Net Premium Estimation by using Forward Selection Linear Model for Motor Vehicle Insurance

Riaman, Sukono, Dwi Susanti, Ellen Marbun
Department of Mathematics, Faculty of Mathematics and Natural Sciences,
Universitas Padjadjaran, Indonesia
riaman@unpad.ac.id, sukono@unpad.ac.id, dwi.susanti@unpad.ac.id, marbunellen@gmail.com

Abdul Talib Bon
Department of Production and Operations,
University Tun Hussein Onn Malaysia, Johor, Malaysia
talibon@gmail.com

Abstract

Protection is fundamental to transfer the risk of loss. The way to protect financial losses caused by car damage is car insurance. Owners of motor vehicles bind themselves to insurance companies, to protect their vehicles, by purchasing insurance premiums. The amount of premium depends on various aspects, including location, year of manufacture, and the contents of the engine cylinder and so on. To claim as a benefit of the premium already paid to the insurer, the amount of net premium can be estimated through the large expectations of claims and the number of possible claims. Therefore, it is necessary to specify the types of dependent variables and independent variables. In this research, modeling is done with Linear Model Selected reverse selection. The model is used to estimate the premiums to be offered. More expensive premiums can cope with some unexpected possibilities.

Keywords:
Net premium, Forward Selection, Number of Claims, Generalized Linear Models

1. Introduction

Humans are faced with various risks in life, such as mental and / or physical health disorders and property damage. If not properly anticipated, such risks may result in losses, especially in financial terms. The risk of damage to assets or property, whether or not moving, can be transferred to an insurance company by paying a certain amount of money. The transfer of risk to the object of this kind of insurance is non-life insurance. One of the non-life insurance products is car insurance (David and Jemna, 2015)

Insurance companies, often referred to as insurers, are obliged to pay claims rights in the form of sum insured in case of loss to the insured party according to mutual agreement or policy (Hoog and Craig, 1995; 2005). In non-life insurance, this claim may occur repeatedly within a period of coverage (Sukono et al., 2017.a; 2017.b). The right of this claim is not provided free of charge, but the policyholder is obliged to pay the premium to the insurer. This premium is divided into two types, namely net premiums and gross premiums. In practice, it is very important for the insurer to estimate the net premium in order to account for the actual premium that will be offered to the prospective policyholder (David and Jemna, 2015; Sukono et al., 2017.a; 2017.b).

In non-life insurance, one of which is motor vehicle insurance, claims can occur repeatedly in a period of coverage. The right of this claim is not provided free of charge, but the policyholder is obliged to pay the premium to the insurer. In practice, it is very important for the insurer to estimate the net premium in order to account for the actual premium that will be offered to the prospective policyholder (gross premium) (Agresti, 2007).

To estimate the net premium of car insurance, the insurer can analyze the model of the number of claims and models of the amount of claims of claim data that have been realized in the past by taking into account the variables that affect. These variables are usually categorical. In addition, the data is usually not normally distributed. In this study, the Generalized Linear Models can be used for such data as the dependent variable, i.e. the amount of the

claim and the number of claims, follows one member of the exponential family distribution (Hoog and Craig, 1995; 2005; Agresti, 2007; Sari, 2016).

2. Methodology

The object to be studied in this research is the net premium of private car insurance. Dependent variables imposed on the object of this study there are two kinds, namely the variable number of claims and variable magnitude of claims.

2.1 Stacked Linear Model

The Stacked Linear Model is the development of a classical linear model that accommodates two main things, namely the distribution of non-normal bound variables and the transformation for linearity. The Stacked Linear Model has three components in it, namely random components, systematic components, and link functions. The random component of the Linear Model consists of identification of the dependent variable *Y* and the selection of the probability distribution. Variables bound to Linear Model Followed by exponential family distribution. The systematic component of the Linear Model Stamped determines the independent variable by including linearity as the equation estimator of the model. The link function is a connecting function between the systematic component and the expected value (average) of the random component. In the modeling of insurance with independent variables that are categorical, the link function used is a log link so the model equation as follows Iskandar et al., 2011; Meyricke and Sherris, 1989):

$$\ln(\mu_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_P X_{ip} \quad (1)$$

$$\mu_i = \exp(\boldsymbol{X}_i^T \boldsymbol{\beta}) \quad (2)$$

where $\boldsymbol{X}_i^T = \begin{bmatrix} 1 & x_{i1} & x_{i2} & \dots & x_{ip} \end{bmatrix}$ and $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 & \beta_{i1} & \beta_{i2} & \dots & \beta_{ip} \end{bmatrix}^T$.

