Zero : Jurnal Sains, Matematika, dan Terapan
E-ISSN : 2580-5754; P-ISSN : 2580-569X
Volume 10, Number 1, 2026
DOI: 10.30829/zero.v10i1.28453
Page: 303 - 314

Comparison of Pure Premiums for Motor Vehicle Insurance Using ZTP-Gamma GLM and Tweedie GLM

[1] **Yushinta Cahya Lestari**
Department of Actuarial Science, Sumatera Institute of Technology, Lampung, 35365, Indonesia

[2] **Tiara Yulita**
Department of Actuarial Science, Sumatera Institute of Technology, Lampung, 35365, Indonesia

[3] **Amalia Listiani**
Department of Actuarial Science, Sumatera Institute of Technology, Lampung, 35365, Indonesia

Article Info

Article history:

Accepted 22 April 2026

Keywords:

Generalized Linear Models;
Motor Vehicle Insurance;
Pure Premium;
Tweedie Distribution;
Zero-Truncated Poisson.

ABSTRACT

The increasing number of motor vehicles has contributed to higher traffic density and a greater risk of accidents, thereby reinforcing the importance of protection through motor vehicle insurance. Therefore, accurately determining the pure premium is essential to maintain risk balance and ensure the sustainability of insurance companies. This study employs Generalized Linear Models, which are an extension of classical linear regression that allow the response variable to follow non-normal distributions, particularly the Zero-Truncated Poisson, Gamma, and Tweedie distributions. Using motor vehicle insurance claim data from 2022 with 386 observations, this research compares two premium modeling approaches, namely the ZTP–Gamma model for estimating claim frequency and claim severity, and the Tweedie GLM for modeling total claims in the calculation of pure premiums for motor vehicle insurance. The analysis shows that the estimated pure premiums for the ZTP–Gamma GLM range from IDR 2,138,532 to IDR 19,939,391, while the estimates for the Tweedie GLM range from IDR 2,153,665 to IDR 20,936,047. The ZTP–Gamma GLM demonstrates better accuracy, with a MAPE value of 23.65% compared to 25.844% for the Tweedie GLM, resulting in an accuracy difference of 2.194%. These findings indicate that the ZTP–Gamma GLM is more effective in producing accurate pure premium estimates.

Corresponding Author:

Tiara Yulita,
Department of Actuarial Science
Sumatera Institute of Technology, Lampung, Indonesia
Email: tiara.yulita@at.itera.ac.id

1. INTRODUCTION

Insurance is an agreement between the policyholder and an insurance company. In general, insurance is divided into two types, namely life insurance and general insurance. Life insurance is insurance that provides a specified payment upon the death of the insured to the family members or beneficiaries entitled to receive it in accordance with the terms of the insurance contract [1]. Meanwhile, general insurance is a form of insurance business that provides protection services against specific risks in the event of incidents that may cause damage, loss, or loss of profit [2]. One form of general insurance is motor vehicle insurance.

Motor vehicle insurance provides financial protection against losses suffered by the insured due to the risk of damage or loss arising from various events related to motor vehicle ownership [3]. The increase in the number

of motor vehicles each year also increases the risk of traffic accidents [4]. Based on data from the Central Bureau of Statistics, the number of traffic accidents in Indonesia in 2022 was recorded at 139,258 cases, representing an increase of 35.61% compared to 2021 (Badan Pusat Statistik). In the following period, data from the Indonesian National Police show that as of August 5, 2024, there had been 79,220 traffic accident cases, indicating a decrease of 28.34% compared to 2023 (Korlantas Polri). This condition emphasizes the importance of motor vehicle insurance due to the high number of traffic accidents. There are two types of coverage in motor vehicle insurance, namely Total Loss Only (TLO) and Comprehensive. TLO coverage provides compensation if the level of damage reaches more than 75% of the sum insured, while Comprehensive coverage covers various types of losses in accordance with the provisions stated in the policy [5].

A request for compensation submitted by the policyholder to the insurance company when a loss covered by the policy occurs is referred to as a claim. The insurance company will pay the claim in accordance with the agreed terms and conditions. Claim payments represent the form of financial protection provided by insurance, thereby helping the insured minimize losses resulting from events covered by the policy. The magnitude of claim payments makes risk management and premium determination very important for insurance companies.

Accurate premium determination is essential to ensure the sustainability of insurance companies and maintain financial stability. Companies can also prepare funds for future claim payments and reduce the risk of policy cancellations by customers, which may disrupt business continuity. Accurate premium setting requires methods capable of capturing the characteristics of claim data, which are generally non-normally distributed and may be either discrete or continuous. One widely used approach is the Generalized Linear Model (GLM) [6].

GLM is a statistical modeling framework that extends classical linear regression by allowing the response variable to follow a non-normal distribution from the exponential family, such as Zero-Truncated Poisson, Gamma, and Tweedie distributions. Claim frequency data are discrete and have a minimum value of one, making the Poisson distribution less appropriate because it still assigns a positive probability to zero values. Therefore, the Zero-Truncated Poisson distribution is used to model claim frequency [7]. Claim severity is continuous, positive, and skewed, making the Gamma distribution a suitable choice [8]. Meanwhile, total claims require an approach capable of handling data that combine discrete and continuous components; thus, the Tweedie distribution is used because it can represent these characteristics [9]. The use of these three distributions enables GLM to provide more accurate premium estimates compared to classical linear regression [10].

