Science World Journal Vol. 16(No 3) 2021
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

Full Length Research Article

MEAN PARAMETER MODELING FOR AN AUTOMOBILE INSURANCE PORTFOLIO USING GENERALIZED ADDITIVE MODELS FOR LOCATION, SCALE AND SHAPE (GAMLSS)

*Chaku Shammah Emmanuel, Adenomon Monday Osagie, Nweze Nwaze Obini

Department of Statistics, Nasarawa State University, Keffi

*Corresponding Author Email Address: chakushammah@nsuk.edu.ng

ABSTRACT

In this study, Generalized Additive Models for location, scale and shape was deployed to model a typical automobile insurance portfolio. The data set used for this study comprises of seven variables, which are: Kilometres, Zone, Bonus, Make, Insured, Claims and Payments, it was compiled by a Swedish Committee on the Analysis of Risk Premium in Motor Insurance. The mean was modeled in terms of the explanatory variables although the GAMLSS has the capacity to model up to four parameters unlike the Generalized Lineal Models (GLMs) and Generalized Additive Models (GAMs). This allows for greater flexibility in modeling. In checking for over-dispersion, the negative binomial was used such that terms were dropped or added. Analysis revealed that all term were important and as such no terms could be dropped. When terms were added, analysis further showed that all the two way interaction terms are needed in the model except for the interaction between Kilometers and Zone. Results from the optimal model check gives the best model as those with separate smoothing terms for both Bonus and Kilometers.

Keywords: GAMLSS, Automobile Insurance, Negative Binomial, Over-dispersion, Parameter Modeling.

1. INTRODUCTION

Financial data sets like prices of assets and commodities in the stock market, claims and payments in the insurance sector, credit worthiness in the banks, etc recorded on different time scales have collective characteristics which help the analyst when carrying out statistical modeling of the financial data. They can be in both qualitative and mostly quantitative form which attaches numerical values or descriptions to the measure or strength of an observation. The science of building and analyzing mathematical models to describe the processes by which money flows into and out of an Insurance company is called Actuarial science. It is a combination of diverse quantitative skills to enable one “make financial sense of the future”. General insurance which includes health insurance, property or personal insurance such as motor and home insurance, not forgetting large commercial risks and liability insurance is now a fast growing area for insurance experts.

The automobile insurance industry in any country is very important, since it helps to remedy the economy, such that financial losses due to the factors of risks are reduced, either through sharing or pooling the risks to a large number of people. In most cases, insurance experts are more concerned with the amount to be paid by the insurance company than the circumstances that gave rise to the need for compensation (Denuit et al 2007). In so doing, they need to have an understanding of the different models for the risk consisting of the total or aggregate amount of claims payable by an insurance company for a fixed period of time. Insurance data sets are known to have relatively huge claim amounts, which may or may not be frequent, hence the knowledge of statistical distributions with relatively heavy tails and highly skewed like the exponential, gamma, Pareto, Weibull, Poisson and lognormal will come in handy (Boland, 2007). Looking at the economic importance of motor insurance in developing countries, attempts have been made by insurance researches to obtain probabilistic models for the distribution of the claim amounts reported by insured drivers. Finally, constructing interpretable models for claims severity can often give one a much added insight into the complexity of the large amounts of claims that may often be hidden in a huge amount of data (Raz & Shaw, 2000).

In pricing automobile insurance policies, two very important factors must be considered, these are, data on the consumers and level of expertise and understanding of the market. In this research, the focus will be on the pricing of insurance using data on the consumers.

Generalized Linear Models (GLM) is widely used in the insurance industry as seen in De Jong et al. (2008) and Ohlsson and Johansson (2010). David (2015) looked at the applications of GLMs for the calculation of the insurance premium. These models are based on several assumptions one of which is the independence of the observations. In so many statistical problems this assumption is violated, which can be caused by personal characteristics such as driving skills, reflexes, concentration or aggressiveness behind the wheels. In majority of studies, the Poisson model is noted to be commonly used to model Claim frequencies while the Gamma distribution model models claim sizes (See Chaku et al, 2017). Spedicato et al (2014) applied GAMLSS triangles on a loss database in order to assess the distribution of unpaid loss reserve in term of best estimate as distributional form. The results obtained were critically compared with those of classical stochastic reserving approach. It was seen that the result yield a lower RMSE than the chain ladder method used when predicting payment. One of the most recent research on motor insurance modeling is by Tzougas (2020) who presented the Poisson-Inverse Gamma regression model with varying dispersion for approximating heavy-tailed and over dispersed claim counts. With the development of an Expectation-Maximization (EM) type algorithm, the empirical analysis examines a portfolio of motor insurance data in order to investigate the efficiency of the proposed algorithm.

