MORTALITA

Vol. 1 No. 2, Desember 2024, Pages 6-11
Mortalita: Journal of Mathematics and Its Applications
e-ISSN 3088-9812
DOI: https://doi.org/ 10.61159/mortalita
https://ejournal.darunnajah.ac.id/index.php/mortalita

Copula-Based Regression Analysis To Estimate the Total Losses on Health Insurance

Dewi Susanawati [1*], Ahmad Fuad Zainuddin[2]

[1] *Universitas Darunnajah,* [2] *School of STEM, Universitas Prasetiya Mulya*
*Email Corresponding Author: dewisusanawati@darunnajah.ac.id

ABSTRACT

The insurance company is a company that received delegation of the risks it has insured, so that this company needs to pay attention to losses incurred as a result of a claim. Estimating losses of claim is an important task for insurance companies to predict their obligations. Total losses in the company's portfolio is defined as the amount of loss policy. Losses in the health insurance policy can be calculated based on two variables: the frequency and severity of claims. In the literature of Statistics, joint distribution is a method of statistical analysis that can combine two different data distribution, it is Copula. This thesis aims to provide a study of Copula for the estimation of loss claims in health insurance, case study is taken from an insurance company XYZ. Further, the authors conducted a regression between the Generalized Linear Model (GLM) of claim frequency and claim severity using Copula-based Regression Model is estimated by Maximum Likelihood Estimation (MLE). In the end of analysis, Copula-based Regression Model can be used to estimate the projected claim load in the budgeting of claim load the insurance company. This will help provide better estimation results compared to the methods currently used.

Keywords: Health insurance, copula, GLM; MLE, regression

1. INTRODUCTION

Based on the Republic of Indonesia Law Number 40 of 2014 on Insurance, the Insurance Business is any business related to insurance services or risk management. In terms of risk management, an insurance company is a company that receives the transfer of risk from the insured, so the insurance company needs to pay attention to the losses incurred as a result of the occurrence of the insured risk [1]. In general, insurance consists of life insurance and general insurance [2]. Health insurance is one of the branches of the general insurance or life insurance business lines that deals with insurance in the health sector. Based on the duration of treatment, health insurance has the concept of life insurance, whereas based on the loss of claims that occur, health insurance has the concept of general insurance. Claims experience data is one of the bases for determining the amount of coverage premium, as in the experience rating method. This method calculates the premium rate by evaluating and measuring past experiences to be used in future estimations [3]. Therefore, a thorough analysis of the existing claim data is necessary, and an estimation of the total claim losses, especially health insurance claims borne by the insurance company, is required. The total loss of claims depends on the size of the claims (severity) and the frequency of claims (frequency) received by the insurance company. Severity and frequency of claims are the two main factors for calculating and predicting the total loss due to claims, therefore, modeling that involves both variables together is necessary [4]. Based on the Actuarial Loss Distribution Approach, it is necessary to first identify the type of distribution for each random variable being analyzed, just as when analyzing the dependence between claim severity and claim frequency [5]. The two random variables are two different data distributions where the frequency of claims is discretely distributed and the severity of claims is continuously distributed, and both have different marginal models, making them difficult to model.

Copula Model

Copula is a method that can combine several marginal distribution functions into a joint distribution function with the assumption that there is dependence between the marginal distributions. Therefore, Copula can be used to determine the dependence between severity and claim frequency so that insurance companies can estimate the total claims that will be the company's burden in the following period. Previous research has been conducted by Czado [4], who reviewed the relationship between severity and frequency using a Gaussian copula-based joint regression model. Next, Krämer [6] discuss a joint regression model for claim frequency and severity using a bivariate copula aimed at accommodating the relationship between claim frequency and severity. Based on the existing issues regarding the estimation of total health insurance claim losses, particularly in Indonesia, this research aims to detect the dependency relationship between claim amounts and claim frequency, determine the appropriate copula method for estimating total claim losses in health insurance, and estimate the total claim losses that occur. Copula is very useful for modeling the relationships of related variables in the fields of finance, actuarial science, and survival analysis [7]. Another advantage of copulas is that the marginal distributions do not have to meet identical conditions and are not strict about assumptions, particularly the normal distribution [4]. A n-dimensional copula denoted by C as a multivariate distribution function F of random variables $X_1, \dots, X_n$ with marginal distributions $F_1, \dots, F_n$ uniformly distributed standard such that $X_i \sim F_i$ where $F_i \sim U\,(0,1); i = 1, \dots, n$ [9]. This copula function is a function that has a domain $[0,1]^n$ and range defined as C: $[0,1]^n \rightarrow [0,1]$. In the bivariate case, a copula is a bivariate cumulative distribution function with density C_φ: $[0,1]^2$ for $\varphi \in R$. The concept of copula was first introduced by Abe Skalar in 1959, which later led to the emergence of Sklar's theorem [8]. Sklar's theorem is the core of copula theory and serves as the foundation for several statistical theories.

