Strathmore UNIVERSITY

Insurance Pricing Using Geographical Ratings

Ivy Eva Wambui Kinyua- 096675

Submitted in partial fulfilment of the requirements for the Degree of Bachelor of Business Science in Actuarial Science at Strathmore University

Strathmore Institute of Mathematical Sciences
Strathmore University
Nairobi, Kenya

February 2021

DECLARATION

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the Research Project contains no material previously published or written by another person except where due reference is made in the Research Project itself.

Ivy Eva Wambui Kinyua [*Name of Candidate*]

[*Signature*]

.....

February 9, 2021 [*Date*]

This Research Project has been submitted for examination with my approval as the Supervisor.

Elphas Okango [*Name of Candidate*]

[*Signature*]

.....

February 9, 2021 [*Date*]

Strathmore Institute of Mathematical Sciences
Strathmore University

Table of Contents

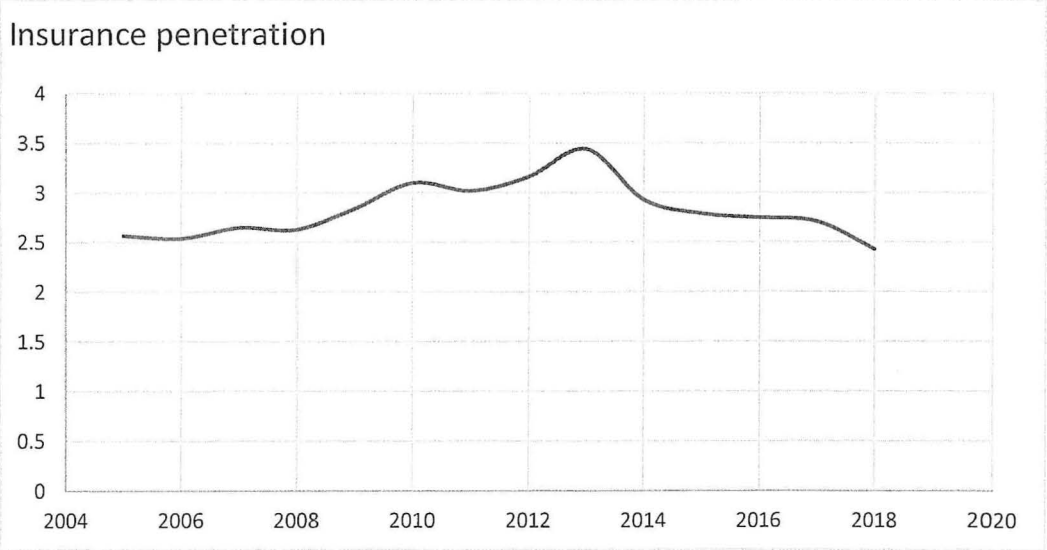
Abstract

With time, insurance industries are expected to improve their pricing models, enabling them to respond to changes and become more efficient. Like any other industry experiencing losses, the Kenyan insurance industry has been on a journey to find solutions to reduce the losses they incur. This research tries to help the Kenyan insurance industry by suggesting the use of geographical ratings as a risk factor in the calculation of premiums.

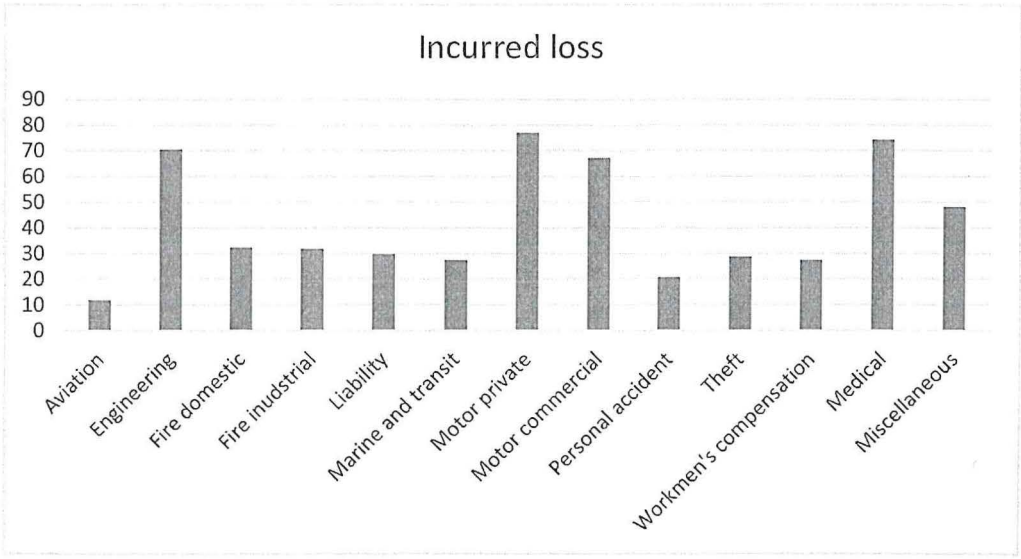
CHAPTER 1: INTRODUCTION

Background information

The Kenyan insurance industry's penetration rate is currently at 2.7% of the Gross Domestic Product (GDP). In clear disparity is South Africa with a penetration rate of 13.8%. The rate was at its peak in 2013 when it was 3.44% and has been declining ever since. According to AKI (Association of Kenya Insurers) journal 2019 report, the decline was due to three reasons; a new method of calculating GDP being introduced (debasing), spur on growth of the economy and competition from new entries. These have led to decreased premium growth hence the low penetration rate. The IRA (Insurance Regulatory Authority) 2018 report mentioned that some of the challenges faced in the insurance industry are mistrust towards insurance products, inappropriate insurance products, poor claims settlement, unsuitable pricing, among others. From my study, fraudulent claims and losses from underwriting are also setbacks to the progress of the insurance industry. According to IRA report on the industry's performance, the underwriting losses were at Ksh556 million in 2017, Ksh1.65 billion in 2018 and Ksh2.97billion in 2019, showing an increase in trend of losses from underwriting. The most loss-making businesses in the insurance industry are the medical, motor private and motor commercial businesses, despite them being the largest. One of the reasons is that they experience the most fraudulent claims in the insurance industry. The motor industry (including both private and commercial) had the highest number of total fraudulent claims from 2016 to 2019 that amounted to Ksh32,470,000. The medical industry had fraudulent claims adding up to Ksh9,253,718, AKI (2019)



Source: IRA 2004-2020 report



Source: IRA 2019 last quarter report

In 2018, Ksh310.49 million of the insurance industry income was fraudulently claimed (IRA 2018). According to the report, the general insurance industry has experienced stable growth in gross written premiums from 2012 to 2018 with the expense and claims ratio being on a slightly upward trend for the past six years. With the claims ratio being higher than the premiums, the insurance industry is still loss making despite the gross written premiums gradually increasing,

One way to curb this problem is through technology and innovation. The industry is slowly embracing the use artificial intelligence, as seen in telematics, which is made effective by use of big data. According to Deloitte Outlook (2019), big data makes algorithms more accurate since they acquire knowledge from the data and provide more detailed results. Insurers can use big data to establish trends on historical data and come up with possible signs of fraud. Use of artificial intelligence can be in processing claims, customer service, underwriting and fraud detection. It can be used to detect fraud through speech recognition, sentiment detection, text analysis and pattern or anomality detection. There is also the use of blockchains which helps in sharing data from one party to another in a secured way. It can automatically collect records of agreements, transactions, and other valuable information.

With correct pricing, the underwriting costs will be reduced, as the premiums to be paid by the policyholders will match the level of risk they are exposed to. Currently, the calculation of premiums for policyholders in motor insurance is determined by charging a premium of 5% of the value of the car (IRA, 2013). The premium rate also differs with the type of insurance cover taken. There are three types of car insurance cover. Third party only which is the most common and most affordable. It covers the policyholder against third party liabilities for property damage and bodily injuries. Third party, fire and theft are an advanced third party only as it covers third party property damage and bodily injuries and fire and theft of the insured vehicle. The comprehensive cover is the most expensive cover. It covers for damages on the car from accident, fire and theft and covers the third party. It, however, does not cover the insured.