2.0 METHODOLOGY

2.1 Data Size and Description

The scope of this research is around the application of GAMLSS models on a Swedish automobile insurance portfolio for modeling claims and losses due to claims. The data covers all third party motor insurance claims in Sweden in 1977. The data size is 2182 and was compiled by a Swedish Committee on the Analysis of Risk Premium in Motor Insurance. The data set comprises of seven variables, which are: Kilometres, Zone, Bonus, Make, Insured, Claims and Payments. The raw data can be obtained electronically from the Statlib data base with web page, https://www.kaggle.com/floser/swedish-motor-insurance. Also, studies will be carried out on some simulated data sets and situations based on some certain algorithms.

The size of the data is 2182 with no missing values. In Sweden all motor insurance companies apply identical risk arguments to classify customers, and so their portfolios and their claim statistics can be combined. The data was compiled by a Swedish Committee on the Analysis of Risk Premium in Motor Insurance. The data parameter and descriptions are given below as

- **KILOMETRES:** Distance (Kilometres) driven by a vehicle, grouped into five categories
  1: < 1000
  2: 1000-15000
  3: 15000-20000
  4: 20000-25000
  5: > 25000

- **ZONE:** Geographical zone of the vehicle, grouped into seven categories

Geographical zone
1: Stockholm, Göteborg, Malmö with surroundings
2: Other large cities with surroundings
3: Smaller cities with surroundings in southern Sweden
4: Rural areas in southern Sweden
5: Smaller cities with surroundings in northern Sweden
6: Rural areas in northern Sweden
7: Gotland

- **BONUS:** No Claims bonus equal to the number of years, plus one, since the last claim- 7 categories
- **MAKE:** The type of vehicle-9 categories
- **INSURED:** The number of policy holders in years
- **CLAIMS:** Number of claims
- **PAYMENT:** Sum of payments

Claims and Payments will be the response variables while Kilometres, Zone, Bonus, Make and Insured are explanatory variables.

2.2 Generalized Additive Model for Location, Scale and Shape (GAMLSS)

The Generalized Additive Model for Location, Scale and Shape (GAMLSS) framework is one which encompasses the properties of the GLMs, GLMM and GAMs. It was introduced by Rigby and Stasinopoulos (2002) to address the limitations GLMs and GAMs. These are:

1. The previous models only allow the modelling of the location parameter (mean).
2. The distribution of the response variable must be from the EFD.

Here, the response distribution must not be an EFD instead, the primary goal are distributions with computable first and second order derivatives which allows flexibility in modeling. The GAMLSS assumes independent response observations y_i for $i = 1, 2, ..., n$ with pdf $f(y_i \mid \theta^i)$ conditional on

$$\theta^i = (\theta_{1i}, \theta_{2i}, \theta_{31}, \theta_{4i}) = (\mu_i, \sigma_i, \nu_i, \tau_i)$$

A vector of $k = 4$ distribution parameters, where each parameter can be a function of the predictor variables. If $y_i = (y_1, y_2, ..., y_n)$ is the length of the response variable, y_i, such that $k = 1, 2, 3, 4$ parameters. Assume that $g_k(.)$ is an unknown link function that is monotonic which relates the distribution parameters to the predictor variables by

$$g_k(\theta_k) = \eta_k = X_k \beta_k + \sum Z_{jk} \gamma_{jk}$$

3.0 ANALYSIS AND RESULTS

3.1 GAMLSS Modeling

In our earlier modeling of the number of claims using the negative binomial distribution, a log-link was used with a constant shape parameter. Generally, it is written as:

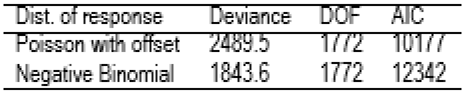
$$y \sim NB(\mu, \kappa), \quad where \log \mu = \log n + x^T \beta, \quad \log \kappa = z^T \gamma$$

Where z is the vector containing explanatory variables for κ and γ and is the corresponding coefficient vector.

