Available online at http://www.iorajournal.org/index.php/ijgor/index

International Journal of Global Operations Research

e-ISSN: 2722-1016
p-ISSN: 2723-1739

Vol. 5, No. 2, pp. 88-92, 2024

Determination of Pure Health Insurance Premiums Using Generalized *Linear Models* (GLM) with Poisson Distribution

Devani Oktaviani[1*], Nabila Zahra[2], Nurfadhlina Abdul Halim[3]

[1.2]*Padjadjaran University, Sumedang Regency, Indonesia*
[3]*Faculty of Science and Technology, Universiti Sains Islam Malaysia, Malaysia*

**Corresponding author email: devani20002@mail.unpad.ac.id , nabila20020@mail.unpad.ac.id*

Abstract

Insurance companies need a method to help companies determine premium prices that are appropriate to the risks they face. In other words, it is also necessary to know the variables that influence premium prices using Generalized Linear Models (GLM) by generalizing the linear regression model to model the relationship between the dependent variable and the independent variable. The aim of this research is to determine the variables that influence premium prices and determine the pure health insurance premium using the GLM method.

Keywords: Pure Premium, Generalized Linear Models, Poisson Distribution

1. Introduction

Insurance is an agreement between the insurance company (insurer) and the policy holder (insured), where the insured pays a premium to obtain coverage for the risk of damage (Gunawan, 2022). Legal liability to third parties that may be suffered by the insured, receiving payments based on the death or life of the insured with benefits the amount of which has been determined and/or is based on the results of fund management. Insurance provides financial protection or financial compensation to someone who experiences certain losses or risks (Fuse Brown, 2017; Prinja et al., 2017).

One of them is health insurance. This provides guarantees to the insured to reimburse any medical costs which include hospital costs, surgical costs and medication costs. Examples of several health insurance companies in Indonesia are Prudential, AXA Mandiri, Allianz Life, and many more (DI, 2020; Fionita, 2018). A method is needed to help companies determine premium prices that are appropriate to the risks they face.

One method that can determine premium prices is Generalized Linear Models (GLM) (Rahmawati et al., 2023). GLM generalizes linear regression models to model the relationship between dependent variables and independent variables by multiplying the conditional expected values of claim frequency and claim costs (Kangwana, 2018; Zhang and Walton, 2019). The aim is to determine the variables that influence premium prices. This study aims to determine the pure health insurance premium using the GLM method.

2. Literature review

Rahmawati et al., (2023) conducted research in her thesis related to pure premium modeling using GLM entitled "Modeling Pure Premiums for Motor Vehicle Insurance Using Generalized Linear Models (GLM)". In this research, the author determines the pure premium for motor vehicle insurance by looking at vehicle risk through vehicle characteristics. The research combines the conditional expected values of claim frequency and claim costs. From the results obtained, the claim frequency follows the Poisson distribution and the claim costs follow the Gamma distribution. This research produces characteristics that influence pure premiums, namely vehicle brand, vehicle characteristics, coverage costs, and type of insurance.

Furthermore, research conducted by Putra et al., (2021) related to determining insurance premiums using the GLM method, the research was entitled "Calculation of Motor Vehicle Insurance Premiums Using Generalized Linear

Models with Tweedie Distribution". The aim of this research is to find variables that can influence premium prices and determine an appropriate premium price model based on the influencing variables. This research produces variables that influence pure premiums, namely the number of children, monthly income, marital status, education, employment, vehicle use, the amount of bluebook paid, and the type of customer's vehicle. Previous studies show that the GLM model can be useful for insurance companies.

3. Research Objects and Methods

3.1. Object

The object used in this research is to determine the pure premium and find the variables that influence it using the Generalized Linear Models (GLM) method. Supported by health insurance claims data obtained from Sumit Kumar Shukla. Data analysis and processing is assisted by SPSS and Microsoft Excel software.

3.2. Research methods

3.2.1. Generalized Linear Models (GLM)

Generalized Linear Models (GLM) is a statistical work used to model the relationship between dependent variables and independent variables (Dunn and Smyth, 2018). GLM is defined as an extension of linear regression using the exponential family distribution. The aim of the GLM model is to estimate response variables (Y)that depend on the explanation of the explanatory variables (X)(Putra *et al.* , 2021). Observation variables Ythat have an exponential family distribution have the following probability function (de Jong and Heller, 2008):

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

where $y$is the response variable, $\theta$is the canonical parameter, $\phi$is the scale parameter, and $S$is a subset of the set of natural numbers or real numbers. Meanwhile $b(\theta)$and $c(y,\phi)$is a known function. In the exponential family distribution applies: $E(y) = b'(\theta)$and $Var(y) = \phi b''(\theta)$(de Jong & Heller, 2008).

