Modelling Motor Insurance Claims Using Arma Model And Appropriate Statistical Distribution of Insurance Companies, Machakos County, Kenya

Wafula Isaac[1], Jonah Masai[2]

[1,2]Machakos University, Kenya

Abstract

In the insurance industry, effective risk management and accurate prediction of claim incidence are crucial for ensuring solvency and optimizing resource allocation. This research paper aims to analyze and compare production and claims data from an insurance company, focusing on both third party and Comprehensive motor insurance claims in Kenyans' insurance company from 2012 to 2022. Advanced statistical techniques, including ARMA (Auto Regressive Moving Average), Moving Average, and AR (Auto Regressive) processes, are employed to model the data and gain insights into the relationship among the claims. Additionally, a statistical distribution is fitted to the claim amount variable, allowing for a more comprehensive understanding of insurance risks and improved risk management strategies. By identifying an appropriate distribution, insurers can accurately estimate claim amounts and better handle unexpected claim payments. Through the utilization of modern programming techniques and statistical software, this paper provides practical solutions for insurers to enhance their actuarial risk control strategies. The ARMA process captures both autoregressive and moving average components, enabling the analysis of temporal dependencies and trends in the data. The Moving Average process focuses on the moving average component. This research contributes to the body of knowledge in the insurance industry and serves as a guide for insurers seeking to improve their understanding of motor insurance claim risk.

Keywords: Augmented dickey- Fuller Test, Autoregressive moving average process, Auto correlation Function.

1. Introduction.

The prediction of future claim amounts based on past or present claims is a challenging task for insurance companies. According to research carried out by Michael Barth and David L. Eccles (2009); claim costs constitute large proportion of insurers expenses in evaluating the premium growth on the loss ratio of property causality insurers. Various statistical techniques have been employed to analyze claims data and assess its impact on productivity. Autoregressive (AR), Moving Average (MA), and Auto Regressive Moving Average (ARMA) processes are commonly used to model the temporal dependencies and trends in the data. These techniques enable insurance companies to understand the dynamics of claim incidence and their effects on productivity over time (Philip J, 2016). Advanced modeling techniques such as generalized additive models, negative binomial regression, and multinomial logit models have been used to evaluate claims processes, predict claim occurrence, and model claim severity (Renshaw A.E, 1994; Edward. W Frees et al., 2008; Daiane Aparecida Zanetti et al., 2006). These models consider various factors such as driver age, gender, vehicle type, and no claim discount to assess their impact on claim incidence and productivity. Previous studies have overlooked the comprehensive examination of the impact of insurance claims on productivity on insurance companies. Therefore, this research aims to address this gap by investigating the factors associated with insurance motor claims and their effects on the productivity of insurance companies by employing statistical techniques like appropriate probability distribution, incorporation of moving average technique, Autoregressive process and ARMA process, as identified by industry professionals and experts. The results obtained from the real-life data for distinctive 11 years indicated the potential strength of ARMA models to provide insurance company's claim amount prediction that could assist the companies to have better planning. The findings of this study, if adopted by various keynote players in the insurance and regulatory sectors, can be highly beneficial in formulating appropriate policies to enhance the management of motor insurance claims and improve the productivity of insurance companies.

2.Methodology.

The secondary data from an insurance company in Kenya is recorded from 2013 to 2020 for distinctive 11 years.

2.1 Modelling Process.

The data for motor claim (Private and Commercial) is not in the form of stationarity. Hence the trends and cyclic behavior have to be taken care of. In particular, any deterministic trends must be removed before applying the modeling process on the data. To Make the data stationary, the Autocorrelation function (ACF) is computed, and these suggest the most suitable model for the data. Hence, the data is analyzed by the ARMA (p, q) model, which is the combination of the Autoregressive and Moving Average model. The modeling process (ARMA) proposed by Box and Jenkins (1976) is as a standard technique in time series modeling and forecasting methods due to its structured modeling basis, and with best performance is implemented in this study to predict the gross outstanding claim amount.

