University of Nairobi

School of Mathematics

ISSN: 2410-1397

Master Project in Actuarial Science

Application of Generalized Linear Models in Pricing Usage-Based Insurance

Research Report in Mathematics, Number 22, 2020

Masese Victor Omaanya

November 2020

Submitted to the School of Mathematics in partial fulfilment for a degree in Master of Science in Actuarial Science

Application of Generalized Linear Models in Pricing Usage-Based Insurance

Research Report in Mathematics, Number 22, 2020

Masese Victor Omaanya

School of Mathematics
College of Biological and Physical sciences
Chiromo, off Riverside Drive
30197-00100 Nairobi, Kenya

Master Thesis

Submitted to the School of Mathematics in partial fulfilment for a degree in Master of Science in Actuarial Science

Submitted to: The Graduate School, University of Nairobi, Kenya

Abstract

Technological advancements and big data adaptations are broadly impacting the insurance industry. Usage Based Insurance is a result of the emerging technologies and big data adaptation as it is based on telematics data. Incorporating telematics data in auto insurance pricing models reduces moral hazard and adverse selection phenomena which arise from information asymmetry. Traditional auto insurance does not consider how and when driving is done which is part of telematics data. Thus, there is need for insurers to redefine their pricing models and risk selection criteria to include telematics data. This study aims to model claim frequency and severity using generalized linear models in order to evaluate the impact of distance driven, speed and time of driving on premium rates which are part of telematics data. Generalized linear models have been applied by insurers in ratemaking, reserving and underwriting general insurance policies for over 50 years. The models allow for response variables with non-gaussian error distributions hence suitable for modelling auto insurance claim frequency and severity. Specifically, the gamma and Poisson generalized models have been employed in this study. The gamma has been used to model claim severity while claim frequency is modelled by the Poisson model. From the insurance portfolio analyzed, speed and distance variables were found to be significant while time was not significant in both models. Coefficient estimates for distance categories were positive indicating positive correlation. Speed band categories had negative estimate values indicating negative correlation. We observe that severity and frequency increase with distance and speed. Similarly, the pure premium increases with distance and speed. These findings are representative of particular auto insurance policies and do not represent a generalized trend. Results of this study can help auto insurance industries to evaluate the risk of driving more precisely and come up with personalized premiums for drivers based on their real time driving factors.

Master Thesis in Mathematics at the University of Nairobi, Kenya.
ISSN 2410-1397: Research Report in Mathematics

DISTRIBUTOR: School of Mathematics, University of Nairobi, Kenya

Declaration and Approval

I the undersigned declare that this dissertation is my original work and to the best of my knowledge, it has not been submitted in support of an award of a degree in any other university or institution of learning.

Signature Date

Masese Victor Omaanya
Reg No. I56/13620/2018

In my capacity as a supervisor of the candidate's dissertation, I certify that this dissertation has my approval for submission.

Signature Date

Dr Davis Bundi
School of Mathematics,
University of Nairobi,
Box 30197, 00100 Nairobi, Kenya.
E-mail: dbundi@uonbi.ac.ke

Signature Date

Dr Carolyne Ogutu
School of Mathematics
University of Nairobi,
Box 30197, 00100 Nairobi, Kenya.
E-mail: cogutu@uonbi.ac.ke

Dedication

This project is dedicated to my mum and dad.

Contents

Figures and Tables

Figures

Tables

Acknowledgments

I wish to thank all the people whose assistance was a milestone in completion of this project. My special regards to Dr. Davis Bundi and Dr. Carolyne Ogutu, my supervisors, for their guidance in this project. I also wish to acknowledge the support and great love of my family, my girlfriend, Elvin; She gave me support and help, discussed ideas and prevented several wrong turns; my mum, Susan Nyaboke; and my dad, James Masese Omesa. They kept me going on and this work would not have been possible without their input. Above all, glory and honor is to the Almighty for protection and providence.

Masese Victor Omaanya

Nairobi, 2020.

