National University of Science and Technology

Think in Other Terms

RESEARCH TITLE

APPLICATION OF GENERALISED LINEAR MODELS IN PRICING COMPREHENSIVE MOTOR INSURANCE

STUDENT NAME

ZIBUSISO VUSUMUZI MASUKU

STUDENT NUMBER

N01415429X

SUPERVISOR

MR. A. TINARWO

Submitted in partial fulfilment of the requirements of the Bachelor of Commerce Honours Degree in Actuarial Science.

JUNE, 2018

BULAWAYO, ZIMBABWE

DECLARATION

I hereby declare that this submission is my own work towards the partial fulfilment of the requirements of the Bachelor of Commerce Honours Degree in Actuarial Science and that, to the best of my knowledge, it contains no material previously published by another person nor material which had been accepted for the award of any other degree, except where due acknowledgment had been made in the text.

Zibusiso V. Masuku (N01415429X)		
Student	Signature	Date

Certified by:		
Mr. A. Tinarwo		
Supervisor	Signature	Date

ABSTRACT

In non-life insurance, a premium is a tariff paid by the insured to the insurance company in exchange of cover for some unknown future risk. Usually, this is done through the 5% method for Comprehensive Motor Insurance policies in Zimbabwe. This paper attempts to present an overview of the Generalised Linear Models in calculating the pure premium. The model is based in part on Poisson Regression and Gamma Regression using the individual policyholder's characteristics. The dataset was obtained from a local Zimbabwean insurer and the statistical analyses were done using RStudio which utilises the R programming language and also using SPSS.

ACKNOWLEDGEMENTS

I render my sincere gratitude to the Almighty God for divine protection and guidance throughout the course.

I also express gratitude to my mother and siblings who supported me immensely during my academic years.

I express my profound gratitude to my supervisor Mr. A. Tinarwo for his help, time and patience. I also express gratitude to the teaching staff and non-teaching staff of the Insurance and Actuarial Science department at the National University of Science and Technology, with special thanks to Dr T. Chowa who helped instil the appreciation of software packages in Actuarial work and to Mr. D. Chimsitu who took me through the Actuarial Statistics courses from which much of the work in this research is based on.

I would also like to extend similar thanks to my classmates and colleagues for their support and assistance with various concepts presented in this paper.

In addition to classmates and teaching staff, I would like to thank the General Insurance team at African Actuarial Consultants who mentored me in Non-life insurance work, with special mention to Mrs Memory Chasara (Chatendeuka), Ashley Chivanganye, Donald Hove, Tafadzwa Chiduza, Aldrin Chari and Sam Mawoyo.

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

CHAPTER I: INTRODUCTION

1.1. BACKGROUND OF THE STUDY

An insurance policy refers to the contractual agreement between a policyholder and an insurance company. The insurance company agrees to assume the policyholder's risks in exchange for a premium. Until such a time that a fortuitous event occurs, the insurance company will then reimburse, repair or recover the losses accumulated by the policyholder.

Unlike life assurance, where actuaries may make use of mortality tables to arrive at a premium, setting a premium for a non-life insurance product is even more complicated. The insurance principle of indemnity guarantees that, should a loss occur, the insured can only be indemnified just enough to get him/her to where he was before the loss. In essence, losing a car tyre to a theft, means the insurer will only reimburse the tyre and nothing more.

Essentially, one cannot profit from an insurance contract. This principle brings more challenges in pricing general insurance in that since the probability that the insured will have a fortuitous event occurring to him/her during the policy term is unknown, so is the value of the claim should such an event occur.

This presents considerable difficulty in pricing the product. Most actuaries agree with Ohlsson and Johansson (2010) when they say that, "…the generally applied principle *is* that the premium should be based on the expected (average) loss that is transferred from the policyholder to the insurer." As such, a risky client should in turn pay more for cover. The vice versa is also true.

Coming closer home, Zimbabwe suffered major hyperinflation in 2007 – 2008 era. A lot of insurance companies' savings and assets were eroded. When the economy was rebooted by the assumption of the US dollar currency, insurance firms had to restart their savings as well.

A number of them tried to undercut competition with outrageously low premiums, until the Insurance and Pensions Commission (IPEC) had to step in to restore sanity and confidence to the industry, which is their government given prerogative (IPEC Act 24:21).

The rate for Road Traffic Act Third Party insurance was then set at $35.60 and 5% was proposed for Comprehensive Motor insurance. A number of Zimbabweans end up opting for the RTA option since Comprehensive Motor insurance generally becomes expensive as the value of the car goes up. In addition, this arrangement ignores the existence of other significant factors in pricing comprehensive insurance.

1.2. STATEMENT OF THE PROBLEM

Since a fair premium is the expected cost of claims arising from the exposure to the policyholder's risk, it is intuitive to have that premium varying with differing levels of risk. The use of the 5% method does not constitute a fair premium to the consumer. For instance, a youthful and newly licensed driver driving a $40,000 Mazda RX-8 sports car will pay the same amount of premium as a mature and more experienced female company secretary who drives, say, a similar priced Toyota Hilux. This pricing arrangement totally ignores:

- That the woman is probably less likely to be involved in an accident as compared the newly licensed driver.
- That the sports car driver has a greater propensity to speed since his vehicle is faster.
- That sports cars are generally more difficult to repair than the more commonplace Toyota Hilux.

1.3. RESEARCH QUESTIONS

- What information do insurers require when underwriting motor insurance policies?
- What is the significance of each factor of the collected information in helping to estimate claims?
- How are the factors used correlated?
- Between the GLMs method and the 5% method, which one is financially viable?

1.4. OBJECTIVES

1.4.1. PRIMARY OBJECTIVE

To determine a premium for Comprehensive Motor insurance using GLM techniques. In order to achieve the primary objective of the study, the following secondary objectives of the research were identified:

1.4.2. SECONDARY OBJECTIVES

- To analyse the significance of other factors, other than vehicle value, in pricing comprehensive motor insurance policies.
- To test for correlation among the collected factors.
- To identify a method which yields lower loss ratios.

1.4.3. ASSUMPTIONS

This research assumes the following:

- The Generalised Linear Model (GLM) assigns full credibility to the data. Results obtained will be regarded as valid irrespective of how large or thin the datasets were.
- GLM assumes that the random component of outcomes has no correlation

In addition to the above, performed calculations also assume that inflation is fixed at 0% of the past 4 years.

1.5. JUSTIFICATION OF THE STUDY

This study aimed at directly benefiting the insurers of motor vehicles and the vehicle owners. Insurers need to profitably insure clients, and their clients only want value for money. This study tried to converge the two ideals, in that GLMs yield actuarially fair premiums and hence result in value for money for clients. Similarly, if the insurers could reduce adverse selection and only accept premiums proportionate to the risk, the insurers would make a profit.

In addition, this study explores the merits of adopting the use of GLMs amongst insurers as a model that has diverse and effective applications in motor insurance. Future innovations in insurance involve regression modelling and applications of this model will help do away with the rudimentary pricing system currently used in Zimbabwe.

1.6. SCOPE OF THE STUDY

This research followed an analytical approach to improving the current pricing methods used by most insurers in Zimbabwe today with respect to comprehensive motor insurance. However, the analysis zoned into one Zimbabwean insurer. The choice for one insurer was actuated by the recognition that different insurers offer varied insurance policies. Although richer conclusions could have been inferred from a larger dataset that combines many insurers'

experience, the biases introduced when attempting to standardize various insurance offerings would have inevitably reduced the credibility of the final results. The data used spanned from 2014 to 2017.

1.7. LIMITATIONS

This research had the following limitations:

1. Data used was obtained from one company.
2. Conclusions of this study are representative and useful for the insurance company business, but they do not present a generalised character, therefore they cannot be applied to all portfolio or insurance companies.
3. The data used was not obtained through a random selection related to the entire population of policyholders.

1.8. ORGANISATION OF THE RESEARCH

The first chapter of this research study is the introduction, which comprises the background of the study, problem statement, and objectives of the study, as well as the scope, limitations and justification of the study. The second chapter includes the literature review, which is, scholarly work done by other people on the topic. Discussion will be made around the GLM techniques, along with their merits. Other pricing techniques will also be discussed including the issues surrounding their application. Consideration will be made with regards to what has to be appreciated when deciding to develop models based on GLMs. The third chapter details the methodologies employed in the study. This research made use of secondary data, and this section helps clarify what kind of data was used and what level of cleaning was applied to ready it for modeling. This section will also outline the models to be used and any software packages on which these analyses were done. The fourth chapter deals with data analysis and presentation. In this section, all preliminary work before the model was developed is discussed. From the splits all the way down to the correlation tests. The fifth and final chapter concludes the study. Under this section, the researcher discusses the findings from the data, and also attempts to answer the questions presented in Section 1.3 of Chapter 1 on Research Questions. It goes to on offer recommendations and any ideas and suggestions for future study related to this topic.

