Generalized Linear Mixed Models for Predicting Non-Life Insurance Claims

Johana Daniella Budhyanto[1], Kie Van Ivanky Saputra[1*], Helena Margaretha[1] and Ferry Vincenttius Ferdinand[1]
[1]Author Affiliation: Department of Mathematics, Universitas Pelita Harapan, Tangerang, Indonesia
*Corresponding author: kie.saputra@uph.edu

Received: 23 May 2024
Revised: 3 July 2024
Accepted: 4 July 2024

ABSTRACT – Generalized linear mixed models (or GLMMs) are an extension of linear mixed models to allow response variables from different distributions. Alternatively, GLMMs are an extension of generalized linear models (GLMs) to include both fixed and random effects (hence mixed models) that can be used as a modeling approach that allows the modeling of nonlinear behaviors and non-Gaussian distributions of residues. These models are very useful for general insurance claim predictions, where the frequency and the severity of claims distributions are usually non-Gaussian. In our research, we shall compare the performance of GLMS and that of GLMMS to estimate the aggregate of claims of auto insurance. The data used are a secondary dataset which is the motor vehicle dataset from Australia named ausprivauto0405. The results of our research suggest that GLMMs approach does not always give the best estimations and even in some cases GLMs outperform GLMMs. The accuracy of the models was compared to choosing the best model for determining pure insurance premiums using R software. More investigation using different models is needed to ensure which model is more appropriate for estimating the aggregate of insurance claims.

Keywords– Aggregate Losses, Auto Insurance, Generalized Linear Models, Generalized Linear Mixed Models, Pure Premium.

I. INTRODUCTION

Insurance is a contract involving two parties: the insured or policyholder and the guarantor or insurance company. The insurance company needs to estimate all future claims related to the protection for policyholders. In return, the insured is required to pay premiums at regular intervals during the coverage period. A claim is an official request made by the insured to the insurance company to compensate for or cover losses included in the terms of the insurance policy. Over time, there has been an increasing trend among practitioners to model losses using Generalized Linear Models (GLM). GLM is an extension of the common linear regression model. This model is very useful in general insurance because the severity (amount) and frequency of claims do not follow a normal distribution.

In generalized linear models (GLMs), the assumption of independence of observations is generally maintained, similar to common linear regression models. However, just like in linear models, this assumption can be problematic when dealing with certain types of data, such as repeated measures, clustered data, or longitudinal data, where observations are naturally grouped and may be correlated. Here is an illustration of the dependence between observations that exists in auto insurance data. Suppose an insurance company has data on five types of vehicles. Each type of vehicle has different specifications, leading to distinct loss characteristics among the types of vehicles. However, the loss characteristics among vehicle owners within the same group of vehicles are likely to be correlated due to the similarity in vehicle specifications. Ignoring this correlation by using standard GLM can lead to inaccurate estimates and suboptimal decisions. By using GLMM, we hope to achieve more precise and reliable predictions, ultimately leading to better risk management and pricing strategies in auto insurance.

GLM has been used to model losses from claim severity or claim frequency, one example is Smyth and Jørgensen [1] in 2002. In GLM modeling, the compound Poisson-gamma is often referred to as a Tweedie distribution. This paper used the fact that the arrival of claims is Poisson distributed and the cost for individual claims is gamma distributed therefore Tweedie's compound Poisson distribution provided a more highly efficient method. Not long after, De Jong and Heller [2] produced an excellent resource for actuaries when they needed to understand generalized linear models (GLMs) for insurance applications. At that time, no text had introduced GLMs in this context or addressed the problems specific to insurance data. The reader will find this book resourceful. Recently, in 2014 Kafková and Křivánková [3] tried to find the best model for an estimation of insurance premium. Their models depend on many risk factors, e.g. the car characteristics and the profile of the driver. They utilized portfolio of vehicle insurance data and performed a generalized linear model (GLM) to predict the relation of annual claim frequency on given risk factors. In 2015, Frees and Lee [4] describe a modeling process for determining rating endorsement, based on GLM techniques. It is common for insurance policies to contain optional insurance coverages, often referred to as endorsements or riders. They consider the Wisconsin Local Government Property Insurance Fund and provide a detailed case study. Through GLM techniques, they provided an approach for handling these optional coverages when it is not known whether a claim is due to an endorsement. The following papers [5-8] utilized the Generalized Linear Model (GLM) methodology to model auto insurance premiums, assuming independence between claim frequency and severity. These studies

DOI: 10.12962/j27213862.v7i2.20447

typically relied on extensive datasets that included detailed information on insurance policies, such as policyholder demographics, vehicle characteristics, and historical claim records. By analyzing this data, the researchers were able to develop models that predict the premium costs based on factors influencing both the likelihood and the cost of claims. Finally, Frees et al. [9] relaxed the independent assumption between claim frequency and severity and used GLM along with copula model to improve loss calculations. To the authors' knowledge, this paper is the first instance in loss modeling that does not rely on the independence assumption between frequency and severity.

As we discussed before, these kinds of studies typically relied on datasets that included detailed information on insurance policies. In this paper, we used the ausprivauto0405 dataset [10] which includes insurance claims from Australian motor insurance policies during the 2004-2005. Several authors have previously utilized this dataset, for instances Giancaterino [11], Oktavia et al. [12], Pernagallo et al. [13] and Gao and Li [14]. GLM approach to predict loss claims has been applied [11, 12] and some also [11, 13] used general additive modeling. Recently, Gao and Li [14] also used this data and used mixed copula approach to model dependency between claim counts and claim amounts. Other than research article, Frees and Huang published an online supplement for working actuaries when modeling insurance price [15]. One of the datasets they used as an example is the ausprivauto0405 dataset.

While generalized linear models (GLMs) are popular and widely used, they do have some limitations. One key limitation is that GLMs require the observations to be independent of each other. In other words, GLMs handle only fixed effects and do not account for random effects that may arise from hierarchical or multi-level data structures. Perhaps, one of the earliest research projects dealing with modeling loss with GLMM was by Yau et al. [16] and Antonio and Beirlant [17]. Study by Yau et al [16] demonstrates the advantage of the GLMM technique and has shown that the GLMM estimation method is, in general, more accurate in terms of the average bias and MSE, especially when the random effect variance is moderate to large. While Antonio and Beirlant [17] discussed generalized linear mixed models as a tool for modelling actuarial longitudinal data. Concepts on model formulation, estimation, inference, and prediction of GLMM are discussed and they used both a maximum likelihood and Bayesian approaches. They also described various applications of GLMMs in the domains of credibility, non-life ratemaking, credit risk modelling and loss reserving. In the last decade, there has been extensive literature discussing applications of GLMM in insurance, for instance Antonio et al. [18], Kim et al. [19], Rohmaniah and Chandra [20], Wang et al. [21], Günther et al. [22] and Lee et al. [23]. The reader can also find applications of GLMM in another field in [24-27].

In this paper, we will model the loss obtained from claims in auto insurance using two different methodologies: the standard Generalized Linear Model (GLM) and the Generalized Linear Mixed Model (GLMM). These models will allow us to explore and compare different assumptions about the relationships between the variables involved in the insurance claims process. We will conduct a detailed analysis of the results obtained from both models, comparing their performance in terms of predictive accuracy. To do this, we will use Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) to evaluate the individual predictions of each model. Additionally, we will compare the aggregate predictions by examining the sum and the average of all claims' predictions. This comprehensive analysis will help identify which model offers more reliable and robust predictions. We hope it can help insurance companies in reserving adequate funds to cover future claims. This will ultimately contribute to better financial planning and risk management within the insurance industry.

