Advances in Social Science, Education and Humanities Research, volume 550

Proceedings of the 1st International Conference on Mathematics and Mathematics Education (ICMMED 2020)

Prediction of Future Insurance Premiums When the Model is Uncertain

Tri Andika Julia Putra[1] Donny Citra Lesmana[1,*] I Gusti Putu Purnaba[1]

[1] *Department of Mathematics, IPB University, Bogor, Indonesia*
[*]*Corresponding author. Email: donnylesmana@apps.ipb.ac.id*

ABSTRACT

It is an important task for an actuary in determining an appropriate premium price for each customer with different risks and characteristics. The purpose of this study is to determine the best model for pure general insurance premiums and variables that can affect the amount of pure premiums. One of statistical analyzes that can be used to model insurance premiums is Generalized Linear Models (GLM). GLM is an extension of the classic regression model that can accommodate the flexibility of its users to use multiple data distributions, but is limited to the exponential family distribution. In the GLM model the premium is obtained by multiplying the conditional expected value from frequency of claims and cost of claims. Based on the research that has been done, it is found that frequency of claims follows the Poisson distribution. Meanwhile, cost of claim follows the Normal distribution. From the two models, it is found that the variables that affect the pure premium are the type of work, the reason for the claim, the location of residence, the marital status and the class of the customer's vehicle. It indicates that the GLM model is a representative model and useful for the insurance company business.

***Keywords:** Premiums of insurance, Generalized linear models, Frequency of claims, Cost of claim.*

1. INTRODUCTION

Recently, the growth of general insurance in Indonesia has increased. It is due to a significant increase in the number of vehicles in the community. The effect is a higher risk of accidents. The basic role of insurance is to provide financial protection, by offering a method of risk transfer through cost coverage in the event of an accident. However, to be able to claim against the risk coverage given, customers are required to make premium payments during the coverage period. Basically, the risk experienced by each individual is different, so it is natural that each insured pays a different premium, according to the high risk faced. The difference in the cost of premiums is based on the heterogeneity of the insurance portfolio which leads to the anti-selection phenomenon [1]. In order to reduce the occurrence of this anti-selection phenomenon, it is necessary to make efforts to determine the general insurance premium rate, by dividing the insurance portfolio into sub-portfolios. This sub-portfolio division is based on certain influencing factors. Therefore, each class will contain policyholders who pay the same premium rates with the same risk profile.

One method that can be used in determining the premium rate is by multiplying the conditional expected value from the frequency of claims with the cost of claims, considering the observed risk characteristics [2]. One of the important roles of an actuary is to evaluate risk in the framework of determining the insurance premium rate. This is because actuaries will propose and apply statistical models that vary from time to time according to the characteristics of the observed data. In this context, the linear regression used to evaluate the impact of the influencing explanatory variables is Generalized Linear Models (GLM) [3]. GLM enables non-linear behavior modeling without having to pay attention to some burdensome assumptions according to existing data [4]. This aspect is very useful for general insurance analysis because the frequency of claims and the cost of claims can of course follow a more diverse distribution. The development of GLM itself has contributed to improving the quality of the risk prediction model and the fair rate-setting process based on the risk characteristics. The purpose of this study is to analyze the variables that affect general insurance premiums and to form the best model for calculating general insurance premiums.

2. DETAILS EXPERIMENTAL

2.1. Generalized Linear Models (GLM)

Linear regression analysis is a modeling with a statistical approach to determine the relationship between one variable and another, namely the response variable (the affected variable) and the explanatory variable (the causal variable). The regression approach in theory gets good results when the response variables are normally distributed and the data variance (data diversity) is constant. Generalized Linear Models (GLM) can overcome when the response variables are not normally distributed and the diversity of data is constant [5]. GLM is defined as an extension of linear regression using an exponential family distribution. The purpose of this GLM model is to estimate the response variable (Y) which depends on the explanation of the explanatory variables (X).

The observation variable Y which has an exponential family distribution has the following probability function [6]:

$$f(y|\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right), \quad y \in S \tag{1}$$

where y is the response variable, θ is the canonical variable, ϕ is the scale variable, and S is the subset of the set of natural numbers ($\mathbb{N}$) or real numbers ($\mathbb{R}$). Meanwhile $b(\theta)$ and $c(y,\phi)$ are known functions. Some examples of members of the exponential family distribution are the Normal, Binomial, Poisson, Geometric, Negative Binomial, Exponential, Gamma, and Inverse Gaussian distributions. The exponential family distribution applies: $E(y) = b'(\theta)$ and $Var(y) = \phi b"(\theta)$ [6].

