Modeling factor as the cause of traffic accident losses using multiple linear regression approach and generalized linear models

H Fitrianti¹, Y P Pasaribu² and P Betaubun³

¹Department of Mathematics Education, Faculty of Teacher Training and Education, Musamus University, Merauke, Indonesia
²Department of Chemistry Education, Faculty of Teacher Training and Education, Musamus University, Merauke, Indonesia
³Department of Civil Engineering, Faculty of Engineering, Musamus University, Merauke, Indonesia

E-mail: hariani_fitrianti@yahoo.co.id

Abstract. Road Traffic and transport as part of the national transportation system that should be developed its potential and role to realize security, safety, order, and smoothness of traffic and Road Transport to support economic development and regional development. Traffic accidents in Indonesia are currently ranked second in the ASEAN region, the number of traffic accidents an average of 28,000 to 30,000 people per year. Large casualties cause high material losses; this can result in poverty levels. The level of poverty experienced by traffic accidents due to natural disadvantages requires care, lost productivity, lost livelihood, stress and prolonged suffering. This study aims to model the factors causing the magnitude of traffic accidents using linear regression model and the GLM model. The results obtained by factor is out of balance and exceeds the speed limit which has a significant effect on the amount of traffic accident loss for both model linear regression model and GLM model, while the disorder factor and the influence factor of alcohol have no significant effect. Multiple linear regression model obtained can be written in the form of equations i.e. $Y = 2367906 + 10716421 X_1 + 987344 X_4$ and the GLM model equations can be written i.e. $ln(\mu) = 17,73500 + 0,05264 X_1 + 0,09291 X_4$. From both models obtained the best model based on model which has the smallest AIC value that is GLM model.

1 Introduction
Traffic and Road Transport as part of the national transportation system that should be developed its potential and role to realize security, safety, order, and smoothness of traffic and Road Transport to support economic development and regional development. Population growth causes an increase in the number of vehicle users, an increase in traffic accidents. The occurrence of a traffic accident is an accidental and unintended accident involving a vehicle and or without other road users causing human casualties as well as loss of property [1][2][3]. The World Health Organization (WHO) predicts that by 2030 traffic accidents will be the fifth largest human killer factor in the world [2][5].

Traffic accidents in Indonesia are currently ranked second in the ASEAN region, the number of traffic accidents an average of 28,000 to 30,000 people per year. Large casualties number cause high material losses; this can result in poverty levels. The level of poverty experienced by traffic accidents...
due to unnecessary requires care, lost productivity, loss of livelihood, stress and prolonged suffering [6][3]. According to Transport State Laboratory study, the death rate due to traffic accidents in Indonesia is much higher when compared to developed countries in Europe and North America [7].

Merauke Resort Police noted in January to October 2016 the case of traffic accidents were 232 cases of accidents consisting of motor vehicle accidents, freight cars, buses, and special vehicles. Cases of the accident caused a severity of 26 deaths, 126 seriously injured, and 298 minor injuries. While the losses were Rp 2,634,800,000. Based on research conducted by Fitrianti [8] traffic accidents case in Merauke Regency the most significant caused by human factor, while environmental factors and vehicle factors have no significant effect on accidents occurring in Merauke Regency. Merauke Regency police report noted that three factors are causing the accidents of traffic, namely human factors, vehicles, and the environment. According to Fitrianti's research [9], the most dominant factor affecting the incidence of accidents in Merauke Regency is the human factor consisting of the factor of careless, disorderly, alcohol use, and exceeding the speed limit. In this study, it examines the factors causing the magnitude of traffic accidents using a multiple linear regression model of generalized linear models (GLM).

The research that has been conducted is among others about the cause of traffic accidents using Ordinal Logistic Regression Bagging [4], modeling the number of accidents using multiple linear regression [10], modeling factors causing traffic accidents using Geographically Weighted Regression [11]. In this research is different from other research that is modeling factor cause a big loss of traffic accidents using multiple linear regression and GLM. GLM has several advantages over the use of conventional regression analysis in predicting accident rates, due to nonnegative accident data events [12].