CHAPTER II: LITERATURE REVIEW

This chapter mainly focuses on discussing and synthesizing the available knowledge with respect to pricing Comprehensive Motor insurance and application of Generalised Linear Models. However, in Zimbabwe the majority of insurance companies use the percentage-based method. This was proposed by the regulator, the Insurance and Pensions Commission of Zimbabwe (IPEC), in order to restore sanity to the pricing regimes and also bolster public confidence in the insurance industry.

Silva and Afonso (2015) cover a number of pricing methods in their comparative study conducted in Brazil. They presented arguments against and in favour of the methods under varying circumstances. Starting with the Zimbabwean setting, it is worth noting that the current percentage method presents considerable advantages to a very young insurance market like that of Zimbabwe. The dollarization era essentially cut off the amount of data that actuaries could consult in terms of pricing Comprehensive Motor insurance. Chidakwa and Munhupedzi (2017) did a study on the Impact of Dollarization on the Zimbabwean economy where they mentioned the presence of excessive inflation that characterized the era preceding 2008. Even at the end of these and other studies, researchers have failed to settle on a definitive USD to ZWD rate prevailing at the time of dollarization. As such, the actuarial data available for exploitation is only from 2009 to date. However, since Zimbabweans generally shy away from Comprehensive Motor insurance, such privileges only exist for large firms such as NicozDiamond and Old Mutual. Smaller companies such as Quality Insurance, Hamilton, and so on, would remain with too thin datasets to create experience-based models.

The percentage-based method has the advantage of being easy to apply. Since the vehicle book value can be easily ascertained by use of evaluators and consulting the open market, one can apply this percentage without challenges. It also helped level the playing field as fewer insurers got any comparative advantage over others since for each vehicle book value, almost each and every insurer had the same value to charge as premium.

In the same vein, this has stifled consumer choice, since it essentially made all insurers the same, without any charging a fairer price. This might be the cause for the low uptake of comprehensive motor insurance. However, we cannot be quick to blame the local regulator for such a move. Studies around Africa and the globe do show that the least economically developed countries generally apply the same method. Nigeria, Rwanda and Kenya charge

around 3.5% to 4.5% of the value of the car as the insurance premium for an annual cover of comprehensive motor insurance.

As long as compulsory Third-Party insurance exists in Zimbabwe, there is little incentive for insurers to pick on more complex rating based, GLM-based or usage-based insurance pricing models. The average insurer in Zimbabwe still enjoys considerable profits from writing Third Party motor insurance. Usually, adverse selection exists on the comprehensive motor insurance front due to the uniform pricing system. Unless signed up for by a company, the individual would usually sign up for comprehensive insurance if their car is expensive, or they consider themselves riskier.

Setting aside local and regional issues, the USA issued the first ever motor insurance policy in 1897 which was written by Traveler's Insurance Company (Kane et al, 2006). As Zimbabwe we could take a leaf out of these more economically developed countries' rulebooks to help with pricing local insurance. The UK introduced their compulsory motor insurance in the 1930s. They later introduced rating systems to help determine the premium. Coutts (1984) explains that the UK and the Netherlands already had rating systems which were consulted and remodeled time and again, however, the emerging models, such as Generalised Linear Models were still in their infancy. They were purely statistical applications which did not provide much benefit as they did not cover practical aspects of premium rating.

Before Generalised Linear Models were more prevalent, rates were used to determine insurance premiums. The whole science, or rather, the art of rating remains an interesting subject until this day. The works of Bailey and Simon (1960) presented two risk classifications. The first was a Class Rating, which simply grouped risks by their characteristics, such as; gender, vehicle use and model, and so on. The other was a Merit Based system which rewarded clients based on past driving behaviour, felonies and misdemeanours committed and so on. Bailey and Simon's work was the first major study to explore the relationship between these two classifications. In modern day pricing, these two systems have merged, with the creation of Bonus-Malus systems which we shall consider in a few sections.

Coutts (1984) conducted a study on a 'points system'. Although based on UK data, he managed to explain that the points system, which is a variation of the standard relativity rating system had a lot of merits and also demerits. However, from the study, he notes that major UK insurers at the time preferred the method for its simplicity and its ease of manipulation. Which is also

the case with the preference of the percentage-based method in Zimbabwe. Ultimately, he called for a detailed breakdown of insurance data, as this allows for an in-depth analysis of claim experience which in turn require statistical modelling, which assists in understanding the underlying structure. This idea reinforces the need for insurers to collect as much information as they can to help price insurance risks.

An Australian study presented to a 2009 Conference Meeting discussed major rating factors. In addition to identifying factors around the driver, vehicle and location as major factors that influence the rating, Henwood and Wang (2009) outlined the existence of interactions between these major factors. They further corroborated Huang and Query (2007, pp.79-83) whose study was focusing on how China could overhaul and rebuild its motor insurance sector. Building on ideas brought forth by Coutts (1984), we note that these analyses are made possible by vast collections of data. Zimbabwean insurers have barely started to collate such data as evidenced by most insurers' insistence on basic data that appears on the Vehicle Registration booklet for underwriting motor insurance. In the UK, they consider the driver as the major and most significant component in pricing motor insurance. The rationale is; in and of itself a car has a degree of causing a loss, either by accident, fire or theft, but only up to a point. For instance, common cars such as the Honda Fit in Zimbabwe are prone to being stolen. In actual usage, the driver is the most significant determinant when considering the occurrence of an accident. However, locally we find few and fewer insurers who collect driver-based information such as occupation, driver's experience and so on. This idea does not tally with Henwood and Wang's findings of the risks brought forward by the driver, before details about the vehicle and location are considered.

Given that some of the most important factors cannot be easily ascertained in the initial underwriting process, most insurers have introduced the Bonus-Malus coefficient. Lemaire (1995) concluded that rates no longer hold if they disregard important factors, whose considerable importance is acknowledged by common sense and experience: individual driving abilities such as accuracy of judgment, swiftness of reflexes, aggressiveness behind the wheel, knowledge of the highway code, and drinking behaviour, are not taken into account in motor insurance rating, *a priori*, as these variables are impossible to measure in a cost-efficient way. It is now possible through the use of on-vehicle devices that employ the technologies of telematics. Back then it would have been difficult to so without excessive costs. Although part of the globalized village, Zimbabwe tends to lend itself to the 1995 environment as mentioned

by Lemaire. The internet is still a fairly pricey commodity, this means, the use of telematics is still considered a dream by many insurers as its application remains costly.

An example is brought forward with two teenage females, driving the same vehicle model in the same city. The two may exhibit very different accident patterns, due to differences in individual behaviour hence the idea of trying to account for these differences *a posteriori*, by adjusting the premium from individual claims experience. The concept of 'Bonus-Malus' is essentially a Latin term which literally means 'good-bad' or more practically to 'reward or penalise'. Slowly, the ideas we have examined up until this point tend toward a premium that is also proportionate to risk. The idea of rating and Bonus-Malus systems are all concepts meant to align the policyholder's premium to the amount of risk they also present. In South Africa, a similar albeit slightly different version exists. On an article on *FA News*, Gari Dombo, the then Managing Director of Alexander Forbes Insurance, explains that, "The Cash Back or No Claim Bonus (NCB) concepts are based on the same principle and *are* intended to deliver the same objective: to reward loyal clients who do not claim or claim infrequently. The short-term insurance industry has historically recognised the need to reward clients for every year they do not claim, while penalising clients that do submit claims. The idea is to encourage better risk management, as well as to discourage clients from submitting nuisance claims." However, from the outset, one can see a problem that this presents. The clients will tend to cover claims themselves in order to enjoy the bonus or prevent an increase in future premiums. While this poses an advantage for the insurer, the insured does not enjoy value for money if they engage in such practices. These practices have become selling points with major insurers such as OUTsurance, Budget and Dial Direct.

Introduced into modern insurance by British actuaries from the City University, John Nelder and Robert Wedderburn in 1972, and later illustrated by McCullagh and Nedler (1989), the Generalised Linear Model (GLM) has been one of the most popular tools used to rate motor insurance for decades. It is a highly valid methodology for vehicle insurance ratemaking and can easily handle a large number of risk combinations being examined and establish complex relationships related to claim experiences (Huang and Query, 2007). Nelder and Verrall (1997) showed how credibility theory can be encompassed within the theory of GLMs. In that vein Schmitter (2004) gave a simple method to estimate the number of claims needed for a GLM tariff calculation.

Furthermore, GLMs can be considered in two variants, the first is an Additive Model, which involves adding the covariates. The other one is a Multiplicative model. As outlined by Ohlsson and Johansson (2010), Goldburd, et. al. (2016) and Huang and Query (2007), the additive model has weaker actuarial applications since it may give spurious results. With sufficient values, one can easily get negative premium values and claim estimates. Besides, it sounds logically insensible to charge, say a $20 penalty for an individual paying $100 and the same amount for an individual paying a $1,000 premium. The multiplicative model would recommend we charge 20% for both individuals, which sounds actuarially fair as 20% for $100 and $1,000 is $5 and $50 respectively. The multiplicative model is usually the most used in actuarial applications and it is also easier to develop rating model from it since rating models also employ multiplicative relativities.