GLM aims to determine the conditional expected value of the response variable using existing observational data (Putra et al. , 2021). Parameters will be determined $\beta_1, \beta_2, \dots, \beta_n$through the log link function of the explanatory average value (μ_i), which can be written as follows:

$$g(\mu_i) = \ln(\mu_i) = \beta_0 + \sum_{j=1}^{k} \beta_k x_{ij} \tag{2}$$

3.2.2. Poisson distribution

The Poisson distribution is a discrete probability distribution used to describe the number of events that occur in a certain time or space interval, when these events occur at a constant rate and are independent of the previous time or space. The characteristics of the Poisson distribution are: The number of results from one experiment does not depend on the number of results from other experiments; The probability of an experimental outcome is proportional to the length of the time interval; The probability of more than one experimental outcome occurring within a short time interval in a small area is negligible.

The Poisson distribution has the following probability density function (de Jong & Heller, 2008) :

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

where $p(Y = y)$is the Poisson distribution probability density function, $y$is the frequency of insured claims, and $\lambda$is the Poisson distribution parameter. The Poisson distribution applies: $E(Y) = \lambda$and $Var(Y) = \lambda$(de Jong & Heller, 2008)

3.2.3. Maximum Likelihood Estimation (MLE)

The exponential family of distributions with functions $f(y_i; \theta, \phi)$has the following log-likelihood function:

$$L(\theta,\phi) = \sum_{i=1}^{n}\left(\frac{y_i\theta - b(\theta)}{\phi} + c(y_i,\phi)\right) = \frac{n(\bar{y}\theta - b(\theta))}{\phi} + \sum_{i=1}^{n} c(y_i,\theta) \tag{4}$$

Then lowered againstθ

$$\frac{\partial L(\theta,\phi)}{\partial\theta}=0\Rightarrow\frac{\partial}{\partial\theta}\frac{n\left(\bar{y}\theta-b(\theta)\right)}{\phi}=0\Rightarrow b'(\theta)=\bar{y} \tag{5}$$

4. Results and Discussion

4.1. Data Implementation

In this study, the data used is health insurance data in the United States. The data was taken from the kaggle.com website and downloaded on October 15 2023. In this data there were 1364 data but only 300 data were used in data analysis because some of the data were empty so not all of them were used in the research. The variables used in this study were age, gender, BMI, blood pressure, history of diabetes, number of children, information about smoking or not and number of insurance claims.

4.2. Variable Identification

In data analysis, variable identification is required to determine the dependent variable and independent variables. The data held by the dependent variable includes the number of insurance claims because it is influenced by the independent variable. Meanwhile, the independent variables include age, gender, BMI, blood pressure, history of diabetes, number of children, and information about smoking or not.

4.3. Generalized Linear Model (GLM) Model Selection

In selecting the right GLM model, analysis of insurance data is required, one of which is the type of distribution and link function that will be used. In this study, the Poisson distribution was used, so the log link function was used as follows:

$$g(\mu_i)=\ln\left(\frac{\mu_i}{1-\mu_i}\right)=x_i^T\beta \tag{6}$$

In this formula $\boldsymbol{\mu}$is the expected value of the Poisson distribution, while $\boldsymbol{g}(.)$it is the log link function. The use of the log link function in the Poisson model provides a better interpretation of the effect of the independent variable on the dependent variable. The coefficient of the independent variable in a log-poisson model represents the change in the log of the expected value of the dependent variable for each additional unit of the independent variable, after controlling for other variables in the model .

Table 1: Case processing summary

	N	Percent
Included	4	1.3%
Excluded	296	98.7%
Total	300	100%

Table 2: Continuous variable information

		N	Minimum	Maximum	Mean	Std. Deviation
Dependent Variable	NUMBER OF INSURANCE CLAIMS	4	1665.00	3578.00	2313.7500	891.85813
Covariate	AGE	4	29	49	35.75	9,069

Table 3: Goodness of fit

	Value	df	Value/df
Deviance	967,455	3	322,485

Scaled Deviance	967,455	3	
Pearson Chi-Square	1031.327	3	343,776
Scaled Pearson Chi-Square	1031.327	3	
Log Likelihood b	-502,796		
Akaike's Information Criterion (AIC)	1007.592		
Finite Sample Corrected AIC (AICC)	1009.592		
Bayesian Information Criterion (BIC)	1006.979		
Consistent AIC (CAIC)	1007.979		

4.4. Calculation of Aggregate Pure Premium

Pure premium is the expected value of annual claim costs stated by the policy holder and is obtained by multiplying the expected value of claim frequency by the expected value of claim size. The estimated frequency of claims with the size of the claim can be stated as follows:

$$E = \left[\sum_{i=1}^{N} C_i\right] = E[n] \times E[C_i] \tag{7}$$

For the size of claims $(C_1, C_2, \ldots, C_N)$to be independent of the number of claims (N). [8]

4.5. Claim Frequency Modeling

The claim frequency data is thought to follow a discrete distribution, however after fitting the distribution to the discrete distribution using the Anderson-Darling (AD) test, the data does not follow a discrete distribution. After testing the distribution of claim frequency data, the p value was obtained , namely 1.132653. Next, the best link function is determined. The best link function is the link function with the smallest AIC results. Based on the tests carried out, it was found that the log link function was the best link function with an AIC value = 10070.65.

