Pakistan Journal of Statistics and Operation Research

Generalized Linear Models for Loss Calculation in General Insurance

Dian Lestari[1*], Raymond Tanujaya[2], Rahmat Al Kafi[3], Sindy Devila[4]

* Corresponding Author

1. Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Indonesia, Depok 16424, Indonesia, dian.lestari@sci.ui.ac.id
2. Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Indonesia, Depok 16424, Indonesia, raymond.tanujaya@sci.ui.ac.id
3. Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Indonesia, Depok 16424, Indonesia, rahmat.alkafi@sci.ui.ac.id
4. Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Indonesia, Depok 16424, Indonesia, sindy@sci.ui.ac.id

Abstract

In most cases, loss in general insurance is calculated based on claim severity and frequency and an assumption of independence. However, in some cases, claim severity depends upon the claim frequency. This paper presents the derivation of aggregate loss calculation by modeling claim severity and frequency as the assumption of independence is eliminated. The authors modeled average claim severity using claim frequency as the covariate to induce the dependence among them. For that purpose, we use the generalized linear models. The calculated loss is obtained after the parameter estimation process.

Key Words: Aggregate Loss; Claim Frequency; Claim Severity; Dependency.

Mathematical Subject Classification: 97M30

1. Introduction

To produce a policy of insurance, an insurer needs a comprehensive calculation. The calculation includes the risk that highly potentially be imposed by the insurer. In this case, how much the aggregate loss required to cover the risk of the policyholders. Aggregate loss is composed of claim frequency and claim severity. Jorgensen and de Souza (1994) developed the loss calculation mathematically within an assumption that claim frequency and severity are independent. Then the model was reviewed deeply by Quijano-Xacur and Garrido (2015). Nevertheless, in practice, we often meet dependency between claim frequency and severity.

Frees and Wang (2006) introduced dependency between claim frequency and severity. Afterward, Frees et al. (2011) modeled average severity used frequency as a predictor for severity claim. Czado et al. (2012) link marginal frequency to severity using copula. Shi et al. (2015) modeled regression of average severities by applying frequency claim as the covariate and make a comparison against mixed copula approached to construct the joint distribution of frequency and severity claims.

Generalized Linear Models (GLM) have commonly been applied to model insurance claims. Montgomery et al. (2012) explained GLM as a unique linear regression method that uniting the usual normal-theory linear regression models and nonlinear models such as logistic and Poisson regression. A fundamental assumption in the GLM is that the response variable distribution is a member of the exponential dispersion family (EDF). Jong and Heller (2008) claimed that GLM is a favorite model because in insurance data, more frequently, the data distribution is a member of EDF

rather than Normal distribution. Garrido et al. (2016) used GLM to simulate the frequency and severity of non-life insurance claims as independent and dependent components.

This paper adopts the model which was developed by Garrido et al. (2016). The model used the assumption that the frequency follows the Poisson distribution and the severity follows Gamma distribution. However, the wide variety of policyholders' characteristics impact the claim frequency and sometimes potentially inflict an over-dispersion. In other words, the variance of data will grow higher than the mean, which leads to an increment of residual. Therefore, the claim frequency distribution on Garrido's model needs to develop. This paper used Negative Binomial distribution as a counting distribution for claim frequency.

2. Loss Modeling

A loss event, which an insurance company experiences, is an accumulation of aggregate loss submitted by policyholders. Claim frequency N of the company is uncertain, and it follows discrete random variable with positive integer values. Each claim mostly has a random amount. The amount of j-th claim is denoted by Y_j, which follows a continuous random variable. Hence, the aggregate loss is denoted as S and given as follows

$$\mathrm{S} = \sum_{j=1}^{N} Y_j.$$

The successive claims are assumed to be under the same distribution and independent. It is important to note that N and Y are members of the EDF, which means the probability density function follow

$$f_{\mathrm{X}}(x, \theta, \phi) = \exp\left[\frac{x\theta - k(\theta)}{a(\phi)} + C(x, \phi)\right],$$

where k, a, and C are the specified functions. θ is the canonical parameter, and ϕ is the dispersion parameter.