Also, of particular importance to premium rating are the Tweedie Models, named after a British statistician who presented a thorough study of the concept in 1984. The Tweedie models are best suited for premium estimation since most of their mass is close to zero and the remaining mass skewed to the right (Goldburd et al, 2016). Which is essentially how most claim distributions would appear when depicted graphically.

Insurers differentiate between different groups of risks using 'risk factors' such as gender or age. A few problems develop when one wants to implement Generalised Linear Models in comprehensive insurance pricing, especially in the modern day global context. In a US based study, Williams and Shabanova (2003) concluded that young males were more likely than young females to be responsible for crash deaths, whereas females in their 50s and older were more likely than same-age males to be responsible. In terms of responsibility for deaths per licensed driver, young drivers, especially males, had the highest rates because of their high involvement rates and high responsibility rates. This clearly shows a relationship between gender/age in claims settlement. If those two are factors that cause crashes, they surely have to be considered in pricing. However, in 2012, the European Court of Justice (ECJ) banned all private European Union (EU) insurers on discriminating risks on gender grounds. Although not yet enforced in Zimbabwe, advocates of human rights and the speeding pace of globalization will soon require that we implement a similar ban as well. By gathering more detailed information, insurers are guaranteed that such a move will cause little to no effect on their pricing methodologies.

As mentioned beforehand, Silva and Afonso (2015) made mention of other models that could be used in pricing motor insurance premiums. They considered the following methods:

- Pure Premium by historically aggregated claims.
- Pure Premium by expected aggregated claims.
- Classic Linear Models
- Generalized Linear Models

The idea of their study was to compare the increase in relation to original pricing and dispersion of pure premiums. To fully explain their ideas, a variable termed *Efficiency* was created, by means of which the increase in price is calculated so that there is a reduction of one percentage point compared to the actual price in their dataset. After conducting their study, they concluded that, "…the modelling techniques of historical pure premium and GLM had the best pricing efficiency (1.60), but the former showed a lower variation than the latter (standard error 5.26% against 17.52%). The comparative advantage of the historical pure premium can be explained by the low variation of the basic indicators over time and the fact that it contains the data for all insured vehicles in Brazil, enabling high adherence to the technique. However, GLM appears to be an interesting pricing alternative for medium portfolios that are seeing growth, or that explore possible niche markets." (Silva and Afonso, 2015)

Zimbabwean insurers could easily apply the first method, since almost all insurers have some form of historical claims data. For the second method which uses expected claims, this would be a challenge, because for some reason insurers in Zimbabwe also use the 5% method in reserving. In this method, set out by the IPEC, a company can set aside reserves of 5% of their Net Written Premiums as Claim Reserves with respect to Incurred but Not Reported (IBNR) claims. Working backwards leads to a fixed value of a premium as well. As mentioned by Silva and Afonso (2015), higher variability is also associated with this method. Making it less efficient than the two methods cited above. At the end of the day, one is better served with applying GLMs as they are more efficient.

Another model that Huang and Query (2007) recommended is combining a Max Model alongside a GLM to improve accuracy in cases where the data has correlated risk factors. Usually, the GLM handles fairly correlated factors quite well, it only struggles to fit values where the factors are highly correlated. Huang and Query (2007) noted that the practice of selecting the most significant and least correlated factors did not work well in China since it

leaves very few factors to work with. They went on to suggest the use of the correlated factors as they narrow the gap amongst relativities.

Policyholders that give incorrect information do not receive a substantial discount from the rate system, and there is enough redundancy to correct it. For example, actuaries in China used the book value of the car, engine capacity, manufacturer or type, mileage, etc. to estimate the risk of the vehicle even though all these factors are all corelated. This does not satisfy the assumptions of GLM. Other factors increase the complexity of distinguishing risks. "For example, consider the case of an experienced driver with a defective car or vice versa. Often major problems can be traced to one specific factor, while the other factors in the model are considered accurate and reliable." (Huang and Query, 2007)

Considering the above issues, they postulated the use of a Max model which solves the problems of correlated variables. It assumes that for the correlated risk factors K_i (i = l, 2, 3, … N), a variable MAX (K_i) is used to improve the accuracy of the model.

More models may be developed from GLMs, such as those mentioned by Goldburd, et. al. (2016). These include:

- Generalised Linear Mixed Models (GLMMs)
- GLMs with Dispersion Modeling (DGLMs)
- Generalised Additive Models (GAMs)
- Multivariate Adaptive Regression Splines (MARS) Models
- Elastic Net GLMs

Although, not further described in this monograph, these models all have their basic techniques built upon GLMs.

Coming to the modern era, one will notice the level of undoubtable influence that GLMs have had on the pricing of motor insurance. Halliday (2015) mentioned GLMs as a basis for more advanced techniques. The emergence of telematics, which is the use of vehicle sensors and positioning systems to monitor vehicle and/or driver behaviour, have enabled insurers re-design their insurance products on the fly based on the data received (Liu et al, 2017). Halliday (2015) demonstrated that even minute details such as braking behaviour, overtaking styles could be used to alter one's premium based on the insurer's perception of his riskiness.

The emergence of Artificial Intelligence has increased the amount of data that insurers can gather from their clients. Already in neighbouring South Africa, companies like Discovery are already employing the use of telematics in vehicle-based insurance. Before we explore the technicalities of these concepts, it is imperative to explore the following definition.

"Artificial intelligence (AI) is intelligence demonstrated by machines, in contrast to the natural intelligence displayed by humans and other animals. In Computer Science, AI research is defined as the study of *intelligent agents*: any device that perceives its environment and takes actions that maximize its chance of successfully achieving its goals." (Poole, et. al., 1998)

Lately, the emergence of all sorts of data collecting devices, colloquially known as the Internet of Things (IoT) and the improvements in technology have allowed actuaries to apply predictive analytics, a privilege that comes with harvesting Big Data. This means the average consumer can finally have premium being determined by individual factors unique only to that individual.

Studies by Lui et al. (2017), Han et al. (2016) and Halliday (2015) have demonstrated ideas and concepts that are already being tested in developed countries on a broader scale. Lui et al. (2017) define a model that considers a driving behaviour model that is based on Usage-based Insurance (UBI). They had built on earlier work by Han et al. (2016) who introduced the concepts of adjusting a GLM estimated static premium by combining it with a dynamic premium that varied with the Telematics data collected from the insureds' vehicles on the fly.

Zimbabwe's best version of usage-based insurance is Champions Insurance Company's offering dubbed 'pay as you drive' insurance. However, this does not compare widely with global offerings in that it is only based on mileage. It has to be mentioned that even the most complicated methods such as the ones that utilise telematics and AI all have at their root, either GLMs or Classical Linear Models.

In conclusion, while the world is already far off in terms of how they apply technology within the insurance sector, Zimbabwean insurers have a lot to gain by shifting to statistical methods in pricing and reserving as compared to the deterministic methods they are currently using. In addition, this means the local insurance industry will be a step closer to catching up with the world.

CHAPTER III: RESEARCH METHODOLOGY

The methodology section of this research paper aims to present pertinent issues related to the GLMs and the usefulness of these models in non-life insurance business. Before the technicalities of GLMs are considered, focus will be drawn to steps taken to prepare the data before the development of an appropriate model.

3.1. DATA PREPARATION

Data preparation is an integral part of the model building process, and usually takes up much time. Since each organisation has varying processes and systems for collating the data. The actuary has to be well versed with the common themes and situations. In each modeling process, the data preparation step is repeated as correcting one error helps discover another.

3.1.1. COMBINING POLICY AND CLAIM DATA

Usually, the data most appropriate for use in building a rating plan is exposure-level premium (policy demographic) and loss (claim) data. The immediate problem with assembling such a dataset is that exposure and claims data tend not to be stored in the same place. In many organizations, a policy-level exposure database is housed within the underwriting area and a claims database is housed within the claims handling area. So, the first task of a modeling project is often to locate these two datasets and merge them.

If best practices are followed, merging these two datasets would not be time-consuming. To achieve this, the Excel `VLOOKUP()` function could be used. Database specific methods are available for organisations that use other storage methods. Using the unique IDs, which were policy numbers in this case, the two sets were then merged.

3.1.2. MODIFYING AND CLEANING THE DATA

Any dataset is likely to have errors. While not easy to present a one-size-fits-all formula for error detection, human judgment can be applied to look for outliers and so on. This could be done in the following steps:

A. Check for Duplicates

In exposure data, one checks for duplicate policy numbers. In claims, the claim and/or policy numbers are used to check for uniqueness. Such checks should be done prior to merging the data so that errors are not imported into the new dataset.