II. METHODOLOGY

The data utilized in this study originates from the **ausprivauto0405** dataset, extracted from the **CASDatasets** package in R [10]. This dataset comprises insurance claims from Australian motor insurance policies during the 2004-2005 period, encompassing a total of 67,856 policies, of which 4,624 policies have at least one claim recorded. In Table 1, one can see all variables in this dataset.

Table 1 Variables used in this research

Variables	Explanations	Variables	Explanations
ClaimAmount	Claim amount per policy	Gender	Gender of the policyholder
ClaimNb	Number of claims per policy	VehBody	Type of vehicles
ClaimOcc	Occurrence of claim per policy	VehValue	Value of vehicles
DrivAge	Driver's age	Exposure	Number of policy years
VehAge	Vehicle's age	ClaimIndAmount-	Claim amount per claim per policy

Initially, a data cleaning process is performed to address missing features. The next step is to perform random partitioning of the data into training and testing sets. Subsequently, a mathematical model is developed using the training set, with model performance evaluated using the testing set.

To construct a predictive model for aggregate claim amounts, assumptions regarding the distribution of response variables are necessary. The frequency of claims is represented by either the ClaimNb or ClaimOcc variables, denoting the claim count and claim occurrence, respectively. ClaimNb is modeled using either a Poisson distribution or a Negative Binomial distribution, while ClaimOcc is modeled using a Bernoulli distribution. On the other hand, the severity of claims is modeled using either the ClaimAmount variable or a newly defined variable, ClaimIndAmount. ClaimAmount denotes the aggregate claim amount per policy, while ClaimIndAmount represents the individual claim

DOI: 10.12962/j27213862.v7i2.20447

amount, computed as ClaimAmount divided by ClaimNb (only if ClaimNb is non-zero). Both severity variables are modeled using either a Gamma distribution or Inverse Gaussian distribution.

Following the distribution assumptions, Generalized Linear Models (GLMs) will be constructed for both the frequency (N) and severity (Y) variables, with respect to the independent variables. Since the model aims to predict either count data or non-negative data, appropriate link functions must be applied to each response variable. For distributions such as Negative Binomial, Poisson, and Gamma, the natural logarithm will serve as the link function. When the response variable follows a binomial distribution, the logit function will be utilized. Finally, for response variables with an Inverse Gaussian distribution, the identity link function will be employed. Additionally, the independent variable Exposure, representing the duration of each individual policy within the insurance company, will be incorporated into the modeling process.

To predict the total loss for each policy in the upcoming period, it is assumed that the frequency and severity variables are independent, as denoted by:

$$\boldsymbol{E}(\boldsymbol{S}) = \boldsymbol{E}(\boldsymbol{N})\boldsymbol{E}(\boldsymbol{Y}) \tag{1}$$

Here, E(S) represents the expected total claim amount for each policy, E(N) denotes the expected number of claims per policy, and E(Y) signifies the expected claim amount per policy. Both E(N) and E(Y) will be modeled using GLMs. The GLM models for E(N) and E(Y) will be combined to obtain the most accurate estimate for E(S).

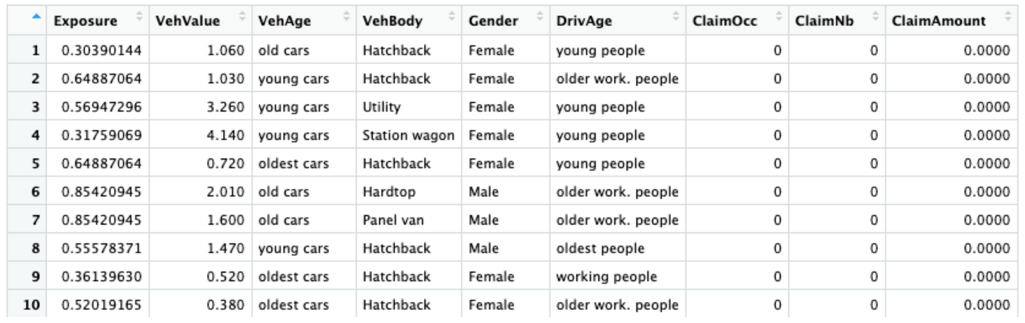
A key objective of this study is to compare the predictions obtained from the standard Generalized Linear Model (GLM) methodology with those obtained from the Generalized Linear Mixed Model (GLMM) methodology. Following the same procedure as described above, we will now employ the GLMM methodology. In this phase, the variable VehBody will be designated as the random effect in our GLMM. VehBody comprises 13 distinct vehicle types and is selected as the random effect due to the potential correlation among the response variables within each VehBody category. The distribution assumptions for each response variable and the procedure for computing aggregate losses will remain consistent, with the only variation being the incorporation of GLMM into the methodology.

In the final stage of analysis, all models developed using the training set will be evaluated using the testing set. We will compare the performance of both GLM and GLMM models. The evaluation will cover the analysis of individual losses as well as aggregate losses incurred by the insurance company.

III. RESULTS AND DISCUSSIONS

A. Data Preparation

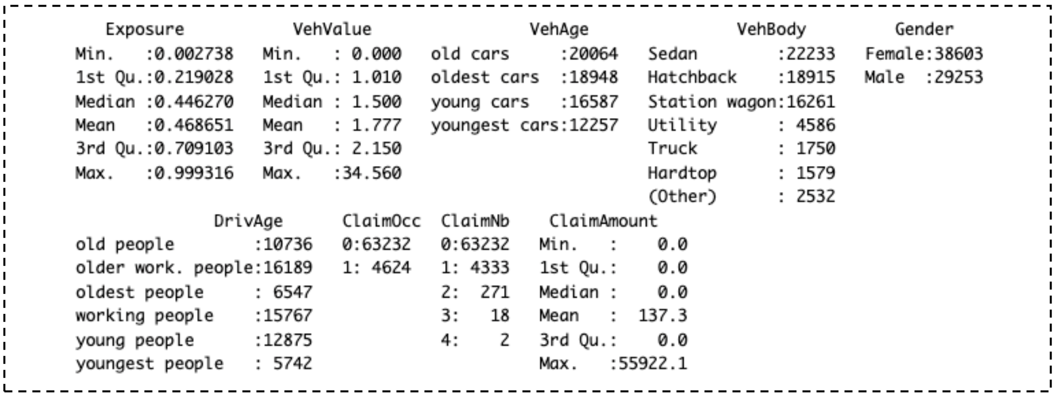
As mentioned before, this dataset comprises insurance claims from Australian motor insurance policies during the 2004-2005 period, encompassing a total of 67,856 policies. The reader can review the first ten entries of this dataset in Figure 1 and a summary of all nine variables can also be seen in Figure 2. There are four categorical variables (VehAge, VehBody, Gender and DriveAge), three numerical variables (Exposure, VehValue and ClaimAmount), and two counting variables (ClaimOcc and ClaimNb). Dependent variables in this dataset are ClaimOcc, ClaimNb and ClaimAmount; we will also define a new dependent variable, ClaimIndAmount, which is ClaimAmount divided by ClaimNb (only if ClaimNb is non-zero).



