means that the Light conditions variable does not have a significant effect on the severity of the accident.

3.3 Binary Logistic Regression Modeling and Odds Ratio

In the study by **[13]**, it is explained that binary logistic regression is an analytical method used to model the relationship between a binary response variable and one or more predictor variables. This model produces the odds of a specific event, expressed in the form of an odds ratio. The odds ratio describes the comparison of event odds between two groups and can be interpreted as a measure of the strength of the relationship between the independent and dependent variables. The study shows that each unit increase in a particular predictor variable can increase the odds of the desired event, measured by the odds ratio. For example, if the odds ratio for a predictor variable is 1.5, this means that an increase of one unit in that variable increases the odds of the event by 50%. The study emphasizes the importance of understanding the odds ratio in the context of logistic regression analysis to gain deeper insights into the relationships between variables.

Table 8. Binary Logistic Regression Modeling and Odds Ratio

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Age band of driver	-.606	.346	3.076	1	.079	.545
	Sex of driver	-.229	.486	.221	1	.638	.796
	Driving experience	-.234	.196	1.424	1	.233	.792
	Light conditions	.636	.346	3.370	1	.066	1.888
	weather conditions	.675	.960	.495	1	.482	1.965
	Constant	-.573	1.614	.126	1	.722	.564

Interpretation of the effect of independent variables on the dependent variable based on Exp(B) values and odds ratios:

a. Age band of driver has an Exp(B) value of 0.545. This means that for each one-unit increase in the driver's age group, the odds of accident severity decrease by 45.5%. With an odds ratio less than 1, this variable indicates that an increase in the driver's age group tends to reduce the risk of accident severity.
b. Sex of driver has an Exp(B) value of 0.796. This shows that male drivers have 20.4% lower odds of accident severity compared to female drivers, but since this variable is not statistically significant, this result should be interpreted with caution.
c. Driving experience has an Exp(B) value of 0.792. This means that each one-unit increase in driving experience reduces the odds of accident severity by 20.8%. However, since this variable is not significant, its effect may not be substantial in predicting accident severity.
d. Light conditions has an Exp(B) value of 1.888. This indicates that poor lighting conditions increase the odds of accident severity by 88.8%. Although this variable approaches statistical significance with a p-value of 0.066, it suggests that poor lighting conditions could be an important risk factor.
e. Weather conditions has an Exp(B) value of 1.965. This means that poor weather conditions increase the odds of accident severity by 96.5%. However, since this variable is not statistically significant (p-value 0.482), the result should be considered with caution.

3.4 Classification Accuracy

The classification table shows the model's ability to correctly classify data into the dependent variable categories and appears in the logistic regression output as the Classification Table, which provides the prediction accuracy percentage for each category, as well as the overall model accuracy.

Table 9. Classification Accuracy

Classification Table[a]

			Predicted		
			Accident severity		Percentage Correct
	Observed		.00	1.00	
Step 1	Accident severity	.00	92	0	100.0
		1.00	17	4	19.0
	Overall Percentage				85.0

The Classification Table shows that the logistic regression model has an overall accuracy of 85%, with excellent predictive ability for the "not severe" (0) category, where 100% of cases are correctly classified. However, the model shows poor performance in predicting the "severe" (1) category, with an accuracy of only 19%, as 17 out of 21 cases that should be classified as severe are instead predicted as not severe. This indicates that the model is more likely to predict the majority category, likely due to data imbalance between the "not severe" and "severe" categories.

4. CONCLUSION

Based on the logistic regression analysis conducted on traffic accident data in Indonesia, it was found that the model has an overall accuracy of 85%. The model is able to predict the "not severe" (0) accident category with an accuracy rate of 100%, but it has weaknesses in predicting the "severe" (1) accident category, with an accuracy of only 19%. This indicates that the model is more effective in identifying minor accident cases but less sensitive to severe accident cases, likely due to the data imbalance between the two categories. Therefore, adjustments to the model or an approach that considers the data distribution are necessary to improve the prediction accuracy for severe accidents.

REFERENCES

[1] M. Syahriza, "Kecelakaan Lalulintas: Perlukah Mendapatkan Perhatian Khusus?," *AVERROUS: Jurnal Kedokteran dan Kesehatan Malikussaleh*, pp. 89-101, 2019.

