*Traffic Accident Research Centre*
*Journal of Indonesia Road Safety - Vol. 2, No. 3, December 2019, p. 149-158*

# MODELING THE PROBABILITY OF SPEEDING BEHAVIOUR AND ACCIDENT INVOLVEMENT USING BINARY LOGISTIC REGRESSION IN EAST JAVA PROVINCE

## *Model Probabilitas Perilaku Speeding dan Keterlibatan Kecelakaan Menggunakan Regresi Logistik Biner di Provinsi Jawa Timur*

**Willy Kriswardhana**
Department of Civil Engineering
University of Jember
Jl. Kalimantan 37 Jember
East Java
willy.teknik@unej.ac.id

**Sonya Sulistyono**
Department of Civil Engineering
University of Jember
Jl. Kalimantan 37 Jember
East Java
sonya.sulistyono@unej.ac.id

**Iin Ervina**
Psychology Study Program
Univ. Muhammadiyah Jember
Jl. Karimata Jember
East Java
iinervina@unmuhjember.ac.id

**Dadang Supriyanto**
Department of Civil Engineering
State University of Surabaya
Jl. Ketintang, Gayungan
Surabaya, East Java
dadang.supriyatno@gmail.com

**Nunung Nuring Hayati**
Urban Planning Study Program
University of Jember
Jl. Kalimantan 37 Jember
East Java
nunung.nuring@unej.ac.id

**Achmad Wicaksono**
Department of Civil Engineering
Brawijaya University
Jl. MT. Haryono No. 147
Malang, East Java
wicaksono68@ub.ac.id

**Sri Wiwoho Mujanarko**
Civil Engineering Study Program
Narotama University
Jl. Arif Rahmad Hakim No. 51
Surabaya, East Java
sri.wiwoho@narotama.ac.id

**Raka Aminul Ramadhani**
Department of Civil Engineering
University of Jember
Jl. Kalimantan 37 Jember
East Java

**Abstract**

Driving at high speed has negative consequences, namely, the high number of accidents. Several factors have been considered as causes of the increasing severity of victims of traffic accidents, such as a human, vehicle, and environmental factors. The risky driving behavior factor is a factor that needs to be considered in traffic safety studies. This study aims to determine the probability model of speeding behavior based on several driver characteristics and their relationship to accident involvement. This study used a binary logistic regression method to determine the probability of driving behavior exceeding the speed limit and accident involvement. The results showed that the younger a person is, the higher the probability of breaking the maximum speed limit. Furthermore, driving experience also shows a similar trend, where the longer the driving experience of someone, the less likely it is to be involved in an accident. Directions for further research are also presented.

**Keywords:** road, accident, probability, road safety, speeding

**Abstrak**

Berkendara dengan kecepatan tinggi mempunyai konsekuensi negatif, yaitu tingginya angka kecelakaan. Beberapa faktor telah dipertimbangkan sebagai penyebab dari peningkatan tingkat keparahan korban kecelakaan lalulintas. Faktor tersebut seperti faktor manusia, kendaraan, dan lingkungan. Faktor perilaku berkendara yang berbahaya, menjadi faktor yang perlu diperhatikan dalam kajian keselamatan lalulintas. Penelitian ini bertujuan untuk mengetahui model probabilitas pada perilaku speeding berdasarkan beberapa karakteristik pengendara, serta hubungannya dengan keterlibatan kecelakaan. Penelitian ini menggunakan metode regresi logistik biner untuk mengetahui probabilitas perilaku berkendara melebihi batas kecepatan

dan keterlibatan kecelakaan. Hasil penelitian menunjukkan bahwa semakin muda usia seseorang, maka semakin tinggi probabilitasnya dalam melanggar batas kecepatan maksimum. Lebih lanjut diperlihatkan bahwa pengalaman mengemudi juga menunjukkan tren yang serupa. Pengalaman mengemudi seseroang, yang lebih lama akan memperkecil kemungkinan dalam keterlibatan kecelakaan. Arahan untuk penelitian selanjutnya juga ditampilkan.