These being the major factors assume that policyholders with the same brand of car and same insurance cover across the country would pay the same amount of premium without considering factors such as geographical, climatic, cultural and socioeconomical differences. Individuals living in areas that are prone to theft cases, natural calamities like floods, places with more traffic and those who drive their cars often and for more kilometres are more likely to claim. Geographical location also shows the behaviour of drivers in different areas.

This research aims at using geographical ratings as a rating factor that will help in coming up with more accurate decisions during problem solving.

Use of geographical ratings has traditionally been used in deriving disease atlases where a map of the primary geographical distribution of diseases is observed. That will help in detecting an outbreak and identifying significant trends in disease rates. We employ the same methodology in drawing insurance pricing in Kenya by looking at claims data in Nairobi, since it contributes to 76.4% of the gross direct premiums in Kenya, IRA report (2018). Understanding how motor insurance may vary with geographical locations is key in policy making and construction of informed, efficient, and profitable insurance products.

Problem statement

Many insurance companies across the globe have improved their pricing models in the underwriting process to incorporate geographical location as a risk factor. In a research carried out by Simon Grima on the analysis of risk factors used to determine insurance premium, one insurer did an analysis on the drivers in Malta compared to their neighbouring state Gozo. Gozo

was seen to have better drivers than Malta hence insurers merited a discount to vehicles in Gozo.

The Insurance Regulatory Authority (IRA) has a directive of developing the insurance industry in Kenya and a goal to promote a competitive and stable industry while providing quality customer service. One way of promoting this is by ensuring that product development in the industry is in line with customer needs and technology trends, for instance, the motor insurance industry has had trends in technology and regulatory developments globally. In the 2012 guidelines on insurance products under pricing, insurers are advised to ensure that there is a process for the product pricing to respond to competitive and other external pressures. Insurers are seen to be conservative since they are using traditional rating factors even though they might not necessarily reflect the risk being insured. It is therefore up to different insurers to improve their pricing models according to global trends in the industry, one of it being revising on the risk factors used to price.

Despite the few trends noted in the Kenyan insurance industry, we still have a lot to improve in terms of pricing models. Improving the underwriting process and pricing models by including geographical ratings would provide valuable data that can be used to detect undesirable business, hence help to reduce losses from having more claim rates than gross written premiums. In general insurance, private motor insurance is said to be the highest loss-making class with a loss of Ksh2.7 billion. Better pricing models will reduce the number of losses faced by the motor insurance industry and increase the reserves held by them.

Research objective

Main objective

Determine how to include geographical ratings for modelling.

Specific objectives

i. To determine how to model claims frequency
ii. To determine how to model claim severity
iii. To use the results of the model to inform pricing.

Research question

i. Is there significant modelling variations of claims frequency and severity?
ii. Can the results of the modelling inform insurance pricing?

Significance of the research

i. It will enhance customer experience as premiums charged will be fair.
ii. Motor insurance products will attract a wider market.
iii. Insurance companies will be able to prepare for claims from accident-prone areas by increasing their reserves.

CHAPTER 2: LITERATURE REVIEW

Introduction

An insurance cover is a policy that protects the policyholder against risk or liability that can lead to loss. Accurately predicting the risk related with each policy allows the insurance company to have different pricing for high and low risk individuals. The risk premium is the ratio between the total claims cost and the exposure to risk. The exposure to risk can be the number of policies, number of policy years and total premium charged. The risk premium shows the risk associated with a group of policies and can be modelled by modelling the severity and frequency of claims. The main reason for splitting the two is because the claim frequency is more stable than claim severity allowing the frequency to be estimated with greater precision, Ohlsson & Johannson (2010)

Theoretical review

Use of geographical ratings has originally and commonly been used in epidemiology (the study of disease occurrence, control, and distribution). According to Trevor C. Bailey (2001), modelling of geographical epidemiology has four areas. There is disease mapping where the geographical distribution of a disease is mapped. This helps in detecting the possibility of an epidemic transpiring in an area and identifying tends in disease movements in different areas. Bayesian hierarchical methods have been used in disease mapping. These are models written in multiple levels that estimate the parameters of the posterior distribution using the Bayesian method. The same models have been used in the insurance industry when modelling premiums where geographical ratings have been included.

The second area is ecological studies which studies the relationship between a disease occurring and the risk factors on groups defined by geographical location. This helps to determine the cause of a disease and come up with preventative measures. Bayesian hierarchical models can also be used to model ecology studies. Thirdly, there is disease clustering studies which identifies areas with high morbidity. This helps study causes of high disease occurrence and preventative measures. Markov Chain Monte Carlo (MCMC) models are used in cases where the data used has unknown parameters involved. The last area is environmental assessment and monitoring which studies the distribution of environmental factors closely connected to health to take prevention measures. There are no specific models for modelling environmental factors since it is a wide field. Most of them, however, use a

Bayesian approach. Image processing and remote sensing techniques are greatly used in this field.

The same methodology has been used in motor insurance where disease mapping can be used in getting the distribution of claims in an area, ecological studies can help determine the relationship between the risk factors to the probability of claiming and get the most relevant risk factors. Disease clustering can help identify areas with high claims and environmental assessment can link environmental conditions to the probability of a claim occurring in a certain area.

Empirical framework

In motor insurance, different models have been used to determine the premiums to be paid by policyholders while including their geographical location. This section, however, only mentions a few of the models with Generalised Linear Model as the main model being used by most researchers.

Boskov and Verrall model

The pure premium will be modelled through the claim frequency and severity. This will be done using multiplicative tariff models, where the expected frequency and severity will be obtained by getting the product of different rating factors such as age, gender, value of the object to be insured and past claims, Oskar Tufvesson (2019). Risk associated with a certain area can be included as a rating factor to represent geographical risk, which is expected to depend on the area's demographic and socio-economic status. The geographical risk among neighbouring areas is expected to be similar which may be a problem for people living directly outside the low-risk areas, as they may be charged higher rates due to higher risks even though they might be of low risk, A Conrad & F J Mostert (2009). This problem can be solved using Bayes method that connects local areas, especially areas with less reliable data, and enables them to 'borrow strength' from their neighbouring areas. This produces smooth estimates for individual local areas. The Bayesian statistical approach treats all unknown parameters as random variables and derives their distribution from the known information. The main advantage of Bayesian framework is that it recognises the magnitude of sample errors and includes the concept of smoothing over neighbouring areas.

In this model, individual claim frequencies are modelled using a Poisson regression model. This is done after getting the posterior distribution of the area under study as a function of Y/X

where Y_i is the observed outcome relating to area i and X_i is the true risk in area i. X is a joint distribution for u and v, where u represents a component with significant geographical factor structure and v represents unexplained variations without a geographical factor. The geographical risk factor is included in u. A type 3 analysis is used to remove least significant variables in the model, and you end up with a linear predictor with relevant variables.

The model as seen in Boskov M. & Verall R. J. (1994) only investigates the frequency risk of claims from different geographical codes. Our aim in modelling the geographical rates is including both the frequency and severity of claims.

Generalised Linear Model (GLM)

GLM in insurance has been used to include premium ratings in modelling the claim frequency and severity. The rating factors important in modelling can be categorized into properties of the policyholders (include age, gender, line of business for a company), properties of the insured objects (include age or model of the car, type of building) and properties of the geographic region that include per capita income and population density of policyholder's residential area, Esbjorn Ohlsson (2010). Limitations on the use of rating factors is that some might be found offensive by the policyholders.

One of the assumptions used in modelling is that the claim frequencies are independent. This might be violated in cases where a catastrophe like floods has caused many claims. Such claims are not categorized as normal claims as they occur at rare occasions. Another assumption is that claims frequency and claims severity are independent at different time periods. This assumption simplifies the model building. Assumption 3 is that claims from the same tariff cell with the same exposure have the same distribution and hence will be charged similar premiums. For non-homogeneity within the same cell, there are bonus systems put in place. Homogeneity is important as it provides repeated observations for statistical analysis. This study also uses multiplicative models in cases where some tariff cells have little or no claims data. These models help to determine the expected pure premium that will vary more smoothly over the cells.

