

Analyzing Motorcycle Accident Frequency Using Generalized Poisson Distributions

Riski Nur Istiqomah Dinnullah^{1,2}, Sobri Abusini¹,
Rahma Fitriani¹, Marjono Marjono¹

¹ University of Brawijaya, Jl. Veteran, Ketawanggede, Malang City, Indonesia

² Universitas PGRI Kanjuruhan Malang, Jl. S. Supriadi 48, Malang City, Indonesia

Abstract – Motorcycle accidents in East Java are more common than accidents in other modes of transportation. In addition to the many motorcycle users today, human, environmental, and road factors are considered the highest causes of these accidents. The study's goal is to find the best model of the Generalized Poisson Family Distribution (GPR), namely Lagrangian Poisson Regression (LPR) and to construct a model that will quantify the frequency of motorcycle accidents in East Java. Akaike Information Criterion (AIC) and Schwarz Bayesian Criterion (SBC) criteria are the model comparison methods used in this research. The selection was also made based on the model's exponential coefficient with a 95% CI to further deepen the selection results obtained. In addition, the paired samples test was performed to determine the degree of dissimilarity between the outcomes produced by the developed model and the actual data. The best performance model is applied to identify the characteristics or factors highly involved in motorcycle accidents. The research uses secondary data from related agencies, namely the East Java Regional Police, especially the traffic accident unit, and East Java BPS, for 38 cities and districts in 2021. The numerical optimization method used is the iteratively reweighted least squares (IRLS) algorithm, assisted by R Studio software.

The study findings show that LPR is the most efficient and exact approach for modeling the frequency of motorcycle accidents. Meanwhile, the percentage of teenagers (X_1), the frequency of motorized vehicles (X_3), and the average annual rainfall (X_5) have a considerable impact on accident occurrence. This research has an important contribution, especially in the field of transportation modeling and designing appropriate strategies to reduce the frequency of motorcycle accidents.

Keywords – Motorcycle accidents, generalized Poisson family distribution, selection model.

1. Introduction

Traffic accidents are an issue in transportation that is becoming more common every year. Globally, fatalities among individuals aged 5 to 29 are predominantly attributed to motor vehicle collisions [1]. They account for the ninth-highest cause of mortality among individuals of all ages. Moreover, WHO reports that traffic accidents are anticipated to be the fifth most prevalent cause of mortality worldwide by 2030 [2], [3].

The issue of traffic incidents is a growing phenomenon in developing countries, such as Indonesia. Indonesia is ranked first, with an accidental death rate of 2.46% of total deaths [4]. From 2014 to 2018, the frequency of road accidents in Indonesia climbed by 3.30% every year [5]. There were 109,215 road events in 2018, with 29,472 fatalities and 13,315 severe injuries [5]. The mode of transportation involved with the highest incidence of accidents is motorcycles. If seen from the statistical percentage of national police data in 2020, 78.9% of motorcycle accidents occurred in Indonesia, of which 41,170 died, 35,660 people were seriously injured, and 160,300 were slightly injured. The areas with a large population, such as East Java Province, dominate Indonesia's vulnerability to traffic accidents. This province has many cities/regencies with metropolitan regions with high traffic density and mobility, so the accident rate is also relatively high in this area.

DOI: 10.18421/TEM131-24

<https://doi.org/10.18421/TEM131-24>

Corresponding author: Riski Nur Istiqomah Dinnullah,
University of Brawijaya, Jl. Veteran, Ketawanggede,
Malang City, Indonesia.


Email: rektorat@ub.ac.id

Received: 19 August 2023.

Revised: 15 November 2023.

Accepted: 08 December 2023.

Published: 27 February 2024.

 © 2024. Riski Nur Istiqomah Dinnullah, Sobri Abusini, Rahma Fitriani & Marjono Marjono; published by UIKTEN. This work is licensed under the Creative Commons Attribution-NonCommercial-NoDeriv 4.0 License.

The article is published with Open Access at <https://www.temjournal.com/>

The motorcycle accident is one of the applied cases involving count data modeling. Because accidents are unusual and stochastic occurrences with a low probability of occurring, the Poisson distribution is the best choice for modeling them [6], [7]. The Poisson distribution serves as the fundamental component upon which a Poisson regression model is built. Meanwhile, equidispersion is a condition that Poisson regression must satisfy, which states that the mean and variance must possess identical values. [8], [9], [10], [11], [12]. Several studies have violations of assumptions, namely cases of overdispersion where the variance value exceeds the mean value. This case of overdispersion has been discussed in researches [13], [14], [15], [16], [17], [18]. Underdispersion is a condition that occurs when the variance is slight in comparison to the mean value. The underdispersion case is an uncommon occurrence [19]. Underdispersion situations have been extensively researched [20], [21], [22]. The principal shortcoming of Poisson regression is that the relationship between the mean and variance is dispersive. The result is a significant deviation value, which diminishes the accuracy of the model acquired.

The generalized Poisson distribution family likely solves the overdispersion issue in Poisson regression. Lagrangian Poisson Regression (LPR) and Generalized Poisson Regression (GPR) are two generalized Poisson distribution families studied in this study. Lagrangian Poisson was introduced by study [23]. The Lagrangian Poisson distribution is a well-studied and popular alternative to the conventional Poisson distribution [24]. The authors of [25] revealed that the Lagrangian Poisson distribution performs at least as well as specific alternatives to the standard distribution when modeling the six data sets obtained in [26] concerning the frequency of injuries in car accidents. Model development is carried out through parameter transformation by transforming the parameters into GPR [27]. The model has been implemented for accident data with overdispersion cases [28]. The findings suggest that GPR is capable of accurately modeling the association between demographic characteristics, driving behavior, drug use, and the frequency of incidents involving senior drivers. The correlation between meteorological conditions and the occurrence of traffic accidents was examined in [29] where the performance of negative binomial regression is compared with GPR, Poisson, and COM-Poisson models. The findings of this research indicate that the GPR model exhibits lower AIC and BIC values in comparison to other regression models. As a consequence, the GPR is deemed more suitable for the examination of traffic accidents.

In this paper, a generalized Poisson distribution family-based regression model is constructed using accident data obtained from the East Java Police Service. The development of these two models was compared using AIC and SBC criteria to determine the most reliable model. The selection was also made based on exponential coefficients of models with 95% CI to deepen further the selection results obtained. In addition, the study use the paired samples test where the statistics gathered, provide insight into the magnitude of differences between the developing model and the observed data. Determination of a good model is also seen based on the value of the smaller mean paired differences.

In addition, the factors contributing to the excessive occurrence of motorcycle accidents are identified by utilizing the performance of the best model. This research examines the combination of drivers attributes, road conditions, vehicle specifications, and environmental factors to assist the government in formulating effective methods to decrease the occurrence of motorcycle accidents in East Java Province.