**Kata kunci:** jalan, kecelakaan, probabilitas, keselamatan jalan, speeding

# INTRODUCTION

The traffic accident that caused death has reached alarming levels. Since 2009, the number of victims killed in traffic accidents in the world has not shown a decline. In 2016, 1.35 million people died on the road. The World Health Organization (WHO) predicts this trend will triple by 2030 (World Health Organization (WHO), 2015). It shows that the progress in the implementation target of Sustainable Development Goals (SDGs) in 2020, which requires a 50% reduction in the number of victims of traffic accidents died, becomes quite difficult.

High speed has positive and negative effects on road users. Driving at high speed directly causes reduced travel time, increases mobility, and is good for road capacity. However, Lai, Carsten, and Tate (2012) stated that driving at high speeds has negative consequences, namely the high number of accidents on urban roads and the increase in CO2 emissions. Furthermore, Wells (2011) states that some motorists claim they are reliable, safe, and do not have negative records. Motorists assume that the behaviour of increasing speed does not affect involvement in traffic accidents.

Several factors have been considered as causes of the increase in the severity of traffic accident victims, such as a human, vehicle, and environmental factors (Goniewicz et al., 2016). The dangerous driving behaviour factor becomes a factor that needs to be considered in the study of traffic safety. In many cities in Southeast Asia, pedestrians, motorbike riders, car riders, and slow vehicles such as pedicabs, use the road in chaotic ways (Kitamura, Hayashi and Yagi, 2018). Besides, the rapid growth of motorized vehicles, the limitations of safety features for protection, and the lack of a safe road environment also have an important role in increasing the level of traffic accidents (Widyastuti, 2012). However, studies on the management of motorist behaviour, especially in the case of speed selection behaviour in developing countries such as Indonesia, are still not comprehensive, compared to studies in developed countries. This study aims to determine the probability model of speeding behaviour based on several characteristics of the driver and its relationship with accident involvement.

# LITERATURE REVIEW

## Logistic Regression

Uncles, Ben-Akiva, and Lerman (2006) state that individual behaviour can be described by discrete variables that can be modelled based on discrete choice models. Logistic regression is one of the analysis techniques in discrete choice modelling. Logistic regression is used as a prediction of the probability of an event that has a logarithmic data function. Logistic regression will produce opportunity ratios associated with predictor variable values.

There are several logistic regressions, namely binary, multinomial, and ordinal logistic regression. Binary logistic regression is usually used for data that has binomial response variables. Multinomial regression is using for data analysis that is an unordered category. Based on bivariate data (X, Y) where X is a numeric variable or one-zero variable, and Y is a one-zero response variable, the logistic regression model has the following general form:

$$Pn(i) = \frac{1}{1+exp-\beta \text{in}} \quad (1)$$

and

$$Pn(j) = \frac{exp-\beta \text{in}}{1+exp-\beta \text{in}} \quad (2)$$

The application of logistics models based on certain data, including bivariate data, aims to estimate or estimate the magnitude of the proportion of Y = 1 in the population concerned. Regarding univariate regression models in general, logistic regression models can also be written in the following form:

$$ln\frac{p}{1-p} = \beta_0 + \beta_1 X \quad (3)$$

Backward elimination is used as an alternative to the model specifications. As in many disciplines, the null-hypothesis significance test is commonly used. The test produces statistical and probability test results (p-value). The null hypothesis is rejected when the p-value is less than or equal to 0.05 or 0.01 in social science (Nickerson, 2000). Theoretically, the p-value is a measure of continuous evidence, but in practice, usually trichotomized will be very significant, slightly significant, and not statistically significant at the conventional level, with limits on $p \leq 0.01$, $p \leq 0.05$ and $p > 0.10$ (Gelman and Stern, 2006). Most statistical handbooks present the rule of thumb for the significance levels of 0.1, 0.05, and 0.01 (Figueiredo Filho et al., 2013).

**The Goodness of Fit Test**

The goodness of fit test was performed using the Hosmer-Lemeshow (H-L Test) statistical test. This test aims to study the suitability of the logistic regression model. The basic principle of this statistical test is the frequency of the predicted results, and the frequency of observations of the dependent variable must have relatively small differences. The smaller the difference, the more feasible the model. Models that are feasible according to this statistical test will have a large probability value (p-value), which is greater than the 5% confidence level or $\alpha = 0.05$ (Coburn, 2009). The formula of the Hosmer-Lemeshow test is:

$$C\hat{} = \sum_{k=1}^{g} \frac{(O_k - E_k)^2}{V_k} \quad (4)$$

with: C ^ = Hosmer-Lemeshow Test (H-L test), Ok = Observation value in k-group, Ek = Expectation value in k-group, Vk = Variance correction factor for k-group.