1 Introduction

1.1 Background

Usage-Based Insurance (UBI) is a type of motor insurance that incorporates driving behavior and distance driven in determining premium rates. Telematics devices such as dongle, blackbox, and smartphones fitted in insured cars are used to monitor and capture data regarding how driving is done. The data is then relayed to insurers and used to determine the risk exposure of a driver. The basic idea of UBI is that premium rates are matched to a driver's risk exposure that is directly monitored as the individual drives [1]. Variables of interest to insurers that are recorded by telematics devices include; distance driven, speed, acceleration, time of day, cornering and braking; all the other variables except distance determine driving behavior. Insurers use the data to determine premium rates and give discounts accordingly [18].

UBI seeks to change the fixed costs linked to driving distance and convert them to costs that vary depending on distance and other rating variables in determining premium rates. Unlike traditional insurance that relies on combined statistics and driving records that depend on past trends, UBI utilizes individual and current driving behaviors, thus making the premium pricing precisely personalized [28]. United States, United Kingdom and Italy were early adopters of UBI solutions as telematics technology that came into existence over 20 years ago. TripSense solution was started by American Insurer Progressive in 1999 and later renamed to MyRate policy. The UBI solution attracted significant publicity thereby expanding telematics ideas to a majority of states in the US. A regulation known as eCall was voted for by the European Parliament in 2015 which requires all new cars in Europe to have telematics devices programmed to dial 112 automatically when a crash occurs, reporting impact data and precise location of the crash [29].

Insurers in the United States offer UBI solutions such as Drivewise by AllState, SmartRide solution by Nationwide, and Snapshot solution by Progressive reward safe and good drivers by considering time of day, speed, mileage and braking. Big data technologies are used to record and relay the data regarding those variables. Nationwide provides discounts ranging from 10% to 40% for good drivers who are determined by analysis of collected actual driving data that is also used to give personalized feedback to drivers. [2].

Policies such as Pay How You Drive (PHYD), Pay As You Speed (PAYS), and Pay As You Drive (PAYD) are modern innovative insurance schemes that are being embraced and

commercialized all over the world [32]. The underlying principle for these schemes is charging personalized premiums depending on how driving is done and the extent of risk exposure rather than using traditional premium rating factors only. In 2016, Aryeh Insurance operating in Israel covered about 200,000 vehicles under PAYD policy which represented about 15% of all vehicles in Israel [16].

Insurers in Switzerland, Germany, Netherlands, Spain and Italy have launched various telematics solutions in those countries. AIG XLNT driver solution by AIG insurance in Ireland, IBM Telematics by IBM corporation in US, Smiles solution by Etiga Insurance based in Singapore, and TD My Advantage solution offered by TD Insurance based in Canada use a driving score to determine insurance premium amounts [2]. Driving data consisting of speed, braking, acceleration, time of day and cornering variables per trip is relayed and analyzed using big data technologies to come up with the driving score. "Drive Safe and Save" solution offered by StateFarm Company in US focuses on similar PHYD parameters. The solution provides roadside assistance, maintenance alerts and stolen car locator services to the insureds as well.

In Africa, Kenya is the second country after South Africa where telematics technology has been employed in auto insurance [9]. Auto Correct solution offered by Heritage Insurance based in Kenya is a program that rewards safe drivers based on driving behavior. Telematics devices fitted in cars send real time driving data to Heritage where it is analyzed to come up with a driver score. Based on the score, Heritage is able to differentiate good and average drivers who are then rewarded with premium cashbacks amounting to a maximum of 15% annually. Additionally, drivers under the program are offered tips on how to improve on driving skill and a review of trips made.