2.2 Kolmogorov-Smirnov Test

The test of the distribution used in this study is Kolmogorov-Smirnov test of one sample. The null hypothesis of this test is that variable Y follows the selected theoretical distribution, in contrast to the alternative hypothesis. Reject the null hypothesis if the value $D_h = maks|F_0(x) - F^*(x)| \geq D_{\alpha,df=n}$ or $pvalue \leq \alpha$ (Goldburd et al., 2016; McCullagh and Nelder, 1989).

2.3 Forward Selection

The forward step regression process begins by regressing the variable magnitude of the claim with each independent variable separately (individual regression). The insignificant independent variables will be set aside, while the significant independent variables are sorted by AIC values for each model. Further, the independent variables in the model with the smallest AIC values are regressed simultaneously with the independent variables in the model with the second smallest AIC value. The insignificant independent variable is set aside, followed by entering the independent variable from the model with the next smallest AIC value. And so on until there are no more variables left (Sukono et al., 2016).

2.4 Likelihood Ratio Test

This test is conducted to determine the significance of independent variables contained in a model. The null hypothesis of this test is that no parameter contributes to the dependent variable, in contrast to the alternative hypothesis. Reject the null hypothesis if the value of $G_h = -2\ln(\hat{\ell} - \check{\ell}) \geq \chi^2_{\alpha,df=p-1}$ or $pvalue \leq \alpha$ (Sukono et al., 2014; Klugman, 2004).

2.5 Akaike's Information Criterion (*AIC*)

The selection of the best model in this research is done by considering the value of AIC owned by model every time forward selection stage is done. The value of AIC is obtained through the following equation.

$$AIC = -2\ell + 2p \quad (3)$$

where $\ell$the value of the log-likelihood function of the model is formed and p is the number of parameters used in the model. Smaller *AIC* values indicate that the model is getting (Sukono et al., 2016).

2.6 Maximum Likelihood Estimation (MLE) Method

To perform the β parameter estimation MLE method is used. Let Y_1, Y_2, ..., Y_n a random sample having a probability density function derived from the Poisson distribution. The maximum likelihood estimator for the β parameter is expressed by $\widehat{\boldsymbol{\beta}}$ which is the completion of the first derivative of the likelihood function.

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} p(y_i) = \prod_{i=1}^{n} \frac{\exp(-\mu_i)(\mu_i)^{y_i}}{y_i!} \quad (4)$$

$$\ln L(\boldsymbol{\beta}) = \sum_{i=1}^{n} [-\exp(\boldsymbol{X}_i^T \boldsymbol{\beta}) + y_i \boldsymbol{X}_i^T \boldsymbol{\beta} - \ln(y_i!)] \quad (5)$$

$$\frac{\partial \ln L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T} = \sum_{i=1}^{n} [-\boldsymbol{X}_i \exp(\boldsymbol{X}_i^T \boldsymbol{\beta}) + y_i \boldsymbol{X}_i] = 0 \quad (6)$$

Because the function in equation 6 is implicit then to get the solution used a numerical iteration procedure, one of which is the Newton-Raphson Method (Wood, 2006; Siegel and Castellan, 1988).

2.7 Wald Test

Unlike the Likelihood Ratio test that can test the significance of several variables simultaneously, this test can only be used to test the significance of a single independent variable only. The initial hypothesis of this test is that the independent variable with the β_j parameter does not contribute to the dependent variable, in contrast to the alternative hypothesis. Reject the initial hypothesis if the value of $z_h \geq \chi^2_{\alpha,df=1}$ or $pvalue \leq \alpha$ (Sukono et al., 2014).