2. Material and Methods

This research uses secondary data from related agencies, namely the East Java Regional Police, especially the traffic accident unit and East Java BPS, for 38 cities/districts in 2021. The response variables are the frequency of motorcycle accidents. Meanwhile, there were five predictor variables involved, namely the percentage of adolescents (X_1), the percentage of low-level education (X_2), the frequency of motorized vehicles (X_3), the length of roads with good road conditions (X_4), and the average annual rainfall (X_5).

Maximum likelihood estimation (MLE) is utilized in order to obtain the LPR and GPR models' parameters. The distribution approach will be ascertained in this paper through the process of maximizing the log-likelihood function. Parameter determination results in a nonlinear solution that cannot be solved exactly. Iteratively reweighted least squares (IRLS) was chosen as the numerical optimization method to solve parameter solutions using the R Studio software.

2.1. Poisson Regression

The p.d.f. of Y that has Poisson distribution is

$$p(y, \lambda) = \frac{e^{-\lambda} \lambda^y}{y!}, y = 0, 1, 2, 3, \dots, \quad (1)$$

$\lambda > 0$ is a distribution parameter [30].

Poisson regression has the following form.

$$\lambda_i = e^{x_i \beta}, \tag{2}$$

where $\mathbf{x} = [1 \ x_{1i} \ x_{2i} \ \dots \ x_{pi}]$ is the predictor variable, and $\boldsymbol{\beta} = [\beta_0 \ \beta_1 \ \beta_2 \ \dots \ \beta_p]^T$ is the regression parameter.

2.2. Dispersion

Equidispersion conditions are assumed by the Poisson distribution [11], [31], where the mean value is equivalent to the variance. In many cases, however, these assumptions must be satisfied. Occasionally, the mean value exceeds the variance value, or conversely. This condition is also known as overdispersion or underdispersion. Model inaccuracies may result from the presence of this dispersion case. If the dispersion-indicating Poisson regression model is still employed, the estimation parameters become consistent and efficient. It causes the parameter estimates to be biased so that they can give wrong results in concluding the observed data [32]. The dispersion condition in the Poisson regression causes the mean value and variance to be related via a parameter denoting dispersion.

$$V(X) = \tau E(X) = \tau \lambda \tag{3}$$

where τ is the dispersion parameter. If $\tau > 1$, there is overdispersion in the data, but if $\tau < 1$ the data experiences underdispersion [33].

2.3. Multicollinearity

Multicollinearity arises when a number of predictor variables exhibit a substantial correlation with both the response variable and other predictor variables [34]. Significant variables become statistically insignificant as a result of the increase in standard error caused by the existence of this correlation. As stated in reference [35], multicollinearity can be detected by utilizing the variance inflation factor (VIF).

$$VIF = \frac{1}{1 - R_i^2} = \frac{1}{Tolerance}. \tag{4}$$

R_i^2 is the coefficient of determination between a predictor variable X_i with other variables.

Furthermore, the interpretation of the multicollinearity case is based on the VIF value. If the value is $VIF > 10$, then multicollinearity exists among the predictor variables within the regression model. It causes a weak estimation of the regression coefficient [36]. Apart from using VIF, multicollinearity in regression can be determined using each predictor variable's tolerance.

Tolerance values that are smaller than 0.10 will indicate the occurrence of multicollinearity.

2.4. Lagrangian Poisson Regression (LPR)

The Poisson distribution underpins the LPR model formation. The LPR model possesses the ability to rectify problems associated with overdispersion or underdispersion.

Definition 1.

For the Lagrangian Poisson distribution, let Y_i represent the count response variable. The following is the definition of the p.d.f.

$$p(y_i; \psi_i, \theta) = \begin{cases} \psi_i (\psi_i + \theta y_i)^{y_i-1} \frac{e^{-\psi_i - \theta y_i}}{y_i!} & \theta > 0 \\ 0 & y_i > k; \theta < 0 \end{cases} \tag{5}$$

with $\psi_i > 0$, $0 \leq \theta \leq 1$, and $\max(1, -\psi_i / 4 \leq \theta < 1)$. Meanwhile, other zeros have $\psi_i + \theta k > 0$ when $\theta < 0$ where k is the largest positive number. The mean and variation of the Lagrangian Poisson distribution are finite when $\theta < 1$ is $E(Y) = \psi_i (1 - \theta)^{-1}$ and $Var(Y) = \psi_i (1 - \theta)^{-3}$ [37].

When $\theta = 0$, equation (1) becomes the Poisson distribution. Next, the systematic components of the Lagrangian Poisson distribution is $\psi_i = \exp(\beta_0 + \sum_{j=1} \beta_j x_{ji})$ and substituted into equation (5), so we get

$$p(y_i; \boldsymbol{\beta}, \theta) = e^{\beta_0 + \sum_{j=1} \beta_j x_{ji}} \left(e^{\beta_0 + \sum_{j=1} \beta_j x_{ji}} + \theta y_i \right)^{y_i-1} \frac{e^{-e^{\beta_0 + \sum_{j=1} \beta_j x_{ji}} - \theta y_i}}{y_i!} \tag{6}$$

2.5. Generalized Poisson Regression (GPR)

GPR extends the Poisson regression model and or more predictor variables and response variables [38]. GPR is a technique utilized to analyze the correlation between two variables. Overdispersion or underdispersion data can handle the use of the GPR model.

Definition 2.

Suppose the response variable count. The generalized Poisson regression model has the p.d.f defined as follows.

$$p(y_i; \sigma_i, \varphi) = \left(\frac{\sigma_i}{1 + \varphi \sigma_i} \right)^{y_i} \frac{(1 + \varphi y_i)^{y_i-1}}{y_i!} e^{-\left(\frac{\sigma_i (1 + \varphi y_i)}{1 + \varphi y_i} \right)}, \tag{7}$$

for $y_i = 1, 2, 3, \dots, n$. If $\varphi = 0$, equation (8) becomes the p.d.f of the Poisson distribution.

Meanwhile, if $\phi > 0$, the generalized Poisson regression model shows overdispersion case, but if $\phi < 0$ generalized Poisson regression shows underdispersion case [29]. Furthermore, the systematic component of the generalized Poisson distribution is $\sigma_i = e^{\beta_0 + \sum_{j=1}^p \beta_j x_{ji}}$ and substituted into equation (7) to obtain

$$p(y_i; \beta, \phi) = \left(\frac{e^{\beta_0 + \sum_{j=1}^p \beta_j x_{ji}}}{1 + \phi e^{\beta_0 + \sum_{j=1}^p \beta_j x_{ji}}} \right)^{y_i} \frac{(1 + \phi y_i)^{y_i - 1}}{y_i!} e^{-\frac{\left(\frac{e^{\beta_0 + \sum_{j=1}^p \beta_j x_{ji}}}{1 + \phi e^{\beta_0 + \sum_{j=1}^p \beta_j x_{ji}}} \right)^{y_i}}{1 + \phi e^{\beta_0 + \sum_{j=1}^p \beta_j x_{ji}}}} \quad (8)$$

2.6. Best Model Selection

Akaike Information Criterion (AIC)

AIC is a criterion utilized in econometrics to select a model. The equation is

$$AIC = -2l(\theta) + 2p \quad (9)$$

$l(\theta)$ is the log-likelihood function, and p represents the quantity of model parameters.