# RESEARCH METHODOLOGY

## Respondent dan Questionnaire

Respondents in this study were people of East Java Province. The preparation of the research variables was developed from a research questionnaire conducted by the Indonesia National Police Traffic Corps in 2019. The questionnaires distributed by the online system with the Google form tool and assisted by *SATLANTAS* in the area of East Java.

## Model Development

The dependent variable in this study is the selection of speed and knowledge of the maximum speed limit. While the independent variables are gender, age, ownership of a driving license, and last education.

Coding on dependent and independent variables used statistical assistance programs. Speed violations and accident involvement were coded in binary form, whereas knowledge of speed limits was coded in ordinal form. The use of ordinal type variables was done to describe the level of knowledge of respondents regarding the permissible speed limits in several types of road sections. The types of roads used are urban 2/2UD, inter-city 2/2UD, urban 2/1, inter-city 2/1, toll roads, and roads in an orderly traffic area.

**Table 1**. Variables

| No. | Variable | Category and Coding |
|---|---|---|
| 1. | Speed violation | 0 = not violating the speed limit |
| | | 1 = violating the speed limit |
| 2. | Accident envolvement | 0 = never |
| | | 1 = ever |
| 3. | Speed limit knowledge | 1 = 1 correct answer |
| | | 2 = 2 correct answers |
| | | 3 = 3 correct answers |
| | | 4 = 4 correct answers |
| | | 5 = 5 correct answers |
| | | 6 = 6 correct answers |
| 4. | Age | Open question |
| 5. | Gender | 1 = Female |
| | | 2 = Male |
| 6. | Driving license ownership | 1 = Yes |
| | | 2 = No |
| 7. | Last education | 1 = Primary school |
| | | 2 = Junior High School |
| | | 3 = Senior High School |
| | | 4 = Diploma |

| No. | Variable | Category and Coding |
|---|---|---|
| | | 5 = Undergraduate |
| | | 6 = Master |
| | | 7 = Doctorate/PhD |
| 8. | Driving experience | Open question |

Analysis of the speed violation probability model requires the goodness of fit test. The model goodness of fit test uses the Hosmer & Lemeshow Test (H-L test) and classification plot. Table 2 shows the results of the test.

**Table 2**. The goodness of fit test

| Test | Parameter | Value |
|---|---|---|
| Hosmer & Lemeshow Test | Sig (p-value) | 0.935 |
| | Chi-square | 2.988 |
| | df | 8 |
| Classification Plot | Percentage Correct | 75.8 |

H-L test shows that the model is valid (sig> 0.05). While the classification plot results show the overall percentage of cases predicted correctly by the model. The overall percentage reaches 75.8%.

**Table 3**. General characteristics of respondents

| No. | Variable | Category | Percentage (%) |
|---|---|---|---|
| 1. | Age (years old) | < 17 | 0.00 |
| | | 17 – 21 | 30.30 |
| | | 22 – 26 | 23.81 |
| | | 27 – 31 | 16.88 |
| | | 32 – 36 | 9.31 |
| | | 37 – 41 | 4.76 |
| | | 42 – 46 | 6.93 |
| | | 47 – 51 | 4.98 |
| | | 52 – 56 | 1.95 |
| | | 57 – 61 | 0.87 |
| | | > 62 | 0.22 |
| 2. | Gender | Female | 41.77 |
| | | Male | 58.23 |
| 3. | Driving license ownership | Yes | 89.39 |
| | | No | 10.61 |
| 4. | Last education | Primary school | 0.216 |

| No. | Variable | Category | Percentage (%) |
|---|---|---|---|
| | | Junior High School | 1.082 |
| | | Senior High School | 42.86 |
| | | Diploma | 9.091 |
| | | Undergraduate | 32.03 |
| | | Master | 10.82 |
| | | Doctorate/PhD | 3.68 |
| 5. | Driving experience (years) | 1 - 5 | 19.70 |
| | | 6 - 10 | 41.13 |
| | | 11 - 15 | 18.83 |
| | | 16 - 20 | 11.26 |
| | | 21 - 25 | 3.68 |
| | | 26 - 30 | 2.81 |
| | | 31 - 35 | 1.52 |
| | | 36 - 40 | 0.65 |
| | | > 40 | 0.43 |