2.8 Collective Model

Let X is a random variable denoting the magnitude of the claim and the random variable N denoting the number of claims. To determine the net premium on the collective model can be used the following formula (Sidi et al., 2017).

$$\text{Net premium} = E[S] = E[N] \,.\, E[X] \quad (7)$$

3. Illustration Analysis

The study used secondary data that is car insurance claim data approved by ABC Company during 2017 calendar year. The data of claims obtained are grouped into two, namely data for model amount of claim and data for model number of claims.

3.1 Selection of Distribution

The data used in determining the model magnitude of claims is the first group of data. Within this first group there is the dependent variable "Claim Amount" and the variables to be used as independent variables, namely "Car Brand", "Car Year", "Territory", "Cause of Loss", and "Nature of Losses". Each of these independent variables is a categorical variable. The results of Kolmogorov test calculations obtained, shows Gaussian Inversion distributed data. Dependent variable magnitude claims have met the assumption of the Linear Model Stamped because Gaussian Inverse is a member of the exponential family distribution. The link function that is used is the log link function.

3.1.1 Forward Selection

Linear Model Regression Process Regarded in SAS program using GENMOD procedure. The order of independent variables of the model with the smallest to largest AIC is "Cause of Loss", "Territory", "Car Brand", "Car Year", and "Nature of Losses". That is, the next regression involves the variables of "Claim Amount", "Cause of Loss", and "Territory". The result variable "Cause Loss" is significant, while variable "Territory" is not significant. The next regression involves "Cause Losses" and "Car Brand". The result of the "Car Brand" makes no significant contribution to the "Claim Amount". The "Cause Loss" variable can still be incorporated into the next process. Both the "Cause Loss" and "Year of Car" variables provide significantly. Finally, the "Causes of Loss", "Car Year", and "Nature of Losses" variables are simultaneously simultaneous. Both the "Cause Loss" and "Car Year" variables are significant, but the "Nature of Losses" is not. Thus, the "Disadvantages" are set aside from the model. In addition, this model has an AIC smaller than the previous AIC model value. That is, this model is the best model for the magnitude of claims. After obtaining the best model for personal car insurance claims using forward selection, systematic system component parameters were estimated using the Maximum Likelihood Estimator (MLE) method

with the help of the SAS program. Significant variables are only "Cause Losses" level 1 and 3, as well as "Year Car" level 12. Therefore, the model of magnitude of the claim can be written as follows.

$$g(\widehat{\mu_1}) = \ln \widehat{\mu_1} = 15{,}2983 - 0{,}6854x_{P1} - 0{,}7116x_{P3} - 0{,}8413x_{T12}$$

where

$\widehat{\mu_1}$: Expectations of estimates of the magnitude of private car insurance claims
x_{P1} : Variable "Cause Loss" category "hit"
x_{P3} : The "Cause Loss" variable of the category "Dispelled"
x_{T12} : Variable "Year Car" category "2015"

3.2 Number of Claims Model

The data used in determining the number of claims model is the second group data. Within this second group there is the dependent variable "Number of Claims" and the variables that will be used as independent variables, namely "Gender", "Car Brand", "Car Year", "CC Car", and "Transmission Type" Territory. Each of these independent variables is a categorical variable. In the insurance field, the policyholder may not file a single claim, or apply only once, or more than once in a given period. Therefore, the variable expressing the number of claims follows one of the discrete distributions. The discrete distributions that belong to the exponential family distribution are Binomial distribution, Poisson distribution, and Negative Binomial distribution. With Microsoft Excel calculated the mean and variance of the variable number of claims. The results show that the mean of this variable is 0.645390071, while the variance is 0.379142374. From these results it can be seen that the mean is greater than the variance, whereas the mean and variance of the Poisson distribution are equal. It is very common to encounter linear modeling for the model of many claims with the Poisson distribution (Agresti, 2007). In SAS, this can be resolved by scale = p in the GENMOD procedure. The link function used is the log link.