Several previous studies have applied the Generalized Linear Models (GLM) approach to claim data modeling and insurance premium calculation. A study conducted in 2023 by Ratna Zafira Hafidzah on motor vehicle insurance showed that claim frequency was modeled using the Zero-Truncated Poisson distribution and claim severity was modeled using the Gamma distribution. However, the data used were obtained from insurance companies in the United States, and therefore may not be directly applicable to the insurance industry in Indonesia. Another study by Tri Andika Julia Putra, Donny Lesmana, and I Gusti Putu Purbana in 2021 used the Tweedie distribution to model total motor vehicle claims. This study employed only the Tweedie distribution approach without comparing it to other models, so the level of accuracy of the model could not be determined with certainty.

This study aims to calculate pure premiums for motor vehicle insurance using Generalized Linear Models with motor vehicle insurance claim data from one insurance company in Indonesia. This study includes five additional predictor variables, namely the sum insured, vehicle age, type of vehicle usage, vehicle type, and vehicle usage region under Comprehensive coverage. This research also seeks to complement previous studies by comparing two premium modeling approaches, namely the ZTP-Gamma model and the Tweedie model. The results of this study are expected to provide insurance companies with options in selecting the most appropriate model for premium calculation, as well as contribute to the development of knowledge in the fields of actuarial science and insurance.

2. RESEARCH METHOD

2.1 Generalized Linear Models

Generalized Linear Models (GLM) extend linear regression by allowing the response variable to follow a distribution from the exponential family and relating its mean to a linear predictor through a link function [11]. For an observation i, the GLM is defined as [12]:

$$g(\mu_i) = \boldsymbol{X_i^T \beta} \tag{1}$$

or equivalently,

$$\mu_i = g^{-1}(\boldsymbol{X_i^T \beta}) \tag{2}$$

where $\mu_i = E(Y_i)$ denotes the expected value of the response variable, $X_i = (1, X_{i1}, \dots, X_{ip})^T$ is the vector of predictor variables, $\beta = (\beta_0, \beta_1, \dots, \beta_p)^T$ is the vector of regression parameters, and $g(\cdot)$ represents the link function connecting the mean response to the linear predictor, which is specified according to the assumed response distribution. In this study, the GLM framework is employed to model insurance claim data exhibiting non-normal characteristics.

2.2 ZTP-GLM for Claim Frequency Modeling

Claim frequency is modeled using a Zero-Truncated Poisson Generalized Linear Model (ZTP–GLM), as the observed claim count data do not contain zero values. For the i-th observation, the expected claim frequency is defined as [7]:

$$\mu_i = \frac{\exp(\boldsymbol{X_i^T \alpha})}{1 - \exp(-\exp(\boldsymbol{X_i^T \alpha}))} \tag{3}$$

where $\mu_i = E(N_i)$ denotes the expected claim frequency, X_i is the vector of predictor variables, and α is the vector of regression parameters. The log link function is employed to ensure that the predicted mean of the claim frequency remains positive.

2.3 Gamma-GLM for Claim Severity Modeling

Claim severity is modeled using the Gamma distribution under the GLM framework [13]. The use of the log link function produces a linear predictor and ensures a positive fitted mean. For the i-th observation, the expected claim severity is defined as [14]:

$$\mu_i = \exp(\boldsymbol{X_i^T \beta}) \tag{4}$$

where $\mu_i = E(Y_i)$ denotes the expected claim severity, X_i is the vector of predictor variables, and β is the vector of regression parameters. The log link function is employed to ensure that the predicted mean remains positive.

2.4 Tweedie-GLM for Aggregate Claim Modeling

Aggregate insurance claims are modeled using a Tweedie Generalized Linear Model (Tweedie GLM), which allows simultaneous modeling of claim frequency and claim severity within a unified framework. The Tweedie distribution is characterized by a power parameter $p \in (1,2)$, corresponding to a compound Poisson–Gamma structure [9]. For the i-th observation, the expected aggregate claim amount is defined as [15]:

$$\mu_i = \exp(\boldsymbol{X_i^T \gamma}) \tag{5}$$

where $\mu_i = E(S_i)$ denotes the expected aggregate claim amount, X_i is the vector of predictor variables, and γ is the vector of regression parameters. The log link function is employed to ensure that the predicted mean remains positive.

2.5 Correlation Analysis

Correlation analysis is performed among predictor variables to examine the degree of association and to identify potential multicollinearity prior to model estimation [16]. Spearman's rank correlation is applied to numerical predictor variables, while Cramer's V is used for categorical predictor variables [17][18]. The response variables are not included in the correlation analysis, as the objective is to assess relationships among explanatory variables before fitting the Generalized Linear Models.

2.6 Goodness-of-Fit Tests

Goodness-of-fit tests are conducted to determine whether the observed response variables follow the assumed theoretical distributions. This step is important to ensure that the selected distributions are appropriate for modeling the data. In this study, goodness-of-fit testing is applied to the response variables prior to GLM estimation [19]. The Kolmogorov–Smirnov and Anderson–Darling tests are employed to evaluate the suitability of the assumed distributions by comparing the empirical distribution of the data with the corresponding theoretical distribution at a given significance level.