For a time, series data X(t), ARMA the ARMA model is expressed as follows,

$X_t = \mu + \alpha_1(X_{t-1} - \mu) + \ldots\ldots + \alpha_p (X_{t-p} - \mu) + e_t + \beta_1 e_{t-1} + \ldots\ldots + \beta_q e_{t-q}$

Where Xt is the actual value, α_i is the coefficient of Autoregressive parameter (p), βj is the coefficient of moving average parameter (q), and e_t is the random error.

While ARIMA (p, d, q) process can be expressed as the AR (p) I (d) MA (q) in terms of the backward shift operator,

$(1 - \alpha B - \ldots\ldots - \alpha_p B_p) (X_t - \mu) = (1 + \beta_1 B + \ldots\ldots + \beta_q B_q)\, e_t$

In this research, we choose to develop an ARIMA model for this claim data, which is at least simple and gives at least correct description of the various patterns of the historical time series data.

2.2 Modelling process for statistical distribution

1. Identification of a model from a family of distribution

The assumption emerging from a given distribution family is selected based on frequency distribution, graphical output, knowledge about nature, and decision made

2. Selection of model parameters

Estimate the parameters based on the claim amount data and distributions using MLE. Usually, choose the parameters that decide a particular case of a distribution family

3. The validity of the distribution

The selected distribution's validity and parameters have a good fit for the data tested using statistical computations by considering the AIC and BIC performance criterion. The secondary data collected was analyzed using mean, median, variance, skewness standard deviation and kurtosis. Table 2.2.1 below shows the descriptive results of all gross outstanding claims to be paid and net retention respectively.

Gross Outstanding		*Net Retention*	
Mean	274747.8617	Mean	244962.9012
Standard Error	29734.45118	Standard Error	25184.16514
Median	115000	Median	100000
Mode	150000	Mode	150000
Standard Deviation	472955.3991	Standard Deviation	400578.6688
Sample Variance	2.23687E+11	Sample Variance	1.60463E+11
Kurtosis	17.59234329	Kurtosis	13.89988351
Skewness	3.682801121	Skewness	3.330865362
Range	3740579	Range	2582320
Minimum	1	Minimum	0
Maximum	3740580	Maximum	2582320
Sum	69511209	Sum	61975614
Count	253	Count	253
CL(95%)	58559.69318	Confidence Level (95.0%)	49598.25807

Table 2.2.1 Descriptive statistics of overall both Gross outstanding and Net retention for insurance policies

From the above table, we can see the summary statistics of the overall Outstanding gross claim amount and net retention. The mean outstanding claim was Rs. 274748 and net retention 244963, and the standard deviation was Rs. 472955 and 400578 respectively, and that of the skewness coefficient is 3.68 and 3.33 respectively, which indicates that the data is asymmetric. It suggests that the claim amount was positively skewed.

Gross Outstanding		*Net Retention*	
Mean	232071.8	Mean	225205.6
Standard Error	27660.58	Standard Error	27600.6
Median	105420	Median	84000
Mode	150000	Mode	150000
Standard Deviation	387248.1	Standard Deviation	386408.3
Sample Variance	1.5E+11	Sample Variance	1.49E+11
Kurtosis	15.87286	Kurtosis	16.2969
Skewness	3.567191	Skewness	3.63759
Range	2582319	Range	2582319

Minimum	1	Minimum	1
Maximum	2582320	Maximum	2582320
Sum	45486064	Sum	44140291
Count	196	Count	196
Largest (1)	2582320	Largest (1)	2582320
Smallest (1)	1	Smallest (1)	1
Confidence Level (95.0%)	54552.3	Confidence Level (95.0%)	54434.01

Table.2.2.2 Summary statistics for Motor private and Commercial claims.

From the above table, we can observe the summary statistics of Outstanding gross claim amount and net retention specifically for motor insurance claims. The average outstanding claim was Rs. 232071.8 and net retention 225205.6, and the standard deviation was Rs. 387246.1 and 386408.3 respectively, and that of the skewness coefficient is 3.56 and 3.64 respectively, which indicates that Motor insurance claims data is asymmetric. It suggests that the claim amount was positively skewed. The first step is to find a parameter of the particular statistical distribution to the claims data. In this case, the Maximum likelihood Estimation is used to determine the most likely probabilities for estimating the model parameter.