As with linear regression models in general, the objective of the GLM model is to determine the conditional expected value of the response variable using existing observational data. In this case, the parameters $\beta_1, \beta_2, \dots, \beta_n$ will be determined through the link function of the mean value of the explanatory variable (μ_i), which can be written as a linear combination of the explanatory variable x_i as follows:

$$g(\mu_i) = \beta_0 + \sum_{j=1}^{n} \beta_j x_{ij} = \boldsymbol{x_i^T \beta} = \eta_i, \quad i = 1,2,\dots,p \tag{2}$$

where $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_n \end{pmatrix}$ is the column vector of size $((n+1)\times 1)$, $\boldsymbol{x_i} = \begin{pmatrix} 1 \\ x_{i1} \\ \dots \\ x_{in} \end{pmatrix}$ is the vector column of size $((n+1)\times 1)$, n is the number of explanatory variables, p is the number of observations, g is a monotonous and differentiable function and is called a link function. This function relates the linear predictor η_i to the value μ_i. In this study, the link function used is the canonical link function g, which is a function that satisfies $g(\mu) = \theta$. Table 1 presents some canonical link functions for several exponential family distributions.

Table 1. The canonical link function of the exponential family distribution

Distribution of y	$g(\mu)$	**Link Function**
Normal	μ	Identity
Binomial	$\ln\frac{\mu}{n-\mu}$	Logit
Gamma $(p = -1)$	μ^p	Negative inverse
Eksponensial $(p = -1)$	μ^p	Negative inverse
Geometrik	$\ln\frac{\mu - 1}{\mu}$	Logit
Poisson	$\ln \mu$	Log
Binomial Negatif	$\ln\frac{k\mu}{1+k\mu}$	Logit
Inverse Gaussian $(p = -2)$	μ^p	Inverse squared

2.2. Estimation of Claim Frequency Model Parameters

In general insurance, many have proven that in GLM, the calculation of the frequency of claims follows the Poisson distribution. The Poisson model as an archetype of modeling the "count of events", or commonly known in the actuarial literature as the frequency of claims [7]. The Poisson model is the main tool for modeling the frequency of claims in general insurance [8]–[11].

Discrete random variable Y is said to be Poisson random variable with parameter λ, if for $\lambda > 0$ satisfies the following probability mass function:

$$f(Y = y) = \frac{e^{-\lambda}\lambda^y}{y!} \tag{3}$$

In general, the expected value and variance of the Poisson distribution are the same, namely λ [12].

$$E(Y = y) = Var(Y = y) = \lambda. \tag{4}$$

By using the canonical link function for the Poisson distribution, namely the log link function then the claim frequency model is obtained as follows:

$$g(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\beta} \quad \leftrightarrow \quad \ln \mu_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$$

$$\leftrightarrow \quad \mu_i = \lambda_i = e^{x_i^T\boldsymbol{\beta}} \tag{5}$$

The standard estimate for this model is the maximum likelihood estimation. The likelihood function is defined as follows:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{e^{-\lambda_i}\lambda_i^y}{y!} = \prod_{i=1}^{n} \frac{e^{e^{-x_i^T\boldsymbol{\beta}}}\left(e^{-x_i^T\boldsymbol{\beta}}\right)^y}{y!}. \tag{6}$$

Based on equation (6), the log-likelihood function is obtained as follows:

$$LL(\boldsymbol{\beta}) = \ln L(\boldsymbol{\beta}) = \sum_{i=1}^{n}[y \ln \lambda_i - \lambda_i - \ln y!]$$

$$= \sum_{i=1}^{n}\left[y\boldsymbol{x}_i^T\boldsymbol{\beta} - \boldsymbol{x}_i^T\boldsymbol{\beta} - \ln y!\right]. \tag{7}$$

The likelihood estimator will be maximum if it meets the requirement that the first derivative of the log-likelihood function is zero and the condition is sufficient that the second derivative of the log-likelihood function is negative. It can be easily verified that the first partial derivative of the log-likelihood function exists and is expressed as follows:

$$\frac{\partial LL(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{p}(y - \lambda_i)x_{ij}$$

$$= \sum_{i=1}^{p}\left(y - e^{x_i^T\boldsymbol{\beta}}\right)x_{ij}, \quad \text{for } j = 1, 2, .., n \tag{8}$$