2.6.1 Kolmogorov-Smirnov Test

The Kolmogorov–Smirnov (KS) test is used to examine whether the observed data follow a specified theoretical distribution. The KS test statistic is defined as [20]:

$$D = \max|S(x) - F(x)| \tag{6}$$

where $S(x)$ denotes the empirical cumulative distribution function and $F(x)$ represents the theoretical cumulative distribution function. The decision rule is based on the associated $p - value$, where the null hypothesis is rejected when the $p - value$ is less than the specified significance level.

2.6.2 Anderson-Darling Test

The Anderson–Darling (AD) test is also applied to assess the goodness-of-fit of the assumed distribution. Compared to the KS test, the Anderson–Darling test places greater emphasis on the tail behavior of the distribution. The AD test statistic is given by [21]:

$$A^2 = -m - \frac{1}{m}\sum_{i=1}^{m}(2i-1)[\ln(F(z_i) + \ln(1 - F(z_{m-i+1}))] \tag{7}$$

where m denotes the sample size and $F(\cdot)$ is the cumulative distribution function of the assumed distribution. The null hypothesis is rejected when the calculated test statistic exceeds the corresponding critical value.

2.6.3 Estimation of the Tweedie Power Parameter

The Tweedie distribution is characterized by a power parameter p, which determines the relationship between the mean and variance. Several special cases of the Tweedie family can be identified based on the value of p, as summarized in Table 1 [22].

Table 1. Power Parameter p

p	Distribution
$p = 0$	Normal
$p = 1$	Poisson
$1 < p < 2$	Tweedie
$p = 2$	Gamma
$p = 3$	Inverse Gaussian

Since the power parameter p cannot be obtained directly, it is estimated using the Maximum Likelihood Estimation (MLE) approach. Candidate values of p are evaluated, and the value that maximizes the log-likelihood is selected as the optimal power parameter for the Tweedie model.

2.7 Parameter Estimation

Parameter estimation for the ZTP, Gamma, and Tweedie models is carried out using the Maximum Likelihood Estimation (MLE) method. The MLE approach estimates model parameters by maximizing the corresponding likelihood function based on the observed data [23]. Due to the complexity of the likelihood functions, closed-form solutions are not available for the models considered in this study. Therefore, parameter estimation is performed numerically using the Fisher scoring algorithm [24]. This iterative optimization procedure updates parameter estimates until convergence is achieved, resulting in stable and efficient parameter estimates for the GLM framework. Parameter significance was evaluated using Wald statistics, and non-significant predictors were excluded to obtain the final model. In addition to point estimation, 95% confidence intervals are constructed for each parameter to assess estimation uncertainty. Based on the asymptotic normality of the MLE, the confidence interval is computed as:

$$\hat{\beta} \pm 1.96 \times SE(\hat{\beta}) \tag{8}$$

where $SE(\hat{\beta})$ denotes the standard error of the estimated parameter.

2.8 Model Goodness-of-Fit (Deviance Test)

Model goodness-of-fit is assessed to evaluate the adequacy of each fitted model in representing the observed data. In this study, the goodness-of-fit of the ZTP–GLM, Gamma–GLM, and Tweedie GLM models is evaluated using the deviance statistic. The deviance statistic compares the fitted model with the corresponding saturated model. A model is considered to provide an adequate fit when the deviance indicates no significant lack of fit at the chosen significance level [25].

2.9 Mean Absolute Percentage Error (MAPE)

The Mean Absolute Percentage Error (MAPE) is used to measure the accuracy of the predicted pure premium by comparing the predicted values with the actual observed values. In this study, MAPE is employed to compare the performance of the ZTP–Gamma GLM and the Tweedie GLM models. The MAPE is defined as [26]:

$$MAPE = \frac{1}{N_e} \sum_{t=1}^{N_e} \left| \frac{A_t - P_t}{A_t} \right| \times 100\% \tag{9}$$

where A_t denotes the actual pure premium value, P_t represents the predicted pure premium value, and N_e is the total number of observations. Lower MAPE values indicate better predictive accuracy and are used as the basis for selecting the preferred model.

2.10 Root Mean Square Error (RMSE)

In addition to the Mean Absolute Percentage Error (MAPE), this study also uses the Root Mean Square Error (RMSE) to evaluate model performance. RMSE measures the square root of the average squared differences between the observed and predicted values. The RMSE is defined as [27]:

$$RMSE = \sqrt{\frac{1}{Ne}\sum_{t=1}^{Ne}(A_t - P_t)^2} \tag{10}$$

where A_t denotes the actual pure premium value, P_t represents the predicted pure premium value, and N_e is the total number of observations. A smaller RMSE value indicates better predictive accuracy.

2.11 Data and Methods

This study employs a quantitative research design with a statistical modeling approach. The analysis is conducted using the Generalized Linear Models (GLM) framework to estimate and compare pure premiums of motor vehicle insurance. Three response variables are considered: claim frequency modeled using a Zero-Truncated Poisson (ZTP) model, claim severity modeled using a Gamma model, and aggregate claims modeled using a Tweedie model.

The data used in this study consist of motor vehicle insurance claim records obtained from a general insurance company in Indonesia for claims occurring in 2022. The sample includes 386 claim observations from policyholders who filed claims during the observation period. Several predictor variables related to policy and vehicle characteristics are considered in the modeling process. The response and predictor variables used in this study are summarized in Table 1.