	Exposure	VehValue	VehAge	VehBody	Gender	DrivAge	ClaimOcc	ClaimNb	ClaimAmount
1	0.30390144	1.060	old cars	Hatchback	Female	young people	0	0	0.0000
2	0.64887064	1.030	young cars	Hatchback	Female	older work. people	0	0	0.0000
3	0.56947296	3.260	young cars	Utility	Female	young people	0	0	0.0000
4	0.31759069	4.140	young cars	Station wagon	Female	young people	0	0	0.0000
5	0.64887064	0.720	oldest cars	Hatchback	Female	young people	0	0	0.0000
6	0.85420945	2.010	old cars	Hardtop	Male	older work. people	0	0	0.0000
7	0.85420945	1.600	old cars	Panel van	Male	older work. people	0	0	0.0000
8	0.55578371	1.470	young cars	Hatchback	Male	oldest people	0	0	0.0000
9	0.36139630	0.520	oldest cars	Hatchback	Female	working people	0	0	0.0000
10	0.52019165	0.380	oldest cars	Hatchback	Female	older work. people	0	0	0.0000

Figure 1 The first ten entries of the dataset

DOI: 10.12962/j27213862.v7i2.20447



```
    Exposure          VehValue              VehAge             VehBody            Gender     
 Min.   :0.002738   Min.   : 0.000   old cars     :20064   Sedan        :22233   Female:38603  
 1st Qu.:0.219028   1st Qu.: 1.010   oldest cars  :18948   Hatchback    :18915   Male  :29253  
 Median :0.446270   Median : 1.500   young cars   :16587   Station wagon:16261                 
 Mean   :0.468651   Mean   : 1.777   youngest cars:12257   Utility      : 4586                 
 3rd Qu.:0.709103   3rd Qu.: 2.150                         Truck        : 1750                 
 Max.   :0.999316   Max.   :34.560                         Hardtop      : 1579                 
                                                           (Other)      : 2532                 
               DrivAge      ClaimOcc  ClaimNb     ClaimAmount    
 old people        :10736   0:63232   0:63232   Min.   :    0.0  
 older work. people:16189   1: 4624   1: 4333   1st Qu.:    0.0  
 oldest people     : 6547             2:  271   Median :    0.0  
 working people    :15767             3:   18   Mean   :  137.3  
 young people      :12875             4:    2   3rd Qu.:    0.0  
 youngest people   : 5742                       Max.   :55922.1  
```

Figure 2 Summary of each variable in the dataset

B. Frequency and Severity models using GLM

In this section, we are going to construct GLMs for both frequency and severity loss predictions. First, we will analyze the frequency models which we will build by using (i) the ClaimNb variable modelled with Poisson and Negative Binomial distributions and (ii) the ClaimOcc variable modelled with Bernoulli/Binomial distribution. The mathematical model of the GLM Poisson regression is given as follows:

$$\text{ClaimNb}_i \sim \text{Poisson}\ (\text{exposure}_i \times \lambda_i),$$

$$\lambda_i = \exp(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i})$$

where X1 represent vehicle's value, X2 represents vehicle's age. X3 represents vehicle's body, X4 represents driver's gender and X5 represents driver's age. The estimations of these coefficients are given in the Table 2. The mathematical model of the GLM Negative Binomial regression is given as follow:

$$\text{ClaimNb}_i \sim \text{Negative Binomial}\ (\text{exposure}_i \times \mu_i),$$

$$\mu_i = \exp(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i})$$

where X1 represent vehicle's value, X2 represents vehicle's age. X3 represents vehicle's body, X4 represents driver's gender and X5 represents driver's age. In Table 2, one can see the estimations of these coefficients along with the coefficients obtained from the GLM Poisson model. Both models give roughly the same coefficients estimations, some variables are also statistically significant in both models. The signs of every coefficient in both models agreed, meaning that each variable gives the same effect to the number of claims. Throughout this paper, we use "***" to indicate a significance level of less than 0.001, "**" to indicate a significance level of less than 0.01, and "*" to indicate a significance level of less than or equal to 0.05.

Table 2 Estimated value and p-value of GLM coefficients (GLM Poisson and GLM Negative Binomial)

Coefficient	Poisson Model Estimate	p-value	Negative Binomial Model Estimate	p-value
Intercept	-1.276935	0.00037	-1.87303	1.20E-06
VehValue	0.010849	0.57978	0.03392*	0.082139
VehAge: oldest	-0.087155*	0.05616	-0.07217	0.121852
VehAge: young	0.080218*	0.06999	0.06843	0.133820
VehAge: youngest	0.032976	0.54066	-0.08265	0.135853
VehBody: Convertible	-1.285732*	0.06192	-1.68452**	0.017997
VehBody: coupe	-0.281752	0.45251	-0.59998	0.135478
VehBody: hardtop	-0.681688*	0.06285	-0.80505**	0.040529
VehBody: hatchback	-0.857344**	0.01596	-0.99943***	0.008981
VehBody: minibus	-0.846507**	0.03057	-1.09349***	0.008780
VehBody: caravan	-0.185297	0.68103	-0.37168	0.439439
VehBody: panel van	-0.774887**	0.04141	-0.80714**	0.047139
VehBody: roadster	-1.183098	0.26534	-1.58475	0.143991
VehBody sedan:	-0.800148**	0.02426	-0.96035**	0.011906
VehBody: wagon	-0.758132**	0.03302	-0.94074**	0.013811
VehBody: truck	-0.850693**	0.02059	-1.00624**	0.010619
VehBody: utility	-0.953084***	0.00809	-1.14275***	0.003100
Gender Male	-0.026113	0.43603	-0.01260	0.715212
DrivAge: older	0.219950***	5.35E-05	0.19794***	0.000398

DOI: 10.12962/j27213862.v7i2.20447

Coefficient	Poisson Model Estimate	p-value	Negative Binomial Model Estimate	p-value
DrivAge: oldest	0.003188	0.96451	-0.01434	0.845080
DrivAge: working	0.252731***	3.61E-06	0.22942***	4.23E-05
DrivAge: young	0.283497***	5.40E-07	0.23730***	4.45E-05
DrivAge: youngest	0.474964***	5.31E-13	0.42191***	5.22E-10

The following is the mathematical model of the ClaimOcc variable (GLM Bernoulli):

$$\text{ClaimOcc}_i \sim \text{Bernoulli}(\text{exposure}_i \times p_i),$$

$$p_i = \frac{1}{1 + \exp[-(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i})]}$$

where all coefficients represent the same variables as the previous model. The value of the coefficients and the p-value are displayed in Table 3. The GLM Bernoulli here only predicts the occurrence of claims while the GLM Poisson and GLM Negative Binomial predict the number of claims in one year. The values and the signs of coefficient in these three models are nearly similar and when one variable is significant in one model, it will also be statistically significant in the other two models.