Jong and Heller, explain that this model uses a linear transformation of its response variable, so this model can be used not only for linear data but also for data that is not normally distributed or where the errors of the data are not homogeneous [9]. The GLM model with response is:

$$f(y) = c(u, \emptyset)\, exp\left\{\frac{y\theta - a(\theta)}{\emptyset}\right\} \text{ and } g(\mu) = x'\beta.$$

Copula can be defined in the form of a joint distribution as follows:

$$H(x, y) = C\big[\big(F(x), G(y)\big)\big] = P[X \leq x, Y \leq y]$$

The copula that links the covariate or predictor of its bivariate distribution is described as follows:

$$C(u, v|r) = H(x, y|r) = C\big(F_{X|r}(x|r), G_{Y|r}(y|r)\big) \text{ for } (x, y)\, \epsilon\, \mathbb{R}^2.$$

where *r* is covariate and $F_{X|\boldsymbol{r}}(x|r)$ and $G_{Y|\boldsymbol{r}}(y|r)$ are conditional marginal distribution function.

So the conditional marginal density function is as follows:

$$h(x, y|r, \alpha, \beta; \theta) = f_{X|r}(x|r; \alpha)\; g_{Y|r}(y|r; \beta)\; c(u, v|r, \alpha, \beta; \theta)$$

for α and β are marginal distribution parameter and θ is copula parameter

For estimation and prediction, a joint distribution function of X and Y is required, defined as

$$f_{X,Y}(x, y) \coloneqq \frac{\partial}{\partial x} P(X \leq x, Y \leq y)$$

The joint distribution function of the continuous variable *X* (claim severity) and the discrete variable *Y* (claim frequency) is defined as follows:

$$\frac{\partial}{\partial x} P(X \leq x, Y = y) = \frac{\partial}{\partial x} P(X \leq x, Y \geq y) - \frac{\partial}{\partial x} P(X \leq x, Y = y - 1)$$

$$\frac{\partial}{\partial x} P(X \leq x, Y = y) = \frac{\partial}{\partial x} C(F_X(x), G_Y(y)|r; \theta) - \frac{\partial}{\partial x} C(F_X(x), G_Y(y-1)|r; \theta)$$

$$\frac{\partial}{\partial x} P(X \leq x, Y = y) = f_X(x)\big(D_1(F_X(x), G_Y(y)|r; \theta) - D_1(F_X(x), G_Y(y-1)|r; \theta)\big)$$

Copula regression

Copula-based regression is a model that describes the causal relationship between predictor variables and response variables. Unlike general regression models such as the OLS regression model, which have the regression parameter β, copula-based regression does not have a regression parameter. Copula-based regression only has parameters for each of its marginal distributions and the parameter θ as the copula parameter. This is because copula-based regression describes data that does not have a specific pattern, so the resulting model cannot be explained by regression parameters.

2. METHODS

The data used in this research is health insurance claim data for 5 years at XYZ insurance company. The claim data used consists of 3.666 insured who submitted claims for 5 years. The components of data used in this research consist of claim severity (claim amount), claim frequency (claim total), deductible and gender. The distribution of claim severity and claim frequency are two different distributions, where claim severity is continuously distributed and claim frequency is discretely distributed. Thus, the Generalized Linear Model (GLM) formed results in two different GLMs. Both GLM models are used as marginal distributions in the combined model using copula.

$$X_i \sim Gamma\ (\mu_i, \delta) \text{ where } \ln(\mu_i) = \boldsymbol{r_i'}\alpha,$$
$$Y_i \sim Poisson\ (\lambda_i) \text{ where } \ln(\lambda_i) = \ln(e_i) + \boldsymbol{s_i'}\beta,$$

for $\boldsymbol{r_i'}$ and $\boldsymbol{s_i'}$ are covariates and e_i is defined by observation time .

The family of copulas used is the Archimedean Copula. The reason for using Archimedean Copula is that Archimedean Copula is a special class of copula for several reasons, including that this copula is easy to construct, there are many variations of copula families included in this class, and it has a varied dependency structure [8]. In general, the form of the Archimedean copula equation is as follows:

$$C\ (u, v) = \ \varphi^{-1}(\varphi(u) + \ \varphi(v)), \ \text{ if } \varphi(u) + \ \varphi(v) \leq \varphi(0)$$

where φ is a generate function of copula *C* with $0 \leq u, v \leq 1$ and $\varphi(0) = \infty, \ \varphi(1) = 0$.