2. Methods

This research was conducted in Merauke Regency of Papua Province. The growth of the local population of the year has increased, this is directly proportional to the number of vehicle users. This research uses multiple linear regression and GLM. The data used were secondary data obtained from the merauke police resort. The variables used were independent variable namely the loss of traffic accidents and independent variables based on research conducted by fitriani factors that cause the number of traffic accidents in the city of Merauke, ie factor of unattended, alcoholic, and exceed the speed limit. As for the modeling steps used were:

2.1. Multiple Linear Regression

Multiple linear regression is an analysis to see the causal relationship between the predictor variables to the response variable. This modeling can be written as follows [11]:

$$Y_i = \alpha_0 + \alpha_1X_1 + \alpha_2X_2 + \alpha_3X_3 + \alpha_4X_4 + e_i$$

(1)

Where:

- $Y$ = Total loss by traffic accidents
- $X_1$ = Careless factor
- $X_2$ = Undisciplined factor
- $X_3$ = Alcohol use factor
- $X_4$ = Exceed speed limit factor
- $e_i$ = Data error i

the assumption of error distribution is normal with mean $\mu$ and its variety $\sigma^2$.

The steps of multiple linear regression modeling are:

a) Prerequisite test analysis namely normality test, multicollinearity test, and heteroscedasticity test.
b) Assess multiple linear regression model parameters.

c) Partial significance test namely partial parameter test using t-test and parameter test simultaneously using an F test.

d) Calculating the Akaike’s Information Criterion (AIC) value

2.2. GLM

GLM is a general linear model, but modeling using GLM model is different from the usual regression model where the GLM model of response variable distribution must come from the exponential family distribution and which has a linear relationship, ie the mean of the response variable to the predictor variable, so the response variable causes the response variable to be heteroskedastic namely the diversity of time-dependent variations. GLM modeling for the analysis of factors causing the magnitude of traffic accidents, ie [9]:

\[ g(\mu) = \ln(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon) \]  

(2)

With \( g(\mu) \) = mean of response variable / mean of response variable

The steps of GLM modeling are:

a) Determine the distribution of the response variable.

b) Assess the distribution variable of the response variable using Maximum Likelihood Estimation (MLE).

c) The matching test of the response variable distribution using Kolmogorov-Smirnov and Anderson Darling.

d) Determine link function \( g(\mu) = x'\beta \)

e) Assess and test parameters using the wald test for partial testing and Likelihood Ratio Test for simultaneous testing.

f) Calculating the AIC value

3. Results and Discussion

3.1. Description of research data

Description of research data aims to describe, brief, explain in general the data of traffic accidents before being used in research. The result of the description of the research data is:

| Numeric essence | Y         | X₁       | X₂       | X₃       | X₄       |
|-----------------|-----------|----------|----------|----------|----------|
| N               | 33        | 33       | 33       | 33       | 33       |
| Mean            | 160315169.7 | 4.878787879 | 4.818181818 | 5.848484848 | 8.545454545 |
| Median          | 148700000 | 2        | 3        | 6        | 8        |
| Standar deviasi | 133917068.9 | 5.829652518 | 4.857047363 | 2.333468611 | 3.961462081 |
| Minimum         | 8800000   | 0        | 0        | 0        | 2        |
| Maksimum        | 594850000 | 23       | 14       | 10       | 18       |
| Skewness        | 1.78      | 1.48     | 0.91     | -0.21    | 0.37     |