3.2.1 Forward Selection

The order of the free variables from the model with the smallest to largest AIC is "Territory", "CC Car", "Transmission Type", "Year of the Car", and "Car Brand". That is, the next regression involves variables "Number of Claims", "Territories", and "CC Car". The two independent variables in this model give meaning to "Number of Claims". In addition, the AIC value of this model is smaller than the AIC model value that involves the "Territory" free variables only. In other words, this model is more suitable than the previous model. Next, the "Transmission Type" variable is inserted into the model to be reordered together with "Number of Claims", "Territories" and "Car Cc". "Transmission Type" must be set aside, while the other two variables in the model indicate a significant effect on the variable "Number of Claims". However, the AIC value of this model is not smaller than the AIC model value involving "Territory" and "Car cc". Therefore, the previous model is still the best model to date. Furthermore, "Year of the Car" is involved in the regression process of "Territory" and "CC Mobil". "Car Year" is not significant. In addition, this model has an AIC value that is higher than the AIC model value that involves "Territory" and "Car cc" only. Therefore, this model is less suitable. Thus, the best model for modeling the number of claims is a model involving only the "Territory" and "cc Car" independent variables. Furthermore, systematic component component parameters are estimated using the Maximum Likelihood Estimator (MLE) method. Significant variables are only category 2 "Territory" and "cc Car" category 2, 4, 6, 7. Therefore, the model number of claims can be written as follows.

$$g(\widehat{\mu_2}) = \ln \widehat{\mu_2} = -0{,}7893 - 0{,}7055x_{W2} + 0{,}4724x_{C2} + 0{,}8823x_{C4} + 0{,}6456x_{C6} + 0{,}6619x_{C7}$$

Where:

$\widehat{\mu_2}$: Expectations of estimates of the number of private car insurance claims
x_{W2} : Variable "Territory" category "Other Territory"
x_{C2} : Variable "cc Car" category "1200cc"
x_{C4} : Variable "cc Car" category "1500cc"
x_{C6} : Variable "cc Car" category "1800cc"
x_{C7} : Variable "cc Car" category "2000cc"

3.3 Estimation of Net Premium Based on Model

Based on each model is obtained expectation of the amount of claims and the number of claims, as follows.

1) Model expectations of personal car insurance claims
   $\widehat{\mu_1} = \exp(15{,}2983 - 0{,}6854x_{P1} - 0{,}7116x_{P3} - 0{,}8413x_{T12})$
2) Expectation model of the number of private car insurance claims
   $\widehat{\mu_2} = \exp(-0{,}7893 - 0{,}7055x_{W2} + 0{,}4724x_{C2} + 0{,}8823x_{C4} + 0{,}6456x_{C6} + 0{,}6619x_{C7})$

***Case* 1**

Suppose a man with the initials A resides in Surabaya insured Honda car in 2013 with 1800cc engine capacity at *XYZ* Company. The car was hit so he filed a claim to the company. Then the net premium assessment is as follows.

$\widehat{\mu_1} = \exp(15{,}2983 - 0{,}6854x_{P1} - 0{,}7116x_{P3} - 0{,}8413x_{T12})$
$\widehat{\mu_1} = \exp\{15{,}2983 - 0{,}6854(1) - 0{,}7116(0) - 0{,}8413(0)\}$
$\widehat{\mu_1} = \exp(14{,}6129)$
$\widehat{\mu_1} = 2219738{,}60$
$\widehat{\mu_2} = \exp(-0{,}7893 - 0{,}7055x_{W2} + 0{,}4724x_{C2} + 0{,}8823x_{C4} + 0{,}6456x_{C6} + 0{,}6619x_{C7})$
$\widehat{\mu_2} = \exp\{-0{,}7893 - 0{,}7055(1) + 0{,}4724(0) + 0{,}8823(0) + 0{,}6456(1) + 0{,}6619(0)\}$
$\widehat{\mu_2} = \exp(-0{,}8492)$
$\widehat{\mu_2} = 0{,}427$

$$Net\ Premiums = \widehat{\mu_1} \times \widehat{\mu_2} = 2219738{,}60 \times 0{,}427 = 947828{,}38$$

Thus, the net premium to be paid by insurance participants in this case is IDR 947.828,00.

***Case* 2**

Suppose a woman with the initials B residing in Bandung insured the Toyota car manual transmission in 2015 with a capacity of 1200cc ABC Company. His car was hissed so he filed a claim to the company. Then the net premium assessment is as follows.