Geographic rating factors are modelled using demographic and socio-economic variables as well as neighbouring geographic rating factors. Claim frequencies were modelled assuming that the number of claims for a single policy follows a Poisson distribution. Since the policies are assumed to be independent, the number of claims will also follow a Poisson distribution. A

linear predictor of the GLM for claims and exposure is obtained. Since the areas will likely be heterogeneous with respect to demographic and socio-economic status, the raw exposure is replaced by a weighted exposure, which will be computed according to the composition of policyholders in each area and their rating factors, Oskar Tufvesson (2019).

When modelling the claims severity, it is not certain on the type of distribution to use. Positive models that are skewed to the right like gamma distribution are, however, used in GLM estimation eliminating the possibility of using a normal distribution. After modelling the claim frequency and severity, some papers like Oskar Tufvesson (2019) have used a Besag, York and Molly model to model the geographical dependence and elastic nets for selecting covariates.

Bonus Malus System (BMS)

The term bonus-malus is a Latin word meaning good-bad, where bonus is rewarded and molus is penalized. This system has commonly been used in the insurance industry, for example policyholders who do not claim in a year are rewarded by being given a discount on the following year's premiums, encouraging them to be more careful while driving. According to David Pascula-Ezama (2015), it partly shows characteristics of the driver that are not easily noticeable like stress levels and risk taking, which might be reflected on the number of claims reported by the policyholder. In some countries like Spain, drivers are required to renew their driving licence regularly where medical tests are taken to determine the driver's physical conditions, or a questionnaire that shows the driver's personality and level of risk taking. This means that premiums are re-adjusted depending on the number of claims reported by the insured. This is done by multiplying the original premium by the BM coefficient determined from the bonus malus scale.

Policies are classified into different homogenous tariff classes where policyholders with similar risk factors are classified in the same class. It makes it easier for the insurance company to charge premiums as policyholders in the same class are charged the same premium. This calls for accuracy from the insurance company by having adequate risk factors that will help to correctly predict future claim rates, especially in a case where the claims rate is higher than the gross written premiums. In motor insurance, these risk factors are personal driver's information like age, gender and place of residence and information on the car like age.

When grouping policyholders, there might be a problem of policyholders with the same data information being grouped in different classes. This will create uncertainty and question the precision on allocation of policyholders to different sets. In this case, a Rough Set (RS) method is introduced to solve such cases by accurately grouping the policyholders. This is done by presenting the data in a table in which rows are labelled policyholder and columns are labelled risk factors. The RS methodology then creates approximations to the classes where the lower approximation represents policyholders who certainly belong to the class and can certainly be grouped while the upper approximation represents the insured who possibly belong to the group and can possibly be classified. This method focuses on variables that are most relevant reducing time, cost and effort of decision making.

Machine Learning (ML)

Machine learning is a field in artificial intelligence that performs classification and prediction methods on data by using limited assumptions on these data sets. According to Julien Antunes Mendes (2017), these techniques have significant advantages over traditional methods (like GLM) by offering different non-linear models that can give more insights and having more complex patterns in the data set that predict more accurately. Since ML is considered a "black box" solution (solutions that have internal mechanisms that are mysterious to the user), one of its limitations is that these solutions are not easily predictable, and they rely heavily on data that might not be available, Kevin Kuo (2020).

The machine learning techniques are Classification and Regression Trees (CART), Random Forests, Gradient Boosted Machines (GBM) and Deep Learners, Giorgio Alfredo (2018). Random Forests have tree models that divide datasets into subsets. The splits in data are defined by 'if then' statements that determine the terminal nodes. New data has different terminal nodes depending on the route taken. These trees can either be used as regressors, to predict continuous responses, or classifiers to predict class probabilities. These trees are known as Classification and Regression Trees. One advantage of using random trees is that they can handle both numeric and categorical predictors without pre-processing. A disadvantage is model instability where a small change in the data can cause a large change in the tree.

The GBM uses both regression and classification trees. The trees are dependent on one another where a new tree corrects the error of previous trees to improve accuracy. An example of a

variant in GBM is eXtreme Gradient Boosting (XGBoost) that has mostly been used by data scientists due to its fast and better performing algorithm.

Deep Learning uses neural networks as inspired by the structure and function of the brain known as artificial neuron networks. These networks are used to solve classification and regression problems. They use many layers to gradually extract higher level features from raw input, and have been used for image recognition, natural language processing and detecting recurring patterns.

Research gap

Recently, Kenya recognized the use of telematics in the insurance industry. Some of the telematics use the GPS tracker to get information on environmental factors such as the weather, location of claim and the conditions of the road. Geographical information captured by telematics is, however, not enough to include geographical location as a rating factor. As the industry grows to embrace telematics, having geographical ratings as a risk factor should also be considered, for the insurer to be able to capture all the geographical factors that are important in the calculation of premiums.

CHAPTER 3: METHODOLOGY

Research Design

The research design used is the diagnostic research design. After concluding that the Kenyan insurance industry is loss making, we try to show how geographical ratings can be included as a key risk factor in calculating the premiums by changing the pricing models. Due to limited data, geographical regions have been divided into 2 namely: Nairobi and outside Nairobi

Population and Sampling

The population for the study will be policy holders in Kenya who have claimed between 2015 and 2019 as recorded by Britam General Insurance. We sampled the data for analysis basis using convenience sampling. We used data that gave information on the policyholders and the vehicles they used. Such data was like city, vehicle type and vehicle make. We got different categories for each column and omitted incomplete data. The remaining part of data was used for analysis.

Data Collection

The type of data used in this research is quantitative discrete. The private motor claims data was obtained from Britam General Insurance by following the company's regulations and procedures. Since the data contains private information, we made confidential agreements on the usage of the data.

The data collection instrument used was electronic mail, where we had to request for the data specifics and explain what it would be used for.

Data Analysis

Claim frequency and severity will be modelled using GLM, Oscar Tufvesson (2019). In the analysis, the main aim is to use geographical ratings.

The claim frequency will be modelled using a Poisson distribution by assuming that policies are independent. The number of claims in each area, i , represented by Y_i will be modelled as follows:

$$Y_i \sim Poi(\mu)$$

Since we are modelling using GLM, a linear predictor will be needed. This is a function of the covariates to be used in the model. The $\beta_i's$ show the correlation between the independent variables and the dependent variable. The higher the value the higher the level of correlation.

The $X_i's$ represent risk factors such as geographical ratings, age, vehicle type, theft, fire, and public liability.

$$\omega_i = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k$$

We then determine the link function that connects the mean of the independent variable to the linear predictor (covariates). The link function of a Poisson distribution will use the log function. In general,

$$g(\mu) = \log(\mu) = \omega_i$$

$$g(\mu)=\log\mu= \alpha+\beta 1X1+\beta 2X2+\beta 3X3+\ldots+\beta kXk$$

Making the mean the subject of the formula,

$$\mu = \exp(\omega_i) = \exp(\alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k)$$

The mean shows the expected number of claims with respect to the explanatory variables.

Therefore, the linear predictor using GLM of the claims frequency is:

$$Y_i = Po{\sim}(E_i \exp(\omega_i^F))$$

Where E_i is the raw exposure. In this research, the exposure will be the total duration of policies in area i. The raw exposure will be replaced by a weighted exposure E^* to account for variation of risk in area i. It is calculated according to the number of policyholders in each area and their rating factors.

The claims severity is modelled using a gamma distribution as has been the standard method, c. It will be modelled conditioned on the number of claims in each area.

$$S_i|Y_i = y_i{\sim}\Gamma(y_i a, a \exp(-\omega_i^S))$$

The linear predictor for the claim severity is ω_i^S and the link function is:

$$E(S_i|Y_i = y_i) = y_i \exp(\omega_i^S)$$

The combined GLM for claim frequency and severity is:

$$\omega_i = \beta_0 + z_i\beta$$

Where β represents the coefficient of the covariates and z_i the collection of covariates in each area. This will then be used to calculate pure premium using a GLM model.