Garrido et al. (2016) described $\bar{Y} = \frac{(Y_1 + Y_2 + \cdots + Y_N)}{N}$ as the average claim severity. Directly, we can see that $\bar{Y}$ contains the claim count component as the covariate. This condition shows the dependency on frequency and severity in this model. On the other hand, the latter explanation state the loss model as the product of the average claim severity and the claim frequency.

Suppose that individual number of claims are mutually independent and member of the EDF, $Y_j \sim \mathrm{EDF}(\mu, \phi)$ for all positive integers j. Based on cumulant-generating function properties, for $N > 0$, $\bar{Y} \sim \mathrm{EDF}\left(\mu, \frac{\phi}{N}\right)$ (Frees et al., 2016).

As the common regression, we will have a set of covariate as explanatory variables. Let $x = \{x_1, x_2, \ldots, x_s\}$ represents the set of s explanatory variables. GLM for N and $\bar{Y}$ is used to describe their relations with the explanatory variables, which are not linear as the general linear model relation has. In general formulation, g_N and g_Y are given by Eq. (1) and Eq. (2), respectively.

$$v = E(N|x) = g_N^{-1}(\boldsymbol{x^T \alpha}), \quad (1)$$

$$\mu_\theta = E(\bar{Y}|N, x) = g_{\bar{Y}}^{-1}(\boldsymbol{x^T \beta} + \theta N), \quad (2)$$

where $\boldsymbol{x^T}$ is $1 \times s$ vector of explanatory variables. $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are $s \times 1$ vector of regression coefficients which explain N and $\bar{Y}$, respectively. $\theta \in \mathbb{R}$ is the parameter that represents the degree of dependence between N and $\bar{Y}$. For some $k \in (1,2, \ldots, s)$, α_k or β_k may zero deliberately if the corresponding explanatory variables are known to not affect the given expected value.

The parameter of μ_θ is the expected value of $\bar{Y}$ given frequency claim and explanatory variables. If $\theta = 0$, then $\mu_\theta = \mu_0 = g_{\bar{Y}}^{-1}(\boldsymbol{x^T \beta})$ which is equivalent to expected values of individual severities with the assumption that the frequency and the severity claims are independent. If $\theta \neq 0$, then

$$E(S|\boldsymbol{x}) = E[E(S|N, \boldsymbol{x})|\boldsymbol{x}] = E[E(N\bar{Y}|N, \boldsymbol{x})|\boldsymbol{x}] = E[NE(\bar{Y}|N, \boldsymbol{x})|\boldsymbol{x}] = E(N\mu_\theta). \quad (3)$$

In this model, a log link is chosen to relate the explanatory variables to the expected frequency and severity claims. Hence, for the mean value of average severity claims, we have

$$\ln(\mu_\theta) = \boldsymbol{x^T \beta} + \theta N \leftrightarrow \mu_\theta = \exp(\boldsymbol{x^T \beta} + \theta N) = \mu e^{\theta N}, \quad (4)$$

where μ denotes the expected value of the average severity claims when the degree of dependence is 0 (i.e., frequency and severity claims are independent). From Eq. (3) and Eq. (4), we obtain

$$E(S|\boldsymbol{x}) = E\left(N\mu e^{\theta N}\middle|\boldsymbol{x}\right) = \mu E\left(Ne^{\theta N}\middle|\boldsymbol{x}\right) = \mu M_N{}'(\theta|\boldsymbol{x}),$$

where M_n is the moment generating function of N based on GLM and $M_N{}'$ is the first derivative of M_n with respect to θ.

It is relatively simple to derive the variance of aggregate claims when $\theta = 0$. But for the dependent model, it is more complicated and does not lead to a simple form. By the law of total variance, we have

$$Var(S|\boldsymbol{x}) = \phi_\theta E\left[NV_Y\left(\mu e^{\theta N}\right)\right] + \mu^2\left[\frac{1}{4}M_N''(2\theta) - \{M_N'(\theta)\}^2\right],$$

where ϕ_θ is the dispersion parameter of severity distribution in EDF representation, V_Y is the variance function of severity, and M_N'' is the second derivative of M_n with respect to θ.