[2] U. S. Lestari, Y. dan R. Adawiyah, "Analisis Kecelakaan Lalu Lintas Dan Penanganan Daerah Rawan Kecelakaan Jalan Ahmad Yani (Ruas KM 37 – KM 82) Kabupaten BanjaR," *JURNAL GRADASI TEKNIK SIPIL*, vol. 6(2), pp. 102-117, 2022.

[3] F. Juwita and F. Maharani, "Metode Accident Rate Dalam Analisis Kecelakaan Lalu Lintas," *Prosiding Seminar Nasional Penelitian dan Pengabdian kepada Masyarakat*, vol. 2(1), pp. 1-9, 2021.

[4] S. F. E. Mubalus, "Analisis Faktor-Faktor Penyebab Kecelakaan Lalu Lintas Di Kabupaten Sorong Dan Penanggulangannya," *SOSCIED*, vol. 6(1), 2023.

[5] R. F. Siregar, N. Paisah and A. Pakpahan, "Analisis Kecelakaan Lalu Lintas (Black Site) Pada Ruas Jalan H.T Rizal Nurdin Kota Padangsidimpuan," *STATIKA*, vol. 5(1), pp. 14-30, 2022.

[6] A. L. Dewi dan Budyanra, "Determinants of Accident Status on Student Commuters of Jabodetabek in 2019," *Jurnal Matematika, Statistika, dan Komputasi*, vol. 18, pp. 102-120, 2021.

[7] T. M. T. Nisva and V. Ratnasari, "Analisis Regresi Logistik Biner pada Faktor-Faktor yang Mempengaruhi Jenis Perceraian di Kabupaten Lumajang," *Jurnal INFERENSI*, vol. 3(1), pp. 2721-3862, 2020.

[8] Y. A. Tampil, H. Komalig and Y. Langi, "Analisis Regresi Logistik Untuk Menentukan Faktor-Faktor Yang Mempengaruhi Indeks Prestasi Kumulatif (IPK) Mahasiswa FMIPA," *JdC*, vol. 6(2), 2017.

[9] A. Henzi, . M. Puke, . T. Dimitriadis and J. Ziegel, "A safe Hosmer-Lemeshow test," *Journal of Statistics in Data Science*, Vols. , 0(1), pp. 1-15, 2023.

[10] H. Hafid, A. S. Ahmar and Z. Rais, "Analisis Pengaruh Profitabilitas dan Ukuran Perusahaan terhadap Audit Delay dengan Menggunakan Regresi Logistik," *Journal of Statistics and Its Application on Teaching and Research*, vol. 5(1), pp. 15-22, 2023.

[11] D. Surjadmodjo and H. Cangara, "Pengaruh Variabel Terhadap Kinerja Usaha: Uji F Simultan dalam Analisis Regresi," *Jurnal BADATI*, vol. 6(1), no. 1, pp. 1-14, 2024.

[12] D. R. Pratiwi and R. Sari, "Analisis Pengaruh Faktor-Faktor Ekonomi terhadap Keputusan Investasi Menggunakan Regresi Logistik," *Jurnal Ilmu Ekonomi*, vol. 12(2), pp. 101-110, 2023.

[13] R. Susetyoko, "Pemodelan Regresi Logistik Biner dan Odds Ratio dalam Analisis Data," *Jurnal Teknologi Informasi-AITI*, Vols. 56-66, p. 15(1), 2023.

[14] W. Sanjaya, "Analisis kepuasan pasien terhadap kualitas pelayanan kesehatan di UPTD Puskesmas Limusnunggal Kota Sukabumi," *Jurnal Ilmu Kesehatan Bhati Husada*, vol. 14(2), pp. 215-225, 2023.

[15] D. S. Sari, "Estimating The Condition Of Traffic Accident Victims By Using Ordinal Logistic RegressiON," *(Doctoral dissertation, Universitas Islam Negeri Sultan Syarif Kasim RiaU)*, 2022.

[16] N. Srimaneekarn, A. Hayter, W. Liu and C. Tantipoj, "Binary Response Analysis Using Logistic Regression in Dentistry," *International Journal of Dentistry*, 2022.