In modeling using GAMLSS, four parameters are used to model in terms of the explanatory variables unlike the GLM that allows only one, i.e. the mean to be modeled in terms of the explanatory variables. Generally, GAMLSS is a framework for modeling univariate regression type of statistical problems. It allows more flexibility than GAMs and GLMs in identifying distributions of the response variable up to including highly skewed and/or kurtotic distributions. Here also, distribution parameters can be modeled flexibly like functions of explanatory variables.

Using a Poisson model, over-dispersion is checked comparing the negative binomial and the Poisson distribution model and we have the below table 3.1.

Table 3.1: GAMLSS Deviance Reduction



Dist. of response	Deviance	DOF	AIC
Poisson with offset	2489.5	1772	10177
Negative Binomial	1843.6	1772	12342

Source: Author's computation

The reduction in deviance is obvious and suggests a reduction in the dispersion as this was measured by dividing the deviance by the degree of freedom.

4.7.1 Modeling Mean (μ) Claims In Terms of the Explanatory Variables

The GAMLSS modeling was done in R and the below output was obtained

Table 3.2: GAMLSS Modeling

```
ModelGamlss2=gamlss(Claims~factor(Kilometres)+factor(Make)+factor(Zone)+cs(Bonus),

family=NBI, data=Insurance)

GAMLSS-RS iteration 1: Global Deviance = 12355.76

GAMLSS-RS iteration 2: Global Deviance = 12296.42

GAMLSS-RS iteration 3: Global Deviance = 12296.14

GAMLSS-RS iteration 4: Global Deviance = 12296.14

GAMLSS-RS iteration 5: Global Deviance = 12296.14
```

Source: Author's computation in R

This results in a deviance of 12296.14. The GAMLSS model was plotted and the below output and plots obtained

Table 3.3. GAMLSS Parameter Estimation

```
> plot(ModelGamlss2)

******************************************************************

       Summary of the Randomised Quantile Residuals

                    mean   =  0.02917698

                 variance   =  0.9821569

          coef. of skewness  =  0.3175908

          coef. of kurtosis  =  4.137316

Filliben correlation coefficient  =  0.9944492

******************************************************************
```

Source: Author's computation using R

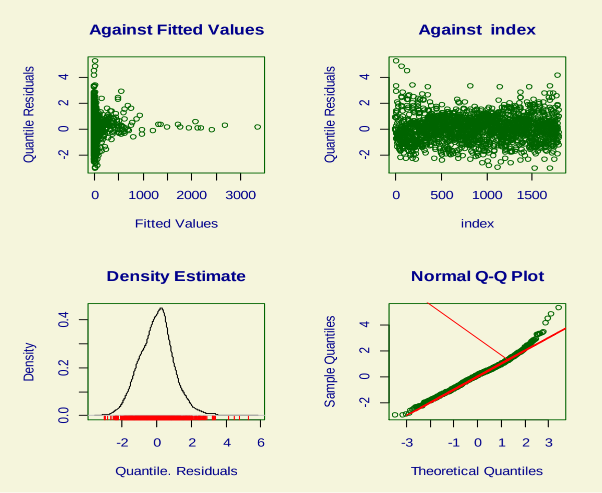


Figure 3.1: GAMLSS Quantile Plots
Source: Author's computation using R

Quantile residuals are hinged on the concept of taking the inverse of the estimated distribution of every observation hoping to get exact standard normal residuals. For discrete distributions (such as Poisson and Binomial), a bit of randomization is introduced to get continuous normal residuals. For situations of large dispersion in GLMs, quantile residuals are the residual of choice when the deviance and Pearson residuals can be non-normal. Quantile residuals are the only useful residuals for Poisson or Binomial data when the response takes only a small number of distinct values (Smyth and Dunn, 1996).
The above plots the residual plots from the negative binomial distribution model. The two top plots are the residuals plotted against the fitted values of mean claims and an index, while the bottom left plot gives a kernel density plot whereas the downright gives a normal Q-Q plot. The residuals are not random. The Q-Q plot shows a string of five outliers, four in the upper tail of the plot and one in the lower tail of the plot. It is seen that the negative binomial distribution model does not provide a reasonable fit to the data.

The *gamlss* package in R is equipped with functions that help with selecting explanatory variable terms. The functions *adterm ()* and *dropterm ()* are used in this regard for addition and removal of a term in models respectively.