Table 3 Estimated value and p-value of coefficients of GLM Bernoulli

Coefficient	GLM Bernoulli Model Estimate	p-value	Coefficient	GLM Bernoulli Model Estimate	p-value
Intercept	-1.18743	0.005492	-	-	-
VehValue	0.01197	0.576743	VehBody: roadster	-1.21808	0.275441
VehAge: oldest	-0.10567**	0.032899	VehBody sedan:	-0.87365**	0.039456
VehAge: young	0.08043*	0.095590	VehBody: wagon	-0.80713*	0.057356
VehAge: youngest	0.04058	0.486980	VehBody: truck	-0.93982**	0.031375
VehBody: Convertible	-1.31929*	0.076933	VehBody: utility	-1.03268**	0.016038
VehBody: coupe	-0.31581	0.478468	Gender Male	-0.02435	0.503672
VehBody: hardtop	-0.69719	0.109318	DrivAge: older	0.21665***	0.000215
VehBody: hatchback	-0.90147**	0.033833	DrivAge: oldest	-0.02294	0.765959
VehBody: minibus	-0.86457*	0.059439	DrivAge: working	0.25682***	1.20E-05
VehBody: caravan	-0.16865	0.749724	DrivAge: young	0.28360***	3.29E-06
VehBody: panel van	-0.78584*	0.079838	DrivAge: youngest	0.51025***	8.66E-13

In the second part of this section, we are going to build a model for severity loss using Gamma distribution and Inverse Gaussian distribution. In total we are going to build four GLM models for severity loss variablewhich are (i) GLM Gamma for ClaimIndAmount (ii) GLM Inverse Gaussian for ClaimIndAmount, (iii) GLM Gamma for ClaimAmount and (iv) GLM Inverse Gaussian for ClaimAmount. The first two models will be paired with either GLM Poisson or GLM Negative Binomial while the last two models will be paired with GLM Bernoulli. Let us discuss the first two models. The following are the mathematical models for GLM Gamma and GLM Inverse Gaussian for ClaimIndAmount:

$$\text{ClaimIndAmount}_i \sim \text{Gamma}(\text{exposure}_i \times \mu_i),$$

$$\mu_i = \exp(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i})$$

and

$$\text{ClaimIndAmount}_i \sim \text{Inverse Gaussian}(\text{exposure}_i \times \mu_i),$$

$$\mu_i = \beta_0 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i}$$

where X1 represent vehicle’s value, X2 represents vehicle’s age. X3 represents the vehicle’s body, X4 represents the driver’s gender and X5 represents the driver’s age. In the GLM Inverse Gaussian, one might notice that the variable vehicle value does not appear as the computation for estimating this value is not convergent. The values of the coefficient of these variables are given in Table 4. In both models, there are three variables that are statistically significant which are VehAge: oldest cars, Gender and DrivAge: young & youngest. These variables positively influence the ClaimIndAmount variable while the other variables do not seem to statistically effect the ClaimIndAmount variable.

Table 4 Estimated values and p-value of GLM coefficients (GLM Gamma and GLM Inverse Gaussian for ClaimIndAmount)

Coefficient	Gamma Model Estimate	p-value	Inverse Gaussian Model Estimate	p-value
Intercept	6.718062	< 2E-16	831.23	0.28365
VehValue	0.025821	0.517770	-	-
VehAge: oldest	0.210023**	0.013621	405.93**	0.01138
VehAge: young	0.005413	0.947901	61.79	0.62824
VehAge: youngest	-0.032306	0.746818	-91.52	0.50650
VehBody: Convertible	0.726317	0.559909	1550.18	0.67158
VehBody: coupe	0.674528	0.346729	881.53	0.35571
VehBody: hardtop	0.549884	0.432586	708.44	0.40076
VehBody: hatchback	0.570082	0.402631	714.80	0.35624
VehBody: minibus	0.704787	0.342865	1128.04	0.33227
VehBody: caravan	-0.632312	0.460149	-490.22	0.53109
VehBody: panel van	0.385558	0.594048	470.37	0.59592
VehBody: roadster	-1.708179	0.374885	-714.89	0.36876
VehBody sedan:	0.454461	0.504274	528.74	0.49304
VehBody: wagon	0.453020	0.506736	484.51	0.53012
VehBody: truck	0.836438	0.234521	1564.64	0.12662
VehBody: utility	0.574651	0.404430	835.17	0.30557
Gender Male	0.149030**	0.016232	175.24*	0.09524
DrivAge: older	0.109599	0.275668	205.91	0.16619
DrivAge: oldest	0.169836	0.200814	287.65	0.20351
DrivAge: working	0.065129	0.518089	61.04	0.65212
DrivAge: young	0.234654**	0.025121	495.00***	0.00657
DrivAge: youngest	0.412275***	0.000705	849.57***	0.00238

Next are the mathematical models for GLM Gamma and GLM Inverse Gaussian for ClaimAmount:

$$\text{ClaimAmount} \sim \text{Gamma}(\text{exposure}_i \times \mu_i),$$
$$\mu_i = \exp(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i}),$$

and

$$\text{ClaimAmount} \sim \text{Inverse Gaussian}(\text{exposure}_i \times \mu_i),$$
$$\mu_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i},$$

where X1, X2, …, X5 represent the same variable as in the previous model. The values of the coefficient of these variables are given in Table 5. Similar to previous analysis, there are three variables that are statistically significant which are VehAge: oldest cars, Gender and DrivAge: young & youngest, which positively influence the ClaimAmount variable.

Finally, we are able to combine the frequency models and the severity models using Eq. (1) to predict the aggregate loss for each claim. There are in total six combinations of aggregate model we can obtain which will be analysed in the section D.

Table 5 Estimated values and p-value of GLM coefficients (GLM Gamma and GLM Inverse Gaussian for ClaimAmount)

Coefficient	Gamma Model Estimate	p-value	Inverse Gaussian Model Estimate	p-value
Intercept	6.855627	< 2E-16	1072.15	0.3077
VehValue	0.022849	0.557256	21.03	0.75535
VehAge: oldest	0.212712**	0.010399	458.75***	0.00855
VehAge: young	0.004958	0.951056	35.59	0.79395
VehAge: youngest	-0.050859	0.602243	-158.05	0.32393
VehBody: Convertible	0.603086	0.619609	1108.8	0.74672
VehBody: coupe	0.653401	0.349938	795.88	0.51107
VehBody: hardtop	0.449294	0.510818	456.18	0.67659
VehBody: hatchback	0.466622	0.482362	503.77	0.62964
VehBody: minibus	0.580010	0.423423	865.21	0.52187
VehBody: caravan	0.710821	0.394516	-804.34	0.44541

Coefficient	Gamma Model Estimate	p-value	Inverse Gaussian Model Estimate	p-value
VehBody: panel van	0.261873	0.710466	230.37	0.83817
VehBody: roadster	1.799608	0.337725	-1007.38	0.35381
VehBody sedan:	0.370545	0.576603	329.77	0.75145
VehBody: wagon	0.372265	0.575830	277.65	0.79021
VehBody: truck	0.743578	0.278470	1420.84	0.25608
VehBody: utility	0.515608	0.442997	720.88	0.50424
Gender Male	0.141212**	0.019509	153.48	0.14680
DrivAge: older	0.114276	0.243787	236.85	0.11590
DrivAge: oldest	0.185875	0.151092	346.05	0.13862
DrivAge: working	0.079074	0.421016	77.13	0.57047
DrivAge: young	0.251975**	0.013665	536.91***	0.00374
DrivAge: youngest	0.425392***	0.000338	921.49***	0.00127