For the Archimedean copula equation involving covariates, the conditional Archimedean copula equation model is as follows:

$$C\ (u, v|r) = \ \varphi^{-1}(\varphi(u|r) + \ \varphi(v|r)), \ \text{ jika } \varphi(u|r) + \ \varphi(v|r) \leq \varphi(0)$$

where $u = F_X(x)$, $v = G_Y(y)$, and r is covariate that affects the random variable *X* and the random variable *Y*.

Joint density function of Archimedean copula as follows [10]:

$$h(x, y) = f(x)g(y)c\big(F_X(x), G_Y(y)\big)$$

In this study, there are predictor variables in the estimation of policy losses, so the copula density function as follows:

$$h(x, y|r, \alpha, \beta; \theta) = f(x)g(y)\ c\left(\big(F(x|r, \alpha)\big), \big(G(y|r, \beta)\big); \theta\right)$$

for α and β are parameter of marginal distribution function and θ is copula parameter.

Kendall's tau correlation function for Archimedean Copula as follows:

$$\tau = 1 + 4\int_0^1 \frac{\varphi(t)}{\varphi'(t)}dt$$

The Archimedean Copula family discussed in this study is limited to the Clayton Copula, with generate function is:

$$\varphi_\theta(t) = \frac{1}{\theta}\big(t^{-\theta} - 1\big)$$

And for bivariate is:

$$C(u,v|r) = exp\left(-\left[(-\ln u)^{\theta} + (-\ln v)^{\theta}\right]^{\frac{1}{\theta}}\right)$$

After obtaining the parameter estimates from the copula-based regression, the prediction of claim frequency and claim severity is obtained by calculating the conditional expectation of the model formed for each copula family using the following equation:

$$E[f(y|x)]$$

where $f(y|x)$ is the conditional density function of the copula-based regression model.

From the prediction of claim frequency and claim severity, the total loss of health insurance policies at XYZ insurance company can be predicted. Policy loss is a continuous random variable that takes positive values and depends on the parameters of the formed model. Policy loss (L) is defined as the product of claim severity (X) and claim frequency (Y), with the following equation:

$$L := X.Y$$

Total loss as the sum of policy losses for *n* individual policy contract with claim severity X_i and claim frequency Y_i for $i = 1,2, \dots, n$, as follows [6]:

$$T := \sum_{i=1}^{n} L_i = \sum_{i=1}^{n} X_i.Y_i$$

This study use software *EasyFit* and *R* for calculating the estimation and prediction [11].

3. RESULTS AND DISCUSSION

Based on the analysis results, it is known that the random variable for claim frequency follows a Poisson distribution and the random variable for claim severity follows a Gamma distribution. The correlation test of the two random variables yielded a correlation value of 0.9071304. This explains that there is a positive correlation between claim frequency and claim severity, so both random variables can be modeled with a copula.

Modeling the Generalized Linear Model (GLM) for the random variables of claim frequency and claim severity shows that the covariates consisting of deductible and gender have a significant impact on claim frequency and claim severity as the response variables. The resulting marginal model is then modeled using copula-based regression with the copula used being a member of the Archimedean copula family. The results of the parameter estimation, both the estimation of the independent marginal model parameters and the estimation of the copula-based regression model parameters, are presented in the following Table 1.

Table 1. The Result of the Regression Parameter Estimation

	Parameter	**Copula Clayton**
α *(Severity* Klaim*)*	*Intercept*	15,43800
	Deductible	0,00000015
	Jenis Kelamin	0,271236
β (Frekuensi Klaim)	*Intercept*	1,130574
	Deductible	0,00000009
	Jenis Kelamin	0,158942
	Δ	0,7541653
	Θ	0,03100265
	Loglikelihood	-83.089,30

The estimation results explain that the frequency of claims and the severity of claims have a positive dependency relationship, which means that the frequency of claims and the severity of claims have a unidirectional dependency relationship. The results of the claim frequency and severity predictions are

presented in Table 2 below.