The data contained many risk factors. The covariates used for analysis were those that describe the policy, claim and policyholder. The covariates were city, fleet type, vehicle type, vehicle make, description (cause of claim), claim number and number of months in policy (exposure) were used for modelling. They were chosen

The final step will be to model the pure premiums. We use a tweedie GLM model since it can access distribution combinations that are not allowed by GLM in R programming. A tweedie GLM assumes that:

Let $\mu_i = E(y_i)$ be the expectation of the i^{th} response. We assume that:

$$\mu_i^q = x_i^T b, var(y_i) = \emptyset \mu_p^i$$

Where x_i is a vector of the independent variables and b is a vector of regression coefficients for some $\emptyset, p \; and \; q$. p is the var power and q the link power of the function. The canonical link for a tweedie family is link power= I – var power, (Dunn, P.K, (2018)).

The following table summarises possible tweedie response distributions:

Var power	**Response distribution**
0	Normal
1	Poisson
(1,2)	Compound Poisson, non-negative with mass at zero
2	Gamma
3	Inverse-Gaussian

CHAPTER 4: ANALYSIS, RESULTS AND DISCUSSIONS

Introduction

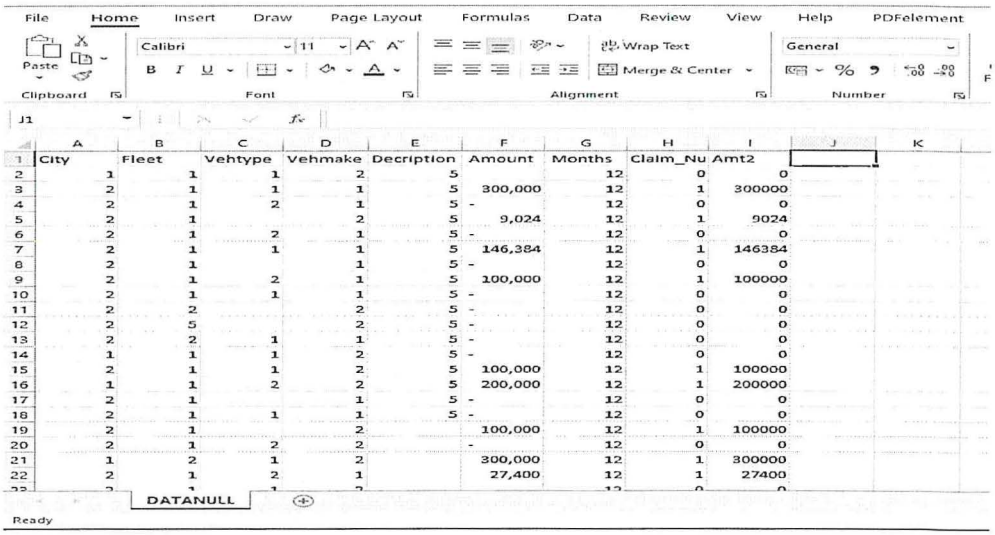
This chapter contains the analysis of the data and results of the modelling. As discussed in chapter one, the research questions to be answered are: Is there significant modelling variations of claims frequency and severity? Can the results of the modelling inform insurance pricing?

Description of the data

As mentioned in chapter 3, the data used for this research is from Britam General Insurance. The data was in excel form and contained different sheets. During data collection, we requested for the policyholders' geographical location. The information was then used to calculate the geographical ratings on excel. The ratings used in this analysis are also used by Britam to calculate premiums, except city (geographical ratings).

The policyholders information that was used for analysis was location (city), the fleet type of their cars, the start and end date of their policies, vehicle type and vehicle make. Under claims data, claims amount, description (cause of loss), date of loss and date of report were used for analysis.

For effective results, the data set was categorised. City was categorised into 1=Nairobi, 2=others, Fleet into 1= PC1(B), 2= PC1(A), 3= PC1(C), 4= PSV and 5= MC1. Vehicle type was categorised into 1=S/Wagon and 2= Saloon, vehicle make into 1=Toyota, 2=others and 3= Motorbike, and the description into 1= Accident, 2= Fire damage, 3=Flood damage, 4= Theft and 5= RTA (Road Traffic Act). The categories were then used for modelling in R. Classifying the data will help in knowing which categories are more prone to accidents.



City	Fleet	Vehtype	Vehmake	Decription	Amount	Months	Claim_Nu	Amt2
1	1	1	2	5		12	0	0
2	1	1	1	5	300,000	12	1	300000
2	1	2	1	5	-	12	0	0
2	1		2	5	9,024	12	1	9024
2	1	2	1	5	-	12	0	0
2	1	1	1	5	146,384	12	1	146384
2	1		1	5	-	12	0	0
2	1	2	1	5	100,000	12	1	100000
2	1	1	1	5	-	12	0	0
2	2		2	5	-	12	0	0
2	5		2	5	-	12	0	0
2	2	1	1	5	-	12	0	0
1	1	1	2	5	-	12	0	0
2	1	1	2	5	100,000	12	1	100000
1	1	2	2	5	200,000	12	1	200000
2	1		1	5	-	12	0	0
2	1	1	1	5	-	12	0	0
2	1		2		100,000	12	1	100000
2	1	2	2		-	12	0	0
1	2	1	2		300,000	12	1	300000
2	1	2	1		27,400	12	1	27400

Figure 1: DATANULL- Snapshot

Data Analysis and Modelling

Background

This section starts by modelling the expected claim amount per claim for both the claims frequency and severity. This will then be used to model the pure premium, that will be compared to Britam's original premiums. As discussed in chapter 3, claim frequency and severity are expected to be positively skewed leading us to using a GLM model.

Data Analysis

This study employs complete case analysis. The bar plot below shows the proportion of data that had claims and those that did not.

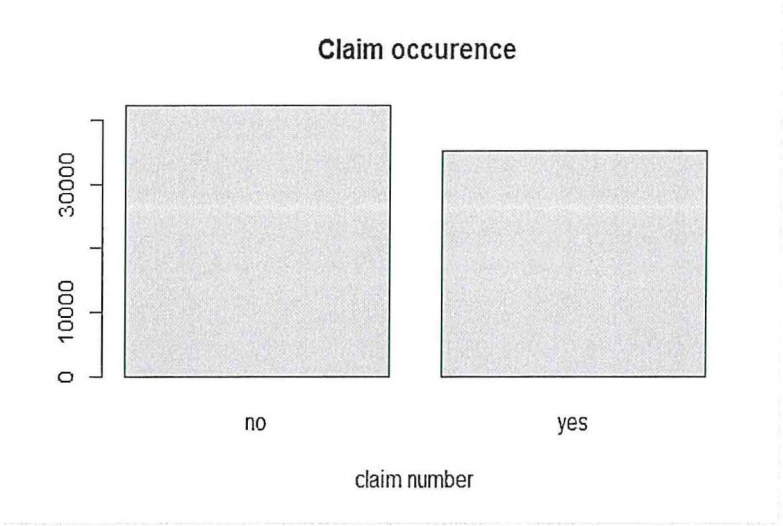
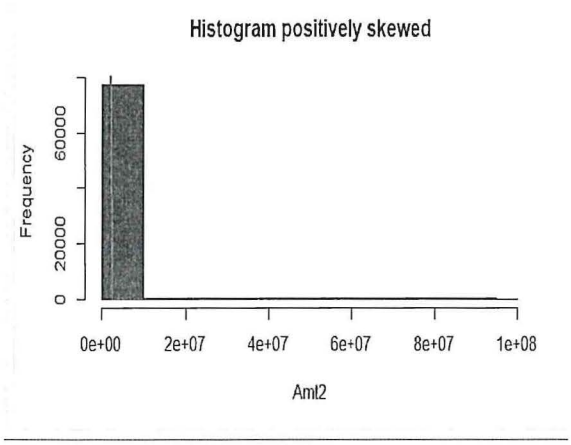


Figure 2: Claim occurrence

From the graph, a larger proportion of the data did not claim. The data cannot be modelled with most of the outcomes being no claims. Outliers beyond the 99th percentile are removed making the data positively skewed as seen below:



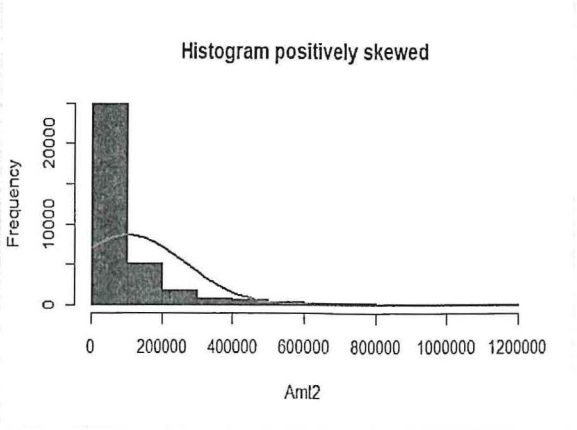


Figure 5 shows the vehicle make which is more prone to accidents while figure 6 shows the highest cause of claim as Road Traffic Act (RTA).