B. Reasonability Checks

In addition to duplicates, numeral values are checked for reasonability. For instance, we do not expect negative ages in our datasets.

C. Check for Missing Values

This step involves checking for missing values and deciding how they shall be imputed. While a lot of subjectivity is involved here, the usual way to add a mean or a modal value. In other cases, where errors are minimal, it could be easier to delete the entries altogether. Where deletion is not possible, probably because it leaves a very thin dataset to work with, a new column could be added to reflect error flags and data corrected accordingly.

D. Splitting the Data

When modeling, it is important to hold two datasets. One is termed the training set and it is the one that is used to help build the model. The other one is the holdout or test set, this one is used to test the effectiveness of the model created with the training set. In this study these two were determined by policy years. However, a rule of thumb is to use the 70:30 ratio, 70% for the training set and 30% for the test set.

3.2. GENERALISED LINEAR MODELS

Generalised linear models (GLMs) make up a rich class of statistical methods, which generalises the Ordinary Linear Models (OLMs). The general Ordinary Least Squares Regression Model takes the following form.

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in} + \varepsilon_i$$

Where y_i is the response variable. α represents the intercept value and/or the mean value can be taken to also mean the $\beta_0 x_{i0}$ term. β_n denotes coefficients to the n covariates and x_{in} refers to the observed values for each covariate and ε_i is the random component, which is Gaussian.

The GLM, takes a slightly similar form, however, with two distinct changes.

- **Probability distribution**

Instead of following a normal distribution, GLMs work with a variety of distributions – specifically, distributions referred to as the Tweedie models, which shall be examined shortly.

- **Model for the mean.**

In linear models the mean is a linear function of the covariates x. In GLMs, some transformation of the mean is a linear function of the x's, with the linear and multiplicative models as special cases (Ohlsson and Johansson, 2010). Usually, this monotone transformation is the natural logarithm function.

GLMs can be used to predict the following variables:

- A binary variable, say, whether or not a claim occurs or not. In this case, the binary value takes either 0 or 1 for which it applies the binomial regression models (logit, probit and log-log complementary models);

- A count variable, say, the model of the frequency of claims, in which case, a Poisson regression model is applied.

- A real positive value, say, the model for the severity of a claim, in which case Gamma and Inverse Normal regression models can be applied.

3.2.1. EXPONENTIAL DISPERSION MODELS

The GLM makes use of Exponential Dispersion Models (EDM) to generalise the Gaussian distribution used in the OLMs. The probability density (mass) function of a distribution in the Exponential Dispersion family, takes the following general form.

$$f(y_i|\theta_i, \varphi) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\varphi} + c(y_i, \varphi\right), \qquad y_i \epsilon S$$

Tweedie models are actually a family of EDMs that demonstrate scale invariance. These have their variance function, denoted as $v(u) = \mu^p$ for some value, p. In specific values of p, these Tweedie models reduce to distributions tabulated overleaf.

	Type	Name	Measure
$p < 0$	Continuous	-	-
$p = 0$	Continuous	Normal	-
$0 < p < 1$	Non-existing	-	-
$p = 1$	Discrete	Poisson	Claim Frequency
$1 < p < 2$	Mixed, non-negative	Compound Poisson	Pure Premium
$p = 2$	Continuous, positive	Gamma	Claim Severity
$2 < p < 3$	Continuous, positive	-	Claim Severity
$p = 3$	Continuous, positive	Inverse Normal	Claim Severity
$p > 3$	Continuous, positive	-	Claim Severity

Fig. 3.1: Overview of Tweedie Models

As per table, this research paper made use of the Poisson Regression to model the occurrence of claims per exposure level and also utilise the Gamma Regression to model the claim severity per exposure.

3.2.2. ESTIMATING CLAIM FREQUENCY

The next step is to utilize a Poisson model to estimate the 'event count' of claims. David (2015) adds that the Poisson model is the main tool for the modelling claim frequency in non-life insurance. Instead of going over the iterative methods used to estimate values of the estimates by hand, the researcher has opted to use the RStudio software package to calculate the necessary coefficients using the underlying R programming language. The syntax of the code is appended at the end of this dissertation. However, due to the nature of the dataset involved, the researcher employed the use of the Over dispersed Poisson (ODP) model.

As Ohlsson and Johansson (2010) stated, "Random variation between customers and insured objects, and effects of explanatory variables that are not included in the model, lead to

overdispersion: the variance of the observations within a tariff cell is larger than the variance of a Poisson distribution."

To overcome this, the researcher called upon a GLM function in R with the parameter `family = "quasipoisson"` instead of the usual `family = "poisson"`

The ODP is similar to the usual Poisson distribution, with the exception of the Φ, the dispersion parameter. In the ODP, Φ can take any value other than 1 given in the 'true' Poisson

3.2.3. ESTIMATING CLAIMS COSTS

In actuarial terminology, Claim Cost refers to Claim Severity. Basing on the Fig. 3.1, the researcher utilised the Gamma Regression model to model the claim amounts per each exposure level. The Gamma distribution is right skewed with a sharp peak and a long tail to the right. It has a lower bound at zero. These characteristics make this distribution a natural fit for Claim Severity. Given the Tweedie Model table (Fig. 3.1), the variance function of a Gamma Distribution is $v(u) = \mu^2$.

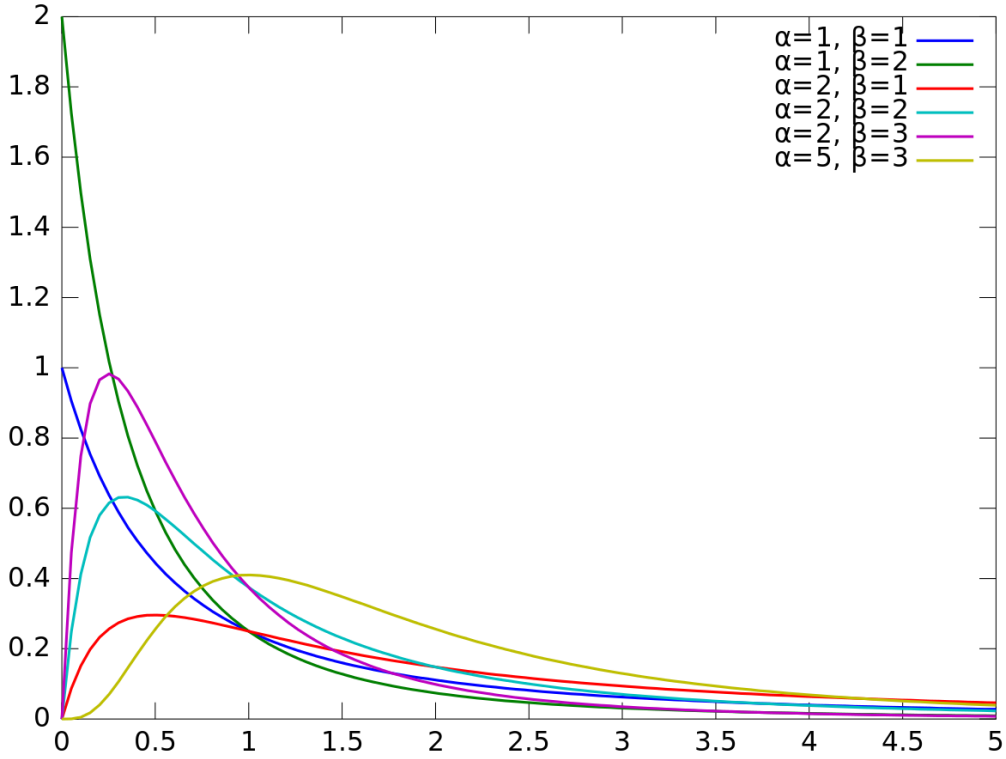

Fig. 3.2: The Gamma Distribution

3.2.4. ESTIMATING PURE PREMIUM

Under this facet, the pure premium can be estimated via two methods. One could model frequency then model severity and combine the two together. One could also apply the Compound Poisson Model to come up with one model that can estimate the pure premium directly. The researcher chose to use separate models in order to demonstrate the various factors that are at play when estimating either claim frequencies or claim severity. As shown by Charpentier and Denuit (2005), the separate approach of frequency and cost of claims is particularly relevant because the risk factors, which influence the two components of the pure premium, are usually different.

Essentially, the separate analysis of the two phenomena provides a clearer perspective on how the risk factors are influencing the premium.

A fair pure premium reflects the expected value of claims. Loading for expenses and claim handling costs, this value is transformed to represent the actual tariff per exposure level.