While the second derivative of the log-likelihood function is as follows:

$$\frac{\partial^2 LL(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{p} -\lambda x_{ij} = \sum_{i=1}^{p} -e^{x_i^T\boldsymbol{\beta}}x_{ij}, \quad \text{for } j = 1, 2, .., n \tag{9}$$

The maximum likelihood estimate of $\widehat{\beta_j}$ is the solution of equation (8) obtained from the condition that the first derivative of the log-likelihood function is zero. The equation is difficult to solve analytically, therefore it needs to be solved numerically using an iterative algorithm. The most commonly used iterative method is Newton-Raphson.

2.3. Estimation of Claim Cost Model Parameters

The cost of claims (or the economic compensation payable) is more difficult to predict than the frequency of claims. In this case the analysis is less clear because there is no distribution for positive real values. In some literature it is stated that the classic method that allows econometric modeling of the claim cost is the Gamma model. But there is a possibility of using the GLM model with a different distribution, because the data for each company is different. The following are some distributions for the continuous random variable that might fit the claim size data.

- Normal Distribution

Positive random variable Y is normally distributed if it follows the probability density function as follows:

$$f(y) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right), \quad -\infty < y < \infty \tag{10}$$

where the expected value and variance are $E(y) = \mu$ and $Var(y) = \sigma^2$ [6].

- Exponential Distribution

The exponential distribution is a continuous distribution which is usually used to model time-lapse cases between two occurrences of an event. A continuous random variable Y is said to have an exponential distribution if it satisfies the following probability density function:

$$f(y) = \lambda\, e^{-\lambda y}, y > 0 \tag{11}$$

where $\lambda > 0$ is the distribution rate parameter [13]. With the expected value and variance are, respectively $E(y) = \frac{1}{\lambda}$ and $Var(y) = \frac{1}{\lambda^2}$.

- Gamma distribution

The gamma distribution is an extension of the exponential distribution. Both are expressed in terms of the gamma function defined by:

$$f(y) = \frac{1}{\Gamma(v)} \left(\frac{vy}{\mu}\right)^v \exp\left(-\frac{vy}{\mu}\right)\frac{1}{y}, \quad y > 0 \tag{12}$$

where the expected value and variance are $E(y) = \mu$ and $Var(y) = \frac{\mu^2}{v}$ [6].

- Inverse Gaussian Distribution

Positive random variable Y has inverse gaussian distribution if it follows the probability density function as follows:

$$f(y) = \frac{1}{\sqrt{2\pi y^3}\sigma} \, exp\left(-\frac{1}{2y}\left(\frac{y-\mu}{\mu\sigma}\right)^2\right), \quad y,\mu,\lambda > 0 \tag{13}$$

where the expected value and variance are $E(y) = \mu$ and $Var(y) = \mu^3\sigma^2$ [6].

2.4. Pure Premium Calculation

In general insurance, pure premium represents the estimated cost of all claims made by policyholders during the coverage period. The calculation of premiums is based on a statistical method that combines all available information about the risks received, thereby providing a more accurate assessment of the rates charged to each insured [1].

Pure premium is the expected value of the amount of annual claims stated by the policyholder and is obtained by multiplying the expected value of the frequency of claims with the expected value of the claim cost. The estimated claim frequency with the claim cost can be stated as follows:

$$E\left[\sum_{i=1}^{N} C_i\right] = E[N] \times E[C_i] \tag{14}$$

where N represents the number of claims and C_i represents the claim size of the policy, where the number of claims (N) is independent of the number of claims (C_i) [14]. As pointed out by Charpentier and Denuit (2005), the separate approach to frequency and size of claims is particularly relevant because the risk factors affecting the two components of the pure premium are usually different. In essence, a separate analysis of the two phenomena provides a clearer perspective on how risk factors affect the premium.

3. RESULTS AND DISCUSSION

3.1. Data Implementation

The empirical analysis in this study used statistical software SPSS version 26 and EasyFit 5.5 Professional. EasyFit software is used to determine the distribution of the response variable while SPSS software is used to determine the estimation of model parameters. The proposed method can also be used in other branches of general insurance (car insurance, building and fire insurance, travel insurance, etc.), taking into account the associated contract features.