Figure 3

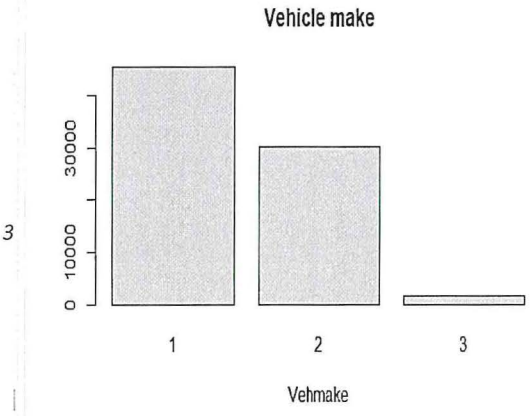


Figure 5: Vehicle make

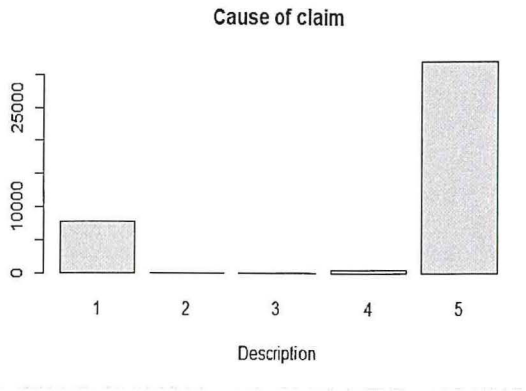


Figure 6: Cause of claim (description)

From figure 5, vehicle make with the highest claim is 1 (Toyota) and the least is 3 (Motorbike). From figure 6, the highest cause of claim is 5 (Road Traffic Act), followed by 1 (Accident) then 4 (Theft).

Modelling

Claim Frequency

Claim frequency is the expected claim count per unit of exposure. It is assumed to follow a Poisson distribution meaning the mean and variance are equal. A GLM model with Poisson distribution and log link function is therefore appropriate for modelling. A challenge in modelling insurance data is the tendency of claims data having more zeros causing a great variability. In modelling claim frequency, Poisson would underestimate the variance of the observed counts.

To check for variability, we run a Poisson regression and test the null hypothesis of no variability in the model against alternative hypothesis of over or under variability.

Variables definition

Claim number will be the target variable. The independent variables will be city, fleet type, vehicle type, vehicle make and description (cause of claim). Claim amount (Amt2) is not used as it is related to the dependent variable. The offset term to be used to model claim number per exposure, exposure being months.

Modelling

Before modelling, goodness of fit test and dispersion tests are done to determine the relationship between claim frequency and the rating factors. On the dispersion test, the regression gives an alpha of -1, which is less than zero. We therefore reject the null hypothesis that alpha= 0, meaning the mean is not equal to variance. The following text is the result of performing the dispersion test:

Overdispersion test

data: rd

z = -1.8959e+15, p-value = 1

alternative hypothesis: true alpha is greater than 0

sample estimates:

alpha -1

Rootogram is used to assess Poisson model fit on the data and determine if overdispersion is a serious concern. As seen below, Poisson model fits the data quite well, hence is fit to model claim frequency.

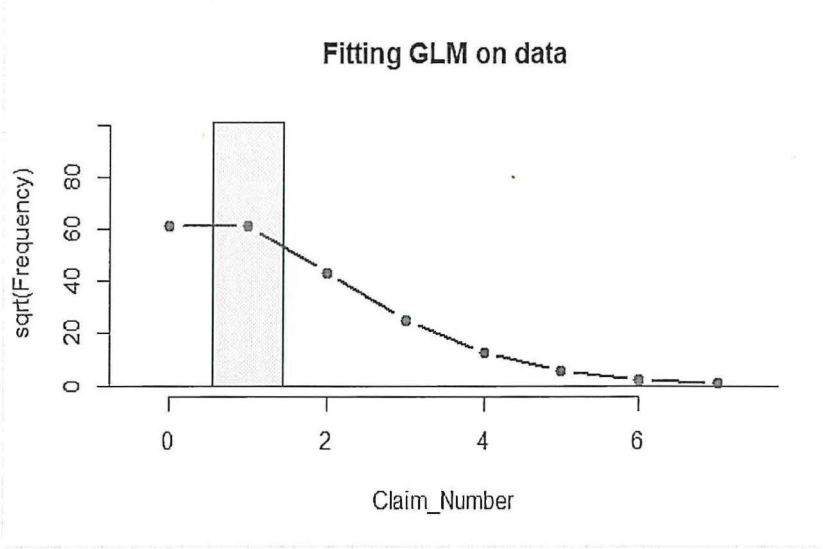


Figure 7: Fitting Poissonglm on Claim_number

GLM was fitted on claim number since that was the target variable in modelling claim frequency. In the original data set, the claim number was 0 for no claims and 1 for a claim. This explains why the bar graph is concentrated around 0 and 1.

Claim Severity

When modelling claim severity, it tends to have an excessive number of zero outcomes (as seen in figure 2), that makes it difficult to model. In this case, a GLM model with gamma distribution will be used. The term 'offset' is used in R to mathematically replace the claim amounts with claim severity. The GLM with Gaussian distribution is a good fit when using the 'offset' function since we need to use log (claim number) as an offset for it to be on the same scale as the linear predictor. For this reason, modelling will be done on GLM with gamma and GLM with Gaussian, both with log as the link function, then finalise on the better model.

Variables definition

Claim amount (Amt2) is used as the target variable. The independent variables will be city, fleet type, vehicle type, vehicle make, description (cause of claim), claim number. The offset term to be used is claim number.

Modelling

Before modelling, the data set is split into 80% train and 20% test. This means that most of the data is used for training and the remaining for testing. It ensures that data in both sets are similar. This is important to minimise inconsistency in the data and help understand the characteristics of the model better.

We model GLM with Gaussian then GLM with gamma, and see the following results:

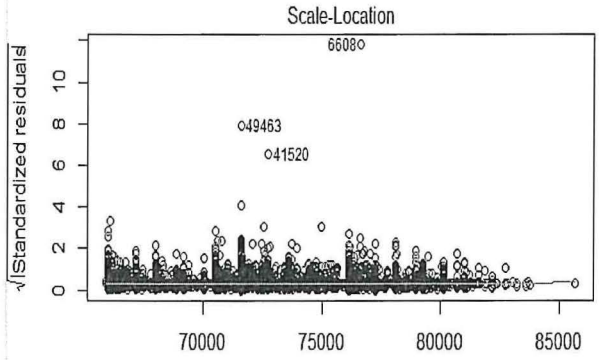


Figure 8: GLM with Gaussian

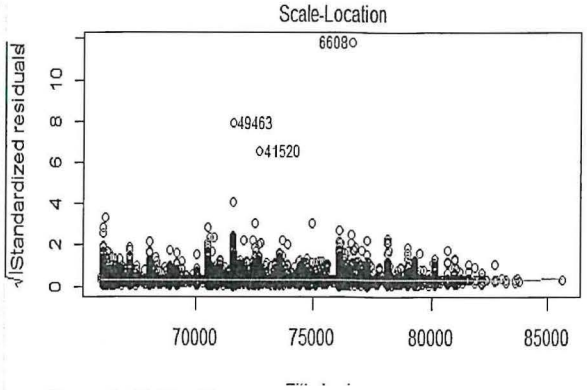


Figure 9: GLM with gamma

The graphs in figure 8 and 9 resemble making it hard to tell which model is a better fit. We therefore look at the AIC values of regression to determine the better fit. GLM with Gaussian has an AIC of 273,748 while GLM with gamma has an AIC of 254,333. Since GLM with gamma has a lower AIC, it is a better model for claim severity.