Given that the first model gives us the expected value of claim frequency and the second model provides us with claims severity, the combination yields the expected claim severity per exposure level.

$$Pure\ Premium = E[Claim\ Frequency]\ \times E[Claim\ Costs]$$

CHAPTER IV: DATA PRESENTATION AND ANALYSIS

4.1. EXPLORING THE DATA

A dataset containing 10,302 observations was used. The data related to an insurer's experience over a period of 4 years (2014 – 2017). The dataset used contained actuarial data for Comprehensive Motor insurance policyholders, including indications of whether each policyholder had been involved in a car accident and, if so, the value of the claims settled in respect of accidents.

There were 10,302 rows with behavioural and demographic information about each policy holder. For each insured, we retained 10 attributes that could potentially be used as predictor variables. We added two response response variables, that is, `claim_freq` and `claim_sev`.

- `claim_freq`: Indicates if a client has been involved in a car accident in the past.
- `claim_sev`: Denotes amount that the insurer settled with respect to claims lodged.

The objective of this research is to predict the pure premium per policy. As mentioned in Chapter 3, of this paper, this was arrived at by first estimating the claim counts and then the claim amounts separately.

The original data set had 5 numeric predictors and two numeric response variables. A summary of the data is given in the table below:

Variable	Min	First Quartile	Median	Mean	Third Quartile	Max
age	17	39	45	44.81	51	81
daily_mileage	8	35.4	53.1	53.78	70.8	228.5
car_value	1500	9200	14400	15659.9233	20890	69740
policy_length	1	1	4	5.33	7	25
car_age	0	1	8	7.78	12	28
claim_sev	0.00	0.00	0.00	4033.98	4647.50	57037.00
claim_freq	0	0	0	0.80	2	5

Fig. 4.1: Summary of Numeric Variables.

The original data set also 5 categoric predictors. 4 of them were binary and only one was multi-categorical. A summary of the data is given in the table below:

marital	sex	car_use	location	education
Not Married	Male	Private	Urban	High School
Married	Female	Commercial	Rural	Undergrad
				Postgrad

Fig. 4.2: Summary of Categorial Variables.

To be able to model the non-numeric variables in the statistical software, the categorical variables were converted to the following values:

Sex	Code
Male	1
Female	2
Marital Status	**Code**
Not Married	1
Married	2
Education	**Code**
Postgraduate	1
Undergraduate	2
High School	3
Car Use	**Code**
Private	1
Commercial	2
Location	**Code**
Urban	1
Rural	2

Fig. 4.3: Summary of Conversion Codes

4.2. CORRELATION AMONG PREDICTORS

As mentioned by Goldburd et. al. (2017), the predictors going into a GLM *may* exhibit correlation among them. Where such correlation is moderate, the GLM can handle that just fine. In fact, determining accurate estimates of relativities in the presence of correlated rating variables is a primary strength of GLMs versus univariate analyses; unlike univariate methods, the GLM will be able to sort out each variable's unique effect on the outcome, as distinct from

the effect of any other variable that may correlate with it, thereby ensuring that no information is double-counted. Fig. 4.4 below shows the correlation matrix between the variables that were used.

4.2.1. CORRELATION

Correlations

	age	marital	sex	daily_mileage	car_use	car_value	policy_length	car_age	location	claim_sev	claim_freq	education
age	1	.088**	-.070**	0.000	-.025*	.168**	-0.004	.163**	-.042**	-.034**	-.038**	-.243**
marital	.088**	1	0.001	0.006	-0.012	-0.007	-0.007	-.030**	0.007	-.045**	-.071**	.041**
sex	-.070**	0.001	1	0.011	-.282**	-.062**	0.007	-.021*	.046**	0.004	-0.014	.050**
daily_mileage	0.000	0.006	0.011	1	0.011	-.023*	-0.013	-.027**	.167**	-0.014	0.008	.052**
car_use	-.025*	-0.012	-.282**	0.011	1	.227**	0.004	-.052**	.021*	.032**	.078**	.095**
car_value	.168**	-0.007	-.062**	-.023*	.227**	1	0.001	.178**	-.087**	-.031**	-.042**	-.272**
policy_length	-0.004	-0.007	0.007	-0.013	0.004	0.001	1	0.010	-0.007	-0.016	-0.017	-0.003
car_age	.163**	-.030**	-.021*	-.027**	-.052**	.178**	0.010	1	-.153**	-0.016	-0.017	-.655**
location	-.042**	0.007	.046**	.167**	.021*	-.087**	-0.007	-.153**	1	-.154**	-.243**	.232**
claim_sev	-.034**	-.045**	0.004	-0.014	.032**	-.031**	-0.016	-0.016	-.154**	1	.494**	.022*
claim_freq	-.038**	-.071**	-0.014	0.008	.078**	-.042**	-0.017	-0.017	-.243**	.494**	1	.026**
education	-.243**	.041**	.050**	.052**	.095**	-.272**	-0.003	-.655**	.232**	.022*	.026**	1

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Fig 4.4: Correlation Matrix Between Variables

As shown by this SPSS extract, the correlations coefficients between the variables are minimal. No values were outside the range [-0.5, 0.5]. This gave more confidence in using the data for actual modelling as no spurious results would be obtained with such minimal correlation.

4.2.2. MULTI-COLLINEARITY

According to Farrar and Glauber (1967), multicollinearity is an interdependency condition. It is defined in terms of a lack of independence, signified here by high inter-correlations within a set of variables. There are two types of multicollinearity:

- **Structural multicollinearity** is a mathematical artefact caused by creating new predictors from other predictors — such as, creating the predictor x^2 from the predictor x.
- **Data-based multicollinearity**, on the other hand, is a result of a poorly designed experiment, reliance on purely observational data, or the inability to manipulate the system on which the data are collected.

In our case, Data-based multi-collinearity is more likely. To overcome this, the researcher employed the use of another statistic, called the variance inflation factor (VIF). The VIF for any predictor is a measure of how much the (squared) standard error for the predictor is increased due to the presence of collinearity with other predictors. It is determined by running a linear model for each of the predictors using all the other predictors as inputs and measuring the predictive power of those models. Appendix A shows the VIF values for each of the 10 explanatory variables.

After running the iterative VIF statistic calculation, all variables demonstrated low VIF values with each other. This is an indicator that the data has no variable that is linearly or otherwise, correlated with another variable. As such, the researcher was confident to use all the variables in the GLM model that was then developed.

4.3. TEST AND TRAIN DATA

Before the model could be developed, the data was split into two datasets.

1. Train – This consisted of 8,161 observations from 2014 - 2016, that were used to help with the model building.
2. Test – This consisted of the remainder 2,141 observations from 2017, that were used to check the model's fit and usefulness.

Factor Name (as per model)	Description
age	This defines the age of the policy holder
car_age	This define the age of the car
car_use	This defines the purpose of the car. This is a binary value with only two options: Private or Commercial
car_val	This is the value of the car based on the market value of the car at the inception of the current policy year
daily_mileage	The average distance travelled per day.
ed	The highest level of education attained by the client.
loc	Denotes clients' primary location, between Urban and Rural
marital	Marital Status of the client
policy_length	The length in years the client has held a policy with the insurer.
sex	Denotes if the gender is male or female,

Fig. 4.5: Description of Factors Used

CHAPTER V: RESEARCH FINDINGS, RECOMMENDATIONS AND CONCLUSION

5.1. RESEARCH FINDINGS

In this section, the results generated by the model are presented and interpretation of the models, based on which the insurance pure premium is determined.

5.1.1. POISSON MODEL

The Poisson Model attempts to estimate the expected claim frequency. This model was built using the train dataset, which is based on the 2014 - 2016 observations. The RStudio software package was used to analyse the data, and the **`glm()`** function in the MASS package was called to analyse the data. Initially, the model was run using a 'true' Poisson and the following results were obtained.

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.101e+00  1.286e-01   8.563  < 2e-16 ***
age           -6.797e-04  1.477e-03  -0.460  0.64532
marital       -2.035e-01  2.510e-02  -8.107 5.18e-16 ***
sex            6.364e-02  2.596e-02   2.451  0.01425 *
ed             8.559e-02  2.073e-02   4.129 3.64e-05 ***
daily_mileage  2.638e-03  4.982e-04   5.295 1.19e-07 ***
use            2.896e-01  2.715e-02  10.666  < 2e-16 ***
car_val     -1.132e-05  1.617e-06  -6.997 2.61e-12 ***
policy_length -8.686e-03  3.035e-03  -2.862  0.00421 **
car_age        5.914e-04  2.810e-03   0.210  0.83331
loc           -1.413e+00  5.113e-02 -27.635  < 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: 13691  on 8160  degrees of freedom
Residual deviance: 12364  on 8150  degrees of freedom
AIC: 20446

Number of Fisher Scoring iterations: 6
```

As mentioned in Chapter 3, the resulting model is over-dispersed. To overcome this, a second iteration was then run, this time, using an Over-dispersed Poisson Model, and the following was the outcome:

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.101e+00  1.655e-01   6.655 3.02e-11 ***
age           -6.797e-04  1.900e-03  -0.358  0.72058
marital       -2.035e-01  3.230e-02  -6.300 3.12e-10 ***
sex            6.364e-02  3.341e-02   1.905  0.05686 .
ed             8.559e-02  2.667e-02   3.209  0.00134 **
daily_mileage  2.638e-03  6.411e-04   4.115 3.91e-05 ***
use            2.896e-01  3.494e-02   8.289  < 2e-16 ***
car_value     -1.132e-05  2.081e-06  -5.438 5.56e-08 ***
policy_length -8.686e-03  3.905e-03  -2.224  0.02616 *
car_age        5.914e-04  3.616e-03   0.164  0.87009
loc           -1.413e+00  6.579e-02 -21.476  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 1.655832)

    Null deviance: 13691  on 8160  degrees of freedom
Residual deviance: 12364  on 8150  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 6
```

In this particular case, the ODP model had a lower Akaike's Information Criterion (AIC). A lower AIC means a model provides a better fit. In conclusion, the ODP provided a better fit compared to the 'true' Poisson model. However, from the ODP model, a few variables were not significant at 95% level and had to be discarded from the model.