An attempt to demonstrate the drop term on the model gives the following results

Table 3.4: Drop Term on Model GAMLSS 2

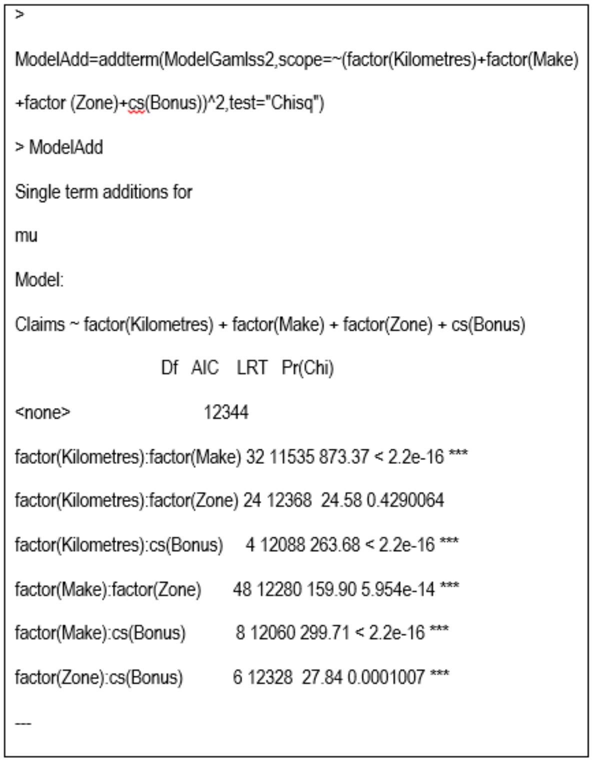
```
> ModelDrop=dropterm(ModelGamlss2,test="Chisq")

> ModelDrop

Single term deletions for

mu

Model:

Claims ~ factor(Kilometres) + factor(Make) + factor(Zone) + cs(Bonus)

                    Df   AIC    LRT   Pr(Chi)

<none>                 12344

factor(Kilometres) 4.0000 13313  976.4 < 2.2e-16 ***

factor(Make)      8.0000 15939 3610.8 < 2.2e-16 ***

factor(Zone)      6.0000 13701 1369.2 < 2.2e-16 ***

cs(Bonus)         3.9996 14076 1740.0 < 2.2e-16 ***

---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Source: Author's computation using R

The Chi square test shows that no terms can be dropped from the above model, which implies that all terms must be included in the model for the mean of claims.

On the other hand, the addterm () function was deployed but not without stating the terms we hope to add up. Here, we suggest an interactive term like: (factor(Kilometres)+factor(Make)+factor(Zone)+cs(Bonus))^2.

Table 3.5: Addterm on Model GAMLSS 2

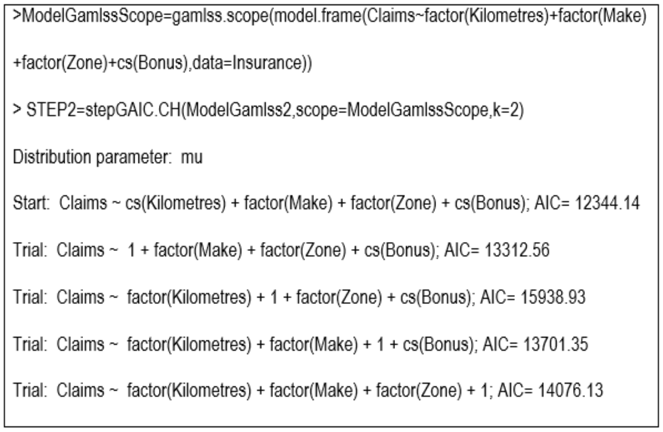


```
>
ModelAdd=addterm(ModelGamlss2,scope=~(factor(Kilometres)+factor(Make)
+factor (Zone)+cs(Bonus))^2,test="Chisq")
> ModelAdd
Single term additions for
mu
Model:
Claims ~ factor(Kilometres) + factor(Make) + factor(Zone) + cs(Bonus)
                          Df  AIC   LRT  Pr(Chi)
<none>                        12344
factor(Kilometres):factor(Make) 32 11535 873.37 < 2.2e-16 ***
factor(Kilometres):factor(Zone) 24 12368  24.58 0.4290064
factor(Kilometres):cs(Bonus)    4 12088 263.68 < 2.2e-16 ***
factor(Make):factor(Zone)      48 12280 159.90 5.954e-14 ***
factor(Make):cs(Bonus)          8 12060 299.71 < 2.2e-16 ***
factor(Zone):cs(Bonus)          6 12328  27.84 0.0001007 ***
---
```

Source: Author's computation using R

The above test reveals that all two way interactions are needed in the model except for the two interaction between Kilometres and Zone. Thus for modeling the mean claims, all two way interaction terms are going to be beneficial.
A further check on the model selection to get the optimal model is done below.

Table 3.6: Optimal Model Check

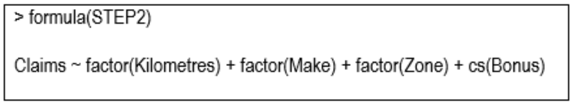