Pure Premiums

The final part of the analysis is to get the pure premium which is a product of the claim frequency and claim severity previously modelled. 'Tweedie' distribution will be used to model since it models directly without the need of any other models. It is a special case of exponential dispersion models mostly used as a distribution in GLM. It has been used to model claims due to its property to cluster data items at zero.

The response variable will be claim amount (Amt2) and exposure (Months) as the offset. The independent variables used in the previous regressions will be used here.

Modelling

Before modelling using tweedie, the variance power ('p') needs to be defined. It ranges between 1 and 2. The function 'tweedie. profile' is used to generate the maximum likely value of p. As seen below, the graph was not a curve. It was, however, able to show the maximum value, pointing above 2. Since the 'p' value should be between 1 and 2, I decided to use 1.8 in this analysis.

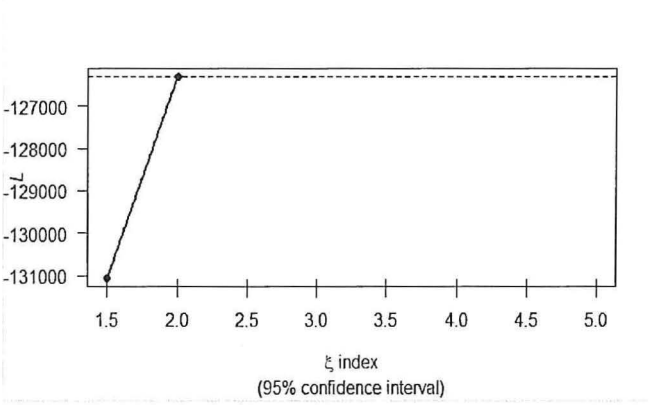


Figure 10: 'p' value for tweedie

Results

The results of the study will be discussed in line with the research question in chapter 1.

i. Is there significant modelling variations of claims frequency and severity?

In this analysis, it was found that there is significant modelling variations of claim frequency and severity that could be used for modelling. The following snap shots are form the analysis and will be used to show that the models are significant for modelling. Significance will be

measured by looking at the deviance, Akaike Information Criterion (AIC) and the number of fisher scoring iterations. The number of fisher scoring iterations is the number of iterations made to fit the model. A low number means that the model is a good fit. AIC is a tool use to evaluate how well the model fits the data set. It is used when comparing two models. The model with a lower AIC is a good fit. Under deviance, we have the null deviance (how well the response variable is predicted by the model) and residual deviance. The residual deviance shows how well the model predicts when independent variables are added. It is also a measure of goodness of fit.

Claim Severity

```
> summary(model_gamma)

Call:
glm(formula = Amt2 ~ Decription + Fleet + City + Vehtype + Vehmake +
    Months + Claim_Number, family = Gamma(link = "log"), data = train,
    offset = log(Claim_Number))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.7728  -1.7235  -0.3860   0.2041   3.6317

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.920991   0.084916 140.386  < 2e-16 ***
Decription2 -2.033537   0.994156  -2.045 0.040832 *
Decription3 -3.780566   1.404743  -2.691 0.007129 **
Decription4  0.577715   0.138843   4.161  3.2e-05 ***
Decription5 -0.123060   0.035209  -3.495 0.000476 ***
Fleet2       0.042714   0.036502   1.170 0.241966
Fleet3      -0.038187   0.293264  -0.130 0.896399
Fleet4      -1.113212   0.811088  -1.372 0.169941
Fleet5       0.069409   0.091454   0.759 0.447900
City2       -0.079926   0.051497  -1.552 0.120681
Vehtype2     0.019108   0.028453   0.672 0.501877
Vehmake2     0.052448   0.028890   1.815 0.069487 .
Months      -0.006772   0.005106  -1.326 0.184770
Claim_Number        NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 1.971976)

    Null deviance: 26197  on 10185  degrees of freedom
Residual deviance: 26074  on 10173  degrees of freedom
  (17638 observations deleted due to missingness)
AIC: 254864

Number of Fisher Scoring iterations: 10
```

Figure 11: Claim severity modelling results

When modelling claim severity, model gamma was used over model Gaussian since it had a lower deviance, and AIC. The number of fisher scoring iterations is 10, which is a significant number to measure the model fit. The residual deviance has reduced by 123 with a loss of 12 degrees of freedom. This reduction in deviance is significant. Looking at the p-values, the

significant rating factors are description 2, description 3, description 4 and description. This is because their p-values are below 0.05.

Claim frequency

```
> summary(poissonglm)

Call:
glm(formula = Claim_Number ~ Decription + Fleet + City + Vehtype +
    Vehmake + Months, family = "poisson", data = train, offset = log(Months))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.25715   0.01680   0.01935   0.02084   0.72768

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.670447   0.062102 -10.796   <2e-16 ***
Decription2  0.018478   0.707953   0.026    0.979
Decription3  0.017155   1.000337   0.017    0.986
Decription4 -0.005048   0.098868  -0.051    0.959
Decription5 -0.003759   0.025073  -0.150    0.881
Fleet2       0.001174   0.025991   0.045    0.964
Fleet3       0.019983   0.208837   0.096    0.924
Fleet4       0.099384   0.577546   0.172    0.863
Fleet5       0.003522   0.065125   0.054    0.957
City2       -0.002894   0.036665  -0.079    0.937
Vehtype2     0.001503   0.020261   0.074    0.941
Vehmake2    -0.002498   0.020572  -0.121    0.903
Months      -0.146828   0.003818 -38.459   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1195.04  on 10185  degrees of freedom
Residual deviance:  104.29  on 10173  degrees of freedom
  (17638 observations deleted due to missingness)
AIC: 20502

Number of Fisher Scoring iterations: 4
```

Figure 12: Claim frequency model results

The deviance residuals are low (below 1) meaning the model is appropriate for claim frequency. The number of fisher scoring iterations 4, meaning the model is a good fit. The residual deviance has reduced by 1090 with a reduction of degrees of freedom of 12. This is a great reduction caused by including the variables in the model to make it a good fit. In this model, only months is a significant rating factor.

When modelling claim frequency, a dispersion test was performed, and it was seen that the data was over dispersed. A GLM model was then fit to see if it was a goof fit. We concluded that despite the presence of overdispersion, it would not be an issue to use GLM. This explains why the residual variance has reduced by 1090 for GLM to be a good fit for the data.

ii. Can the results of the modelling inform insurance pricing?