The ODP model generated 5 variables significant at 0.01% and 1 significant at 1% and another significant at 5% percent. The 7 variables in order of significance are, **marital, daily_mileage, use, car_value, loc, ed** and **policy_length**. The variables **age, sex** and **car_age** were dismissed since they were not significant at 5% level although, sex was found to significant at 10% level.

With the new significant variables, the model was also re-run and the following was the output:

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.193e+00  1.115e-01  10.698  < 2e-16 ***
marital       -2.050e-01  3.209e-02  -6.389 1.76e-10 ***
ed             8.760e-02  2.033e-02   4.309 1.66e-05 ***
daily_mileage  2.630e-03  6.408e-04   4.104 4.09e-05 ***
use            2.717e-01  3.351e-02   8.107 5.95e-16 ***
car_value     -1.133e-05  2.068e-06  -5.478 4.42e-08 ***
policy_length -8.652e-03  3.904e-03  -2.216   0.0267 *
loc           -1.410e+00  6.577e-02 -21.436  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 1.655862)

    Null deviance: 13691  on 8160  degrees of freedom
Residual deviance: 12371  on 8153  degrees of freedom
AIC: NA
```

```
Number of Fisher Scoring iterations: 6
```

In addition to the new model's output, the researcher also added an extra test. The test for influence and leverage is important in statistics since it shows the researcher if there are any observations that are influencing the outcome of the model in their direction. The researcher employed the use of the Cook's Distance in order to determine if there were any values disproportionately influencing the model.

The following plot was generated by the RStudio program by calling the **`cooks.distance()`** function on the new model crated above.

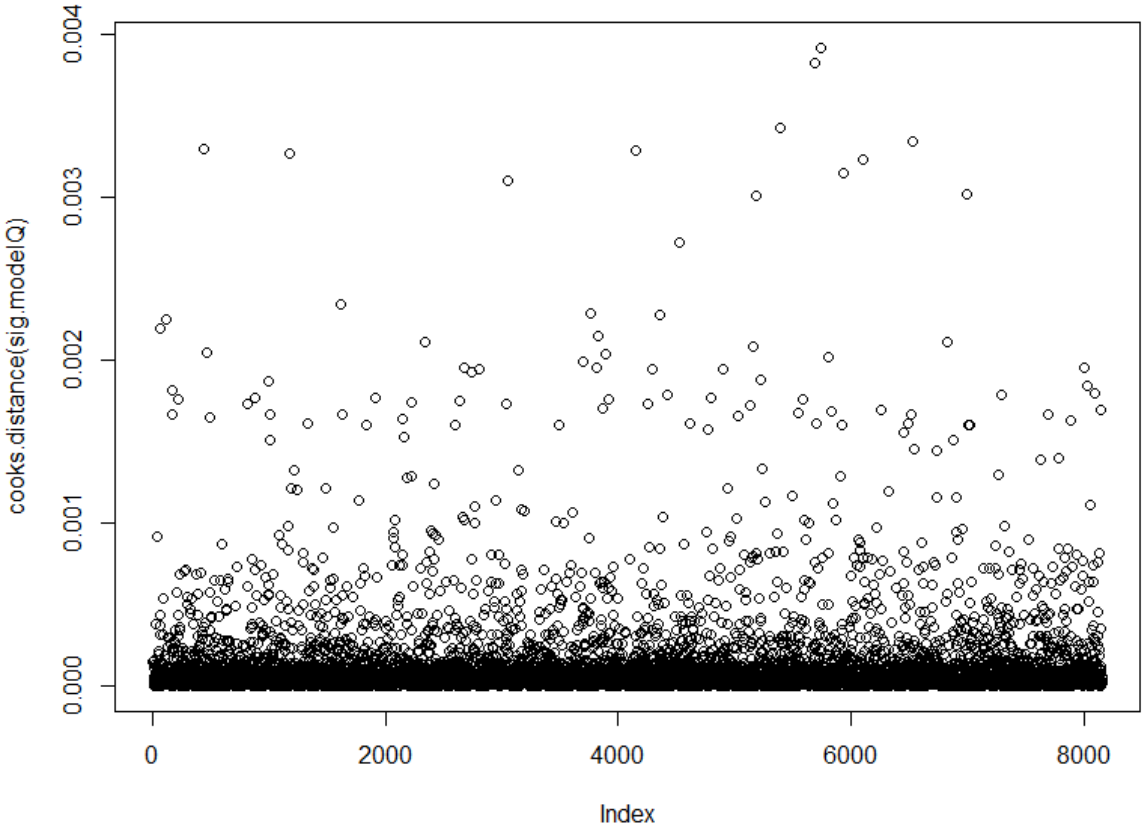

Fig. 5.1: Cook's Distances on the Values of the Poisson Model

A common rule of thumb is that an observation with a value of Cook's Distance over 1.0 has too much influence. Given the highest value in our dataset has a Cook's Distance value of just under 0.004, it was safe to conclude that all given observations had little influence.

5.1.2. GAMMA MODEL

The same train data used to build the Poisson Model was used to build the Gamma model. However, the $0 observations were removed since the Gamma model only accepts positive numbers as inputs. The Gamma model, also surprisingly generated the following:

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    9.391e+00  2.046e-01  45.904   <2e-16 ***
age           -1.368e-03  2.219e-03  -0.617    0.538
marital       -1.261e-02  3.996e-02  -0.316    0.752
sex           -1.563e-03  4.173e-02  -0.037    0.970
ed             9.899e-03  3.293e-02   0.301    0.764
daily_mileage -1.659e-03  7.996e-04  -2.075    0.038 *
use           -5.919e-02  4.363e-02  -1.357    0.175
car_value      8.925e-09  2.590e-06   0.003    0.997
policy_length -3.397e-03  4.856e-03  -0.699    0.484
car_age        1.640e-03  4.432e-03   0.370    0.711
loc            9.502e-02  8.033e-02   1.183    0.237
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 1.213275)

    Null deviance: 3083.3  on 3151  degrees of freedom
Residual deviance: 3072.7  on 3141  degrees of freedom
AIC: 64645

Number of Fisher Scoring iterations: 6
```

In this model, all variables save for **daily_mileage** were insignificant at 5% level in explaining claim amounts. This one variable was used to create a revised model.

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    9.339879   0.046976 198.820   <2e-16 ***
daily_mileage -0.001600   0.000794  -2.015    0.044 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 1.209153)

    Null deviance: 3083.3  on 3151  degrees of freedom
Residual deviance: 3078.4  on 3150  degrees of freedom
AIC: 64634