Prior to model estimation, exploratory analyses are conducted to examine relationships among predictor variables using correlation analysis. The suitability of the assumed distributions for claim frequency, claim severity, and aggregate claims is then evaluated using goodness-of-fit tests, including the Kolmogorov–Smirnov and Anderson–Darling tests.

In this study, no specific outlier detection or treatment procedures are applied. Extreme claim values are retained in the dataset because large claim amounts are inherent in insurance data and reflect the heavy-tailed nature of claim severity distributions. Removing or modifying such observations could potentially distort the underlying risk characteristics captured by the model.

Model parameters for the ZTP–GLM, Gamma–GLM, and Tweedie GLM are estimated using the Maximum Likelihood Estimation (MLE) method, with numerical optimization performed via the Fisher scoring algorithm. The pure premium is obtained from the fitted models, where the ZTP–Gamma GLM computes the premium as the product of the expected claim frequency and expected claim severity, while the Tweedie GLM directly estimates the expected aggregate claim amount.

Model adequacy is assessed using deviance-based goodness-of-fit measures, and the predictive performance of the ZTP–Gamma GLM and Tweedie GLM is compared using the Mean Absolute Percentage Error (MAPE). All data processing and statistical analyses are performed using RStudio, with additional support from EasyFit 5.5 and Microsoft Excel. In this study, no specific outlier detection or treatment procedures are applied. All observations are included in the modeling process, and parameter estimation is conducted based on the original data.

The data used in this study are limited to insurance claims recorded in a single observation year (2022). Therefore, the analysis does not capture multi-year trends or incorporate exposure measures such as policy-years. In addition, the sample consists only of policyholders who filed claims during the observation period, and no random sampling procedure is applied. As a result, potential selection bias may arise, since policies without claims are not included in the analysis.

Table 2. Description of Response and Predictor Variables

Symbol	Variable Description
N	Claim frequency
Y	Claim severity
S	Aggregate claims
X_1	Sum insured
X_2	Vehicle age
X_3	Vehicle usage type
X_4	Vehicle type
X_5	Area of vehicle usage

3. RESULT AND ANALYSIS

3.1 Descriptive Statistics

This subsection presents descriptive statistics of the response variables used in the study. Table 3 summarizes the minimum, maximum, mean, and standard deviation of claim frequency, claim severity, and aggregate claims. The results indicate that the claim-related variables exhibit right-skewed distributions, as

reflected by the large dispersion between minimum and maximum values. These characteristics support the use of non-normal modeling approaches such as the ZTP, Gamma, and Tweedie models within the GLM framework.

Table 3. Descriptive Statistics of Response Variables

	Claim Severity	Claim Frequency	Aggregate Claims
Minimum	Rp266.067	1	Rp266.067
Maximum	Rp10.618.479	3	Rp20.223.592
Mean	Rp3.797.715	1	Rp4.454.017
Standard Deviation	Rp2.231.046	1	Rp3.221.345

In addition, the claim frequency variable shows limited variation, with a mean of approximately one claim and a maximum value of three claims. This pattern reflects the nature of the dataset, which only includes policyholders who submitted claims during the observation period. Although the variability is relatively small, the Zero-Truncated Poisson model remains appropriate for modeling count data that exclude zero observations.

3.2 Correlation Analysis Results

This subsection presents the results of the correlation analysis among the predictor variables used in this study. Correlation analysis is conducted to determine the degree of relationship between each predictor variable prior to model estimation. Spearman's rank correlation coefficient is used to measure the relationship between numerical variables, while Cramer's V is applied to measure the relationship between categorical variables. The correlation results among the predictor variables are presented in Table 4.

Table 4. Correlation Results among Predictor Variables

Variable Pair	Correlation Coefficient
X_1 and X_2	-0.418
X_1 and X_3	-0.012
X_1 and X_5	-0.004
X_2 and X_3	0.075
X_2 and X_4	0.095
$X_2$and X_5	0.082
$X_3$and X_4	0.055
X_3 and X_5	0.066
X_4 and X_5	0.044

Based on the results presented in Table 3, the correlation values among the predictor variables are relatively low. This indicates that there is no strong correlation among the predictor variables used in the study. Therefore, the predictor variables can be included simultaneously in the Generalized Linear Models without causing multicollinearity problems.

3.3 Distribution Fit Results

This subsection presents the results of the distributional fit analysis for the response variables used in this study, namely claim severity, claim frequency, and aggregate claims. Distributional fit analysis is conducted to determine whether the assumed probability distributions are suitable for modeling the response variables prior to parameter estimation. The suitability of the distribution for claim severity is evaluated using the Kolmogorov–Smirnov test, while the suitability of the distribution for claim frequency is examined using the Anderson–Darling test. In addition, the Tweedie distribution is evaluated for aggregate claims based on the estimated value of the power parameter p.

3.3.1 Kolmogorov-Smirnov Test Results for Claim Severity

The Kolmogorov–Smirnov test is used to assess the suitability of the Gamma distribution for modeling claim severity. The test results are presented in Table 5.