For $N \sim \mathcal{NB}(r,p)$ and $\bar{Y} \sim \mathcal{G}\left(\mu_\theta, \frac{\phi_\theta}{N}\right)$, where $\mathcal{NB}$ denotes a Negative Binomial distribution with number of failures given by r and success probability given by p, $0 < p < 1$, and $\mathcal{G}$ denotes a Gamma distribution. Hence the expected value and the variance of the aggregate claims are given by Eq. (5) and Eq. (6), respectively

$$E(S|\boldsymbol{x}) = \mu\left(\frac{r(1-p)}{p}\right)e^{\theta}\left[\frac{p}{(1-(1-p)e^{\theta})}\right]^{r+1}, \tag{5}$$

$$\begin{aligned} Var(S|\boldsymbol{x}) = {} & \phi_\theta\mu^2(p-1)re^{2\theta}\left[(p-1)e^{2\theta}+p\right]^{r-1} \\ & + \frac{1}{4}\left[e^{2\theta} - (p-1)\left\{(p-1)e^{2\theta}+p\right\}^{r-2}r\left(pre^{2\theta} - re^{2\theta} + p\right)\right] \\ & - (p-1)^2r^2e^{2\theta} - \left[(p-1)e^{\theta}+p\right]^{2r-2}. \end{aligned} \tag{6}$$

Eq. (5) and Eq. (6) are derived based on the moment generating function of N, that is $\left[\frac{p}{(1-(1-p)e^{\theta})}\right]^{r+1}$.

3. Parameter Estimation

As mentioned in the previous section, there are s explanatory variables. Suppose there are m policyholders, $i \in \{1,2,\dots,m\}$, let S_i be the total claim size and $\bar{Y} = \frac{S}{N}$ when $N > 0$. Based on GLM structure for claim frequency and severity components of the aggregate claims, $E(\bar{Y}|N_i)$ and $E(N_i)$ can be expressed as

$$v_i = e^{x_i\alpha},$$
$$\mu_{\theta i} = e^{x_i\beta + \theta n_i},$$

where v and μ_θ respectively denote the expected value of claim frequency and severity. Denote f_N and $f_{\bar{Y}|N}$ respectively as the marginal density function of frequency and conditional density function of severity. The likelihood function of m joint density functions is given by Eq. (7)

$$L(\boldsymbol{\alpha},\boldsymbol{\beta},\theta;\boldsymbol{n},\boldsymbol{y}) = \prod_{i=1}^{m} f_{\bar{Y}|N}(y_i|n_i)f_N(n_i). \tag{7}$$

Hence for general EDF distribution, the log likelihood of m joint density functions is given by

$$\begin{aligned} l(\boldsymbol{\alpha},\boldsymbol{\beta},\theta;\boldsymbol{n},\boldsymbol{y}) = {} & \sum_{i=1}^{m}\left(\frac{n_i\alpha_i - k_N(\alpha_i)}{a_{iN}(\phi_N)}\right) + \sum_{i=1}^{m}C(n_i,\phi_N) + \sum_{i=1}^{m}\left(\frac{y_i\beta_i - k_Y(\beta_i)}{a_{iY}(\phi_Y)}\right) \\ & + \sum_{i=1}^{m}C(y_i,\phi_Y), \end{aligned}$$

The estimated value of $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1,\hat{\alpha}_2,\dots,\hat{\alpha}_s)^T$, $\hat{\boldsymbol{\beta}} = \left(\hat{\beta}_1,\hat{\beta}_2,\dots,\hat{\beta}_s\right)^T$, and $\hat{\theta}$ can be obtained by solving the following system:

$$\frac{\partial}{\partial\alpha_i} = \sum_{i=1}^{m}\left(\frac{(n_i - v_i)x_{ik}}{\ln(1+v_i)}\right) = 0, \qquad i = 1,2,\dots,s$$

$$\frac{\partial}{\partial\beta_i} = \sum_{i=1}^{m}\left(\frac{n_i}{\phi_\theta}\frac{x_{ik}}{\mu_{\theta i}}(y_i - \mu_{\theta i})\right) = 0, \qquad i = 1,2,\dots,s$$

$$\frac{\partial}{\partial\theta} = \sum_{i=1}^{m}\left(\frac{n_i}{\phi_\theta}\frac{n_i}{\mu_{\theta i}}(y_i - \mu_{\theta i})\right) = 0,$$

under the assumption that $N_i \sim \mathcal{NB}(r_i, p_i)$ and $\bar{Y}_i|N_i \sim \mathcal{G}\left(\mu_{\theta i}, \frac{\phi_\theta}{N_i}\right)$.