# RESULT AND DISCUSSION

## Speed Violation Probability Model

The probability of speed violation is obtained from the logistic regression analysis. The use of logistic regression produces a model of the probability of speed violation based on several variables. In this case, the variable is considered significant if it has a significance level of 5% (sig< 0.05). The results of logistic regression analysis on the probability of speed violation are shown in table 4.

**Table 4**. Output model of speed violation probability

| Variable | β | S.E. | Sig. | Exp(β) |
|---|---|---|---|---|
| Age* | -0.047* | 0.019 | 0.013* | 0.954* |
| Gender | 0.168 | 0.229 | 0.465 | 1.182 |
| Driving license ownership | -0.589 | 0.396 | 0.137 | 0.555 |
| Last education | -0.054 | 0.162 | 0.741 | 0.948 |
| Driving experience | 0.035 | 0.024 | 0.147 | 1.036 |
| Constant | 0.384 | 0.834 | 0.625 | 1.468 |

Where: S.E. = Standard Error and Sig. = p-value = significance level

This model shows the relationship between the dependent variable (in this case, the speed violation variable) and the independent variable. The variable that influences speed violation is age. The negative coefficient indicates that increasing the unit of the independent variable causes a decrease in the dependent variable.

The conversion from odd-ratio to the prediction of the probability statement is done so that the interpretation of the results of the model can be made more easily. Sensitivity analysis is also carried out to describe the results of the probability model better.

$$\text{Logit (p)} = ln\frac{p}{1-p} = 0.384 - 0.047\ age$$

therefore, the probability model is:

$$P_j = \frac{exp^{0.384-0.047(\text{age})}}{1+exp^{0.384-0.047(\text{age})}}$$

For the age variable, one unit increase will decrease the probability value by 0.047 log-odds. As an illustration, the probability for someone with the age of 17 becomes:

$$\text{Logit (p)} = ln\frac{p}{1-p} = 0.384 - 0.047\ x\ 17 = -0.415$$

then, it produces a probability:

$$P_j = \frac{exp^{-0.415}}{1+exp^{-0.415}} = 40\%$$

The probability of someone at the age of 17 years violating driving speed is 40%. While the probability of someone at 40 years old (using the same method as the probability at 17 years old) is 18%. This shows that the greater a person's age, the smaller the probability of breaking the driving speed. This is in line with research by Teo and Gan (2016), which states that young drivers prefer high speed. The sensitivity of the effect of age on the probability of speed violation is shown in Figure 1.


**Figure 1**. Age sensitivity to the probability of speed violation

## Accident Involvement Probability Model

Several variables can influence involvement in accidents. The results of logistic regression analysis on the probability of accident involvement are shown in Table 5.

**Table 5**. The output of the probability of accident involvement

| Variable | β | S.E. | Sig. | Exp(β) |
|---|---|---|---|---|
| Age* | -0.164 | 0.064 | 0.01 | 0.849 |
| Constant | -0.645 | 0.253 | 0.011 | 0.525 |
| Driving experience | -0.164 | 0.064 | 0.055 | 0.849 |
| Constant | -0.645 | 0.253 | 0.000 | 0.431 |

The probability of accident involvement is influenced by two variables, namely age and driver experience. The negative coefficient on both variables shows that increasing the unit of the independent variable causes a decrease in the dependent variable. The negative coefficient value at age (-0.164) shows that the smaller a person's age, the higher the probability of being involved in an accident.

$$\text{Logit (p)} = ln\frac{p}{1-p} = -0.645 - 0.164\ age$$

therefore, the probability model is:

$$P_{\text{j}} = \frac{exp^{-0.645-0.164(\text{age})}}{1+exp^{-0.645-0.164(\text{age})}}$$