Table 5. Kolmogorov-Smirnov Test Results for Claim Severity

Kolmogorov-Smirnov Test	Calculated statistic	Critical value	p-value	Decision
Gamma distribution	0.038366	0.06923	0.6208	Fail to reject H_0

Based on the results presented in Table 5, the p-value obtained from the Kolmogorov–Smirnov test indicates that the Gamma distribution is suitable for modeling the claim severity data. Therefore, the Gamma distribution is used in the subsequent GLM modeling for claim severity.

3.3.2 Anderson-Darling Test Results for Claim Frequency

The Anderson–Darling test is applied to evaluate the suitability of the Zero-Truncated Poisson distribution for modeling claim frequency. The results of the Anderson–Darling test are presented in Table 6.

Table 6. Anderson-Darling Test Results for Claim Frequency

Anderson-Darling Test	Calculated statistic	Critical value	p-value	Decision
ZTP distribution	0.09298	2.492	0.902	Fail to reject H_0

Based on the results shown in Table 6, the p-value obtained from the Anderson–Darling test indicates that the Zero-Truncated Poisson distribution is appropriate for modeling claim frequency. Thus, the Zero-Truncated Poisson distribution is used in the GLM framework for claim frequency.

3.3.3 Tweedie Distribution Parameter Results for Aggregate Claims

The suitability of the Tweedie distribution for modeling aggregate claims is evaluated based on the estimated value of the power parameter p. The estimated value of the Tweedie parameter is presented in Table 7.

Table 7. Estimated Tweedie Power Parameter for Aggregate Claims

p	Description
1.5469	Tweedie distribution

Based on the estimated value of the Tweedie power parameter presented in Table 7, the parameter p lies within the interval $1 < p < 2$. This result indicates that the Tweedie distribution is appropriate for modeling aggregate claims, as it accommodates data with a mixed discrete–continuous structure.

3.4 Parameter Estimation Results

This subsection presents the results of parameter estimation for the fitted Generalized Linear Models. Parameter estimation is conducted to identify predictor variables that significantly affect claim frequency, claim severity, and aggregate claims. The estimated parameters of the ZTP–GLM, Gamma–GLM, and Tweedie GLM models are obtained using the Maximum Likelihood Estimation method. Only statistically significant predictors retained after the refitting process are reported.

3.4.1 Parameter Estimation Results for Claim Severity

The parameter estimation results for the claim severity model using the Gamma distribution are presented in Table 8. The 95% confidence intervals show that all retained predictors have intervals that do not include zero, confirming their statistical significance at the 5% level.

Table 8. Final Parameter Estimation Results for Claim Severity

Variable	Coefficient	Std. Error	95% CI Lower	95% CI Upper	p-value
Intercept	15.2408079	1.291×10^{-1}	14.9878	15.4938	2×10^{-16}
X_1	2.847×10^{-9}	2.979×10^{-10}	2.263×10^{-9}	3.431×10^{-9}	2×10^{-16}
X_{41}	-7.360×10^{-1}	9.329×10^{-2}	-0.9188	-0.5532	3.27×10^{-14}
X_{42}	-3.698×10^{-1}	1.21×10^{-1}	-0.6070	-0.1326	0.00241
X_{51}	-1.241×10^{-1}	5.755×10^{-2}	-0.2369	-0.0113	0.03170
X_{52}	-1.696×10^{-1}	6.884×10^{-2}	-0.3045	-0.0347	0.01420

Based on the results presented in Table 8, the significant predictor variables influence the expected claim severity. These results form the basis for estimating the expected claim severity in the ZTP–Gamma GLM framework.

3.4.2 Parameter Estimation Results for Claim Frequency

The parameter estimation results for the claim frequency model using the Zero-Truncated Poisson distribution are presented in Table 9. The 95% confidence intervals show that all retained predictors have intervals that do not include zero, confirming their statistical significance at the 5% level.

Table 9. Final Parameter Estimation Results for Claim Frequency

Variable	Coefficient	Std. Error	95% CI Lower	95% CI Upper	p-value
Intercept	−3.9087	0.5763	−5.0382	−2.7792	1.19×10^{-11}
X_{31}	3.5738	0.6110	2.3762	4.7714	4.95×10^{-9}
X_{32}	4.4477	0.5962	3.2791	5.6163	8.61×10^{-14}

Based on the results presented in Table 9, several predictor variables significantly affect claim frequency. These variables are retained in the final ZTP–GLM model and used to estimate the expected claim frequency.

3.4.3 Parameter Estimation Results for Aggregate Claims

The parameter estimation results for the aggregate claim model using the Tweedie distribution are summarized in Table 10. The 95% confidence intervals show that all retained predictors have intervals that do not include zero, confirming their statistical significance at the 5% level.

Table 10. Final Parameter Estimation Results for Aggregate Claims

Variable	Coefficient	Std. Error	95% CI Lower	95% CI Upper	p-value
Intercept	15.260	1.27×10^{-1}	15.0111	15.5089	2×10^{-16}
X_1	2.490×10^{-9}	3.079×10^{-10}	1.887×10^{-9}	3.093×10^{-9}	8.44×10^{-15}
X_{31}	4.287×10^{-1}	7.074×10^{-2}	0.2900	0.5674	3.28×10^{-9}
X_{32}	8.75×10^{-1}	8.057×10^{-2}	0.7171	1.0329	2×10^{-16}
X_{41}	-7.294×10^{-1}	8.53×10^{-2}	−0.8966	−0.5622	3.09×10^{-16}
X_{42}	-3.267×10^{-1}	1.142×10^{-1}	−0.5505	−0.1029	0.00447
X_{51}	-6.080×10^{-2}	6.1×10^{-2}	−0.1804	−0.0588	0.031957
X_{52}	-1.570×10^{-1}	7.34×10^{-2}	−0.3009	−0.0131	0.03309

Based on the results shown in Table 10, the retained predictor variables significantly affect aggregate claims. The final Tweedie GLM model is subsequently used to estimate the pure premium directly from aggregate claims.