Number of Fisher Scoring iterations: 5
```

Similarly, the Cook’s Distance statistics was also employed and the following plot was generated.

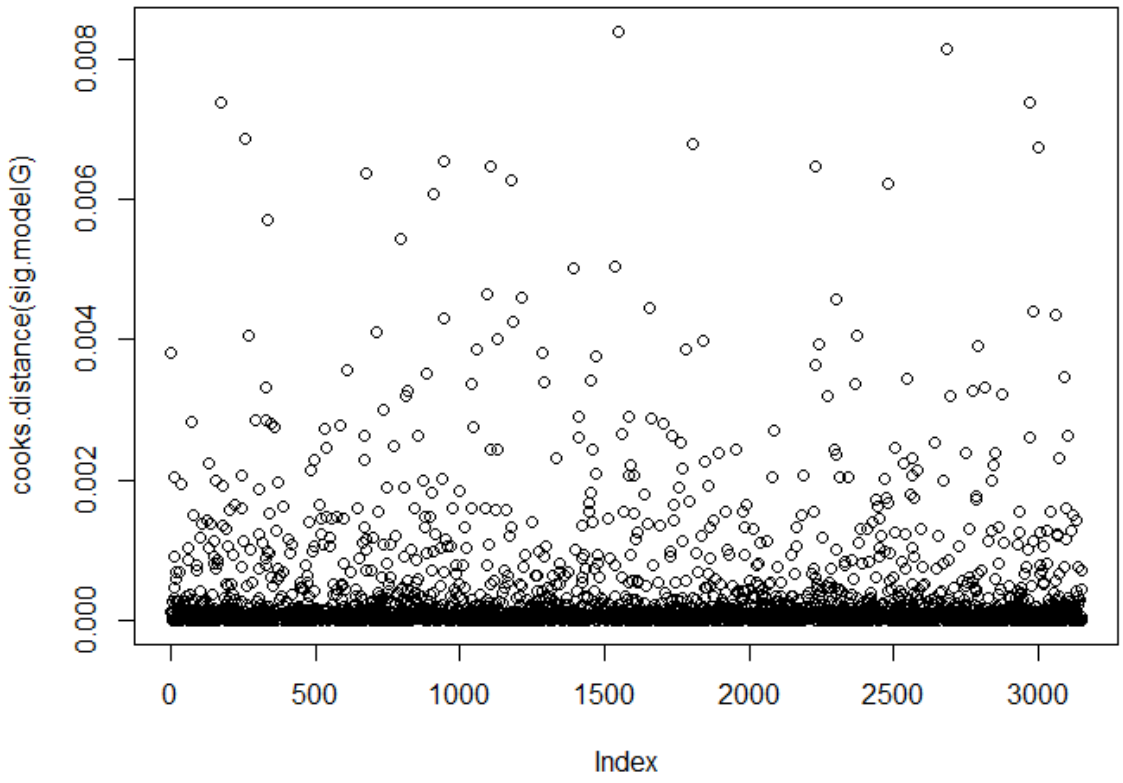

Fig. 5.2: Cook's Distances on the Values of the Gamma Model

A conclusion similar to Poisson distribution findings was drawn, the Cook's Distances were not large enough to make some observations influential. However, as noted by David (2015), "The influence factors of the claims cost are different from the factors corresponding to the frequency of claims, a fact that confirms the assumption suggested by the actuary literature regarding the isolated analysis of these two phenomena."

Given the two models obtained, the pure premium can then be estimated using claims data.

5.1.3. THE PURE PREMIUM MODEL

Using the values of the two models and the test dataset, the results were as follows:

Claims	Premium Written	Estimated Pure Premium	Adjusted Estimate
$8,611,461	$8,772,981	$ 7,749,431	$ 11,042,939

Fig. 5.3: Estimated Values of Premiums

Currently, insurers use the 5% method to price Comprehensive Motor Insurance. This was proposed by the regulator to reduce undercutting by some insurers. However, it is not a hard and fast rule as some companies adopt a slightly lower percentage (e.g., Zimnat quoted $195.30 for a Sum Assured of $4,000, resulting in 4.88%), and some companies use other non-percentage-based methods.

From the results quoted above, the initial Loss Ratio was 98.16%. This is incredibly high for any insurer. However, the expected claims for this model were 10% lower than the claims incurred during the period under consideration. An actuarial rule of thumb is that the pure premium component of any premium should adequately cover the claims expected to arise from that particular group of policies. The model used fell short by 10%.

In addition, to get the Adjusted Gross Premium, the researcher employed the following ratios and key statistics that the company under scrutiny uses as loadings:

- Commission 12.5%
- Expenses 30%

This results in loadings of 42.5%.

Using the ratio

$$Premiums = Expected\ Claims\ \times\ (1 + Loadings\ for\ Expenses\ )$$

Under these bases, the Adjusted Premium is $11,042,939. With this new estimated amount, the Loss Ratio loses 20.18 percentage points to 77.98%. As expected, applying GLMs yields lower loss ratios for the insurers.

OTHER OBSERVATIONS

Further analysis of the data revealed that 52.4% of the current clients will benefit from applying GLMs in pricing. That is, after adjusting for commission and expense loadings, these clients will pay less than what they pay right now – which is the 5% of their car's value. The rest of the clients will be adversely affected by the pricing change, and this might cause a mass exodus of the clients to other insurers.

While GLMs are robust, better suited for the future and better at estimating a fair premium, one need not ignore the competition.

5.2. RECOMMENDATIONS

The insurers in Zimbabwe are losing out when they apply the 5% method. It might fetch clients and business in the long run, but soon enough, world trends are pointing to Artificial Intelligence and the next big thing in insurance. A company that already uses GLMs can easily shift and add more variables in the pricing mix, such as telematics, usage-based insurance, and so on, thus bringing more value to their clients.

In addition, local insurers could also consider collecting more information from their clients. For instance, the claim amount was only significantly correlated with mileage. In the presence of more information, more predictors could have been used. As such I recommend that the insurers collect, also the following:

- Driving Experience of the client
- Number of dependents who use the car.
- Annual or Monthly Income, and so on.

These variables have been noted to have a significant effect on the occurrence of claims (Yao, 2013). Also, insurers could also add Bonus-Malus systems, No Claims Discounts, Claim-free cash backs, excesses and deductibles to deter adverse selection and lower premiums for their clients with respect to Comprehensive Motor insurance.

5.3. SUGGESTIONS FOR FURTHER STUDY

Generally, Zimbabweans hardly opt for Comprehensive Motor insurance, they prefer instead the basic Third-Party that is mandated by the Road Traffic Act. One could further study why this is so. Could it be that insurers charge higher premiums for Comprehensive Motor insurance? Could it be the lack of incentivised pricing for Comprehensive Motor insurance?

Similarly, one could also employ further models, such as Generalised Additive Models, Generalised Linear Mixed Models, and other suitable models in predicting the relationships between the response variables of claim frequency and Claim amounts with their explanatory variables.

In addition, the researcher noted that most Zimbabwean insurers have their income largely coming from motor Third-Party insurance. Assuming, that the government implements the South African equivalent of the Road Accident Fund in Zimbabwe and does away with the compulsory motor Third-Party insurance policies, this could hurt some companies

considerably. One may study the implications or effects of adopting the Road Accident Fund to local insurers.

5.4. CONCLUSION

Non-life insurance pricing consists of establishing a premium or a tariff paid by the insured to the insurance company in exchange for the risk transfer (David, 2015). Premiums are obtained by multiplying the expected frequency of occurrence and the expected claim amount per occurrence. The result is a pure premium, which is loaded with various components to come up with a final premium.

This paper considers an analysis of the effect of applying Generalised Linear Models in an environment where an already more prevalent method exists. The values reached are contrasted with the prevalent method. Using policyholder's personal and demographical indicators, the frequency of claims is estimated through Poisson regression model. With similar inputs, the Gamma model is used to estimate the expected claim cost given the characteristics of the individual.

Essentially, the final price of the premium is reached by considering the risk factors, such as, marital status, level of education, use of car, average mileage driven on a daily basis, urbanicity, car value and car age.

For example, the model demonstrates that the higher value of the car, the lower the likelihood of a claim. This is representative of the insurance business as people with expensive cars tend to drive more carefully.

In conclusion, the data used was extracted from a non-random sample from one company. The results may not be representative of the whole population as various insurers use different underwriting methods and other insurers might have access to different information better suited to help with pricing for their clients.

REFERENCES

Anderson, D., Feldblum, S., Modlin, C., Schirmacher, D., Schirmacher, E. and Thandi, N. (2007) *A Practitioner's Guide to Generalized Linear Models: A Foundation for Theory, Interpretation and Application*. 3rd edition. Arlington, Virginia: Casualty Actuary Society.

Bailey, R.A. and Simon, L.J. (1960) Two Studies in Automobile Insurance Ratemaking. *ASTIN Bulletin*, 1(04) 192–217.

Brisard, E. (2014) *Pricing of Car Insurance with Generalized Linear Models*. Master's Thesis. Brussels, Belgium: Vrije Universiteit Brussel.

Chidakwa, A. and Munhupedzi, R.N. (2017) Investigating the Impact of Dollarization on Economic Growth: A Case of Zimbabwe. *Export Journal of Finance*, 5(1) 12–20.

Coutts, S. (1984) Motor Premium Rating. In: F. Vylder, M. Goovaerts, J. Haezendonck (eds.) *Premium Calculation in Insurance*. Dordrecht: Springer Netherlands, 399–448.

Coutts, S.M. (1984) Motor Insurance Rating, An Actuarial Approach. *Journal of the Institute of Actuaries*, 111(1) 87–148.

David, M. (2015) Auto Insurance Premium Calculation Using Generalized Linear Models. *Procedia Economics and Finance*, 20 147–156.

Denuit, M. and Charpentier, A. (2004) *Mathématiques de l'assurance non-vie. T. 1: Principes fondamentaux de théorie du risque*. Collection économie et statistiques avancées. Paris: Economica.

Farrar, D.E. and Glauber, R.R. (1967) Multicollinearity in Regression Analysis: The Problem Revisited. *The Review of Economics and Statistics*, 49(1) 92.

Goldburd, M., Khare, A. and Tevet, D. (2016) *Generalized Linear Models for Insurance Rating*. Arlington, Virginia: Casualty Actuary Society.

Halliday, C. (2015) *Motor Insurance Pricing: Current and Future Innovation*.

Han, C., Yao, D. and Zheng, S. (2016) *The Automobile Insurance Pricing Model Combining Static Premium with Dynamic Premium*.

Henwood, N. and Wang, B. (2009) *Insights into Comprehensive Motor Insurance Rating*.

Huang, D. and Query, J.T. (2007) *Designing a New Automobile Insurance Pricing System in China: Actuarial and Social Considerations*. In: 2007.

Johnson, P. (1966) *Actuarial Aspects of Motor Insurance*. In: 4 November 1966.

Kane, J.N., Anzovin, S. and Podell, J. (2006) *Famous first facts: a record of first happenings, discoveries, and inventions in American history*. 6th ed. New York: H.W. Wilson.

Lemaire, J. (1995) *Bonus-Malus Systems in Automobile Insurance*. Huebner International Series on Risk, Insurance, and Economic Security. Vol. 19. Dordrecht: Springer Netherlands.

Liu, Z., Shen, Q. and Ma, J. (2017) A driving behavior model evaluation for UBI. *International Journal of Crowd Science*, 1(3) 223–236.

McCullagh, P. and Nelder, J.A. (1998) *Generalized linear models*. Monographs on statistics and applied probability 37. 2nd ed. Boca Raton: Chapman & Hall/CRC.

Nelder, J.A. and Verrall, R.J. (1997) Credibility Theory and Generalized Linear Models. *ASTIN Bulletin*, 27(01) 71–82.

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

Pinquet, J. (1997) Allowance for Cost of Claims in Bonus-Malus Systems. *ASTIN Bulletin*, 27(01) 33–57.

Poole, D.L., Mackworth, A.K. and Goebel, R. (1998) *Computational intelligence: a logical approach*. New York: Oxford University Press.

Schmitter, H. (2004) The Sample Size Needed for the Calculation of a GLM Tariff. *ASTIN Bulletin*, 34(01) 249–262.

Siddig, M.H.M.A. (2016) Application of the Generalized Linear Models in Actuarial Framework. *ArXiv e-prints*,

Silva, Y.R. and Afonso, L. (2015) A Comparative Study of Pricing Methods of Automobile Insurance in Brazil. *Revista Brasileira de Educação*, 19 25–44.

Werner, G. and Modlin, C. (2016) *Basic Ratemaking*. 5th edition. Arlington, Virginia: Casualty Actuary Society.

Williams, A.F. and Shabanova, V.I. (2003) Responsibility of drivers, by age and gender, for motor-vehicle crash deaths. *Journal of Safety Research*, 34(5) 527–531.

Yao, J. (2013) *Generalized Linear Models for Non-life Pricing - Overlooked Facts and Implications*. A Report from GIRO Advanced Pricing Techniques (APT) Working Party. Institute and Faculty of Actuaries.

Zhou, J. (2011) *Theory and Applications of Generalized Linear Models in Insurance*. PhD Thesis. Concordia University.

Best advice: Cash Back vs No Claim Bonus (2010) FA News.

APPENDIX A

R CODE USED FOR DATA ANALYSIS

For Claim Number Estimation:

```
#Load readxl
library("readxl")

#Import the Exposure.xlsx file
exposure <- read_xlsx("train/exposure.xlsx")

#Perform GLM Calcs Using Poisson
summary(modelP <- glm(claim_freq ~ age + marital + sex + ed + daily_mileage
      + use + car_value + policy_length + car_age + loc , family = "poisson",
      data = exposure))

#Perform GLM Calcs Using Quasi-poisson, to allow for overdispersion
summary(modelQ <- glm(claim_freq ~ age + marital + sex + ed + daily_mileage
      + use + car_value + policy_length + car_age + loc , family =
      "quasipoisson", data = exposure))

#View a model based only on significant values
summary(sig.modelQ <- glm(claim_freq ~  marital + ed + daily_mileage + use
      + car_value + policy_length + loc , family = "quasipoisson", data =
      exposure))

#Plot Cook's Distance for the final model
plot(cooks.distance(sig.modelQ))
```

For Claim Severity Estimations:

```
#Load readxl

library("readxl")

#Import the claims.xlsx file
claims <- read_xlsx("train/claims.xlsx")

#Perform GLM Calcs Using Gamma Regression
summary(modelG <- glm(claim_sev ~ age + marital + sex + ed + daily_mileage
      + use + car_value + policy_length + car_age + loc , family = Gamma(link
      = "log"), data = claims))

#View a model based only on significant values
summary(sig.modelG <- glm(claim_sev ~  daily_mileage, family = Gamma(link =
      "log"), data = claims))

#Plot Cook's Distance for the final model
plot(cooks.distance(sig.modelG))
```

APPENDIX B

VIF TABLES PER FACTOR

Coefficients[a]

Model		Collinearity Statistics	
		Tolerance	VIF
1	marital	.998	1.002
	sex	.912	1.096
	education	.518	1.929
	daily_mileage	.972	1.029
	car_use	.846	1.182
	car_value	.861	1.162
	policy_length	1.000	1.000
	car_age	.571	1.752
	location	.920	1.086

a. Dependent Variable: age

Coefficients[a]

Model		Collinearity Statistics	
		Tolerance	VIF
1	age	.924	1.083
	sex	.908	1.101
	education	.507	1.971
	daily_mileage	.972	1.029
	car_use	.844	1.184
	car_value	.849	1.178
	policy_length	1.000	1.000
	car_age	.571	1.752
	location	.920	1.087

a. Dependent Variable: marital

Coefficients[a]

Model		Collinearity Statistics	
		Tolerance	VIF
1	age	.918	1.089
	education	.508	1.968
	daily_mileage	.972	1.029
	car_use	.921	1.086
	car_value	.850	1.176
	policy_length	1.000	1.000
	car_age	.571	1.751
	location	.921	1.085
	marital	.988	1.012

a. Dependent Variable: sex

Coefficients[a]

Model		Collinearity Statistics	
		Tolerance	VIF
1	age	.936	1.068
	daily_mileage	.972	1.029
	car_use	.861	1.162
	car_value	.887	1.128
	policy_length	1.000	1.000
	car_age	.923	1.083
	location	.944	1.059
	marital	.990	1.010
	sex	.911	1.097

a. Dependent Variable: education

Coefficients[a]

Model		Collinearity Statistics	
		Tolerance	VIF
1	age	.914	1.094
	car_use	.844	1.185
	car_value	.849	1.178
	policy_length	1.000	1.000
	car_age	.571	1.752
	location	.944	1.060
	marital	.988	1.012
	sex	.908	1.101
	education	.506	1.975

a. Dependent Variable: daily_mileage

Coefficients[a]

Model		Collinearity Statistics	
		Tolerance	VIF
1	age	.914	1.094
	car_value	.849	1.178
	car_age	.571	1.752
	location	.920	1.087
	marital	.988	1.012
	sex	.908	1.101
	education	.506	1.975
	daily_mileage	.972	1.029
	car_use	.844	1.185

a. Dependent Variable: policy_length

Coefficients[a]

Model		Collinearity Statistics	
		Tolerance	VIF
1	age	.927	1.079
	marital	.988	1.012
	sex	.909	1.100
	education	.529	1.891
	daily_mileage	.972	1.029
	car_use	.908	1.102
	policy_length	1.000	1.000
	car_age	.571	1.752
	location	.921	1.086

a. Dependent Variable: car_value

Coefficients[a]

Model		Collinearity Statistics	
		Tolerance	VIF
1	age	.917	1.091
	car_value	.913	1.096
	policy_length	1.000	1.000
	car_age	.571	1.751
	location	.920	1.087
	marital	.988	1.012
	sex	.990	1.010
	education	.516	1.937
	daily_mileage	.972	1.029

a. Dependent Variable: car_use

Coefficients[a]

Model		Collinearity Statistics	
		Tolerance	VIF
1	age	.915	1.093
	car_value	.850	1.177
	marital	.988	1.012
	sex	.909	1.100
	education	.519	1.925
	daily_mileage	.997	1.003
	car_use	.844	1.184
	policy_length	1.000	1.000
	car_age	.571	1.752

a. Dependent Variable: location

Coefficients[a]

Model		Collinearity Statistics	
		Tolerance	VIF
1	age	.914	1.094
	car_value	.849	1.178
	location	.920	1.087
	marital	.988	1.012
	sex	.908	1.101
	education	.819	1.221
	daily_mileage	.972	1.029
	car_use	.844	1.184
	policy_length	1.000	1.000

a. Dependent Variable: car_age