The optimal model is determined by the minimal AIC value [39], [40].

Schwarz Bayesian Criterion (SBC)

SBC is the best model selection criterion, which is formulated as follows:

$$SBC = -2l(\theta) + (\log n)p \quad (10)$$

n denotes the quantity of data. Optimal is the regression model with the smallest SBC value. BIC is another name for SBC, which is Bayesian information criterion. [41].

3. Result and Discussion

This area is one of the provinces in Unitary State Republic of Indonesia and occupies the easternmost region of Java Island. East Java Province comprises an area of 46,428.57 km², comprising both land and sea regions. With 29 regencies and 9 cities, the region is comprised of a total of 38 regencies/cities.

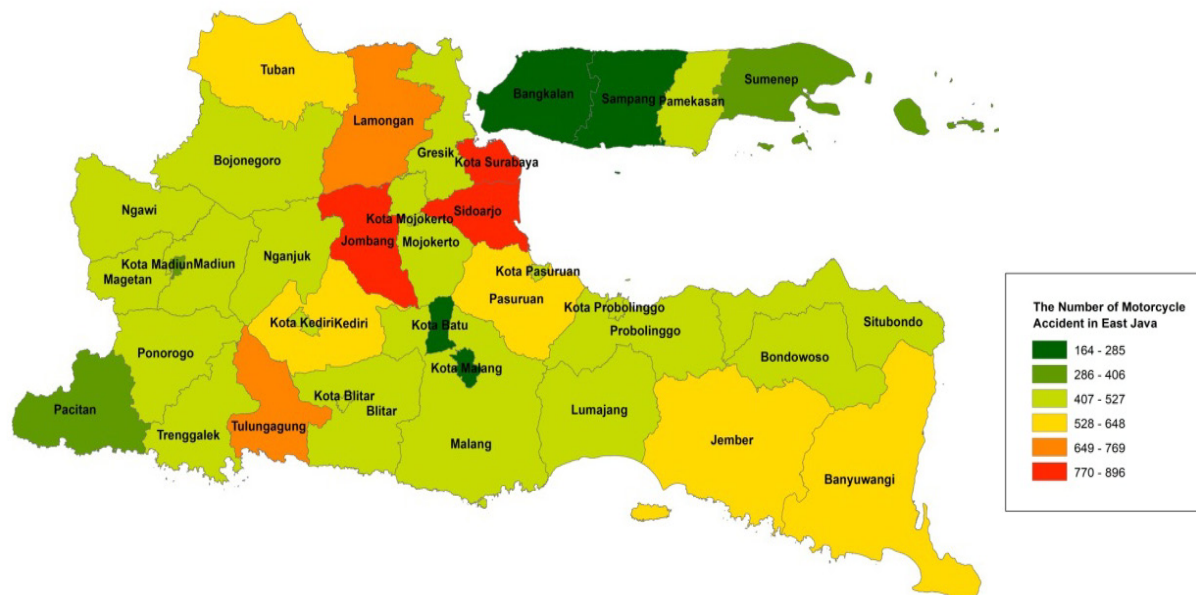


Figure 1. The distribution of accident areas

3.1. The Frequency of Motorcycle Accidents in East Java

The frequency of traffic accidents is classified into six distinct categories in Figure 1. The categories are as follows: very low (164-285 incidents), low (286-406 incidents), moderate (407-527 incidents), moderately high (528-648 incidents), high (649-769 incidents), and very high (770-896 incidents). Areas with the highest incident frequency are marked in red, and the lowest is marked with solid green. In addition, the frequency of accidents shows a pattern of areas with the same category that tend to be clustered.

Almost all regencies/cities located in the east and south of East Java Province are included in areas with a moderate number of motorcycle accidents. Jombang, Sidoarjo, and Surabaya City are three East Java regions with a very high number of accidents. Meanwhile, Bangkalan and Sampang Regencies have a low number of accidents.

3.2. Goodness of Fit Test

For modeling accidents, Poisson distribution is the appropriate choice. This study uses the Kolmogorov-Smirnov test to ascertain data fit to the poisson distribution.

As shown in Table 1, at 0.104, Asymp. Sig. (2-tailed) exceeds the significance level. The data conforms to a Poisson distribution. Tests for the assumptions of non-multicollinearity and equidispersion are conducted after the goodness of fit test has determined that the data results are appropriate poisson distribution.

Table 1. Goodness-of-fit test for Poisson distribution

One-Sample Kolmogorov-Smirnov Test		Y
N		38
Poisson Parameter ^{a,b}	Mean	510.13
Most Extreme Differences	Absolute	.197
	Positive	.197
	Negative	.189
Kolmogorov-Smirnov Z		1.216
Asymp. Sig. (2-tailed)		.104

a. Test distribution is Poisson.
b. Calculated from data.

3.3. Non-Multicollinearity Detection

Non-multicollinearity detection is a procedure utilized to determine whether or not the independent variables comprising a model exhibit correlation. It is presumed that the independent variables exhibit no correlation within the regression model. Multicollinearity can be established through the examination of the correlation matrix values and VIF for each independent variable. If the value of the correlation matrix is relatively modest and the value of the VIF is less than 10, then it is determined that the model does not exhibit multicollinearity.

Table 2. Variance inflation factor (VIF)

Variabel	X ₁	X ₂	X ₃	X ₄	X ₅
VIF	1.091	1.132	1.205	1.212	1.274

According to Table 2, VIF values for each independent variable are all below 10. Therefore, based on the absence of multicollinearity in the observed data, it is possible to proceed with the data analysis using the regression model. In addition, the correlogram demonstrates multicollinearity.

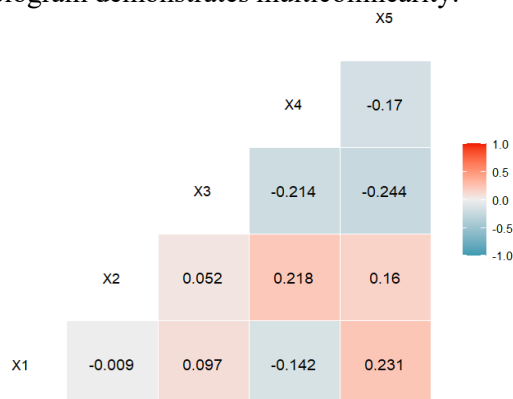


Figure 2. Correlogram with correlation coefficient

The correlation value between independent variables ranges from -0.244 to 0.231, as shown in Figure 2. It indicates that the predictor variables do not exhibit a significant correlation.