Based on table 1, it is found that the magnitude of a traffic accident in Merauke Regency contains the average of traffic accident losses that occur amounts to 160315169, a careless factor of 4.8 is 4.8, undisciplined driver factor is 4.8, alcohol use driver factor is 5.8, and exceeds speed limit factor is 8.5. Minimum magnitude of traffic accidents loss is 8800000, careless factor, undisciplined driver factor and alcohol use driver factor is 0, and exceeds speed limit is 2. Maximum traffic accident loss is 594850000; careless factor is 23, undisciplined driver factor is 14, alcohol use driver factor is 10, and exceeds speed limit is 18. Another thing that can be obtained from table 1 is to mean is greater than
median this shows that the data tends to go to the right and this is also supported from the value of skewness obtained is 1.7. A positive value of skewness means that the chance of a large traffic accident cost has a small chance of happening. While the spread of data for the loss of traffic accidents is 133917068, the careless factor is 5.8, undisciplined driver factor is 4.8, alcohol use factor is 2.3, and exceeds speed limit is 3.9.

3.2. Modeling Factors as the cause of Traffic Accidents Loss Magnitude

a) Results of multiple linear regression modeling

Multiple linear regression modeling to see the influence of careless factor, undisciplined factor, alcohol use factor, and exceeds the speed limit on the magnitude of traffic accident losses with the model equation in equation (1). After that, the equation parameter (1) was estimated using Ordinary Least Square (OLS) method [11]. The result of parameter estimation can be seen in Table 2.

| Parameter | Coefisien | parameter estimation results | Standard error |
|-----------|-----------|-----------------------------|---------------|
| Full model | 21130683 | 61099367 |
| $x_1$ | 10766203 | 4404484 |
| $x_2$ | 249652 | 5423916 |
| $x_3$ | 490740 | 10611596 |
| $x_4$ | 9664261 | 5774581 |

Based on the results obtained in Table 2, then conducted testing of significant parameters. Significant tests were performed partially using the t test [11], a significant variable partially if $p-value<\alpha \ast 0.1$. Then conducted testing simultaneously using F test [11], variables are significant simultaneously if $p-value<\alpha \ast 0.1$. Test results can be seen in Table 3.

| Parameter | $t$-value | $p$-value of $t$ test | $F$-value | $p$-value of F test |
|-----------|-----------|----------------------|-----------|---------------------|
| Full model | 0.3 | 0.732 | | |
| $x_1$ | 2.4 | 0.021 | 3.8 | 0.01 |
| $x_2$ | 0.046 | 0.964 | 3.8 | 0.01 |
| $x_3$ | 0.046 | 0.963 | | |
| $x_4$ | 1.674 | 0.05 | | |

Results of Table 3 obtained simultaneous testing of significant variable influence to response variable, whereas partial testing was only variable $x_1$ and $x_4$ which influences namely the careless factor and exceeds the speed limit factor to the magnitude of the loss of traffic accidents.

b) Results of GLM Modeling

Modeling using GLM is against predictor variables and the same response variables used in multiple linear regression models, its purpose is to see which model is most appropriate to model the factors causing the magnitude of a traffic accident. In the initial phase, the GLM model was done to determine the distribution of response variables, where the distribution must be derived from the distribution of family exponential. The data loss is a continuous data thus the selected distributions for the response variables are Gamma distribution, Gaussian Inversion, Lognormal distribution, and normal distribution. The distribution test was performed using Kolmogorov-Smirnov (KS), and Anderson Darling (AD) [13] for that to be selected from the four distributions were the best distribution based on the value of AIC.
Table 4. Testing Results And Distribution Variable Response Estimates

| Distribution  | Parameter Estimation Results | Test Statistics | Critical Point (α * 0.1) | AIC |
|---------------|------------------------------|-----------------|---------------------------|-----|
| Gamma         | shape = 1.4570852511         | KS = 0.12469790 | 0.21                      | 554 |
|               | rate = 0.0009091079          | AD = 0.38827670 | 1.93                      |     |
| Invers gaussian | mu = 1602.0714               | KS = 0.2480339  | 0.21                      | 563 |
|               | lambda = 967.0847            | AD = 1.9289735  | 1.93                      |     |
| Lognormal     | meanlog = 6.9989969          | KS = 0.2480339  | 0.21                      | 558 |
|               | Sd log = 0.9879575           | AD = 0.8738010  | 1.93                      |     |
| Normal        | Mean = 1603.152              | KS = 0.1530297  | 0.21                      | 571 |
|               | sd = 1318.724                | AD = 1.3558950  | 1.93                      |     |