For the age variable, one unit increase will decrease the probability value by 0.164 log-odds. As an illustration, the probability of someone in the age category 1 (<17 years) becomes:

$$\text{Logit (p)} = ln\frac{p}{1-p} = -0.645 - 0.164\ x\ 1 = \ -0.809$$

then, it produces a probability:

$$P_{\text{j}} = \frac{exp^{-0.809}}{1+exp^{-0.809}} = 31\%$$

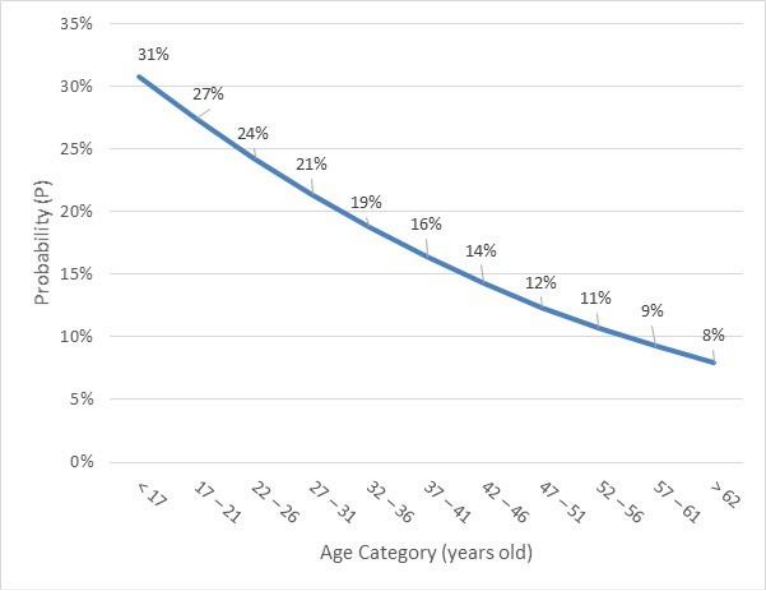


**Figure 2**. The sensitivity of age to the probability of accident involvement

The highest probability of accident involvement is in the age category <17 years (31%). It shows that the level of maturity also influences the likelihood of being involved in an accident.

The same calculation with the age variable was applied to the driving experience variable. The results of sensitivity to these variables are shown in Figure 3.

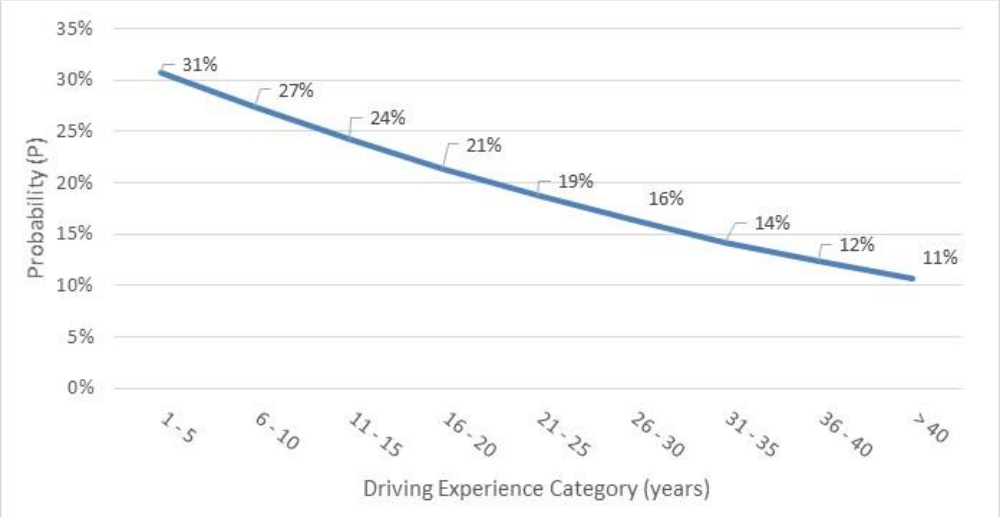


**Figure 3**. The sensitivity of driver experience to probability accident involvement

The highest probability of accident involvement is in drivers with 1-5 years of driving experience (31%). The analysis shows that a long driving experience also reduces the likelihood of someone being involved in an accident.