3.4. Equidispersion Assumption Test

The poisson regression assumptions are satisfied when both the mean and variance possess an identical value. The equidispersion assumption can be examined by calculating the dispersion value.

Table 3. Overdispersion test

z	dispersion	p-value
3.5296	26.05193	0.0002081

As shown in Table 3, the data regarding the total number of traffic incidents are overdispersed, with a dispersion value of 26.05193 being greater than one. In addition, there is an overdispersion issue that can be known based on the fact that the p-value is below the level of significance ($\alpha = 5\%$).

The variance increases in direct proportion to the mean value, as illustrated in Figure 3. It signifies the failure to satisfy the equidispersion assumption. The study employs LPR and GPR to instances of data overdispersion.

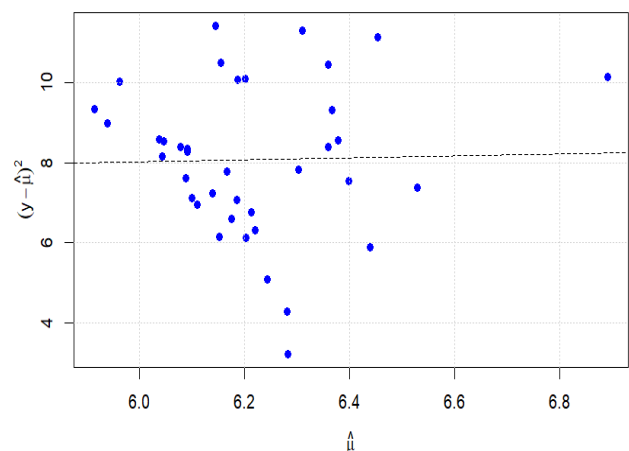


Figure 3. Plot visualization of means against variances

3.5. Estimation of Lagrangian Poisson Regression (LPR)

The LPR parameters were estimated by employing the Maximum Likelihood Estimation (MLE) technique. In accordance with Equation (7), the likelihood function is derived.

$$L(\beta, \theta) = \prod_{i=1}^n \left[e^{-\beta_0 + \sum_{j=1}^k \beta_j x_{ji}} \left(e^{-\beta_0 + \sum_{j=1}^k \beta_j x_{ji}} + \theta y_i \right)^{y_i - 1} \frac{e^{-\left(\beta_0 + \sum_{j=1}^k \beta_j x_{ji} + \theta y_i \right)}}{y_i!} \right] \quad (11)$$

Equation (11) provides the foundation for forming the log-likelihood function as follow.

$$l(\boldsymbol{\beta}, \theta) = \sum_{i=1}^n \left[\beta_0 + \sum_{j=1}^p \beta_j x_{ji} + \ln \left(e^{\beta_0 + \sum_{j=1}^p \beta_j x_{ji}} + \theta y_i \right)^{y_i - 1} - e^{\beta_0 + \sum_{j=1}^p \beta_j x_{ji}} - \theta y_i - \ln(y_i!) \right] \quad (12)$$

Furthermore, equation (12) is maximized for parameter $\beta_0, \beta_1, \dots, \beta_p, \theta$.

Table 4. LPR Model

Variables	Parameter	Standard Error	T value	P value
(Intercept)	4.5111	0.1160	38.8900	0.0000
X_1	0.1029	0.0383	2.6900	0.0072
X_2	0.0585	0.0423	1.3800	0.1665
X_3	0.1065	0.0357	2.9900	0.0028
X_4	-0.0321	0.0450	-0.7100	0.4757
X_5	-0.0968	0.0450	-2.1500	0.0315
AIC	= 485.689		Chi Square	= 21.53
SBC	= 497.152		P value	= 0.000644

The frequency of motorcycle accident model with LPR is given as follows.

$$\hat{\psi} = e^{4.5111 + 0.1029X_1 + 0.0585X_2 + 0.1065X_3 - 0.0321X_4 - 0.0968X_5} \quad (13)$$

Table 4 shows that the percentage of adolescents (X_1), the percentage of low-level education (X_2), the frequency of motorized vehicles (X_3), the length of

The solution will be established by the use of a numerical technique known as the Iteratively Reweighted Least Squares (IRLS), with the assistance of R Studio software for optimization purposes.

roads with good road conditions (X_4), and the average annual rainfall (X_5) significant effect simultaneously. It can be seen from the p-value on the Chi-Square test < level of significance ($\alpha=5\%$). In addition, at a significance level of 5%, the percentage of adolescent age (X_1), the frequency of motorized vehicles (X_3), and the average annual rainfall (X_5) each have a significant effect partially.

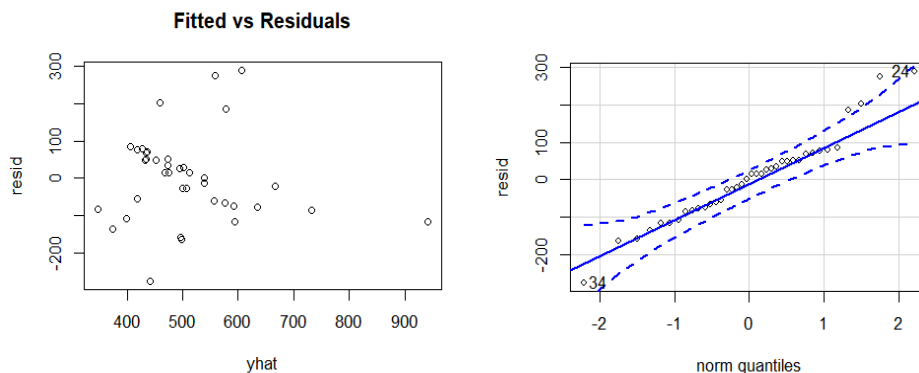


Figure 4. The plot of fitted vs residuals and normal Q-Q for the LPR model

Figure 4 shows that the residuals generated by the Lagrangian Poisson regression spread randomly and are non-patterned, so the regression model can accurately analyze how many motorcycle accidents occur in East Java. Meanwhile, based on the normal Q-Q plot graph, the Lagrangian Poisson regression model with the parameter values that have been obtained provides a good fit for the data and is significant.