From Table 4, we obtained a suitable response variable distribution based on KS and AD test, ie Gamma distribution and normal distribution having test statistic smaller than a critical point with a significance level of 10% with a number of samples that is 33 that is 0.21. For the distribution model of the corresponding response variable, ie the Gamma distribution and the normal distribution will be selected one distribution, where the best distribution selection for the response variable is using Akaike's Information Criterion (AIC) criterion [13]. The best model is the model that has the smallest AIC value, so a good distribution model for the response variable is the Gamma distribution. This is also supported by research conducted by Haryadi [14], that the linear model with a normal distribution is not sufficient to predict traffic accidents. For more clear description it can be seen in Figure 1 distribution closest to the data that is Gamma distribution. Thus, the distribution to be used for the response variable is Gamma distribution. For more clear description it can be seen in Figure 1.

![Figure 1. The plot of Variable Response Distribution](image)

Based on figure 1 it is seen that the distribution closest to the data is Gamma distribution, so the distribution of the response variable to be used in GLM modeling is Gamma distribution. After obtaining the response variable distribution, furthermore the estimation of the GLM model parameters written in equation (2) using the Maximum Likelihood Estimate (MLE) method [13]. To estimate the parameters in this study, it was done using the aid of software R, with the results can be seen in Table 5.
Table 5. GLM Model Parameter Estimation Result

| Parameter | Parameter Estimation Results | Standard error |
|-----------|-----------------------------|----------------|
| Full model |                            |                |
| Coefficient | 17.823589 | 0.438896 |
| $x_1$       | 0.056436  | 0.031639  |
| $x_2$       | 0.002983  | 0.038962  |
| $x_3$       | -0.029462 | 0.076226  |
| $x_4$       | 0.098704  | 0.041481  |

After obtaining the parameters from equation (2) the result in Table 5, then tested the significance of these parameters. Testing of significance was done to see which factor is the most dominant influence on the big loss of traffic accident. The significance test was performed partially using the Wald test to see each predictor variable against the response variable [15]. Testing simultaneously using Likelihood Ratio (LR) test to know the influence of predictor variable together to the response variable [16]. The test results of significance can be seen in table 6.

Table 6. Test Results Significance of GLM Model For All Variables

| Parameter | Wald Test | $a$ p-value of Wald test | LR | $a$ p-value of LR test |
|-----------|-----------|--------------------------|----|------------------------|
| Full model |           |                          |    |                        |
| Coefficient | 40.610 | $<2E^{-16}$ *** | 12.17 | 0.02 |
| $x_1$      | 1.784    | 0.0853                   |    |                        |
| $x_2$      | 0.077    | 0.9395                   |    |                        |
| $x_3$      | -0.387   | 0.7020                   |    |                        |
| $x_4$      | 2.380    | 0.0244                   |    |                        |

Predictable variables were said to have a significant effect on the response variable when testing the parameters of each variable obtained $p-value<\alpha = 0.1$. Based on Table 7 significant variables are $x_1$ dan $x_4$. While for testing simultaneously significant variables if $p-value<\alpha = 0.1$. The results obtained by predictor variables simultaneously have a significant effect on predictor variables.

After GLM modeling and regression were obtained, the assumption of error was then performed. In equations (1) and (2) it has been assumed that the distribution of error is the normal distribution and therefore it will be tested that assumption which results can be seen in table 7.