Gamma distribution

The Gamma distribution has two parameters: shape (α) and scale (β). We then estimate the number of two parameters and since we are fitted the distribution to the data, and performed maximum likelihood estimation to find the best-fitting parameters

Log Normal distribution

The log-normal distribution also has two parameters: mean (μ) and standard deviation (σ) of the natural logarithm of the variable. We fitted the distribution to data and then estimated these two parameters and calculated the maximum likelihood estimation to find the best fit.

Table 2.2.2; provides the parameters of two distributions fitted to the claims results. The estimated parameters of the distribution were used to find AIC, and BIC values of the two distributions in table 2.2.4

Gamma	Log Normal
Likelihood estimator	Likelihood estimator =110.8785
Shape =2.727767	Log mean =13.59803
Scale =0.004959023	Log std dev =2.276893

Table 2.2.3 Parameter Estimation for Motor commercial.

Gamma	Log Normal
Shape =2.727767	Log mean =14.65953
Scale =0.004959023	Log std dev=2.019248

Table 2.2.5: Parameter Estimation for Motor Private

The chosen distributions will be tested if they accurately describe the data or not by Akaike Information Criterion, Bayesian Information Criterion and log-like hood. From the results it indicates that the Log-normal and Gamma distributions will suit the claims distribution.

Table 2.2.6, summarizes the findings Performance Criteria, and goodness of fit test values for the individual distributions. Based on the results, one can choose the model with the lowest AIC or BIC values for modeling Motor insurance claims data respectively.

Table 2.2.6 and **Table 2.2.7Shows** Performance criterion for Motor commercial and Motor Private Respectively

	Gamma Distribution	**Log- Normal**
AIC	**226605.5**	**225.757**
BIC	**226609.4**	**229.6488**

Table 2.2.6 for Motor Commercial

	Gamma Distribution	**Log- Normal**
AIC	**411950.4**	**104.6871**
BIC	**411952.6**	**106.8844**

Table 2.2.7 for Motor Private

In this case, the lognormal distribution had shown minimum performance criteria values for both Motor private and commercial respectively. As a result, it was concluded that the lognormal distribution was the best-fitting distribution compared to gamma distribution for the claims data, as it had the lowest AIC and BIC value of 225.57 ,229.6488 for Motor commercial and 104.6871 and 106.8844 for Motor Private respectively. That is, of the two distributions considered, the lognormal distribution was calculated to be the most appropriate model to be fit for the claims data.

3.0 THE RESULTS FOR ARMA MODEL

Firstly, the data which is not stationarity is brought down to stationarity by differencing the process. Thus, the original process can then be regained by integrating the differenced series, and the original series will be modeled by an ARMA (p, d, q). The Augmented Dickey-Fuller (ADF) unit root test is important ($p = 0.00001 < 0.01$), and we conclude that the data evaluated is stationary. Now, we computed the optimal lag for the autoregressive parameter (p) and moving average parameter(q) based on autocorrelation function (ACF)

3.1 SUMMARY RESULTS FOR POSSIBLE ARMA MODELS

Variable coefficients	AIC	BIC
ARMA (0 0 0)	232.0575	232.0034
ARMA (0 0 1)	234.0094	233.9012
ARMA (0 0 2)	235.3166	235.1544
ARMA (1 0 0)	233.9956	233.8874
ARMA (1 0 1)	235.9883	235.826
ARMA (10 2)	236.3422	236.1258

ARMA (2 0 0)	235.9031	235.7408
ARMA (2 0 1)	237.0127	236.7964
ARMA (2 0 2)	238.2774	238.007

Table 3.1 Motor commercial

From table above, all good possible output of ARIMA models is estimated, and more coefficients are analyzed. The best output models of the 9 entries are compared using AIC, and BIC (Minimum value). Comparing all the models evaluated by ARIMA, ARIMA (2, 0, 0) model is better approximate model to use be of minimum values for AIC and BIC. Hence, we consider the model ARIMA (2, 0, 0) is the best model to predict the future claim's own damage amount in Motor commercial outstanding claims amount.