3.6. Generalized Poisson Regression (GPR) Estimation

The maximum likelihood estimation (MLE) was also employed to compute the parameter for the GPR model. Furthermore, equation (9) is the basis for forming the likelihood function.

$$L(\boldsymbol{\beta}, \varphi) = \prod_{i=1}^n \left[\frac{e^{\beta_0 + \sum_{j=1}^p \beta_j x_{ji}}}{1 + \varphi e^{\beta_0 + \sum_{j=1}^p \beta_j x_{ji}}} \right]^{y_i} \frac{(1 + \varphi y_i)^{y_i - 1}}{y_i!} e^{-\beta_0 - \sum_{j=1}^p \beta_j x_{ji}} \quad (14)$$

Furthermore, the log-likelihood function is obtained.

$$l(\boldsymbol{\beta}, \varphi) = \sum_{i=1}^n \left[\left(y_i \left(\beta_0 + \sum_{j=1}^p \beta_j x_{ji} \right) + (y_i - 1) \ln(1 + \varphi y_i) - y_i \ln \left(1 + \varphi \exp \left(\beta_0 + \sum_{j=1}^p \beta_j x_{ji} \right) \right) - \frac{\exp \left(\beta_0 + \sum_{j=1}^p \beta_j x_{ji} \right) (1 + \varphi y_i)}{1 + \varphi \exp \left(\beta_0 + \sum_{j=1}^p \beta_j x_{ji} \right)} - \ln(y_i!) \right) \right] \quad (15)$$

Furthermore, the log-likelihood function is maximized to get parameter values. As with LPR, parameter determination is carried out using the

iteratively reweighted least squares (IRLS) algorithm optimization assisted by R Studio software.

Table 5. GPR model

Variables	Parameter	Standard Error	T value	P value
(Intercept)	6.2179	0.0410	151.8300	0.0000
X1	0.0733	0.0446	1.6400	0.1006
X2	0.0911	0.0423	2.1600	0.0311
X3	0.1548	0.0551	2.8100	0.0049
X4	-0.0656	0.0430	-1.5300	0.1272
X5	-0.1000	0.0452	-2.2100	0.0271
AIC	= 487.749		Chi Square	= 19.47
SBC	= 499.212		P value	= 0.00157

Based on Table 5, the frequency of motorcycle accident model with GPR is given as follows.

$$\hat{\sigma} = e^{6.2179+0.0733X_1+0.0911X_2+0.1548X_3-0.0656X_4-0.1000X_5} \quad (16)$$

The chi-square test shows that the p-value is below the predetermined level of significance ($\alpha=5\%$). The percentage of low-level education (X_2), the frequency of motorized vehicles (X_3), and the average annual rainfall (X_5) influence partially on the frequency of motorcycle accidents. Besides that, the percentage of

adolescents (X_1), the percentage of low-level education (X_2), the frequency of motorized vehicles (X_3), the length of roads with good road conditions (X_4), and the average annual rainfall (X_5) have an effect significant simultaneously.

Figure 5 shows that the normal Q-Q plot points are almost linear, so the GPR is still suitable for modeling data. However, the residuals generated by the GPR model are not randomly distributed. It suggests that the GPR model's analysis of the model representing the frequency of motorcycle incidents is less precise.

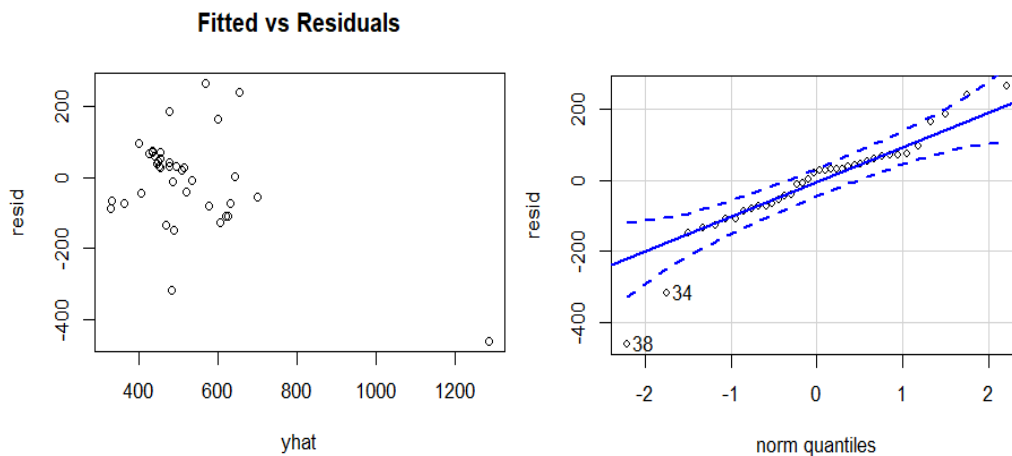


Figure 5. The plot of fitted vs residuals and normal Q-Q for the GPR model

3.7. Best Model Selection

AIC and SBC are used to determine the best model. The good model is based on AIC and SBC with the smallest value. In addition, the frequency of significant parameters is also considered—the AIC and SBC values from the LPR and GPR models are given as follows.

Table 6 shows that LPR has smaller AIC and SBC values than GPR. Thus, The LPR is the most effective model for determining the frequency of motorcycle collisions.

Table 6. Model selection

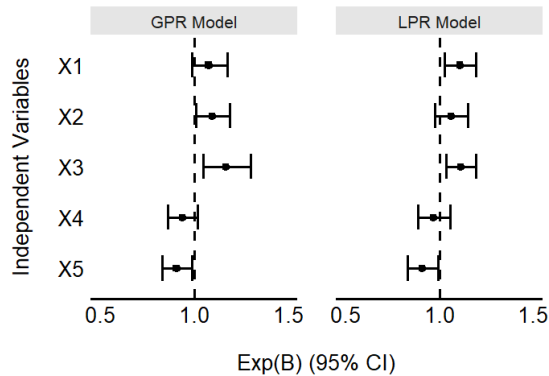
Step	Model	AIC	SBC
1	LPR	485.689	497.152
2	GPR	487.749	499.212

Best model selection can also be made using exponential coefficients of models with 95% CI to make the results of the comparison of the two models even more convincing.

Table 7. Exponential coefficients of models with 95% CI

Variables	LPR Model			GPR Model		
	Exp(B)	95% CI		Exp(B)	95% CI	
		LCI	UCI		LCI	UCI
(Intercept)	91.0258	72.5170	114.2587	501.6449	462.9544	543.5689
X1	1.1084*	1.0283	1.1947	1.0760	0.9859	1.1744
X2	1.0603	0.9759	1.1520	1.0954*	1.0083	1.1900
X3	1.1124*	1.0373	1.1930	1.1674*	1.0479	1.3005
X4	0.9684	0.8867	1.0577	0.9365	0.8607	1.0189
X5	0.9077*	0.8311	0.9914	0.9049*	0.8281	0.9887

* level of significance ($\alpha=5\%$)



A marginal distinction can be observed between the LPR and GPR models, as indicated by the exponential coefficient value accompanied by a 95% confidence interval in Table 7 and Figure 6. If we focus on the 95% confidence interval, the LPR model produces an exponential coefficient value smaller than the GPR model. It indicates that the LPR model exhibits a reduced standard error in comparison to the GPR model. Thus, the LPR model exhibits superior predictive capability to the GPR model in determining the frequency of motorbike accidents in East Java.