C. Frequency and Severity models using GLMM

In this section, we will now use GLMM to construct both frequency and severity models. The distribution assumptions for each response variable are the same as the distribution assumptions in the GLM section. In the first part of this section, we are going to build the frequency loss models using variables ClaimNb and ClaimOcc as the response variables. The variable VehBody will be designated as the random effect in our GLMM, which comprises 13 distinct vehicle types. The following is the mathematical model for ClaimNb using GLMM Poisson:

$$\mathrm{ClaimNb}_{i,j} | u_{i,0} \sim \mathrm{Poisson}\ (\mathrm{exposure}_{i,j} \times \lambda_{i,j} | u_{i,0}),$$

$$\boldsymbol{\lambda_{i,j} | u_{i,0} = \exp(\beta_0 + u_{i,0} + \beta_1 X1_{i,j} + \beta_2 X2 + \beta_4 X4_{i,j} + \beta_5 X5_{i,j})}$$

where X1, X2, X4, X5 are the same variable as previously, $u_{i,0}$ represents the random effect within VehBody category and $u_{i,0} \sim \mathrm{Normal}(0,\sigma_u^2)$. Notice that we have two subscripts; i represents the VehBody category and j represents the observation. The coefficient estimations for this model are given in Table 6. One can see that we have two separate estimations: one for fixed effect coefficients and one for random effect coefficients. We also provide prediction intervals for random intercepts. In both GLM and GLMM, variables VehAge and DrivAge give statistically significant influence on the number of claims. However, for VehBody both models give a different result as there are only Hatchback, Coupe and Utility in GLMM that produce statistically significant coefficients while there are ten VehBody classes producing statistically significant coefficients in GLM. Next, we are going to build GLMM Negative Binomial for ClaimNb:

$$\mathrm{ClaimNb}_{i,j} | u_{i,0} \sim \mathrm{Negative\ Bin}\ (\mathrm{exposure}_{i,j} \times \mu_{i,j} | u_{i,0}),$$

$$\boldsymbol{\mu_{i,j} | u_{i,0} = \exp(\beta_0 + u_{i,0} + \beta_4 X4_{i,j})}$$

where X4 represents gender, $u_{i,0}$ represents the random effect within VehBody category and $u_{i,0} \sim \mathrm{Normal}(0,\sigma_u^2)$. In this model we only have variable gender as the computation of the coefficients is not convergent when other independent variables are used. The coefficient estimations for this model are given in Table 7. One can see that in this model, there is no significant variable which is different from the result obtained in the GLM Negative Binomial previously.

Table 6 Fixed effect and random effect coefficient estimations in the GLMM Poisson

Fixed Effect Estimations		Random Effect Estimations			
Coefficient	GLMM Poisson Model Estimate	VehBody Class	Estimation	Lower	Upper
Intercept	-2.038411	Bus	0.06330	-0.17539	0.302003
VehValue	0.011976	Convertible	-0.03062	-0.26764	0.206404
VehAge: oldest	-0.081770*	Coupe	0.22708*	0.04354	0.410623
VehAge: young	0.076673*	Hardtop	0.04970	-0.10050	0.199899
VehAge: youngest	0.027530	Hatchback	-0.08898*	-0.14713	-0.03083
Gender Male	-0.027850	Minibus	-0.03255	-0.22709	0.161989
DrivAge: older	0.219236***	Caravan	0.07979	-0.15220	0.311788
DrivAge: oldest	0.002299	Panel van	-0.00610	-0.18821	0.176022
DrivAge: working	0.251408***	Roadster	-0.00804	-0.25124	0.235163
DrivAge: young	0.281729***	Sedan	-0.03635	-0.08980	0.017096
DrivAge: youngest	0.472538***	Wagon	0.00259	-0.05721	0.062391
-	-	Truck	-0.05532	-0.20611	0.095473
-	-	Utility	-0.15067*	-0.26346	-0.03789

DOI: 10.12962/j27213862.v7i2.20447

Table 7 Fixed effect and random effect coefficient estimations in the GLMM Negative Binomial

Fixed Effect Estimations		Random Effect Estimations			
Coefficient	GLMM NB Model Estimate	VehBody Class	Estimation	Lower	Upper
Intercept	-1.81084	Bus	3.40E-04	-0.02400	0.024684
Gender Male	-0.04075	Convertible	-1.18E-04	-0.02446	0.024225
-	-	Coupe	2.42E-03	-0.02188	0.026729
-	-	Hardtop	1.02E-03	-0.02323	0.025273
-	-	Hatchback	-2.38E-03	-0.02567	0.020912
-	-	Minibus	-3.69E-04	-0.02468	0.023938
-	-	Caravan	5.23E-04	-0.02382	0.024862
-	-	Panel van	6.62E-05	-0.02423	0.024365
-	-	Roadster	-5.55E-05	-0.02440	0.02429
-	-	Sedan	-2.39E-03	-0.02551	0.020730
-	-	Wagon	4.34E-03	-0.01909	0.027776
-	-	Truck	-4.98E-04	-0.02474	0.023745
-	-	Utility	-2.90E-03	-0.02699	0.021184

For the last model for frequency loss modeling, we will look at variable ClaimOcc that will be modeled using GLMM Bernoulli. The following is the mathematical model of ClaimOcc using GLMM Bernoulli:

$$\text{ClaimNb}_{i,j}|u_{i,0} \sim \text{Bernoulli}\ (\text{exposure}_{i,j} \times p_{i,j}|u_{i,0}),$$

$$p_{i,j}|u_{i,0} = \frac{1}{1 + \exp[-(\beta_0 + u_{i,0} + \beta_1 X1_{i,j} + \beta_2 X2_{i,j} + \beta_4 X4_{i,j} + \beta_5 X5_{i,j})]}$$

where $p_{i,j}|u_{i,0}$ represents the probability there will be a claim in the following period given a polis coming from a certain vehicle's body. Other variables represent the same meaning as previous models. The values of parameters can be seen in Table 8. As expected, in this model, significant variables are the same as the ones obtained in the GLM Bernoulli previously. However, for GLMM Bernoulli produces statistically significant on VehBody types: Hatchback, Coupe and Utility. This fact also occurs in GLMM Poisson.

Table 8 Fixed effect and random effect coefficient estimations in the GLMM Bernoulli

Fixed Effect Estimations		Random Effect Estimations			
Coefficient	GLMM NB Model Estimate	VehBody Class	Estimation	Lower	Upper
Intercept	-2.00317	Bus	0.05661	-0.19236	0.305577
VehValue	0.01360	Convertible	-0.02832	-0.27557	0.218932
VehAge: oldest	-0.09899**	Coupe	0.21823*	0.02218	0.414282
VehAge: young	0.07642	Hardtop	0.07098	-0.08977	0.231726
VehAge: youngest	0.03399	Hatchback*	-0.07941	-0.14190	-0.016930
Gender Male	-0.02663	Minibus	-0.01911	-0.22391	0.185691
DrivAge: older	0.21603***	Caravan	0.07918	-0.16337	0.321733
DrivAge: oldest	-0.02400	Panel van	0.01252	-0.18102	0.206057
DrivAge: working	0.25569***	Roadster	-0.00750	-0.26053	0.245530
DrivAge: young	0.28180***	Sedan	-0.05500	-0.11322	0.003221
DrivAge: youngest	0.50751***	Wagon	0.00640	-0.05846	0.071254
-	-	Truck	-0.07425	-0.23608	0.087588
-	-	Utility	-0.16828*	-0.28921	-0.047350

After frequency loss random variables, we are now ready to model severity random variables using GLMM. As before, we are going to assume that this response variable is either Gamma distribution or Inverse Gaussian distribution. The variable VehBody will still be employed as the random effect in our GLMM assumption. As before, we are going to build four GLMMs which are (i) GLMM Gamma for ClaimIndAmount (ii) GLMM Inverse Gaussian for ClaimIndAmount, (iii) GLMM Gamma for ClaimAmount and (iv) GLMM Inverse Gaussian for ClaimAmount. The following is the first GLMM which models ClaimIndAmount response variable using Gamma distribution:

$$\text{ClaimIndAmount}_{i,j}|u_{i,0} \sim \text{Gamma}\ (\text{exposure}_{i,j} \times \mu_{i,j}|u_{i,0}),$$

$$\mu_{i,j}|u_{i,0} = \exp(\beta_0 + u_{i,0} + \beta_4 X4_{i,j} + \beta_5 X5_{i,j})$$

where the values of parameters can be seen in Table 9. Note that there are only variable gender and DrivAge as independent variables as the computation of coefficients is not convergent when other independent variables are used. As previously observed, variable DrivAge gives statistically significant estimations and for random effect intercept, it is only VehBody: Sedan that gives a statistically significant influence. This observation is not obtained in the GLM.