4. Results and Discussion

As an illustration of how this model works to calculate aggregate loss under the assumption that claim frequency and average severity are dependent where the claim frequency follows negative binomial distribution and the average severity follows gamma distribution, we present a fictive portfolio involving 1000 policyholders. We also generated two information (x_1 and x_2) as covariates for claim frequency, as x_3 and x_4 for claim severity. They are generated following half-normal distribution.

x_{i1} and x_{i2} denote the first and second explanatory for claim frequency variables for policyholder i as x_{i3} and x_{i4} explain claim severity variables, with $i = 1,2,\ldots,1000$. Therefore, based on the policyholders, the expected claim frequency and claim severity constructed by

$$v = E(N|x) = g_N^{-1}(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2),$$

$$\mu_\theta = E(\bar{Y}|N, x) = g_Y^{-1}(\beta_0 + \beta_1 x_3 + \beta_2 x_4 + \theta N).$$

After doing calculation in R used glm function with log-links, we find

$$\hat{v}_i = \exp(2.6949 - 0.1489x_{i1} + 0.0742x_{i2}),$$

$$\hat{\mu}_{\theta i} = \exp(2.4464 - 0.1751x_{i3} - 0.063x_{i4} + 0.0812N_i).$$

Standard error and p-value test for $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}$ respectively are shown in Table 1 and Table 2. From the table, by looking at the t-test, the null hypothesis H_0 the predictor well explains the response variable is rejected for α_2. So, we are unable to say α_2 well significant to describe N. For the same reason, we can say that β_2 is not substantial to explain $\bar{Y}$. It may happen because this model simulation just used a generated observation or the using of the unappropriated link function. On the other hand, we have a good estimation for α_1, β_1, and θ. It means we have N well explain the value of $\bar{Y}$. By inspection, $\theta > 0$ implies average severity positively correlated to claim frequency.

Table 1: Standard error and p-value for $\widehat{\boldsymbol{\alpha}}$.

	Estimated Value	Standard Error	p-value
α_0	2.6949	0.1256	$< 2 \times 10^{-16}$
α_1	-0.1489	0.0577	0.01
α_2	0.0742	0.0548	0.176

Table 2: Standard error and p-value for $\widehat{\boldsymbol{\beta}}$ and θ.

	Estimated Value	Standard Error	p-value
β_0	2.4464	0.1383	$< 2 \times 10^{-16}$
β_1	-0.1751	0.073	0.0168
β_2	-0.063	0.0689	0.3603
θ	0.0812	0.0288	0.0049

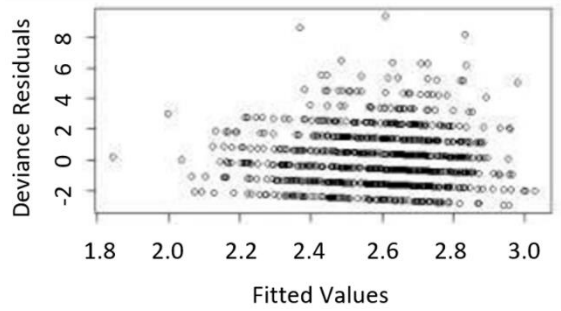

Figure 1: Deviance residuals plot versus fitted values for claim frequency

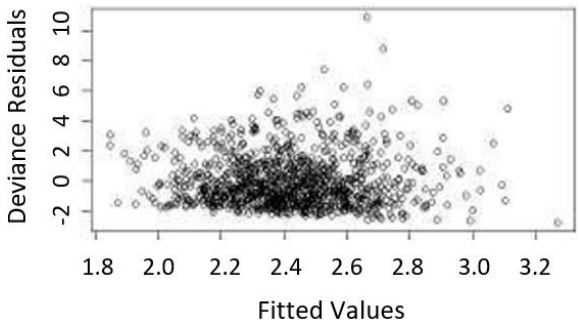

Figure 2: Deviance residuals plot versus fitted values for average severity

Fig. 1 and Fig. 2 given above show the plots of the fitted values versus the deviance residuals. The plot in Fig. 1 shows deviance residuals on fitted values for claim frequency while Fig. 2 for average severity. The plot in Fig. 1 looked more tenuous than the plot in Fig. 2. It indicates claim frequency's residual more vary than the average severity residual. For average severity, we see an adequate model because the residual points are centered near zero.