Figure 6. Exponential coefficients of models with 95% CI

Table 8. Paired samples test actual-LPR

Paired Samples Test		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	Actual - LPR	-7.526	118.671	19.251	-46.532	31.480	-3.391	37	.008

Table 9. Paired Samples Test Actual-GPR

Paired Samples Test		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	Actual - GPR	-11.658	122.214	19.826	-51.829	28.513	-3.588	37	.013

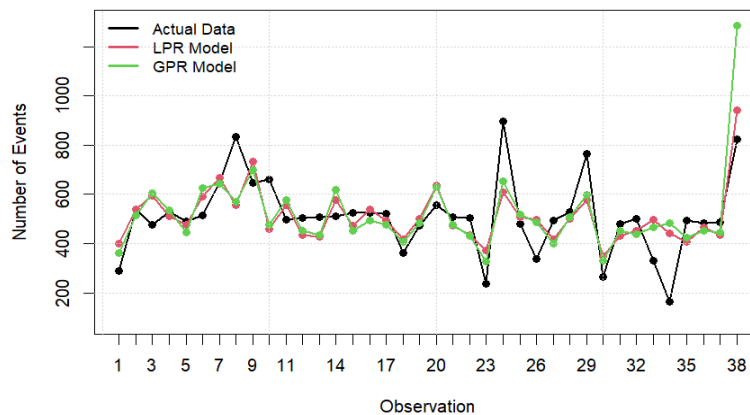


Figure 7. The observations graph on LPR and GPR models against actual data

The value of Sig. (2-tailed) in Tables 8 and Tables 9 are 0.008 and 0.013. The values are smaller than the level of significance ($\alpha=2.5\%$). It means that a discrepancy exists in the average number of accidents between the LPR and GPR models with actual data.

In comparison to the GPR model, the mean paired differences value for the LPR model is less. This indicates that the LPR model more closely approximates the actual data in comparison to the GPR model. The LPR model's accuracy is consistent with findings from research [42], which indicate that the LP is a significantly more effective and user-friendly model for modeling count data. Figure 3 is given to visualize the closeness of the LPR and GPR models to the actual data.

Based on the best model, namely the LPR model, it is known that the percentage of adolescent age (X_1), the frequency of motorized vehicles (X_3), and the average annual rainfall (X_5) have a considerable impact on accident occurrence in East Java. Equation 4 shows that variable X_1 has a coefficient value of $\exp(0.1029)$. It means that the frequency of motorcycle accidents in East Java will increase by $\exp(0.1029) = 1.1083 \approx 1$ event if there is an increase in the percentage of adolescents aged by 1%. Lack of experience and maturity is one of the main reasons for the high risk of accidents among young drivers compared to other age groups. Critical mistakes contributing to 75% of teenage drivers' accidents include failing to check for, detect, and respond to hazards, driving faster than the specified speed limit, and hesitating. Teenagers are more likely to make critical decision-making errors when compared to adult drivers [43], [44].

The coefficient of the variable X_3 is $\exp(0.1065)$. It signifies that for 1 unit increase in the quantity of motorized vehicles in East Java, there would be a commensurate rise of $\exp(0.1065) = 1.109 \approx 1$ motorcycle accident. In accordance with the circumstances in Indonesia, the annual growth rate of motorized vehicles has increased yearly. The occurrence of traffic accidents escalates in direct proportion to the growth in the quantity of cars present on the road, resulting in congestion across various sections of the road, especially during peak hours [45]. This study is in line with research [46], a significant proportion of fatal accidents are caused by a large number of motorized vehicles operating in unfavorable conditions, including tire explosions, which can increase the likelihood of fatal and property-damaging accidents.

The coefficient of the variable X_5 is $\exp(0.1065)$. It means that the frequency of motorcycle accidents in East Java will decrease by $\exp(-0.0968) = 0.907 \approx 1$ event if there is an increase in average annual rainfall of 1 mm³.

Authors of study [47] claim that rainfall is a protective factor in the lower traffic accident rates observed during heavy rains. Possible explanations for this observation are reduced traffic levels during heavy rain [48], [49] or driving slower to reduce risk [50]. Another study [51] shows that postponing driving during periods of heavy precipitation yields an approximation of the decrease in accident risk observed the day following the precipitation event [51].

4. Conclusion

The generalized Poisson distribution family can be done to solve overdispersion issues in applied data. Based on the AIC and SBC values, LPR has a lower value than GPR, so the LPR model is the most effective approach for assessing the frequency of motorcycle accidents in East Java. The examination of the LPR model's standard error value, which is relatively lower in comparison to the GPR model, can provide insight into its effectiveness. The LPR model is smaller than the GPR model, as indicated by the exponential coefficient value accompanied by a 95% confidence interval. Mean paired differences value is less for the LPR model than for the GPR model in the paired samples test. It suggests that the LPR model is more accurate in representing the real data compared to the GPR model. Therefore, this study determines that the LPR model is the most optimal approach for constructing models to predict the frequency of motorcycle accidents in East Java. Furthermore, the percentage of adolescent age (X_1), the frequency of motorized vehicles (X_3), and the average annual rainfall (X_5) are the variables that have a significant effect on the occurrence of accidents.

Acknowledgments

We are grateful to the Ministry of Education, Culture, Research, and Technology for financing the research for our doctoral dissertation in 2023.

References:

- [1]. World Health Organization. (2018). Road Traffic Injuries. Retrieved from: <https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries> [accessed: 20 July 2023].
- [2]. Atubi, A. (2012). Determinants of road traffic accident occurrences in Lagos State: Some lessons for Nigeria. *International Journal of Humanities and Social Science*, 2(6), 252-259.
- [3]. Gebru, M. K. (2017). Road traffic accident: Human security perspective. *International Journal of Peace and Development Studies*, 8(2), 15-24.
- [4]. Andarini, D., Camelia, A., & Ibrahim, M. M. (2021). Factors related to road accidents in Palembang, South Sumatera, Indonesia. *International Journal of Public Health Science (IJPHS)*, 10(3), 638-645.
- [5]. BPS-Statistics Indonesia. (2019). *Land transportation statistics 2018*. Keselamatanjalan. Retrieved from: <https://keselamatanjalan.files.wordpress.com/2020/07/bps-statistik-transportasi-darat-2018-1.pdf> [accessed: 03 August 2023].
- [6]. Zhao, J., Zhang, F., Zhao, C., Wu, G., Wang, H., & Cao, X. (2020). The properties and application of Poisson Distribution. *Journal of Physics: Conference Series*, 1550(3), 032109.
- [7]. Kumar, P., Jain, J. K., & Singh, G. (2022). Accident prediction modelling for expressways: A review. *IOP Conf. Series: Materials Science and Engineering*, 1236(1), 012011.
- [8]. Harris, T., Yang, Z., & Hardin, J. W. (2012). Modeling underdispersed count data with Generalized Poisson Regression. *The Stata Journal*, 12(4), 736-747.
- [9]. Obubu, M., & Nwokolo, P. C. (2016). Prevalence of breast cancer in Delta State, Nigeria. *World Journal of Probability and Statistics*, 2(2), 1-9.
- [10]. Osuji, G. A., Obubu, M., & Obiora-Ilouno, H. O. (2016). Uterine fibroid on women's fertility and pregnancy outcome in Delta State, Nigeria. *Journal of Natural Sciences Research*, 6(2), 27-33.
- [11]. Berliana, S. M., Puhadi, Sutikno, & Rahayu, S. P. (2019). Multivariate Generalized Poisson Regression model with exposure and correlation as a function of covariates: Parameter estimation and hypothesis testing. *AIP Conference Proceedings*. 2192(1), 090001.
- [12]. Kurnia, A., & Sadik, K. (2020). Analysis of overdispersed count data by Poisson model. *European Journal of Molecular and Clinical Medicine*, 7(10), 1400-1409.
- [13]. Gelfand, A. E., & Dalal, S. R. A. (1990). A note on overdispersed Exponential families. *Biometrika*, 77(1), 55-64.
- [14]. Kokonendji, C. C., Dossou-Gbété, S., & Demétrio, C. G. B. (2004). Some discrete exponential discrete models: Poisson-Tweedie and Hinde-Demétrio Classes. *Statist. Oper. Research Trans*, 28(2), 201-214.
- [15]. Lord, D., Washington, S. P., & Ivan, J. N. (2005). Poisson, Poisson-Gamma and Zero-Inflated Regression Models of motor vehicle crashes: Balancing statistical fit and theory. *Accident Analysis and Prevention*, 37(1), 35-46.
- [16]. Kokonendji, C. C., Demétrio, C. G. B., & Zocchi, S. S. (2007). On Hinde-Demétrio Regression models for overdispersed count data. *Statist. Methodology*, 4(3), 277-291.
- [17]. Maher, M., & Mountain, L. (2009). The sensitivity of estimates of regression to the mean. *Accident Analysis and Prevention*, 41(4), 861-868.
- [18]. Kumar, C. N., Paridaa, M., & Jain, S. S. (2013). Poisson Family Regression techniques for prediction of crash counts using Bayesian Inference. *Procedia-Social and Behavioral Sciences*, 104, 982-991.
- [19]. Zhu, F. (2012). Modeling overdispersed or underdispersed count data with Generalized Poisson Integer-Valued GARCH models. *J. Math. Anal. Appl*, 389(1), 58-71.
- [20]. Ridout, M. S., & Besbeas P. (2004). An empirical model for underdispersed count data. *Statistical Modelling*, 4(1), 77-89.
- [21]. del-Castillo, J., & Perez-Casany, M. (2005). Overdispersed and underdispersed Poisson generalizations. *Journal of Statistical Planning and Inference*, 134(2), 486-500.
- [22]. Hayati, M., Sadik, K., & Kurnia, A. (2018). Conway-Maxwell Poisson distribution: Approach for over-and-under-dispersed count data modelling. *IOP Conf. Series: Earth and Environmental Science*, 187(1), 012039.
- [23]. Consul, P. C. (1989). *Generalized Poisson Distribution: Properties and Applications*. Marcel Dekker.
- [24]. Kaviyarasu, V., & Devika, V. (2017). Designing of Special Type of Double Sampling Plan for Compliance Testing through Generalized Poisson Distribution. *International Journal of Pure and Applied Mathematics*, 117(12), 7-17.
- [25]. Famoye, F., & Consul, P. C. (1990). Interval estimation and hypothesis testing for the generalized Poisson distribution. *American Journal of Mathematical and Management Sciences*, 10, 127-158.
- [26]. Gossiaux, A., & Lemaire, J. (1981). Methodes d'ajustement de distributions de sinistres. *Bulletin of the Association of Swiss Actuaries*, 81, 87-95.
- [27]. Consul, P. C., & Famoye, F. (1992). Generalized Poisson Regression. *Communications in Statistics - Theory and Methods*, 21(1), 89-109.
- [28]. Famoye, F., Wulu, J. T., & Singh, K. P. (2004). On the Generalized Poisson Regression model with an application to accident data. *J. Data Sci*, 2(3), 287-295.
- [29]. Maxwell, O., Mayowa, B. A., Chinedu, I. U., & Peace A. E. (2018). Modelling count data: A Generalized Linear Model framework. *American Journal of Mathematics and Statistics*, 8(6), 179-183.