Table 2. The Prediction of Claim Frequency and Claim Severity

Individu	Claim Frequency		Claim Severity	
	Year 1	Clayton	Year 1	Clayton
1	3	5	9.605.345	9.605.362
2	1	3	15.231.951	15.231.968
3	1	3	20.939.809	20.939.826
4	2	4	11.529.187	11.529.204
5	1	3	483.500	483.517
6	1	3	5.007.180	5.007.198
7	1	3	8.326.800	8.326.818
8	1	3	18.253.200	18.253.217
9	1	3	4.663.710	4.663.728
10	1	3	8.924.995	8.925.013
…	…	…	…	…
696	3	5	4.772.636	4.772.654

Based on Table 2, it can be seen that there is a difference between the first-year claim data and the predicted data using copula-based regression. The prediction was made for each individual whose claim had been paid during the first year. The predicted policy loss was obtained by multiplying the predicted claim frequency by the predicted claim severity. The predicted policy loss results are presented in the following Table 3.

Table 3. The Results of Loss Policy Prediction

Year 1	*Clayton*
28.816.034	48.026.810
15.231.951	45.695.904
20.939.809	62.819.478
23.058.373	46.116.816
483.500	1.450.551
5.007.180	15.021.594
8.326.800	24.980.454
18.253.200	54.759.651
4.663.710	13.991.184
8.924.995	26.775.039
…	…
14.317.908	23.863.270

The difference between the first-year claim data and the prediction results from the copula-based regression model is due to the claim data used being volatile over the 5-year period. Furthermore, the risk factors used in this study are limited to only deductible and gender. The limitation of these risk factors is due to the limited data that can be analyzed at XYZ insurance company.

Although there is a difference in claim data that occurred in the first year and the prediction results from the copula-based regression model, the copula-based regression model can be used to estimate the projected claim load in the budgeting of claim load for XYZ insurance company. This will help provide better estimation results compared to the methods currently used by XYZ insurance company.

So that there is no loss of profit in determining the premium and budget for the following year. The method used by XYZ insurance company is relatively simple, namely calculating the claim ratio and assuming the growth rate of policies in the following period.

CONCLUSION

Based on the research results on inpatient health insurance claim data at XYZ Insurance Company, the following is the results of the correlation test between the random variables of claim frequency and claim severity show a dependent relationship between claim frequency and claim severity. The correlation between the two random variables is shown by a Kendall's tau correlation value of 0.9071304, indicating that the frequency of claims and the severity of claims have a positive dependent relationship. The XYZ insurance company can apply copula-based regression in determining claim loss estimates to obtain accurate estimation results, thereby minimizing the significant difference between predicted claim burden and actual claim burden. The estimated loss from health insurance policy claims can be used as one of the references to evaluate premium rates in the subsequent period. Futhermore, subsequent research can include risk factors such as the type of occupation and the location of the insured's residence as covariates that influence the number of health insurance claims.

REFERENCE

[1] UURI, *Undang-Undang Republik Indonesia Nomor 40 Tahun 2014 Tentang Perasuransian*, Jakarta: Republik Indonesia, 2014.

[2] K. Iskandar, N. Fuad, F. Wirasadi and K. Sendra, Dasar-Dasar Asuransi: Jiwa, Kesehatan, dan Anuitas, Jakarta: Asosiasi Ahli Manajemen Asuransi Indonesia (AAMAI), 2011.

[3] M. Nadjib and dkk, Dasar-Dasar Asuransi Kesehatan, Jakarta: PAMJAKI, 2005.

[4] C. Czado, R. Kastenmeier, E. Brechmann and A. Min, "A Mixed Copula Model for Insurance Claims and Claim Sizes," *Scandinavian Actuarial Journal*, vol. 4, no. 1, pp. 278-305, 2011.

[5] S. A. Klugman, H. H. Panjer and G. E. Wilmot, Loss Model From Data to Decision 3rd Edition, New York: John Wiley & Sons, Inc, 2008.

[6] N. Kramer, E. C. Brechmann, D. Silvestrini and C. Czado, "Total Loss Estimation Copula-Based Regression Models," *North Holland Publication Co.*, vol. 53, no. 3, pp. 829-839, 2012.

[7] X. Zhao and X. Zhou, "Copula-Based Dependence Between Frequency and Class in Car Insurance with Excess-Zeros," *Operations Research Letters*, vol. 42, no. 1, pp. 273-277, 2014.

[8] R. B. Nelsen, An Introduction to Copula 2nd Edition, New York: Springer Science & Business, Inc., 2006.

[9] P. D. Jong and G. Z. Heller, Generalized Linear Models for Insurance Data, New York: Cambridge University Press, 2008.

[10] E. W. Frees, Copula and Regression, Madison: University of Wisconsin, 2009.

[11] A. Charpentier, Computational Actuarial Science with R, New York: CRC Press, 2015.