Rate making is a process of determining the level of premium to charge for each and every risk covered by insurers. The rate making process is aimed at charging a fair premium to cover future losses, expenses, and make provision for the cost of capital [15]. Simplifying the rate structure enables clients to easily understand factors in play and encourage behaviors that minimize losses among those covered [30]. For instance, drivers will easily comprehend the impact of hard braking or high speed on their premium rates, hence they will be more cautious on the road [33]. A comprehensive driver's risk profile in UBI is developed by incorporating driving behavior specific to each driver [10]. Insurance companies that have adopted the advanced use of telematics to build driver risk profiles and come up with UBI solutions include Allstate with Drivewise solution, Progressive with Snapshot solution, Esurance with Drivesense program, Travellers with IntelliDrive program. Variables of interest monitored by insurers in these UBI solutions are not restricted to distance driven and time but also include how the actual driving is done; speed, acceleration, braking and cornering that determine driving behavior [2].

Limitations to adopting UBI technology for most insurers include; costs that come along with handling and collecting of telematics data. Secondly, most of the potential customers are not willing to sacrifice their privacy for the incentives offered in form of premium reductions and bonuses [18]. In addition, policyholders' accumulated driving behavior is not transferable to newly acquired cars since the telematics devices are linked to the car and not the policyholder which as well contradicts ownership of the data [7].

1.2 Statement of the problem

According to East Africa Insurance Outlook Report 2019/2020, there is a need for auto insurance companies to redefine their pricing models, risk selection criteria and underwriting techniques by integrating big data such as data from telematics devices in their models [9]. Traditional auto insurance is based on actuarial calculations of combined historical data to give rating factors such as driver record, type of car, previous claims, and driver characteristics. It does not consider how and when driving is done, instead it assigns an average premium rate to specific drivers considering only the traditional rating factors hence leading to adverse selection and moral hazard phenomena in auto insurance industry. Adverse selection and moral hazard are a result of information asymmetry between insurers and insureds [6]. Insureds take advantage of the information asymmetry to unfairly claim or get covered at a premium that does not match their risk exposure, a phenomenon referred to as moral hazard. Insurers on the other hand are discouraged from covering medium risks or charge the risks exorbitantly due to information asymmetry leading to adverse selection phenomenon. Therefore, usage-based insurance through use of telematics devices can reduce the issue of moral hazard and adverse selection by enabling information symmetry. This lowers fraudulent claims and offers fair premiums to the insured based on their risk profiles

1.3 Objectives

The overall objective of this study is to apply Generalized Linear Models (GLMs) in evaluating Usage-Based Insurance (UBI) premium rates. Variables of interest included are average speed, distance driven and time of day.

The specific study objectives are;

(i) To model claim frequency using the Poisson GLM and assess the impact of the variables of interest on the frequency of auto insurance claims.
(ii) To model claim severity using the gamma GLM and assess the impact of the variables of interest on the severity of auto insurance claims.
(iii) To compute the pure premium rate.

1.4 Justification

Due to increased driving behavior monitoring and improved client segmentation, UBI addresses issues of adverse selection and moral hazard that result from information asymmetry between policyholders and insurers. Closer matching of car insurance policies to their actual risks lowers cross-subsidization thereby increase actuarial fairness compared to classifying drivers into broad groups [15]. UBI enables better quantification of a driver's risk exposure for premium rate adjustment. By use of telematics technology to monitor driving of the insured, insurers can easily differentiate safe drivers from risky drivers who appear safe on paper [27]. Drivers in high-risk groups and the young are usually charged higher premium rates, which sometimes do not reflect their actual risk exposure. With telematics technology, a policyholder's true accident risk can be determined based on how they drive furthermore, the cost of car insurance determination is potentially reduced when using telematics risk factors.

1.5 Limitations of the study

Data collected is not segmented according to the different road types used by insureds. Driving on certain roads such as highways that allow higher speed limits compared to other types of roads render speed and acceleration non-uniform factors. In consequence, drivers' risk exposure is not proportionate to speed and acceleration factors.

2 Literature review

In a study to unravel the predictive capability of telematics data in pricing auto insurance, it is concluded that premium depends on individual attributes such as car age and type of road while gender is identified as a redundant variable [2]. Black box technology, which is a telematics technology was used to collect data from Belgian young drivers that was used in estimating risk exposure. Compositional predictors and generalized additive models were applied to evaluate the impact of telematics factors on frequency and severity of the expected car claims. Similar results were achieved in a study analyzing pay as you drive model where road crashes were linked to road categories in Netherlands [32].