In this paper, the data used is the United States auto insurance portfolio containing 9134 policies downloaded from the website http://dyzz9obi78pm5.cloudfront.net/app/image/id/560ec66d32131c9409f2ba54/n/Auto_Insurance_Claims_Sample.csv, accessed at 15 October 2020. The elements included in the policy are the factors that are considered in this study. Thus, apart from the frequency and cost of claims variables, other variables are considered as risk factors, which are known at the beginning of policy formation by the insurer and are used to adjust the risk profile of each insured. This explanatory variable reflects the characteristics of the insured which includes, among others: city code, type of insurance (premium, basic, extended), education level, type of work, gender, income per month, location of residence (urban, rural), marital status, length of time. as a customer, the type of policy, the reasons for making a claim, the channel to be a policy participant, the vehicle class and the size of the vehicle.

3.2. Result

3.2.1. Model of Claim Frequency

In SPSS software, there are tools that can be used to estimate model parameters by providing information in the form of data distribution and link functions to be used. It is used to fit the regression model in the GLM framework. The Type III analysis, which is produced using this procedure, allows the evaluation of the contribution of each variable taking into account all other explanatory variables. At the initial stage, the partial likelihood ratio test will be carried out. This test aims to see the overall variables that affect the model. The results of the Type III analysis are presented in Table 2. In the likelihood ratio column, calculated for each variable, twice the difference between the log-likelihood model which includes all explanatory variables and the log-likelihood of the model obtained by removing one of the variables that has no significant effect.

Table 2. Partial likelihood ratio test to claim frequency

Parameters	Type III	
	Likelihood Ratio	**Sig.**
Intercept	337.164	0.000
City code	9.174	0.057
Type of insurance	11.094	0.004
Education level	3.282	0.350
Type of work	27.859	0.000
Gender	0.019	0.891
Income per month	20.497	0.015
Location of residence	1.580	0.454
Marital status	1.495	0.473
Old as a customer	10.517	0.310
Policy type	16.994	0.009
Reason for claim	88.409	0.000
Policy channels	2.689	0.442
Vehicle class	15.486	0.008
Vehicle size	1.254	0.534

The next stage is the Wald test. This test is conducted to see the impact of each risk factor on the observed response variables. This statistical test follows an asymptotic Chi-Square distribution with degrees of freedom df, which represents the number of parameters associated with the analyzed variable. The results of the Type III analysis are presented in Table 3. Column Sig. shows the probabilities associated with the test used.

Table 3. Wald test of the frequency of claims

Parameters	**Type III**	
	Wald Chi-Square	**Sig.**
Intercept	199.295	0.000
Type of insurance	4.635	0.099
Type of work	11.543	0.009
Income per month	9.443	0.397
Policy type	9.855	0.275
Reason for claim	40.239	0.000
Vehicle class	6.989	0.221

It can be observed in Table 2 that the variables city code, education level, gender, location of residence, marital status, length of time as a customer, channel to get a policy and vehicle size have no significant effect on the frequency of claims. This can be seen in the p-value which is greater than the real level $\alpha = 0.05$. As a result, this variable is excluded from the model and the analysis will continue in the same way until the optimal factor combination ($p - value < 0.05$) is obtained which can explain the variation in the frequency of claims. From the results of the analysis presented in Table 3, it can be seen that the frequency of claims is influenced by the type of work of the customer and the reason the customer makes a claim.

3.2.2. Model of Claim Cost

The next stage in determining the insurance premium is an estimate of the cost of claim based on the risk factors considered by the insurance company. In this study, it was found that the cost of claim follows the Normal distribution

Let $c_{i1}, c_{i2}, \dots, c_{in_i}$ be the cost of claims made by the th respondent i, where n_i is the number of claims made by the insured i, and c_i is the amount of claims that are assumed to be mutually independent. The random variable C_i is normally distributed if it follows the probability density function as follows:

$$f(c_i) = \frac{1}{\sqrt{2\pi}\sigma_i}\, exp\left(-\frac{1}{2}\left(\frac{c_i - \mu_i}{\sigma_i}\right)^2\right) \tag{15}$$

where the expected value and variance are $E(c_i) = \mu_i$ and $Var(c_i) = \sigma_i^2$.