The results of the modelling can be used to inform pricing. We were able to come up with a pure premium model whose results were also significant. The from the analysis were compared with the original premiums to see if the premiums from this analysis are more informed than the ones charged by Britam.

```
> summary(tweedie_model)

Call:
glm(formula = Amt2 ~ Decription + Fleet + City + Vehtype + Vehmake +
    Months, family = tweedie(var.power = 1.8, link.power = 0),
    data = train, offset = log(Months))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-10.4347   -5.0857   -1.4304    0.6268   13.7636

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.381341   0.085696 132.811  < 2e-16 ***
Decription2 -2.013415   1.218324  -1.653 0.098441 .
Decription3 -3.767087   2.070184  -1.820 0.068836 .
Decription4  0.583598   0.131758   4.429 9.55e-06 ***
Decription5 -0.128743   0.035448  -3.632 0.000283 ***
Fleet2       0.042260   0.036899   1.145 0.252125
Fleet3      -0.021396   0.299252  -0.071 0.943004
Fleet4      -1.083497   0.920799  -1.177 0.239346
Fleet5       0.074521   0.092034   0.810 0.418126
City2       -0.089347   0.051770  -1.726 0.084402 .
Vehtype2     0.018897   0.028825   0.656 0.512102
Vehmake2     0.046259   0.029236   1.582 0.113618
Months      -0.162876   0.005179 -31.452  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Tweedie family taken to be 21.10381)

    Null deviance: 236451  on 10185  degrees of freedom
Residual deviance: 203849  on 10173  degrees of freedom
  (17638 observations deleted due to missingness)
AIC: NA

Number of Fisher Scoring iterations: 5
```

Figure 13: Pure premium model results

We look at the significance of tweedie GLM model in modelling pure premium. From figure 13, it is seen that the residual deviance has reduced by 32602, which is not a large number compared to the deviance values. Significant rating factors are description 4, description 5 and months. The degrees of freedom have also reduced by 12. The number of fisher scoring iterations is also a significant number, 4. The model was a good fit to model pure premium.

CHAPTER 5: DISCUSSION, CONCLUSION AND RECOMMENDATION

Discussion

This section contains a discussion on the analysis done. The aim of the analysis was to improve the insurance pricing model that would help in charging more accurate premiums for the forecasted claims.

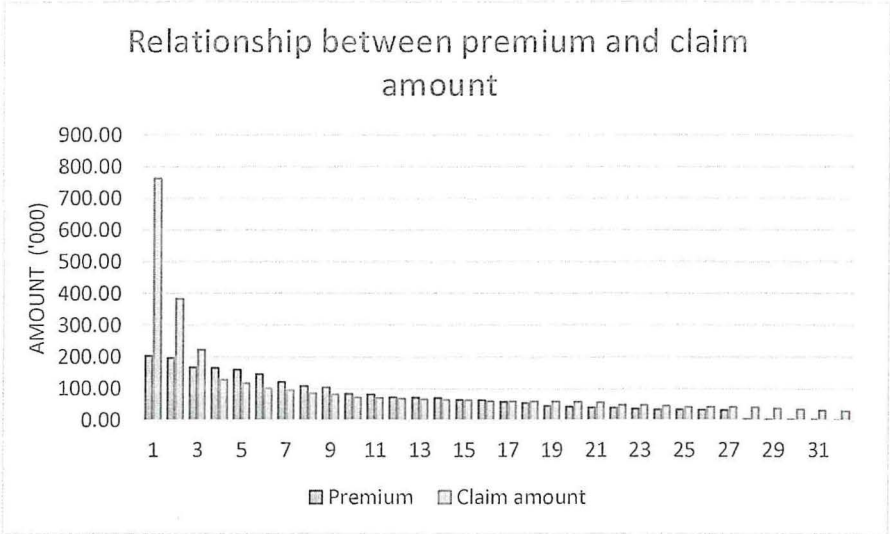


Figure 14: Relationship between original premiums and claims

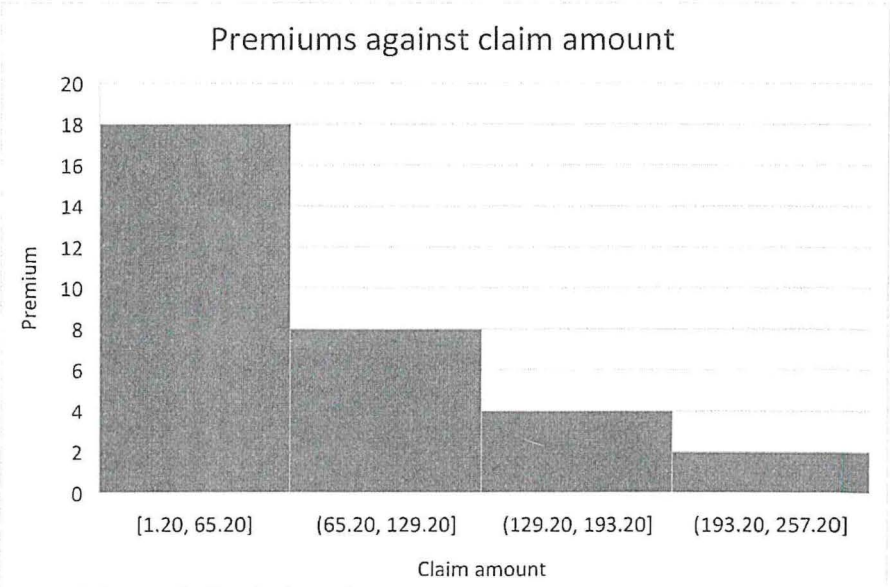


Figure 15: Bar graph of original premiums against claims

The above graphs show the relationship between premiums charged by Britam against the claim amount they pay. Since the data is large, the data used for plotting was on above average basis. Duplicates were also removed for effectiveness. In the first graph, the claim amounts are seen to be higher than the premiums charged. This can cause losses for the company since the premiums paid do not match the claims the company pays.

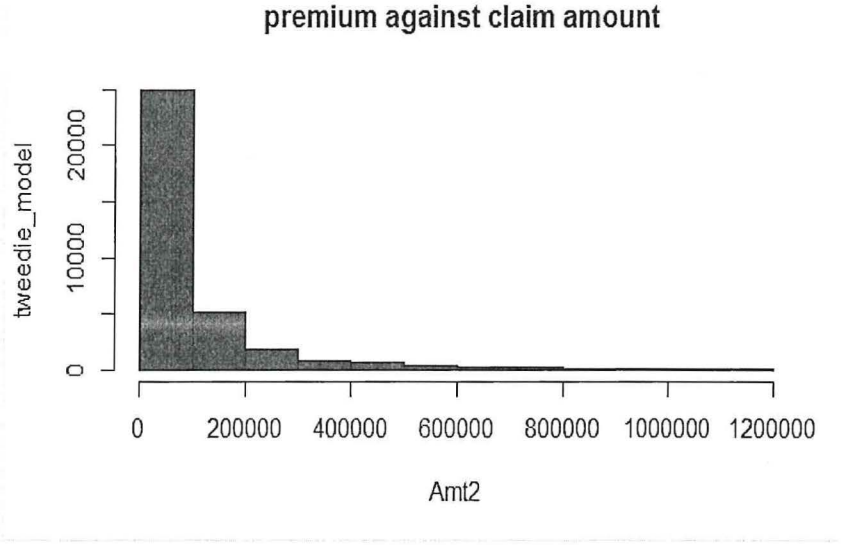


Figure 16: Premiums against Claim amount

Figure 14 shows the premium model generated in R against the claims amount. Compared to the bar graph on Britam's premiums against claim amount, we see that the premium generated in R is a better fit on the claim amount.

Linkages to other studies

The results of this study are consistent with the observations of Oscar Tufvesson (2019), who found out that including geographical ratings as a rating factor improves the pricing model. The covariates used in this literature were selected using elastic nets, which filtered them depending on their significance in modelling. They also used a Besag, York and Mollie model to model geographical ratings. In our analysis, this was not necessary since the geographical ratings were provided in the data and the rating factors used were the ones used by Britam.

Limitations of the study

Since we modelled the premiums using a tweedie model, R was not able to plot the model on the data set for us to know if it is a good fit. We resulted in using a bar plot to get the fit of the model.

When modelling the claim frequency, the data was seen to be over dispersed by 1. We tried to fit negative binomial, zero inflation and hurl models as they are appropriate for over dispersed models. The data set, however, did not support the above models. This resulted in using a Poisson model. It was not a perfect fit, but it was a good fit.

Conclusions

The findings suggest that including geographical rating in insurance pricing results to more accurate allocation of premiums compared to not using it. This will be a good step in reducing the losses faced by insurance companies.