3.5 Model Goodness-of-Fit Results

This subsection presents the results of the goodness-of-fit evaluation for the fitted Generalized Linear Models. The goodness-of-fit of each model is assessed using the deviance test to examine the suitability of the fitted models in representing the observed data. The deviance test results for the Gamma GLM, Zero-Truncated Poisson GLM, and Tweedie GLM models are summarized in Table 11.

Table 11. Deviance Test Results for Fitted Models

Model	Deviance	df	p-value
Gamma GLM	96.1445	382	1.0000000
ZTP GLM	84.0913	384	1.0000000
Tweedie GLM	276.3033	381	0.9999801

Based on the results presented in Table 11, the deviance values obtained for all fitted models are accompanied by p-values greater than the significance level of 0.05. This indicates that there is no evidence of lack of fit for the Gamma GLM, Zero-Truncated Poisson GLM, and Tweedie GLM models. Therefore, all models are considered suitable for modeling claim severity, claim frequency, and aggregate claims, respectively. Overall, the goodness-of-fit results confirm that the fitted models adequately represent the observed data and can be used for subsequent pure premium estimation and model comparison.

The extremely high p-values observed in some models, particularly the Gamma GLM, may be influenced by the relatively limited variability in the dataset and the moderate sample size. In such cases, deviance-based goodness-of-fit tests may yield values close to 1 without necessarily indicating model misspecification. Therefore, the results should be interpreted together with other evaluation measures and the underlying characteristics of the data.

3.6 Pure Premium Estimation Results

Pure premium estimation is conducted as a continuation of the modeling stage based on the significant parameters and goodness-of-fit results obtained previously. In this study, pure premium estimation is carried out using two approaches, namely the ZTP–Gamma GLM and the Tweedie GLM.

3.6.1. Pure Premium Estimation Using ZTP-Gamma GLM

Pure premium estimation using the ZTP–Gamma GLM is performed through a two-stage approach by combining the claim severity model and the claim frequency model. The pure premium is obtained by multiplying the expected claim severity and the expected claim frequency derived from the fitted models. Based on the Gamma GLM and Zero-Truncated Poisson GLM, the pure premium is expressed as:

$$E\left[\sum_{i=1}^{N} Y_i\right] = E[Y] \times E[N] \tag{11}$$

where $E(Y_i)$ denotes the expected claim severity and $E(N_i)$ denotes the expected claim frequency. The expected claim severity obtained from the Gamma GLM is given by:

$$E(Y_i) = \exp\left(\beta_0 + \beta_1 X_1 + \sum_{j=1}^{2} \beta_{4j} X_{4j} + \sum_{j=1}^{2} \beta_{5j} X_{5j}\right) \tag{12}$$

while the expected claim frequency obtained from the Zero-Truncated Poisson GLM is given by:

$$E(N_i) = \left(\frac{\exp\left(\alpha_0 + \sum_{j=1}^{2} \alpha_{3j} X_{3j}\right)}{1 - \exp\left(-\exp\left(\alpha_0 + \sum_{j=1}^{2} \alpha_{3j} X_{3j}\right)\right)}\right) \tag{13}$$

Substituting Equations (12) and (13) into Equation (11) yields the following pure premium formulation.

$$E\left[\sum_{i=1}^{N} Y_i\right] = \exp\left(\beta_0 + \beta_1 X_1 + \sum_{j=1}^{2} \beta_{4j} X_{4j} + \sum_{j=1}^{2} \beta_{5j} X_{5j}\right) \times \left(\frac{\exp\left(\alpha_0 + \sum_{j=1}^{2} \alpha_{3j} X_{3j}\right)}{1 - \exp\left(-\exp\left(\alpha_0 + \sum_{j=1}^{2} \alpha_{3j} X_{3j}\right)\right)}\right) \tag{14}$$

3.6.2. Pure Premium Estimation Using Tweedie GLM

Pure premium estimation using the Tweedie GLM is conducted based on the aggregate claim model. In this approach, the pure premium is directly estimated from the expected value of total claims obtained from the fitted Tweedie GLM. The pure premium formulation based on the Tweedie GLM is expressed as:

$$E[S] = \exp\left(\gamma_0 + \gamma_1 X_1 + \sum_{j=1}^{2} \gamma_{3j} X_{3j} + \sum_{j=1}^{2} \gamma_{4j} X_{4j} + \sum_{j=1}^{2} \gamma_{5j} X_{5j}\right) \tag{15}$$

3.7 Model Comparison Using MAPE and RMSE

This subsection presents the comparison of the ZTP–Gamma GLM and the Tweedie GLM in estimating the pure premium of motor vehicle insurance. The comparison is conducted using two accuracy measures, namely the Mean Absolute Percentage Error (MAPE) and the Root Mean Square Error (RMSE). MAPE measures the average percentage deviation between the actual and predicted pure premiums, while RMSE evaluates the magnitude of prediction errors in the original scale. Lower values of both measures indicate better predictive performance. The MAPE and RMSE values for the ZTP–Gamma GLM and the Tweedie GLM are presented in Table 12.