## CONCLUDING REMARKS

The probability of driving behaviour exceeding the speed limit is influenced by age. The younger person's age causes, the higher the probability of breaking the maximum speed limit. Zhang, Yau, and Chen (2013) stated that young drivers and men who violated the maximum speed limit would be more likely to be involved in an accident. The results of this study also showed that drivers with age categories under 17 were more likely to be involved in accidents.

Furthermore, driving experience also shows a similar trend, where the longer the driving experience is, the less likely it is to be involved in an accident. Tao, Zhang, and Qu (2017) stress the importance of personality traits and driving experience in understanding risky driving behaviour and accident risk. Future studies can further explore the behavior of speeding, which emphasises more on the psychological aspects. Furthermore, studies on accident involvement need to be emphasized more on a person's perception of risky behaviour in driving and how much they are willing to accept that risk.

## ACKNOWLEDGEMENT

The authors would like to thank the Indonesia National Police Traffic Corps for their support in funding this research. This article is part of a research on the Causes of Traffic

Accidents initiated by the Traffic Corps, in collaboration with researchers from fourteen provinces in Indonesia.

## REFERENCES

Coburn, T. C. (2009) 'Statistical and Econometric Methods for Transportation Data Analysis,' *Technometrics*. DOI: 10.1198/tech.2004.s238.

Figueiredo Filho, D. B. *et al.* (2013) 'When is statistical significance not significant?', *Brazilian Political Science Review*, 7(1), pp. 31–55. Available at: http://www.bpsr.org.br/index.php/bpsr/article/view/154.

Gelman, A. and Stern, H. (2006) 'The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant,' *The American Statistician*, 60(4), pp. 328–331. DOI: 10.1198/000313006X152649.

Goniewicz, K. *et al.* (2016) 'Road accident rates: strategies and programmes for improving road traffic safety,' *European Journal of Trauma and Emergency Surgery*, 42(4), pp. 433–438. DOI: 10.1007/s00068-015-0544-6.

Kitamura, Y., Hayashi, M. and Yagi, E. (2018) 'Traffic problems in Southeast Asia featuring the case of Cambodia's traffic accidents involving motorcycles,' *IATSS Research*. Elsevier Ltd, 42(4), pp. 163–170. DOI: 10.1016/j.iatssr.2018.11.001.

Lai, F., Carsten, O. and Tate, F. (2012) 'How much benefit does Intelligent Speed Adaptation deliver? - An analysis of its potential contribution to safety and environment', *Accident Analysis and Prevention*. DOI: 10.1016/j.aap.2011.04.011.

Nickerson, R. S. (2000) 'Null hypothesis significance testing: A review of an old and continuing controversy,' *Psychological Methods*, 5(2), pp. 241–301. DOI: 10.1037//1082-989X.5.2.241.

Tao, D., Zhang, R. and Qu, X. (2017) 'The role of personality traits and driving experience in self-reported risky driving behaviors and accident risk among Chinese drivers,' *Accident Analysis and Prevention*. DOI: 10.1016/j.aap.2016.12.009.

Teo, S. H., and Gan, L. M. (2016) 'Speeding driving behavior: Age and gender experimental analysis,' *MATEC Web of Conferences*. DOI: 10.1051/matecconf/20167400030.

Uncles, M. D., Ben-Akiva, M. and Lerman, S. R. (2006) 'Discrete Choice Analysis: Theory and Application to Travel Demand.', *The Journal of the Operational Research Society*. DOI: 10.2307/2582065.

Wells, H. (2011) 'Risk and expertise in the speed limit enforcement debate: Challenges, adaptations and responses,' *Criminology and Criminal Justice*. DOI: 10.1177/1748895811401976.

Widyastuti, H. (2012) *Valuing motorcycle casualties in developing countries using willingness-to-pay method: stated-preference discrete choice modeling approach*, *PQDT - UK & Ireland*.

World Health Organization (WHO) (2015) *Global Status Report on Road, Report*.

Zhang, G., Yau, K. K. W. and Chen, G. (2013) 'Risk factors associated with traffic violations and accident severity in China,' *Accident Analysis and Prevention*. DOI: 10.1016/j.aap.2013.05.004.