We are still using the same variable, which is ClaimIndAmount modeled using GLMM Inverse Gaussian:

$$\text{ClaimIndAmount}_{i,j}|u_{i,0} \sim \text{Inverse Gaussian }(\text{exposure}_{i,j} \times \mu_{i,j}|u_{i,0}),$$

$$\boldsymbol{\mu_{i,j}|u_{i,0} = \beta_0 + u_{i,0} + \beta_2 X2_{i,j} + \beta_4 X4_{i,j} + \beta_5 X5_{i,j}}$$

where the values of parameters can be seen in Table 10. In this scenario, variablesVehAge, Gender and DrivAge are all significant. For random effect intercept, we do not find statistically significant VehBody variables. Next are the final two GLMMs which are GLMM Gamma for ClaimAmount and GLMM Inverse Gaussian for ClaimAmount. Consider the following model:

$$\text{ClaimAmount}_{i,j}|u_{i,0} \sim \text{Gamma }(\text{exposure}_{i,j} \times \mu_{i,j}|u_{i,0}),$$

$$\boldsymbol{\mu_{i,j}|u_{i,0} = \exp(\beta_0 + u_{i,0} + \beta_4 X4_{i,j} + \beta_5 X5_{i,j})}$$

where the values of parameters β_j can be seen in Table 11. Similar to the model of ClaimIndAmount, only variable Gender and variable DrivAge are used. As previously observed, variable DrivAge gives statistically significant estimations and for random effect intercept, VehBody: Sedan and Truck give a statistically significant influence to the amount of total claim. The final model of GLMM is the following:

$$\text{ClaimAmount}_{i,j}|u_{i,0} \sim \text{Inverse Gaussian }(\text{exposure}_{i,j} \times \mu_{i,j}|u_{i,0}),$$

$$\boldsymbol{\mu_{i,j}|u_{i,0} = \beta_0 + u_{i,0} + \beta_1 X1_{i,j} + \beta_2 X2_{i,j} + \beta_4 X4_{i,j} + \beta_5 X5_{i,j}}$$

where all independent variables are used in this equation. Notice that almost all variables are significant here, which was not obtained in the GLM. In Table 12, one can see the value of the coefficients and also the value of the random effect intercepts. In our final step, we shall combine the frequency models and the severity models using Eq. (1) to predict the aggregate loss for each claim. We will also mix GLM for frequency and GLMM for severity (and the other way around) to find the best possible combination.

Table 9 Fixed effect and random effect coefficient estimations in the GLMM Gamma for ClaimIndAmount

Fixed Effect Estimations		Random Effect Estimations			
Coefficient	**GLMM NB Model Estimate**	**VehBody Class**	**Estimation**	**Lower**	**Upper**
Intercept	8.53781	Bus	-0.50831	-5.78712	4.770513
Gender Male	0.04338	Convertible	-0.08698	-7.31079	7.136832
DrivAge: older	0.11089**	Coupe	0.192755	-1.79423	2.179746
DrivAge: oldest	0.21496***	Hardtop	-0.40618	-1.90424	1.091875
DrivAge: working	0.14607	Hatchback	0.277009	-0.20806	0.762076
DrivAge: young	-0.06748	Minibus	-0.25727	-2.81427	2.29972
DrivAge: youngest	1.19158***	Caravan	-1.35027	-5.54655	2.846014
-	-	Panel van	-0.50836	-2.66172	1.645004
-	-	Roadster	-0.57690	10.1986	9.044853
-	-	Sedan*	1.024874	0.57476	1.474991
-	-	Wagon	0.061528	-0.4401	0.563152
-	-	Truck	1.766746	0.17650	3.3570
-	-	Utility	0.163042	-0.89274	1.21883

Table 10 Fixed effect and random effect coefficient estimations in the GLMM Inverse Gaussian for ClaimIndAmount

Fixed Effect Estimations		Random Effect Estimations			
Coefficient	**GLMM NB Model Estimate**	**VehBody Class**	**Estimation**	**Lower**	**Upper**
Intercept	1525.78	Bus	-5.3364	-261.809	251.1357
VehAge: oldest	410.96***	Convertible	5.6984	-251.124	262.5207
VehAge: young	33.98	Coupe	12.9538	-235.875	261.7823
VehAge: youngest	-68.27*	Hardtop	2.56240	-237.978	243.1028
Gender Male	207.57***	Hatchback	17.7452	-144.312	179.8023
DrivAge: older	143.74***	Minibus	15.5536	-236.765	267.8721
DrivAge: oldest	261.77***	Caravan	-23.5854	-278.430	231.2594
DrivAge: working	87.81***	Panel van	-15.7239	-265.229	233.7810
DrivAge: young	482.50***	Roadster	-3.20070	-260.304	253.9025

Fixed Effect Estimations		Random Effect Estimations			
Coefficient	GLMM NB Model Estimate	VehBody Class	Estimation	Lower	Upper
DrivAge: youngest	860,89***	Sedan	-107.3980	-249.069	34.27233
-	-	Wagon	-139.0340	-293.423	15.35461
-	-	Truck	81.8874	-163.360	327.1350
-	-	Utility	21.8250	-206.700	250.3495

Table 11 Fixed effect and random effect coefficient estimations in the GLMM Gamma for ClaimAmount

Fixed Effect Estimations		Random Effect Estimations			
Coefficient	GLMM NB Model Estimate	VehBody Class	Estimation	Lower	Upper
Intercept	8.56653	Bus	-0.48320	-5.69532	4.728927
Gender Male	0.04013	Convertible	-0.10399	-7.22703	7.019049
DrivAge: older	0.10949*	Coupe	0.27336	-1.69111	2.237824
DrivAge: oldest	0.21191***	Hardtop	-0.42324	-1.90445	1.057977
DrivAge: working	0.14972	Hatchback	0.26404	-0.21563	0.743707
DrivAge: young	-0.05976	Minibus	-0.27457	-2.80222	2.253080
DrivAge: youngest	1.18466***	Caravan	-1.33241	-5.47800	2.813193
-	-	Panel van	-0.52902	-2.65789	1.599854
-	-	Roadster	-0.56910	-10.0357	8.897519
-	-	Sedan*	1.00578	0.56067	1.450883
-	-	Wagon	0.05571	-0.44033	0.551750
-	-	Truck	1.74183*	0.16948	3.314171
-	-	Utility	0.16824	-0.87573	1.212223

Table 12 Fixed effect and random effect coefficient estimations in the GLMM Inverse Gaussian for ClaimAmount

Fixed Effect Estimations		Random Effect Estimations			
Coefficient	GLMM NB Model Estimate	VehBody Class	Estimation	Lower	Upper
Intercept	1592.162	Bus	-3.82812	-257.689	250.0328
VehValue	4.926	Convertible	4.469663	-249.721	258.6604
VehAge: oldest	449.115***	Coupe	19.09163	-227.798	265.9815
VehAge: young	17.618	Hardtop	-3.58631	-242.426	235.2535
VehAge: youngest	-115.381***	Hatchback	3.06473	-158.742	164.8711
Gender Male	196.253***	Minibus	12.08509	-237.997	262.1673
DrivAge: older	163.425***	Caravan	-23.3178	-275.659	229.0239
DrivAge: oldest	306.685***	Panel van	-18.1483	-265.473	229.1758
DrivAge: working	109.733***	Roadster	-3.30986	-257.719	251.099
DrivAge: young	536.887***	Sedan	-105.94	-249.528	37.64719
DrivAge: youngest	- 931.585***	Wagon	-133.779	-290.623	23.06512
-	-	Truck	82.68164	-160.559	325.9227
-	-	Utility	36.00448	-191.84	263.8493

D. Results

In this section, we are going to predict the total claim of each policy using each model derived from previous sections. The frequency of claim has been modeled by either GLM and GLMM and so is the severity of claim. Thus, we shall have four combinations i.e. (i) GLM for both frequency and severity of claim, (ii) GLM for frequency of claim and GLMM for severity of claim, (iii) GLMM for frequency of claim and GLM for severity of claim and (iv) GLMM for both frequency and severity of claim. Also, recall that GLM/GLMM Bernoulli will be paired with ClaimAmount while other frequency random variables will be paired with ClaimIndAmount.