From the parameter estimation, we have all required components to estimate loss for the *i*-th policyholder, make use of $\hat{v}_i$ and $\hat{\mu}_{\theta i}$. From Eq. (4), we obtain

$$\widehat{E(S_i)} = E(\widehat{S_i \mu_{\theta i}}) = \hat{\mu}_{\theta i} \widehat{E(N_i)} = e^{5.1413+0.1489x_{i1}+0.742x_{i2}-0.1751x_{i3}-0.063x_{i4}+0.0812N_i}.$$

Furthermore, using an individual estimated loss, we can calculate the pure premium of a policy insurance.

5. Conclusion

This paper has described an aggregate claim model. The model is standing with an assumption that claim severity and claim frequency is dependent by modeling average claim severity conditioning to the claim frequency. This assumption makes the model more flexible to use on real data rather than the independent one. Even though the model is not a new one, this paper presents a new condition on this model, that is, the distribution of claim frequency follows the Negative Binomial distribution. Based on their characteristic, the Negative Binomial distribution would be better to overcome data with a heavier tail than the Poisson distribution model. Poisson regression frequently impacts an overdispersion. Negative Binomial distribution will help overcome this matter, especially when the data do not fit the Poisson distribution well.

On the contrary, it fitted the Negative Binomial better. In the simulation section, we have seen that not all estimated parameters are significant to explain claim severity and claim frequency. However, this model can estimate the degree of dependency very well because it is not rejected by the statistical test even though $\bar{Y}$ and N's observation are generated separately. Theoretically, it is quite clear that the model under the Negative Binomial distribution is more complicated than the Poisson one. Nevertheless, it is essential to develop this model by applying another distribution, either the discrete or the continuous type distribution. It is also interesting to consider another way to see dependency between claim frequency and the average severity, not only on the linear form described in this paper.

Acknowledgment: This work is fully funded by Publikasi Terindeks Internasional (PUTI) 2020 research grant from DRPM Universitas Indonesia (Project ID: NKB-1965/UN2.RST/HKP.05.00/2020).

References

1. Czado, C., Kastenmeier, R., Brechmann, E. C. and Min, A. (2012). A mixed copula model for insurance claims and claim size. Scandinavian Actuarial Journal, 2012(4), 278-305.
2. Frees, E. W. and Wang, P. (2006). Copula credibility for aggregate loss models. Insurance: Mathematics and Economics, 38(2), 360-373.
3. Frees, E. W., Gao, J. and Rosenberg, M. A. (2011). Predicting the frequency and amount of health care expenditures. North American Actuarial Journal, 15(3), 377-392.
4. Garrido, J., Genest, C. and Schulz, J. (2016). Generalized linear models for dependent frequency and severity of insurance claims. Insurance: Mathematics and Economics, 70(C), 205-215.
5. Jorgensen, B. and de Souza, M. C. P. (1994). Fitting Tweedie's compound Poisson model to insurance claims data. Scandinavian Actuarial Journal, 1994(1), 69-93.
6. Jong, P. D. and Heller, G.Z. (2008). Generalized Linear Models for Insurance Data. Cambridge University Press. New York.
7. Montgomery, C. Douglas, Peck, E. A. and Vining, G. G. (2012). Introduction to Linear Regression Analysis 5th ed. John Wiley & Sons, Inc. New Jersey.
8. Quijano-Xacur, O. A. and Garrido, J. (2015). Generalised linear models for aggregate claims: to Tweedie or not?. European Actuarial Journal, 5, 181-202.
9. Shi, P., Feng, X. and Ivantsova. A. (2015). Dependent frequency-severity modeling of insurance claims. Insurance: Mathematics and Economics, 64(C), 417-428.