Table 12. MAPE and RMSE Results for Pure Premium Estimation

Model	MAPE	RMSE
ZTP-Gamma GLM	23.650%	Rp2.075.831
Tweedie GLM	25.844%	Rp2.251.805

Based on the results shown in Table 12, the ZTP–Gamma GLM produces lower MAPE and RMSE values compared to the Tweedie GLM. The MAPE difference of 2.194% and the lower RMSE value indicate that the ZTP–Gamma GLM provides more accurate and stable pure premium estimates. Therefore, the ZTP–Gamma GLM demonstrates superior predictive performance in this study. Although the MAPE values fall within the moderate accuracy range, they remain within the category of reasonable forecasting performance according to commonly used forecasting accuracy criteria. In actuarial applications, prediction errors in the range of 20%–50% are often considered acceptable due to the inherent variability and uncertainty present in insurance claim data.

4. CONCLUSION

This study aims to estimate and compare pure premiums for motor vehicle insurance using Generalized Linear Models, specifically the Zero-Truncated Poisson–Gamma (ZTP–Gamma) GLM and the Tweedie GLM, based on insurance claim data from 2022. The results show that both models are capable of capturing variations in pure premiums according to insured risk characteristics. Factors such as vehicle usage type, vehicle type, and usage region significantly influence premium levels. In particular, vehicles used for commercial purposes tend to

have higher pure premiums, while jeep-type vehicles and vehicles operated in region I are associated with higher premium estimates.

Based on model accuracy evaluation using the Mean Absolute Percentage Error (MAPE) and the Root Mean Square Error (RMSE), the ZTP–Gamma GLM demonstrates better predictive performance than the Tweedie GLM. This finding suggests that separating claim frequency and claim severity may provide more accurate pure premium estimates for zero-truncated insurance claim data.

From a practical perspective, these findings provide useful insights for insurance companies in Indonesia in determining more accurate pure premium estimates based on policyholder risk characteristics. The application of the ZTP–Gamma modeling approach may support more data-driven pricing strategies and improve risk classification in motor vehicle insurance.

This study is subject to several limitations. The modeling framework is based on Generalized Linear Models and does not incorporate exposure measures or credibility adjustments. In addition, comparisons with alternative machine learning approaches are not considered. Furthermore, the sample consists of 386 claim observations, which may limit the stability of parameter estimates, particularly for rare events. Future research may extend this study by incorporating exposure information, exploring alternative modeling techniques, and applying Generalized Linear Mixed Models to account for potential dependence among insurance claims.

5. REFERENCES

[1] D. Ekawati and Faedinah, "Penentuan Cadangan Premi Asuransi Jiwa Bersama Dwiguna dengan Metode Canadian," *J. Math. Theory Appl.*, vol. 2, no. 1, pp. 2–5, 2020.

[2] N. Nafis, Saryadi, and A. Wijayanto, "Pendahuluan Kajian Teori," *J. Ilmu Adm. Bisnis*, vol. 12, no. 3, pp. 748–757, 2023.

[3] C. K. Waha, A. J. Rindengan, and T. Manurung, "d ' CartesiaN Model Distribusi Data Klaim Asuransi Mobil untuk Menentukan Premi Murni," *J. Mat. dan Aplilkasi*, vol. 8, no. 2, pp. 108–113, 2019.

[4] H. S. Rotua Tinambunan, B. Waskito, M. B. Rizhaldi, and A. F. K.R. Uno, "Asuransi Kecelakaan Kendaraan Bermotor Roda Dua Sebagai Moda Transportasi Umum Berbasis Online," *J. Huk. Ius Quia Iustum*, vol. 26, no. 3, pp. 627–649, 2020, doi: 10.20885/iustum.vol26.iss3.art10.

[5] S. F. Sari and M. S. Ubay, "Tweedie Generalized Linear Models Pada Penentuan Premi Asuransi Kendaraan Bermotor," *J. Stat. Teor. dan Apl. Biomed. Ind. Bus. Soc. Stat.*, vol. 12, no. 1, pp. 1–9, 2018, [Online]. Available: biastatistics.statistics.unpad.ac.id

[6] P. D. England. and R. J. Verrall., "Stochastic claims reserving in general insurance," *Br. Actuar. J.*, vol. 8, no. 3, pp. 443–544, 2002.

[7] R. Z. Hafidzah, "Pemodelan Data Klaim Asuransi Kendaraan Bermotor Menggunakan Generalized Linear Models (GLM)," Universitas Indonesia, 2023.

[8] J. A. Nelder. and R. W. M. Wedderburn., "Generalized linear models," *J. R. Stat. Soc. Ser. A*, vol. 135, no. 3, pp. 370–384, 1972.

[9] T. A. J. Putra, D. C. Lesmana, and I. G. P. Purnaba, "Penghitungan Premi Asuransi Kendaraan Bermotor Menggunakan Generalized Linear Models dengan Distribusi Tweedie," *Jambura J. Math.*, vol. 3, no. 2, pp. 115–127, 2021, doi: 10.34312/jjom.v3i2.10136.