Table 7. Testing Results And Estimation of Distribution Parameters Error

| Model       | Parameter Estimation Results | Statistics Test | Critical Point ($\alpha = 1$) |
|-------------|------------------------------|-----------------|-----------------------------|
| GLM         | Mean = -0.1739721           | KS = 0.10854181 | 0.21                        |
|             | Sd = 0.7193780              | AD = 0.40888626 | 1.93                        |
| Multiple    | Mean = -1.254216E-14        | KS = 0.1487734  | 0.21                        |
| Linear      | Sd = 1.060045E+03           | AD = 1.1868123  | 1.93                        |
| Regression  |                             |                 |                             |

The results of table 7 based on Kolmogorov-Smirnov and Anderson darling test obtained statistical test greater than the critical point means the error distribution can be assumed to a normal distribution with the mean of 0.17 and standard deviation 0.7 for GLM model and multiple linear regression model mean that 1.254216E-14 and the standard deviation is 1.060045E+03. For more clear description it can be seen in Figure 2 it can be seen that the normal distribution forms the same pattern as the error, so the error distribution can be assumed to be a normal distribution.
GLM Model regresi linier berganda Model

Figure 2. Dissemination Plot Error For GLM Model and Multiple Linear Regression Model

Based on figure 2 it can be seen that the normal distribution forms the same pattern as the error, so the error distribution can be assumed to be a normal distribution.

3.3. Modeling of Multiple Linear Regression and GLM Model Based on Significant Variables

In the previous results obtained significant variables for multiple linear regression model and the GLM model in table 3 and table 6 is the careless factor and factor variables exceed the speed limit. Next will be modeled again with multiple linear regression model and the GLM model for the significant variables using the same stages. So the one will be modeled that is an only careless factor and exceed the speed limit to the amount of traffic accident loss. The results of the parameter estimation can be seen in table 8, while for testing significance in table 8.

Table 8. Estimation Results of GLM Model and Multiple Linear Regression for Significant Variables

| Parameter | GLM Significant Model | Model of Significant of Multiple Linear Regression |
|-----------|-----------------------|--------------------------------------------------|
|           | Parameter estimation  | Standard error                                  | Parameter estimation | Standard error                                  |
| Coefficient | 17.73500              | 0.34917                                          | 23676906             | 47404872                                          |
| $x_i$       | 0.05264               | 0.02528                                          | 10716421             | 3431870                                           |
| $x_d$       | 0.09291               | 0.03720                                          | 9871344              | 5050310                                           |

After obtained parameter estimation in table 8, then tested partially and simultaneously on the variable $x_i$ dan $x_d$. Test results can be seen in table 9.

Table 9. Significant Test Results Variable $x_i$ dan $x_d$

| Parameter | GLM Significance Model | Significance model of multiple linear regression |
|-----------|-----------------------|--------------------------------------------------|
|           | Wald test of wald test | $p$-value | LR test of LR test | $p$-value | t-test | $p$-value | F-test | $p$-value of F test |
| Coefficient | 50.792                | <2e-16    | 6.48               | 0.004     | 0.499  | 0.62110   | 8.214  | 0.001429               |
Based on table 9 can be seen the value of testing simultaneously and partially both multiple linear regression model and GLM model obtained value \( p\text{-value} < \alpha \cdot 0.05 \). So it can be concluded that the careless factor and exceeds speed limit significantly influence both partially and simultaneously to the magnitude of the loss of traffic accidents.

### 3.4. Selecting Best Model

The best model selection was made by the AIC method, where the best model is the model that has the smallest AIC value [13]. The results can be seen in table 10.