By using the canonical link function for the normal distribution, that is, the identity link function then the model for the size of the claim is obtained as follows:

$$g(\mu_i) = \boldsymbol{x_i^T\beta} \quad \leftrightarrow \quad \mu_i = \boldsymbol{x_i^T\beta} \tag{16}$$

The standard estimate for this model is the maximum likelihood estimation. The likelihood function is defined as follows:

$$L(\boldsymbol{\beta}) = \prod_{i|n_i>0}\prod_{k=1}^{n_i}\frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{1}{2}\left(\frac{c_{ik}-\mu_i}{\sigma_i}\right)^2\right). \tag{17}$$

Based on equation (12), the log-likelihood function is obtained as follows:

$$\begin{aligned} LL(\boldsymbol{\beta}) = \ln L(\boldsymbol{\beta}) &= \frac{n_i}{2}\ln 2\pi + \sum_{i|n_i>0}\sum_{k=1}^{n_i}\left(\ln\frac{1}{\sigma_i} - \left(\frac{c_{ik}-\mu_i}{\sigma_i}\right)^2\right) \\ &= \frac{n_i}{2}\ln 2\pi + \sum_{i|n_i>0}\sum_{k=1}^{n_i}\left(\ln\frac{1}{\sigma_i} - \left(\frac{c_{ik}-\boldsymbol{x_i^T\beta}}{\sigma_i}\right)^2\right). \end{aligned} \tag{18}$$

The likelihood estimator will be maximum if it meets the requirement that the first derivative of the log-likelihood function is zero and the sufficient condition is that the second derivative of the log-likelihood function is negative. It can be easily verified that the first partial derivative of the log-likelihood function exists and is expressed as follows:

$$\begin{aligned} \frac{\partial LL(\boldsymbol{\beta})}{\partial\beta_j} &= \sum_{i|n_i>0}\sum_{k=1}^{n_i}\frac{(c_{ik}-\mu_i)\boldsymbol{x_i^T}}{\sigma_i^2} \\ &= \sum_{i|n_i>0}\sum_{k=1}^{n_i}\frac{(c_{ik}-\boldsymbol{x_i^T\beta})\boldsymbol{x_i^T}}{\sigma_i^2}, \\ &\text{for } j = 1,2,..,n \end{aligned} \tag{19}$$

While the second derivative of the log-likelihood function is as follows:

$$\frac{\partial^2 LL(\boldsymbol{\beta})}{\partial\beta_j} = -\sum_{i|n_i>0}\sum_{k=1}^{n_i}\left(\frac{\boldsymbol{x_i^T}}{\sigma_i}\right)^2, \quad \text{for } j = 1,2,..,n \tag{20}$$

The maximum likelihood estimate of $\widehat{\beta_j}$ is the solution of equation (14) obtained from the condition that the first derivative of the log-likelihood function is zero. The equation is difficult to solve analytically, therefore it needs to be solved numerically using an iterative algorithm. The most commonly used iterative method is Newton-Raphson.

Table 4. Partial likelihood ratio test to the cost of the claim

Parameters	**Type III**	
	Likelihood Ratio	**Sig.**
Intercept	1120.368	0.000
City code	1.017	0.907
Type of insurance	0.809	0.667
Education level	4.256	0.235
Type of work	41.064	0.000
Gender	2.735	0.098
Income per month	4.410	0.882
Location of residence	9.364	0.009
Marital status	12.756	0.002
Old as a customer	11.095	0.269
Policy type	7.246	0.299
Reason for claim	34.852	0.000
Policy channels	7.099	0.069
Vehicle class	45.326	0.000
Vehicle size	3.561	0.169

Based on the results of the analysis presented in Table 4, it shows that the cost of claim is not influenced by city code variables, type of insurance, education level, gender, income per month, length of time as a customer, type of policy, channel to policy template, and vehicle size. The factors that influence the cost of claim are different from those that are related to the frequency of claims. This is in line with the assumptions put forward by the actuarial literature regarding the isolated analysis of the two phenomena.

Table 5. Wald test of claim cost

Parameters	**Type III**	
	Wald Chi-Square	**Sig.**
Intercept	2352.558	0.000
Type of work	42.891	0.000
Location of residence	7.253	0.027
Marital status	11.227	0.004
Reason for claim	30.279	0.000
Vehicle class	46.994	0.000

The cost of claim is a fundamental component that is considered in determining the insurance premium. Slightly different from the claim frequency model, the variables that significantly influence the cost of claims ($p - value < 0.05$) based on the results of the analysis presented in Table 5 are the type of work, location of residence, marital status, reasons for making claims and vehicle class.