```
>ModelGamlssScope=gamlss.scope(model.frame(Claims~factor(Kilometres)+factor(Make)
+factor(Zone)+cs(Bonus),data=Insurance))
> STEP2=stepGAIC.CH(ModelGamlss2,scope=ModelGamlssScope,k=2)
Distribution parameter:  mu
Start:  Claims ~ cs(Kilometres) + factor(Make) + factor(Zone) + cs(Bonus); AIC= 12344.14
Trial:  Claims ~  1 + factor(Make) + factor(Zone) + cs(Bonus); AIC= 13312.56
Trial:  Claims ~  factor(Kilometres) + 1 + factor(Zone) + cs(Bonus); AIC= 15938.93
Trial:  Claims ~  factor(Kilometres) + factor(Make) + 1 + cs(Bonus); AIC= 13701.35
Trial:  Claims ~  factor(Kilometres) + factor(Make) + factor(Zone) + 1; AIC= 14076.13
```

Source: Author's computation using R

The first model has the least AIC and as such means that the best model is between the model with smoothing term for Bonus and the one without. The final model is obtained as

Table 3.7: Model with Smoothing Term



```
> formula(STEP2)
Claims ~ factor(Kilometres) + factor(Make) + factor(Zone) + cs(Bonus)
```

Source: Author's computation using R

Conclusion

The modeling of mean in terms of response variables reveals that all terms are critical in the model. When terms are added, analysis further showed that all the two way interaction terms are needed in the model except for the interaction between Kilometers and Zone. Results from the optimal model check gives the best model as those with separate smoothing terms for both Bonus and Kilometers.

REFERENCES

Boland, P. J. (2007). Statistical and probabilistic methods in actuarial science. CRS Press, Ireland: 35 – 47.

Chaku, S. E., Nwankwo, C. H., and Adehi, M. U. (2017) Modeling Claim Frequency and Loss Due to Claims of Automobile Insurance, Nigerian Statistical Association, 41stAnnual and 1st International Conference Proceedings, 2017, Vol. 2: 69-84

Chaku, S. E., Maijama'a, B., Isah, S. H., and Abubakar, M. A. (2019). Applying Generalised Additive Models (GAMs) On Insurance Data Using R. 6th International Conference on Mathematical Analysis and Optimization. Theory and Applications, Conference Proceedings: 95-107.

David, M. (2015) Auto insurance premium calculation using generalized linear models. Procedia Economics and Finance 20,147 – 156.

De Jong, P., & Heller, G. Z. (2008). Generalized linear models for insurance data. Cambridge University Press.

Denuit, M., Marechal, X., Pitrebois, S., and Walhin, J. F. (2007). Actuarial modeling of claim counts: Risk classification, credibility and bonus-malus systems. John Wiley & Sons.

Ohlsson, E.,& Johansson, B. (2010). Non-life insurance pricing with generalized linear models. EAA lecture notes. Heidelberg; New York: Springer.

Raz, O., & Shaw M. (2007). An approach to preserving sufficient correlations in open resource coalitions. 10th International work-shop on software specification and design (IWSSD-10) IEEE Computer Society.

Rigby, A., & Stasinopoulos, M. (2002). The implementation of generalized additive models for location, scale and shape. Statistical modelling in Society: Proceedings of the 7th International Workshop on Statistical Modeling, 75 – 83.

Spedicato, G. A., Clemente, G. P., & Shewe, F. (2014). The uuse of GAMLSS in assessing the distribution of unpaid claims reserves. Casualty Actuarial Society E-Forum. https://www.researchgate.net/publication/265215661.

Tzougas, G. (2020). EM estimation for the Poisson-Inverse Gamma regression model with varying dispersion: An application to insurance ratemaking. Risks 2020, 8, 97. Doi:10.3390/risks8030097. www.mdpi.com/journal/risks.