$\widehat{\mu_1} = \exp(15{,}2983 - 0{,}6854x_{P1} - 0{,}7116x_{P3} - 0{,}8413x_{T12})$
$\widehat{\mu_1} = \exp\{15{,}2983 - 0{,}6854(0) - 0{,}7116(1) - 0{,}8413(1)\}$
$\widehat{\mu_1} = \exp(13{,}7454)$
$\widehat{\mu_1} = 932290{,}74$
$\widehat{\mu_2} = \exp(-0{,}7893 - 0{,}7055x_{W2} + 0{,}4724x_{C2} + 0{,}8823x_{C4} + 0{,}6456x_{C6} + 0{,}6619x_{C7})$
$\widehat{\mu_2} = \exp\{-0{,}7893 - 0{,}7055(0) + 0{,}4724(1) + 0{,}8823(0) + 0{,}6456(0) + 0{,}6619(0)\}$
$\widehat{\mu_2} = \exp(-0{,}3169)$
$\widehat{\mu_2} = 0{,}728$

$$Nett\ Premiums = \widehat{\mu_1} \times \widehat{\mu_2} = 932290{,}74 \times 0{,}728 = 678707{,}65$$

Thus, the net premium to be paid by insurance participants in this case is IDR 678,707.65. Keep in mind that two examples of such cases are some of the possible cases that are expected to occur within one calendar year of insurance. In both cases researchers suggested the company to apply a net premium of IDR 947,282.00 to avoid future losses.

4. Conclusions

Based on the analysis and discussion that has been done then it can be concluded that: The modeling in this study begins with the selection of distribution and link functions of each dependent variable. Furthermore, the regression process is done by forward selection method. The tests used to test the significance of the independent variables are the Likelihood Ratio test and the Wald test. The best model

selection is done by looking at the value of AIC owned by model in every regression process. The independent variables that contribute significantly to the variable magnitude of claims are the cause of the loss and the year of the car, while for the number of claims are the region and the car cc. After going through these processes, the model for the number of private car insurance claims, namely $\ln\widehat{\mu_1} = 15{,}2983 - 0{,}6854x_{P1} - 0{,}7116x_{P3} - 0{,}8413x_{T12}$, while the model for the number of private car insurance claims is $\ln\widehat{\mu_2} = -0{,}7893 - 0{,}7055x_{W2} + 0{,}4724x_{C2} + 0{,}8823x_{C4} + 0{,}6456x_{C6} + 0{,}6619x_{C7}$. Based on the two case studies in this study, the estimated net premium of ABC corporate private car insurance is IDR 947.282,00. This price is the highest price among the premiums generated through case studies. This price is chosen to anticipate future company losses.

Acknowledgments

The Authors would like to thank to DRPMI, directorate of research and community service, Padjadjaran University, who gave funding for this research and dissemination of this paper, via Riset Fundamental Unpad (RFU)

References

Agresti, A. An Introduction to Categorical Data Analysis (2nd Ed.). New Jersey: John Wiley & Sons, Inc. 2007.

David, M. and Jemna, D. Modeling The Frequency of Auto Insurance Claims by Means of Poisson and Negative Binomial Models. Scientific Annals of the "Alexandra Ioan Cazu" University of Iaşi Economic Sciences, (Online), Vol. 62(2), 2015. pp. 151-167, (http://www.degruyter.com, Downloaded on July 10, 2017).

Goldberg, M., Khare, A., and Tevet, D. Generalized Linear Models for Insurance Rating. Virginia: Casualty Actuarial Society. 2016.

Hogg, R. V. and Craig, A. T. Introduction to Mathematical Statistics (5th Ed.). New Jersey: Prentice Hall. 1995.

Hogg, R. V. and Craig, A. T. Introduction to Mathematical Statistics (6th Ed.). Delhi: Dorling Kindersley (India) Pvt. Ltd. 2005.

Iskandar, K., Fuad, N., Wirasadi, F., and Sendra, K. Dasar-dasar Asuransi: Jiwa, Kesehatan, dan Anuitas. Jakarta: Asosiasi Ahli Manajemen Asuransi Indonesia (*Fundamentals of Insurance: Life, Health and Annuity. Jakarta: Indonesian Association of Insurance Management Experts*). 2011.