[10] M. Denuit., X. Marechal., S. Pitrebois., and J. F. Walhin., *Actuarial Modelling of Claim Counts*. Wiley, 2007.

[11] W. Romadhona and V. Syavera, "Jurnal Sains Ekonomi dan Edukasi Generalized Linear Model Menggunakan Distrbusi Lognormal dan Gamma : Aplikasi Terhadap Indeks Demokrasi Indonesia di Jawa Barat," vol. 2, no. 1, pp. 117–126, 2025.

[12] A. C. Cameron. and P. K. Trivedi., *Regression Analysis of Count Data*. Cambridge University Press, 2013.

[13] R. Sardiani, W. Somayasa, and L. Gubu, "Estimasi Parameter Distribusi Gamma pada Data Tersensor Tipe II Progresif Menggunakan Metode Likelihood Maksimum," *J. Nas. Has. Penelit. Bid. Multidisiplin*, vol. 1, no. 2, pp. 77–92, 2024, [Online]. Available: http://journal.unnes.ac.id/sju/index.php/ujm

[14] A. M. Ramlan and Irmasari, "Pembelajaran Distribusi Gamma dalam Masalah Biomedis: Tinjauan pada Pengaruh Dosis Beracun pada Tikus," *J. Res. Sci. Math. Educ.*, vol. 3, no. 2, pp. 111–119, 2024, doi: 10.56855/jrsme.v3i2.1069.

[15] S. A. Klugman., H. H. Panjer., and G. E. Willmot., *Loss Models From Data To Decisions*, 5th ed. Society Of Actuaries, 2019.

[16] F. Jabnabillah and N. Margina, "Analisis Korelasi Pearson Dalam Menentukan Hubungan Antara Motivasi Belajar Dengan Kemandirian Belajar Pada Pembelajaran Daring," *J. Sintak*, vol. 1, no. 1, pp. 14–18, 2022.

[17] D. Mustofani and H. Hariyani, "Penerapan Uji Korelasi Rank Spearman Untuk Mengetahui Hubungan Tingkat Pengetahuan Ibu Terhadap Tindakan Swamedikasi Dalam Penanganan Demam Pada Anak," *Unisda J. Math. Comput. Sci.*, vol. 9, no. 1, pp. 9–13, 2023, doi: 10.52166/ujmc.v9i1.4272.

[18] A. A. Hutagalung, A. Rahayu, D. Anggitasyah, F. A. Yunisa, Q. P. Andini, and R. F. Sari, "Mengukur Tingkat Efektivitas Google Drive Dengan Uji Chi Square Dan Cramer (C) Dalam Pengarsipan Dokumen Amdal," *J. Garuda Pengabdi. Kpd. Masy.*, vol. 1, no. 1, pp. 16–21, 2023, [Online]. Available: https://journal.aira.or.id/index.php/gabdimas/article/view/600

[19] D. Febriani, S. Sugito, and A. Prahutama, "Analisis Metode Bayesian Menggunakan Non-Informatif Prior Uniform Diskrit Pada Sistem Antrean Pelayanan Gerbang Tol Muktiharjo," *J. Gaussian*, vol. 10, no. 3, pp. 337–345, 2021, doi: 10.14710/j.gauss.v10i3.32783.

[20] R. Amelia, W. Somayas, Alfian, and Ruslan, "Uji Goodness Of Fit Untuk Distribusi Geometrik Menggunakan Uji Statistik Kolmogorov-Smirnov," *J. Mat. Komputasi dan Stat.*, vol. 4, no. 1, pp. 491–497, 2024.

[21] G. D. Ahadi and N. N. L. E. Zain, "Pemeriksaan Uji Kenormalan dengan Kolmogorov-Smirnov, Anderson-Darling dan Shapiro-Wilk," *Eig. Math. J.*, vol. 6, no. 1, pp. 11–19, 2023, doi: 10.29303/emj.v6i1.131.

[22] B. Jørgensen., "Exponential dispersion models," *J. R. Stat. Soc. Ser. B*, vol. 49, no. 2, pp. 127–162, 1987.

[23] T. Yulita, M. Patricia, and A. S. E. Hidayat, "Penentuan Premi Murni Dari Data Klaim Asuransi Kendaraan Roda Empat Dengan Jenis Perlindungan Comprehensive," *Var. J. Stat. Its Appl.*, vol. 6, no. 1, pp. 75–86, 2024, doi: 10.30598/variancevol6iss1page75-86.

[24] Y. Widyaningsih, H. N. Rizka, and T. Siswantining, "Performance comparison between maximum likelihood estimation and variational method for estimating simple linear regression parameter," *ITM Web Conf.*, vol. 61, p. 01010, 2024, doi: 10.1051/itmconf/20246101010.
[25] B. Jørgensen., *The Theory of Dispersion Models*. Chapman & Hall, 1997.
[26] M. A. Kurniawan and A. Solichin, "Peramalan Persediaan Sparepart Menggunakan Metode Double Exponential Smoothing Pada Pt. Mayora Indah Tbk," *J. Commun. Educ.*, vol. 15, no. 1, 2021, doi: 10.58217/joce-ip.v15i1.227.
[27] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 3rd ed. OTexts, 2021.