3.2.3. Pure Premium Model

At this stage of general insurance pricing, the variable described is themultiplication of the estimated frequency and the estimated claim cost:

$$E\left[\sum_{k=1}^{N_i} C_{ik}\right] = E[N_i] \times E[C_{ik}]. \qquad (21)$$

The results obtained are summarized in Table 6. The calculated value is the pure insurance premium set for the i^{th} policy, which is based on the variable vector x_i. Considering this relationship in the insurance portfolio analyzed, the pure premiums for each policyholder category are determined based on the claims frequency model and the claim cost model, which includes all statistically relevant rate variables that explain variations in the frequency and cost of claims. was found that the cost of claim follows the Normal distribution

The following shows the results of the claim frequency model and the claim cost based on the results the goodness of fit test obtained.

$$\begin{aligned} E[N_i] = \mu_f = \exp(-1.098 + \beta_{4i}x_{4i} \\ + \beta_{11i}x_{11i}) \end{aligned} \qquad (22)$$

where N_i is the klam frequency for customer-i, β_{4i} is the model parameter for the variable type of work, customer-i, x_{4i} is the data on the type of work of the customer-i, β_{11i} is the model parameter for the variable reason customer makes a claim, customer-i, x_{11i} is data on the reason the customer-i made a claim.

$$\begin{aligned} E[C_{ik}] = \mu_c = 3790836673 + \gamma_{4i}x_{4i} \\ + \gamma_{7i}x_{7i} + \gamma_{8i}x_{8i} \\ + \gamma_{11i}x_{11i} + \gamma_{13i}x_{13i} \end{aligned} \qquad (23)$$

where C_{ik} is the cost of the customer's claim i with the magnitude-k, γ_{4i} is the model parameter for the variable type of work, customer-i, x_{4i} is the data on the type of work of the customer-i, γ_{7i} is the model parameter for the variable location where the customer-i lives, x_{7i} is the data on the location of the residence of the customer-i, γ_{11i} is the model parameter for the variable reasons for making a claim, customer-i, x_{11i} is data on the reason for the i^{th} customer to claim, γ_{13i} is the model parameter for the i^{th} customer vehicle class variable, x_{13i} is i^{th} customer vehicle class data.

From equations (17) and (18), the pure premium model is obtained as follows:

$$\begin{aligned} E\left[\sum_{k=1}^{N_i} C_{ik}\right] &= E[N_i] \times E[C_{ik}] \\ &= \left(e^{-1.098+\beta_{4i}x_{4i}+\beta_{11i}x_{11i}}\right) \\ &\quad \times (3790836673 \\ &\quad + \gamma_{4i}x_{4i} + \gamma_{7i}x_{7i} \\ &\quad + \gamma_{8i}x_{8i} + \gamma_{11i}x_{11i} \\ &\quad + \gamma_{13i}x_{13i}). \end{aligned} \qquad (24)$$

Table 6. Estimation of Model Parameters

Parameters	**Estimased[a]**	**Std.Error[a]**	**Estimased[b]**	**Std.Error[b]**
Intercept	-1.098	0.2509	3790836673	128761676.1
Type of work (employed)	0.065	0.0588	83410237.67	735311057.97
Type of work (medical leave)	-0.264	0.1044	-242916870	109016710.2
Type of work (retired)	-0.014	0.1475	-903826490	167507934.3
Location of residence (rural)	-	-	194626097.4	90333515.05
Location of residence (sub urban)	-	-	385263.173	79194133.52
Marital status (divorced)	-	-	-38295247.1	92301077.85
Marital status (married)	-	-	180204001.0	698664960.70
Reason for claim (collision)	0.181	0.0729	226044667.1	81885903.45
Reason for claim (hail)	-0.194	0.0790	-99247290.0	84365468.68
Reason for claim (other)	-0.108	0.0998	-141784689	107087092.0
Vehicle class (four door car)	-	-	215779437.6	71345794.57
Vehicle class (luxury car)	-	-	-345691539	213498201.2
Vehicle class (luxury SUV)	-	-	-37751352.8	133146263.8
Vehicle class (sports car)	-	-	725056065.7	133146263.8
Vehicle class (SUV)	-	-	375414605.8	86504458.71

[a] the results of the claim frequency model

[b] the result of the claim size model

4. CONCLUSIONS

General insurance pricing consists of determining the premium or rate paid by the insured to the insurance company as compensation for risk transfer. The calculation of insurance premiums is based on multiplying the estimated frequency and cost of claims.