Klugman, S. A. Loss Models: From Data to Decisions (2nd Ed.). New Jersey: John Wiley & Sons, Inc. 2004.

McCullagh, P. and Nelder, J.A. Generalized Linear Models in Monograph on Statistics and Applied Probability (2nd Ed.). New York: Chapman and Hall. 1989.

Meyricke, R. and Sherris, M. The Determinants of Mortality Heterogeneity and Implications for Pricing Annuities. Insurance: Mathematics and Economics, (Online), Vol. 53, 2013. pp. 379-387, (http://sciencedirect.com, Dowloaded on September 10, 2016).

Sari, D. P. Prediksi Premi Murni dengan Generalized Linear Models (GLM) pada Asuransi Kendaraan Bermotor. Skripsi tidak diterbitkan (*Prediction of Pure Premium with Generalized Linear Models (GLM) on Motor Vehicle Insurance. Unpublished thesis*). Jatinangor, Indonesia: Universitas Padjadjaran. 2016

Side, P., Mamet, M., Sukuma, & Spain, S. Evaluation model for risk insurance premiums of building damage caused by flood: Case study in Cit arum watershed, southern Bandung, Indonesia. *Journal of Engineering and Applied Sciences, 12*(17), 2017. pp. 4420-4425. doi:10.3923/jeasci.2017.4420.4425

Siegel, S. and Castellan, N. J. Nonparemetric Statistical for the Behavioral Sciences. New York: McGraw-Hill. 1988.

Sukono, Sholahuddin, A., Mamat, M., & Prafidya, K. Credit scoring for cooperative of financial services using logistic regression estimated by genetic algorithm. *Applied Mathematical Sciences, 8*(1-4), 2014. pp. 45-57. doi:10.12988/ams.2014.310600

Sukono, Hidayat, Y., Suhartono, Sutijo, B., Bon, A. T. B., & Supian, S. Indonesian financial data modeling and forecasting by using econometrics time series and neural network. *Global Journal of Pure and Applied Mathematics, 12*(4), 2016. pp. 3745-3757. Retrieved from www.scopus.com

Sukono, Nahar, J., Putri, F. T., Subiyanto, Mamat, M., & Supian, S. Estimation of outstanding claims reserving based on inflation risk on car insurance companies by using the bootstrap method. *Far East Journal of Mathematical Sciences, 102*(4), 2017.a. pp. 687-706. . doi:10.17654/MS102040687

Sukono, Suyudi, M., Islamiyati, F., & Supian, S. Estimation model of life insurance claims risk for cancer patients by using Bayesian method. Paper presented at the *IOP Conference Series: Materials Science and Engineering, 166*(1), 2017.b. doi:10.1088/1757-899X/166/1/012022

Wood, S. N. 2006. Generalized Additive Models. Boca Raton: Chapman and Hall.

Biographies

Riaman is the staff of the Department of Mathematics, University of Padjadjaran, with the field of research are: actuarial mathematics, financial mathematics, survival model analysis and reliability.

Sukono is a lecturer in the Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran. Currently serves as Head of Master's Program in Mathematics, the field of applied mathematics, with a field of concentration of financial mathematics and actuarial sciences.

Dwi Susanti & Ellen Marbun. is a lecturer & Student at the Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, the field of applied mathematics, with a field of concentration of financial mathematics and actuarial sciences.

Abdul Talib Bon is a professor of Production and Operations Management in the Faculty of Technology Management and Business at the Universiti Tun Hussein Onn Malaysia since 1999. He has a PhD in Computer Science, which he obtained from the Universite de La Rochelle, France in the year 2008. His doctoral thesis was on topic Process Quality Improvement on Beltline Moulding Manufacturing. He studied Business Administration in the Universiti Kebangsaan Malaysia for which he was awarded the MBA in the year 1998. He's bachelor degree and diploma in Mechanical Engineering which his obtained from the Universiti Teknologi Malaysia. He received his postgraduate certificate in Mechatronics and Robotics from Carlisle, United Kingdom in 1997. He had published more 150 International Proceedings and International Journals and 8 books. He is a member of MSORSM, IIF, IEOM, IIE, INFORMS, TAM and MIM.