Even thought the models used were appropriate for the data set, there are other models that would be suitable for other data sets. When modelling claim frequency with an over dispersed data set, a Poisson model might not be a good fit. Other models that can be used are negative binomial, zero inflated models and hurdle models. These models are appropriate in a data set with excessive zero counts, Ajay Tiwari (2020). A negative binomial model will be used because of its extra variation in the distribution. As this term tends to zero, the negative binomial tends to Poisson equating mean to the variance. The zero inflated models have two parts, a Poisson or negative binomial model and a logit model for modelling the excess zeros. A hurdle model has two processes, one where zero counts are generated and another one where positive values are generated. To choose the appropriate model, fit these models to the data and choose the best fit.

Recommendation

This study will be helpful to students interested in finding solutions on how to help the Kenyan insurance industry make profits. This study will also be relevant to the IRA and insurers in Kenya as they can get an idea of how to get the insurance industry back to making profits.

Appendix

This section shows the codes used in the analysis.

```
## get the working directory

getwd ()

## set the working directory

setwd ("C:\\Users\\Ivy Eva\\Desktop\\Notes\\4.2\\Project- simulation and results")

## name the data as 'Ndata'

Ndata=read.csv("DATANULL.csv")

## access variables present in the data

attach (Ndata)

## compactly display the internal structure of the data

str (Ndata)

summary (Ndata)

## delete incomplete data

na. omit (Ndata)

## install packages needed for analysis

install. packages("tweedie")

install. packages("pscl")

install. packages("vcd")

install. packages("macros")

install. packages ("countreg", repos="http://R-Forge.R-project.org")

## load packages needed for analysis

library(ggplot2)

library(dplyr)

library(class)
```

```
library (MASS)
library(caret)
library(countreg)
library(magrittr)
library(tweedie)
library(statmod)
library(pscl)
library(vcd)
## bar plot to see which categories are more prone to claim
counts<- table (Ndata$Claim_Number)
names <-c ("no", "yes")
barplot (counts, main= "Claim occurence", xlab= "claim number", names.arg = names)
counts<- table (Ndata$Vehmake)
barplot (counts, main = "Vehicle make", xlab = "Vehmake")
counts<- table (Ndata$Decription)
barplot (counts, main = "Cause of claim", xlab = "Description")
x <- df$Amt2
x1<-hist (x, breaks=10, col="blue", xlab="Amt2",
      main="Histogram positively skewed")
xfit<-seq(min(x), max(x), length=40)
yfit<-dnorm (xfit, mean=mean(x), sd=sd(x))
yfit <- yfit*diff (x1$mids [1:2]) *length(x)
lines (xfit, yfit, col="red", lwd=2)
## converting to factor for modelling
```

```
names <- c (1:5,7)

Ndata [, names] <- lapply (Ndata [, names], factor)

str (Ndata)

newdata = subset (Ndata, Claim_Number ==1)

df <- newdata [newdata$Amt2 < quantile (newdata$Amt2, 0.99),]

## defining the variables

Ndata$Amount =as. numeric (Ndata$Amount)

Ndata$Months =as. numeric (Ndata$Months)

Ndata$Claim_Number =as. numeric (Ndata$Claim_Number)

Ndata$Amt2=as.numeric(Ndata$Amt2)

Ndata$numclaims= as. numeric (Ndata$numclaims)

## dividing data into train and test

data_partition <- createDataPartition (df$Amt2, times = 1, p = 0.8, list = FALSE)

str(data_partition)

train <- df[data_partition,]

test <- df[-data_partition,]

## modelling claim severity

model_gamma = glm (Amt2 ~ Decription + Fleet + City + Vehtype +Vehmake
+Months+Claim_Number, data=train, offset, log (Claim_Number),
family=Gamma(link="log"))

lm. good2<-lm (Amt2~ Decription+Fleet+City+Vehtype+Vehmake+Months)

plot (lm. good2, which= 3)

summary(model_gamma)


##modelling claim frequency
```

```
##dispersion test

install. packages("AER")

library (AER)

rd<glm (Claim_Number~Decription+Fleet+City+Vehtype+Vehmake+Amt2+Months, data=train, family = poisson(link="log"))

dispersiontest (rd, trafo=1)

remove (rd)

poissonglm <- glm (Claim_Number ~ Decription + Fleet + City + Vehtype +Vehmake + Months, data=train, family = "poisson", offset= log (Months))

mode(poissonglm)

summary(poissonglm)

##plot graph of poissonglm on data set using rootogram

countreg: rootogram (poissonglm, style = "standing", scale= "sqrt", plot =TRUE, main ="Fitting GLM on data", xlab= "Claim count")

##modelling pure premium

##estimating the value of p

est_p <-tweedie. profile (Amt2 ~ Decription + Fleet + City + Vehtype+Vehmake +Months, data=train, link. power=0, do. smooth = TRUE, do. plot= TRUE, eps = 1/6)

tweedie_model <- glm (Amt2 ~ Decription + Fleet + City + Vehtype +Vehmake +Months, data=train, family = tweedie (var. power=1.8, link. power=0), offset=log (Months))

lm. good3<- lm (Amt2 ~ Decription+Fleet+City+Vehtype+Vehmake +Months)

plot (lm. good3, which= 3)

summary(tweedie_model)
```

REFERENCES

A Conrad, F J Mostert & J M Mostert (2009)- The Underwriting Process of Motor Vehicle Insurance.

AKI insurance industry report (2017)- *Association of Kenya Insurers*.

AKI insurance industry report (2019)- *Association of Kenya Insurers*.

AKI Journal December (2019)- *Association of Kenya Insurers*.

Client Centric Analysis: Turning Big Data into Actionable Intelligence- https://www.aig.com/content/dam/aig/america-canada/us/documents/brochure/aig-cca-overview-final-brochure.pdf

David Pascual-Ezama (2015)- Risk Factor Selection in Auto Mobile Insurance Policies- *A Way to Improve the Bottom Line of Insurance Companies.*

Dunn, P.K., & Smyth, G. K, (2018) - Generalized linear models with examples in R. Springer, New York, NY (Chapter 12)

East Africa Insurance Outlook Report (2019)- *Deloitte*

Esbjorn Ohlsson & Bjorn Johansson (2010)- Non-Life Insurance Pricing with Generalize Linear Models.

Giorgio Alfredo Spedicato, Christophe Dutang & Leonardo Petrini (2018)- Machine Learning Methods to Perform Pricing Optimization: A Comparison with Standard GLM.

Insurance Regulatory Authority (2012)- Guidelines on Insurance Products for Insurance Companies and Intermediaries.

Insurance Regulatory Authority (2013)- Guideline on Valuation of Technical Liabilities for General Insurers.

Insurance Regulatory Authority (2018)- Insurance Industry Annual Report.

Insurance Regulatory Authority (2019)- Insurance Industry Report for the period January.

Insurance Risk Pricing- Tweedie Approach:

https://towardsdatascience.com/insurance-risk-pricing-tweedie-approach-1d71207268fc

Julien Antunes Mendes, Samuel Mahy & Xavier Marecha (2017)- Machine Learning Application to Non-Life Pricing: *Frequency modelling: An Educational Case Study.*

Kevin Kuo & Daniel Lupton (2020)- Towards Explainability of Machine Learning Models in Insurance Pricing.

Modelling Insurance Claim Frequency

https://medium.com/swlh/modeling-insurance-claim-frequency-a776f3bf41dc

Modelling Insurance Claim Severity:

https://medium.com/swlh/modeling-insurance-claim-severity-b449ac426c23

N. Brouhns, M. Denuit, B. Masuy & R. Verrall (2002). Rate Making by Geographical Area *A Case Study Using the Bascov and Verrall Model*.

Oscar Tufvesson, Johan Lindstrom & Erik Lindstrom (2019)- Spatial Statistical Model of Insurance Risk: *A Spatial Epidemiological Approach to Car Insurance.*

Simon Grima, Andre Farrugia (2019)- An Analysis of the Risk Factors Determining Motor Insurance Premium in a Small Island State: The Case of Malta.

Trevor C. Bailey (2001)- Spatial Statistical Models in Health