#### Table 10. AIC Value Calculation Result

| MODEL                      | AIC     |
|----------------------------|---------|
| Full model GLM             | 1310.383|
| Full model Linear Multiple Regression | 1325.255|
| Model GLM                  | 1306.559|
| Model regression           | 1321.263|

Table 10 shows the model that has the smallest AIC value is the GLM significance model. Therefore the best model is the model with the predictor variable that is a careless factor and the factor exceeds the speed limit to the response variable that is the loss of traffic accident. Thus the model equation can be written based on the estimated parameters obtained in table 8 that is:

\[
\ln(\mu) = 17.73500 + 0.05264X_1 + 0.09291X_4
\]  

### 4. Conclusion

From the analysis and discussion, we can conclude a good model in modeling the cause of the magnitude of traffic accident losses based on the smallest AIC value criteria that are 1306.559 is the GLM model. Factors that significantly influence the amount of traffic accidents costs on the GLM model obtained are the drivers being off guard and drivers exceed the speed limit with a significant level of 5%. During undisciplined factor and alcohol, use has no significant effect on the amount of traffic accident. Model factor causes the amount of loss using GLM i.e. \( \ln(\mu) = 17.73500 + 0.05264X_1 + 0.09291X_4 \).

### References

[1] Dewan Perwakilan Rakyat 2009 *Undang - Undang RI Nomor 22 Tahun 2009 tentang lalu lintas dan angkutan jalan & Peraturan Pemerintah RI Nomor 55 Tahun 2012 tentang kendaraan beserta penjelasannya* (Surabaya: Kesindo Utama) 345-346

[2] Atubi A2012 Determinants of road traffic accident occurrences in lagos state: some lessons for nigeria *International Journal of Humanities and Social Science* 2 252-9

[3] Wangdi C G, Duba M S, Wilkinson T E, Tun M Z and Tripathy J P 2018 Burden, pattern and causes of road traffic accidents in Bhutan, 2013-2014: A Police Record Review *International Journal of Injury Control and Safety Promotion* 25 65-69

[4] Fitriah W W, Mashuri M and Irhamah 2012 Faktor-faktor yang mempengaruhi keparahan korban kecelakaan lalu lintas di kota surabaya dengan pendekatan bagging regresi logistik ordinal *Jurnal Sains dan Seni ITS* 1 253-8

[5] Gebru M K 2017 Road traffic accident: Human security perspective *International Journal of Peace and Development Studies* 8 15-24
[6] Sharma S M 2016 Road Traffic Accidents in India International Journal of Advanced and Integrated Medical Sciences 1 57-64.

[7] Transport Research Laboratory 1995 Costing road accident in developing countries, overseas road note 10 (United Kingdom: OverseasCentre, Crowthorne, Beshire) 567-578

[8] Fitrianti H 2017 Model faktor-faktor penyebab banyaknya kecelakaan lalu lintas pada kendaraan sepeda motor dengan pendekatan Generalized Linear Models (GLM) Magistra 4 94-104

[9] Fitrianti H 2017 Generalized Linear Models (GLM) kejadian kecelakaan lalu lintas di Kota Merauke Prosiding SNM 262-271

[10] Fraticasari S Y, Ratnawati D K and Wihandika RC 2018 Optimasi pemodelan regresi linier berganda pada prediksi jumlah kecelakaan sepeda motor dengan algoritme genetika Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer 2 1932-39

[11] Dewi P L A and Ismaini Z 2016 Pemodelan faktor penyebab kecelakaan lalu lintas berdasarkan metode Geographically Weighted Regression di Jawa Timur Jurnal Sains dan Seni ITS 5 58-64.

[12] Fitrianti H 2016 Menentukan premi murni menggunakan Generalized Linear Models dan Model Copula Magistra 3 69-81

[13] Authors 2010 Di Tikungan karena Pengaruh Konsistensi Alinyemen Horisontal dalam Desain Geometri Jalan Raya Media Teknik Sipil, 10 85-92.

[14] Haryadi B 2011 Eksplorasi model tingkat kecelakaan lalu-lintas jalan tol dengan teknik GLM (Generalized Linear Modeling) Jurnal Teknik Sipil & Perencanaan 1 91-100

[15] Cahyandari R 2014 Pengujian overdispersi pada model regresi poisson Statistik 14 69-76.

[16] De Jong P and Heller G Z 2008 Generalized Linear Models For Insurance Data (Cambridge: Cambridge University) 789-790