In Table 13, one can see in the first group that the combination between GLM Bernoulli and GLM Gamma gives the smallest RMSE and MAE with RMSE = 1021,55 and MAE = 250,824. The smallest RMSE and smallest MAE in the second group are obtained by the combination of GLM Bernoulli and GLMM Inverse Gaussian. The RMSE value obtained is 1021,64 while the MAE obtained is 251,7. In the next two groups, the combination between GLMM Poisson and GLM Gamma provides the smallest RMSE and MAE with RMSE = 1021,8 and MAE = 252,618 while the combination between GLMM Poisson and GLMM Inverse Gaussian finds the the smallest RMSE and MAE with RMSE = 1021,64 and MAE = 251,704.

Of all models, the combination between GLM Bernoulli and GLM Gamma produces the smallest RMSE and MAE. We can conclude that generalization of GLM to GLMM does not improve our predictions as the smallest error was obtained using GLM methodology on both frequency and severity models. However, we have interesting observations related to the distribution of observed variables. When using GLM on frequency models, it seems that Binomial Distribution is more favorable other than Poisson or Negative Binomial distributions. As for the severity, Gamma distribution is more helpful than Inverse Gaussian. However, if GLMM is used, Poisson Distribution for frequency random variable and Inverse Gaussian for severity random variable are more accurate.

In the previous analysis, the accuracy of model is measured using the difference between the actual value of the total claim and the predicted value of the total claim. Each difference is calculated for each policy and will be averaged. This measurement is not wrong but is rather imprecise, because it is very difficult to predict the total claim in the individual level. Next, we are going to measure the accuracy of the model using the aggregate total of claim of all policies. In Table 14, we have chosen one combination of models according to the smallest RMSE dan MAE of each group. We then shall calculate the total claim of all policies along with its averages, which will be compared with the actual values. Moreover, we shall also illustrate the distribution of predicted values of the chosen models in each group using boxplot in Figure 3.

From Table 14, it shows that the combination between GLMM Poisson and GLM Gamma gives the closest measurement with the actual value. It indicates that the model is able to predict the total reserve of the insurance company. This amount of money must be available in the following year such that all claims can be paid to the policy holder. The average of all claim amount is also provided and can be used to calculate premium charged to policy holders in the next period. However, as shown in Figure 3, even though the sum and the average are very close to the actual value, the distribution of actual claim amount is very different from each distribution of the predicted claim amount obtained from each model. The actual data is left skewed as there are so many zero claims. The same conclusion also applies here, in which that GLM and GLMM produce generally the same measurement. Thus, generalization of GLM to GLMM does not improve our predictions.

Table 13 Accuracy metrics of all combinations of GLM/GLMM

	Frequency Model	Severity Model	RMSE	MAE
(i)	GLM Poisson	GLM Gamma (Ind)	1,021.77	252.458
	GLM Poisson	GLM Inverse Gaussian (Ind)	1,021.97	252.916
	GLM Negative Binomial	GLM Gamma (Ind)	1,021.73	258.338
	GLM Negative Binomial	GLM Inverse Gaussian (Ind)	1,021.81	258.740
	GLM Bernoulli	GLM Gamma	1,021.55	250.824
	GLM Bernoulli	GLM Inverse Gaussian	1,021.68	251.247
(ii)	GLM Poisson	GLMM Gamma (Ind)	1,723.60	947.049
	GLM Poisson	GLMM Inverse Gaussian (Ind)	1,544.50	958.627
	GLM Negative Binomial	GLMM Gamma (Ind)	1,624.28	899.205
	GLM Negative Binomial	GLMM Inverse Gaussian (Ind)	1,021.90	253.339
	GLM Bernoulli	GLMM Gamma	1,021.74	259.345
	GLM Bernoulli	GLMM Inverse Gaussian	1,021.64	251.704
(iii)	GLMM Poisson	GLM Gamma (Ind)	1,021.80	252.618
	GLMM Poisson	GLM Inverse Gaussian (Ind)	1,022.59	255.399
	GLMM Negative Binomial	GLM Gamma (Ind)	1,032.62	380.324
	GLMM Negative Binomial	GLM Inverse Gaussian (Ind)	1,022.03	253.135
	GLMM Bernoulli	GLM Gamma	1,022.86	256.123
	GLMM Bernoulli	GLM Inverse Gaussian	1,032.91	381.263
(iv)	GLMM Poisson	GLMM Gamma (Ind)	1,736.10	950.504
	GLMM Poisson	GLMM Inverse Gaussian (Ind)	1,021.94	253.459
	GLMM Negative Binomial	GLMM Gamma (Ind)	1,618.33	944.081
	GLMM Negative Binomial	GLMM Inverse Gaussian (Ind)	1,022.73	256.410
	GLMM Bernoulli	GLMM Gamma	2,542.14	1754.60
	GLMM Bernoulli	GLMM Inverse Gaussian	1,032.61	382.038

Table 14 Comparison between aggregate claim in the actual data and aggregate claim resulted from predictive models

		Sum of total claim	Mean of total claim
-	Actual Data	1.874.898	138.154
(i)	GLM Bernoulli – GLM Gamma	1.845.701	136.003
(ii)	GLM Bernoulli – GLMM Inverse Gaussian	1.861.801	137.190
(iii)	GLMM Poisson – GLM Gamma	1.876.832	138.297
(iv)	GLMM Poisson – GLMM Inverse Gaussian	1.891.640	139.388

IV. CONCLUSIONS AND SUGGESTIONS

In this paper, we discuss statistical models to predict claim counts and claim amounts. We use vehicle value, vehicle age, type of vehicle, gender of the driver, and driver's age as covariates for our GLMs. Specifically, we employ Poisson and Negative Binomial distributions for claim counts and the Bernoulli distribution for claim occurrence. For claim amounts, we use the Gamma and Inverse Gaussian distributions. A significant contribution of this paper is the inclusion of random effects in the model for both claim counts and claim amounts, allowing for underlying dependency patterns. Using the same datasets, we compare the model estimation and results of the GLMs with those of GLMMs.

For claim count prediction using GLMs, the Bernoulli distribution proves to be more favorable than either the Poisson or Negative Binomial distributions. Regarding claim amounts, the Gamma distribution is more effective than the Inverse Gaussian distribution. However, when using GLMMs, the Poisson distribution for claim counts and the Inverse Gaussian distribution for claim amounts yield more accurate results. Although generalizing GLMs to GLMMs does not significantly improve our predictions, we hope our modeling framework for claim counts and claim amounts will become a valuable tool for determining pure premiums for future claims.