This paper discusses about the analysis of Generalized Linear Models (GLM) to determine the best pure premium model according to the characteristics of policyholders. Therefore, as a first step, the claim frequency is estimated by means of a GLM model with a Poisson distribution. In the next analysis stage, using the GLM model with a normal distribution, an estimate of the average cost of claims that is appropriate for each group of policyholders is determined. In the end, the results show that for new customers, insurance premiums are determined by considering a series of risk factors, such as type of work, reasons for making claims, location of residence, marital status and class of customer vehicles.

The conclusions of this study are representative and useful for insurance company business, but are not in general so that they cannot be applied to all portfolios or insurance companies. On the one hand, this aspect is justified by the data used and the risk factors considered during the analysis process, meaning that each insurer can use different information about the insured to their advantage. On the other hand, the data used were not obtained through random selection regarding the entire population of policyholders.

ACKNOWLEDGMENTS

The authors are very thankful to the Directorate of Research and Community Service, Directorate General of Higher Education, Ministry of National Education, Republic of Indonesia through Hibah Penelitian Tesis, No. 2822/IT3.L1/PN/2020, who supporting and assisting to provide research grant.

REFERENCES

[1] M. David, “Auto insurance premium calculation using generalized linear models,” *Procedia Econ. Financ.*, vol. 20, no. 15, pp. 147–156, 2015, doi: 10.1016/S2212-5671(15)00059-3.

[2] M. V Wuthrich and M. Merz, *Stochastic Claims Reserving Methods in Insurance*, 1.1. ETH

Zurich, University Tubingen, 2006.

[3] S. A. Klugman, H. H. Panjer, and G. E. Willmot, *Loss Models From Data ro Decisions*, 5 th., vol. 6, no. 1. Hoboken, New Jersey: John Wiley & Sons, Inc., 2019.

[4] A. Guisan and T. C. Edwards, "Generalized linear and generalized additi v e models in studies of species distributions : setting the scene," *Ecol. Modell.*, vol. 157, no. 2–3, pp. 89–100, 2002.

[5] A. Guisan and N. E. Zimmermann, "Predictive habitat distribution models in ecology," *Ecol. Modell.*, vol. 135, no. 2–3, pp. 147–186, 2002.

[6] G. Z. Jong, P D, Heller, *Generalized Linear Models for Insurance Data*. Cambridge: Cambridge University Press, 2008.

[7] E. A. Antonio K, Valdez, "Statistical concepts of a priori and a posteriori risk classification in insurance," *Adv. Stat. Anal.*, vol. 96, no. 2, pp. 187–224, 2011, doi: 10.1007/s10182-011-0152-7.

[8] G. Dionne and C. Vanasse, "Automobile insurance ratemaking in the presence of asymmetrical information," *J. Appl. Econom.*, vol. 7, no. 2, pp. 149–165, 1992, doi: 10.1002/jae.3950070204.

[9] G. Dionne and C. Vanasse, "A generalization of actuarial automobile insurance rating models : The negative binomial distribution with a regression component," *ASTIN Bull.*, vol. 19, no. 2, pp. 199–212, 1988.

[10] M. Denuit and S. Lang, "Non-life rate-making with Bayesian GAMs," *Insur. Math. Econ.*, vol. 35, no. 3, pp. 627–647, 2004, doi: 10.1016/j.insmatheco.2004.08.001.

[11] C. Gourieroux and J. Jasiak, "Heterogeneous INAR (1) model with application to car insurance," *Insur. Math. Econ.*, vol. 34, no. 2, pp. 177–192, 2004, doi: 10.1016/j.insmatheco.2003.11.005.

[12] S. Ross, *Introduction to Probability Models*, 10th ed. Orlando, Florida (US): Academic Pr., 2007.

[13] C. O. Omari, S. G. Nyambura, J. Martha, and W. Mwangi, "Modeling the Frequency and Severity of Auto Insurance Claims Using Statistical Distributions," *J. Math. Financ.*, vol. 8, no. 1, pp. 137–160, 2018, doi: 10.4236/jmf.2018.81012.

[14] M. Charpentier, Arthur and Denuit, *DE L ' ASSURANCE*. Paris: Economia, 2005.