- [30]. Salah, J., Rehman, H. U., & Al Buwaiqi, I. (2023). Inclusion Results of a Generalized Mittag-Leffler-Type Poisson Distribution in the k-Uniformly Janowski Starlike and the k-Janowski Convex Functions. *Statistics*, 11(1), 22-27.
- [31]. Al-Eid, M., & Shoukri, M. M. (2021). Inference procedures on the Generalized Poisson Distribution from multiple samples: Comparisons with nonparametric models for Analysis of Covariance (ANCOVA) of count data. *Open Journal of Statistics*, 11(3), 420-436.
- [32]. Fauwziyah, F., Astutik, S., & Pramoedyo, H. (2022). Geographically Weighted Negative Binomial Regression Modeling using Adaptive Kernel on the frequency of Maternal Deaths during Childbirth. *Mathematics and Statistics*, 10(5), 1133-1139.
- [33]. Hilbe, J. M. (2011). *Negative Binomial Regression*. Cambridge University Press.
- [34]. Shrestha, N. (2020). Detecting multicollinearity in regression analysis. *American Journal of Applied Mathematics and Statistics*, 8(2), 39-42.
- [35]. Daoud, J. I. (2017). Multicollinearity and regression analysis. *Journal of Physics: Conf. Series*, 949, 012009.
- [36]. Belsley, D. A. (1991). *Conditioning Diagnostics: Collinearity and Weak Data in Regression*. John Wiley & Sons.
- [37]. Consul, P. C., & Famoye, F. (2006). *Lagrangian Probability Distributions*. Birkhäuser: New York.
- [38]. Haris, M., & Arum. P. (2022). Negative Binomial Regression and Generalized Poisson Regression models on the frequency of traffic accidents in Central Java. *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, 16(2), 471-482.
- [39]. Grover, G., Vajala, R. & Swai, P. K. (2015). On the assessment of various factors effecting the improvement in CD4 count of aids patients undergoing antiretroviral therapy using Generalized Poisson regression. *Journal of Applied Statistics*, 42, 1291-1305,
- [40]. Ikbal, N. A. M., Halim, S. A., & Ali, N. (2022). Estimating Weibull Parameters Using Maximum Likelihood Estimation and Ordinary Least Squares: Simulation Study and Application on Meteorological Data. *Mathematics and Statistics*, 10(2), 269-292.
- [41]. Stasinopoulos, M., Bob, R., & Calliope. A. (2008). *Instruction on How to Use the GAMLSS package in R (2nd ed)*. Storm Research Center Metropolitan University.
- [42]. Janardan, K. G., Kerster, H. W., & Schaeffer, D. J. (1979). Biological Applications of the Lagrangian Poisson Distribution. *BioScience*, 29(10), 599-602.
- [43]. McDonald, C. C., Curry, A. E., Kandadai, V., Sommers, M. S., & Winston, F. K. (2014). Comparison of teen and adult driver crash scenarios in a nationally representative sample of serious crashes. *Accident Analysis and Prevention*, 72, 302-308.
- [44]. Duddu, V. R., Kukkapalli, V. M., & Pulugurtha, S. S. (2019). Crash risk factors associated with injury severity of teen drivers. *IATSS Research*, 43(1), 37-43.
- [45]. Sun, L. L., Liu, D., Chen, T., & He, M. T. (2019). Road traffic safety: An analysis of the cross-effects of economic, road and population factors. *Chin. J. Traumatol*, 22(5), 290-295.
- [46]. Ali, S. I. A., Elturki, F. A. A., & Jibrel, S. N. (2020). Analysis of increment of road traffic accidents in Libya: Case study city of Tripoli. *IOP Conf. Series: Materials Science and Engineering*, 800, 012003.
- [47]. Sangkharat, K., Thornes, J. E., Wachiradilok, P., & Pope, F. D. (2021). Determination of the impact of rainfall on road accidents in Thailand. *Heliyon*, 7(2), e06061.
- [48]. Bergel-Hayat, R., Debbarh, M., Antoniou, C., & Yannis, G. (2013). Explaining the road accident risk: weather effects. *Accident Analysis and Prevention*, 60, 456-465.
- [49]. Yannis, G., & Karlaftis, M. G. (2010). *Weather Effects on Daily Traffic Accidents and Fatalities: A Time Series Count Data Approach*. Nrso. Retrieved from: <https://www.nrso.ntua.gr/geyannis/wp-content/uploads/geyannis-pc102.pdf> [accessed: 11 August 2023].
- [50]. Brodsky, H., & Hakkert, A. S. (1988). Risk of a road accident in rainy weather. *Accident Analysis and Prevention*, 20(3), 161-176.
- [51]. Brijs, T., Karlis, D., & Wets, G., (2008). Studying the effect of weather conditions on daily crash counts using a discrete time-series model. *Accident Analysis and Prevention*, 40(3), 1180-1190.

Appendix

The Source Code of Comparing Two Models (*Exponential Coefficients of Models with 95% CI*)

```

# Compare Two Models
# -----
VGAM::lrtest(lpr.fit, gpr.fit) # Likelihood ratio test

# Exp of Coefficients - lpr.fit
(exp.lpr = round(exp(coef(lpr.fit,matrix = TRUE)), 4))
(exp.confint.lpr = round(exp(confint(lpr.fit, matrix = TRUE)), 4))
# Exp of Coefficients - gpr.fit
(exp.gpr = round(exp(coef(gpr.fit,matrix = TRUE)), 4))
(exp.confint.gpr = round(exp(confint(gpr.fit, matrix = TRUE)), 4))

options(scipen=100, digits=6)
options(max.print = 500000000)

# Data for 2 forest plot (Estimates are Made up) ####
DV <- c("LPR Model", "LPR Model", "LPR Model", "LPR Model", "LPR Model",
        "GPR Model", "GPR Model", "GPR Model", "GPR Model", "GPR Model") # Heading for Facet Wrap
IV <- c("X1", "X2", "X3", "X4", "X5", "X1", "X2", "X3", "X4", "X5") # Independent variable names
ES <- c(exp.lpr[-1,1], exp.gpr[-1,1]) # b Estimate (could be standardized estimate, Odds Ratio,
Incident Rate Ratio, etc.)
LCI <- c(exp.confint.lpr[-c(1:2),1], exp.confint.gpr[-c(1:2),1]) # Lower 95% confidence interval
UCI <- c(exp.confint.lpr[-c(1:2),2], exp.confint.gpr[-c(1:2),2]) # Upper 95% confidence interval

A <- data.frame(DV, IV, ES, LCI, UCI)
A$IV <- factor(A$IV, levels=c("X5", "X4", "X3", "X2", "X1"))

# 2 Forest Plots ####
ggplot(data=A, aes(x=IV, y=ES, ymin=LCI, ymax=UCI)) +
  geom_pointrange()+ # Makes range for ggplot values based on the data and AES specified in first
  line
  geom_hline(yintercept=1, lty=2, size =1) + # add a dotted line at x=0 after flip
  geom_errorbar(aes(ymin=LCI, ymax=UCI), width=0.5, cex=1)+ # Makes whiskers on the range (more
  aesthetically pleasing)
  facet_wrap(~DV)+ # Makes DV header (Can handle multiple DVs)

coord_flip() + # flip coordinates (puts labels on y axis)
geom_point(shape = 15, size = 2) + # specifies the size and shape of the geompoint
ggtitle("")+ # Blank Title for the Graph
xlab("Independent Variables") + # Label on the Y axis (flipped specification do to coord_flip)
ylab("Exp(B) (95% CI)") + # Label on the X axis (flipped specification do to coord_flip)
scale_y_continuous(limits = c(0.5,1.5), breaks = c(0.5,1.0,1.5))+ # limits and tic marks on
X axis (flipped specification do to coord_flip)
theme(line = element_line(colour = "black", size = 1), # My personal theme for GGplots
      strip.background = element_rect(fill="gray90"),
      legend.position = "none",
      axis.line.x = element_line(colour = "black"),
      axis.line.y = element_blank(),
      panel.border= element_blank(),
      panel.grid.major = element_blank(),
      panel.grid.minor = element_blank(),
      panel.background = element_blank(),
      panel.spacing = unit(2, "lines"), # added to theme to add space inbetween facet_wrap plots
      axis.ticks = element_blank(),
      axis.title.x = element_text(family="Times New Roman",colour = "Black", margin = margin(t =
20, r = 0, b = 0, l =0)),
      axis.title.y = element_text(family="Times New Roman",colour = "Black", margin = margin(t = 0,
r = 20, b = 0, l = 0)),
      plot.title = element_text(family="Times New Roman", colour = "Black", margin = margin(t = 0,
r = 0, b = 20, l = 0)),
      axis.text=element_text(family="Times New Roman",size=16, color = "Black"),
      text=element_text(family="Times New Roman",size=16), plot.margin = margin(t = 2, r = 2, b = 2,
l = 2, unit = "cm"))

```