4.0 CONCLUSIONS AND RECOMMENDATIONS.

General insurance companies require an accurate insurance pricing process that makes adequate provision for expenses, and profits. Claims incurred by the company form a large part of the cash outgo of the company therefore reducing the company's productivity. An insurance company is required to model its claims in order to predict future claims experience are prepare adequately for claims when they fall due. In this project, selected probability distributions (Gamma and Log Normal distribution respectively) which are used as distributions for modeling outstanding claims made on Motor insurance policies. Therefore, an effort was made to determine an optimal statistical distribution that perfectly fits the insurance claims data by statistical computing and visual representation utilizing the R programing. This research demonstrated that the assumptions taken before analyzing motor insurance claims data could significantly impact the outcome the objective's beliefs resulted in selecting a family of distributions comprised of Lognormal and Gamma distributionS. As per our general analysis, it showed that the lognormal distribution could have a good fit for the data. The empirical results show that the proposed distributions successfully model the insurance claims.

4.1 RECOMMENDATIONS

We expect that the approach can see broader use in Actuarial Sciences and other similar fields, thus increasing the sample distributions in the study and ensuring the precision of the results. This analysis will help the insurance companies to predict the incidence of claims and help the companies in predicting unexpected claims through the use of ARMA model and Lognormal distribution.

REFERENCES

1. Marlin P (1984). Fitting the Lognormal Distribution to Loss Data Subject to Multiple Deductibles. The Journal of Risk and Insurance, 51(4), 687-701.
2. Achieng, O.M. and NO, I. (2010) Actuarial Modeling for Insurance Claims Severity in Motor Comprehensive Policy using Industrial Statistical Distributions. International Congress of Actuaries, Cape Town, Vol. 712, 7-12.
3. Krzysztof Burnecki, Grzegorz kukla and Rafal Weron (2000). Property Insurance Loss distribution.
4. Physica A: Statistical Mechanics and its Applications, 287(1), 269-278.
5. Omari C, Nyambura S and Mwangi J (2018). Modeling the frequency and severity of Auto Insurance Claims using Statistical Distributions. Journal of Mathematical Finance, 8(1), 137-160.
6. Ramin Kazemi, Abdollah Jalilian and Akram Kohansal (2019). Fitting Skew Distributions to Iranian

7. Auto Insurance Claim Data. Applications and Applied Mathematics: An International Journal, 12(2), 790-802.
8. Eling, Martin (2012). Fitting insurance claims to skewed distributions: Are the skew-normal and skew-student good models. Insurance: Mathematics and Economics, Elsevier, 51(2): 239-248.
9. Edward. W Frees and Emiliano A. Valdez (2008). Hierarchical Insurance Claims Modeling. Journal of the American Statistical Association, 103:484,1457-1469.
10. Daiane Aparecida Zanetti, Carlos A.R. Diniz and Jose Galvoa Leite (2006). A lognormal model for Insurance Claims Data. RevStat- Statistical Journal, 4(2),131-142.
11. Simon Fontaine, Yi Yang, Wei Qian, Yuwen Gu & Bo Fan (2020). A Unified Approach to Sparse Tweedie Modeling of Multisource Insurance Claim Data, Technometrics, 62:3, 339-356
12. Claudia Czado, Rainer Kasten Meier, Eike Christian Breachmann, and Aleksey Min (2012). A mixed copula model for insurance claims and claim sizes. Scandinavian Actuarial Journal, 4(1), 278-305.
13. Susanne G and Claudia Czado (2007). Spatial modeling of claim frequency and claim size in non-life insurance. Scandinavian Actuarial Journal, 2007:3, 202-225.
14. V V Haragopal and K Navatha (2017). Fitting Distributions to Big Data: Example of Large Claim Insurance Data. Visleshana, Big Data Analytics, Vol.1.