By using the log link function, the following claim frequency model is obtained:

$$g(\mu_i) = X_i^T \beta \rightarrow \ln \mu_i = x_i^T \beta$$
$$\mu_i = e\text{^}x_i^T \beta \tag{8}$$

From the claim frequency model in equation 3, the goodness of the model will be tested. The test used to see the goodness of the model is the partial likelihood ratio test. The test was carried out using SPSS software, the results of the overall model significance test showed that the model was significant and suitable for use because p-value = 2.2 × 10−16 < 0.05. Meanwhile, for partial testing, a Wald test is carried out. Based on the tests carried out, the variables age, gender, BMI, blood pressure, history of diabetes, number of children, and whether or not they smoke significantly influence the frequency of claims.

4.6. Modeling the Size of the Claim

As with the frequency of claims, data on the size of claims is thought to follow a continuous distribution, however after fitting the distribution to the continuous distribution using the chi-square test, the data does not follow a continuous distribution.

Table 4: Continuous Distribution

Parameter Estimates										
			95% Wald Confidence Interval		Hypothesis Test				95% Wald Confidence Interval for Exp(B)	
Parameter	B	Std. Error	Lower	Upper	Wald Chi-Square	dr	Sig.	Exp(B)	Lower	Upper
(Intercept)	7.747	.0104	7.726	7.767	555394.371	1	.000	2313.750	2267.088	2361.372
(Scale)	1[a]									

Dependent Variable: Number of Insurance Claims
Model: (Intercept)
a. Fixed at the displayed value

4.7. Calculation of Aggregate Premium

Based on the estimated premium parameters obtained, the pure premium model is obtained as follows:

$$E = \left[\sum_{i=1}^{N} C_i\right] = \exp((\beta_0 + \gamma_0) + (\beta_1 + \gamma_1)x_i^1 + \beta_4 x_4^i + (\beta_5 + \gamma_5)x_5^i + \beta_7 x_7^i + \sum_{j=1}^{9} \beta_{8,j} x_{8,j}^i + \beta_{10} x_{10}^i + (\beta_{11} + \gamma_{11})x_{11}^i + \sum_{j=1}^{6} \beta_{12,j} + \gamma_{12,j})x_{12,j}) \quad (9)$$

5. Conclusion

In determining the pure premium for health insurance, it is influenced by several risk factors, such as age, gender, BMI, blood pressure, history of diabetes, number of children and smoking. To determine the premium, the company can determine the amount of the premium through the Generalized Linear Model (GLM) model following equation 2.

References

de Jong, P., & Heller, G. Z. (2008). *Generalized Linear Models for Insurance Data*. Cambridge University Press. https://doi.org/10.1017/CBO9780511755408

DI, A. (2020). The Efficiency Analysis of Sharia Life Insurance in Indonesia and Family Takaful in Malaysia Using Data Envelopment Analysis Method Case Study On Al Abrar's Shariah Financial Service Cooperative. *Jurnal Ekonomi Syariah Teori dan Terapan*, *7*(2), 319-331.

Dunn, P. K., & Smyth, G. K. (2018). *Generalized linear models with examples in R* (Vol. 53, p. 16). New York: Springer.

Fionita, I. (2018, October). The Influence of Service Quality, Product Quality Toward Brand Image with Customer satisfaction as Intervening Variable (Case Study at Agency Customers of PT. Prudential Life Assurance). In *Proceeding International Conference on Information Technology and Business* (pp. 145-157).

Fuse Brown, E. C. (2017). Consumer Financial Protection in Health Care. *Wash. UL Rev.*, *95*, 127.

Gunawan, M. (2022). Implementation of Legal Principles of Agreement Between Policyholders and Insurance Companies. *LITERACY: International Scientific Journals of Social, Education, Humanities*, *1*(3), 208-218.

Kangwana, G. O. (2018). *Application of Generalized Linear Models in Medical Insurance Rate Making* (Doctoral dissertation, University of Nairobi).

Prinja, S., Chauhan, A. S., Karan, A., Kaur, G., & Kumar, R. (2017). Impact of publicly financed health insurance schemes on healthcare utilization and financial risk protection in India: a systematic review. *PloS one*, *12*(2), e0170996.

Putra, T. A. J., Lesmana, D. C., & Purnaba, I. G. P. (2021). Penghitungan Premi Asuransi Kendaraan Bermotor Menggunakan Generalized Linear Models dengan Distribusi Tweedie. *Jambura Journal of Mathematics*, *3*(2), 115-127.

Rahmawati, T., Susanti, D., & Riaman, R. (2023). Determining Pure Premium of Motor Vehicle Insurance with Generalized Linear Models (GLM). *International Journal of Quantitative Research and Modeling*, *4*(4), 207-214.

Shukla, S. K. (n.d.). Insurance Claim Analysis: Demographic and Health Impact on Risk and Severity of Insurance Claim. https://www.kaggle.com/datasets/thedevastator/insurance-claims-data

Zhang, Y., & Walton, N. (2019). Adaptive pricing in insurance: Generalized linear models and gaussian process regression approaches. *arXiv preprint arXiv:1907.05381*.