For future research, we suggest implementing zero-inflated models, as more than 90% of the data consists of zero claims, and utilizing Generalized Additive Models (GAMs) to capture potential non-linear relationships and interactions among covariates.

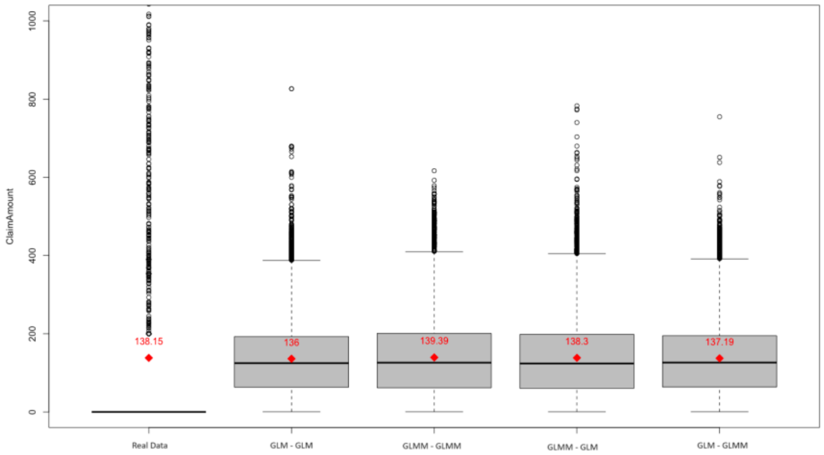


Figure 3 Boxplot of actual claim and predicted claim using predictive model from Table 14

REFERENCES

[1] G. K..Smyth, and B. Jørgensen, “Fitting Tweedie's compound Poisson model to insurance claims data: dispersion modelling,” ASTIN Bulletin: The Journal of the IAA, vol. 32, pp. 143-157, 2002.

[2] P. De Jong, and G. Z. Heller, Generalized linear models for insurance data, Cambridge University Press, 2008.

[3] S. Kafková and L. Křivánková, “Generalized linear models in vehicle insurance,” Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis, vol. 62, pp. 383-388, 2014.

[4] E. W. Frees and G. Lee, “Rating endorsements using generalized linear models,” Variance, vol. 10, pp. 51-74, 2015.

[5] M. David, “Auto insurance premium calculation using generalized linear models,” Procedia Economics and Finance, vol. 20, pp. 147-156, 2015.

[6] E. Šoltés, S. Zelinová and M. Bilíková, “General linear model: an effective tool for analysis of claim severity in motor third party liability insurance,” STATISTICS, vol. 13, pp. 13-31, 2019.

[7] T. A. J. Putra, D. C. Lesmana and I. G. P. Purnaba, “Penghitungan Premi Asuransi Kendaraan Bermotor Menggunakan Generalized Linear Models dengan Distribusi Tweedie,” Jambura Journal of Mathematics, vol. 3, pp. 115-127, 2021.

[8] T. Rahmawati, D. Susanti and R. Riaman, “Determining Pure Premium of Motor Vehicle Insurance with Generalized Linear Models (GLM),” International Journal of Quantitative Research and Modeling, vol. 4, pp. 207-214, 2023.

[9] E. W. Frees, G. Lee and L. Yang, “Multivariate frequency-severity regression models in insurance,” Risks, vol. 4, pp. 4, 2016.

[10] C. Dutang and A. Charpentier, “Insurance Datasets,” R package version 1.0-11, 2021

[11] C. G. Giancaterino, "GLM, GNM and GAM Approach on MTPL Pricing," Journal of Mathematics and Statistical Science, vol. 2, pp 427-481, August 2016

[12] R. Oktavia, R. Zuhra, H. Hafnani, N. Nurmaulidar and I. Syahrini, "Application of Poisson and negative binomials models to estimate the frequency of insurance claims," Jurnal Natural, vol. 23, pp. 21-27, 2023.

[13] G. Pernagallo, A. Punzo and B. Torrisi, “Women and insurance pricing policies: a gender-based analysis with GAMLSS on two actuarial datasets,” Scientific Reports, vol. 14, pp. 3239, 2024.

[14] G. Gao and J. Li, “Dependence modeling of frequency-severity of insurance claims using waiting time,” Insurance: Mathematics and Economics, vol. 109, pp. 29-51, 2023.

[15] E. W. Frees and F. Huang, “Online Supplement to: The Discriminating (Pricing) Actuary,” Available at SSRN: https://ssrn.com/abstract=3892473, July 23, 2021.

[16] K. Yau, K. Yip and H. K. Yuen, “Modelling repeated insurance claim frequency data using the generalized linear mixed model,” Journal of Applied Statistics, vol. 30, pp. 857-865, 2003

[17] K. Antonio and J. Beirlant, “Actuarial statistics with generalized linear mixed models,” Insurance: Mathematics and Economics, vol. 40(1), pp. 58-76, 2007.

[18] K. Antonio, E. W. Frees and E. A. Valdez, “A multilevel analysis of intercompany claim counts,” ASTIN Bulletin: The Journal of the IAA, vol. 40, pp. 151-177, 2010.

[19] Y. Kim, Y. K. Choi and S. Emery, “Logistic regression with multiple random effects: a simulation study of estimation methods and statistical packages,” The American Statistician, vol. 67, pp. 171-182, 2013.

[20] S. A. Rohmaniah and N. E. Chandra, “Perhitungan Premi Asuransi Jiwa Menggunakan Generalized Linear Mixed Models,” Jurnal Ilmiah Teknosains, vol. 4, pp. 80-84, 2018

[21] N. Wang, L. Qian, N. Zhang, and Z. Liu, “Modelling the aggregate loss for insurance claims with dependence,” Communications in Statistics-Theory and Methods, vol. 50, pp. 2080-2095, 2021.

[22] C. C. Günther, I. F. Tvete, K. Aas, J. A. Hagen, L. Kvifte and Ø. Borgan, “Predicting Future Claims Among High Risk Policyholders Using Random Effects,” In D. Silvestrov and A. Martin-Löf (Eds.), Modern Problems in Insurance Mathematics, pp. 171-185, 2014.

[23] W. Lee, J. Kim, and J. Y. Ahn, “The Poisson random effect model for experience ratemaking: Limitations and alternative solutions,” Insurance: Mathematics and Economics, vol. 91, pp. 26-36, 2020.

[24] F. A. Farisa, S. N. H. Salby, A. A. Rahman, and P. Purhadi, “Modeling the Number of Pneumonia in Toddlers in East Java Province in 2021 with Generalized Poisson Regression,” Inferensi, vol. 62, pp. 91-96, 2023.

[25] S. Ully, Analisis Dana Pinjaman Pegawai Negeri Sipil Menggunakan Metode Generalized Linear Mixed Model (GLMM) pada Data Longitudinal, Doctoral dissertation, Universitas Andalas, 2021.

[26] P. M. Caçola, and M. D. Pant, “Using a Generalized Linear Mixed Model Approach to Explore the Role of Age, Motor Proficiency, and Cognitive Styles in Children's Reach Estimation Accuracy,” Perceptual and Motor Skills, vol. 119, pp. 530-549, 2014.

[27] A. M. Gad and R. B. El Kholy, “Generalized linear mixed models for longitudinal data,” International Journal of Probability and Statistics, vol. 1, pp. 41-47, 2012.

DOI: 10.12962/j27213862.v7i2.20447