Analysis of PAYD model was carried out to evaluate the impact of large-scale implementation of the model in Netherlands where seven strategies were analyzed. Experimental design was applied while the strategies were differentiated into road type, age of car and time of day categories. From the study it was concluded that depending on the type of PAYD differentiation, risk exposure varies greatly [3]. However, differentiating the premium to reflect risk exposure will appeal to more drivers and therefore reduce road crashes as a result [11]. With the implementation of PAYD model in Netherlands, total road crashes are expected to reduce by 5% or more every year thereby reducing the number of fatalities and injured people by 60 and 1,000 respectively [32]. Road category differentiation was found crucial in the Netherlands as inter-urban roads were found riskier than motorways hence should attract higher premiums. Driving during the night was also found to be riskier than day driving since 33% of single car crashes happen during the night compared to only 13% during the day [32].

Generalized linear models have also been used in a French auto insurance portfolio to determine pure premium by factoring characteristics of policyholders that are observable while on the road. Variables of interest in the study included profession and age of the insured, age of insurance contract, coefficient of bonus-malus, type and purpose of car. Claim frequency and severity were analyzed separately in the study thus confirming isolation of the two phenomena as stated by Actuarial literature [31]. Poisson GLM was used to model frequency whereas severity was modelled by gamma GLM. A decrease in frequency of claims was found to be in line with both increase in age of insurance contract and insured's age while an increase in bonus-malus coefficient increases claim frequency [6]. Similar findings that age of insured is a significant factor were echoed in a study where claim frequencies were predicted using tree-based models [25]. Profession of insured, car type and car purpose were not significant in the frequency model. Claim

severity on the other hand was influenced by car type, profession and age of insured. A decrease in cost of claims was recorded as the age of insured increased.

Tree-based methods have been applied as alternatives to GLMs in predicting the expected claim frequencies for non-life insurance [24]. Unlike GLMs, advanced tree-based methods such as bagging, random forest and regression trees are not limited to prior information about data structure hence can be used when little or nothing is known about the structure of the data [5]. In a study involving both simulated data and collision data from AXA Winterthur in Switzerland, tree-based methods performed better than GLMs in predicting claim frequency. Tree based algorithms have the ability to incorporate non-significant or correlated predictor variables without affecting the outcome unlike GLMs thus they provide competitive options in predictive capacity terms as determined from the study [25].

A study carried out to examine the validity of applying Bonus Malus (BM) coefficient to group car insureds sufficiently concluded that the coefficient increases the predictive power of a priori risk factors. Automobile policies data from an insurance company in Spain was used in the study that employed Rough Set (RS) theory method to achieve the study's objective. BM coefficient, driving region, age of insured and characteristics of the insured car such as type, purpose and age were used as predictor variables. Driving region, age of both car and insured were found to be the most relevant predictor variables in grouping car policies while BM coefficient slightly increased the predictive power of a priori risk variables [23]. Purpose or use of the insured car was not significant in classifying car policies according to the study which concluded from empirical analysis of the data that predictor variables used by insurers in classifying policies were significant.

The existence of several insurers in Ghana enabled carrying out a study involving car policyholders in the country to determine factors that insureds consider before choosing an insurer. Random Utility Theory (RUT) and Discrete Choice Experiment (DCE) approach were deployed in analyzing the choice behavior of policyholders. Claims settlement, cost of premium, proximity of insurer and customer satisfaction were the factors used in the analysis to achieve the study's objective. Parameters for choice consideration were estimated using the probit model. Results of the study showed that car owners highly value insurers who pay claims promptly while charging moderate premium levels and are closer to them when choosing an insurer. However, paying claims promptly was the most important attribute influencing choice amongst alternative insurers. Customer satisfaction levels can be substituted by prompt claims payment, near proximity and moderate premiums [17].

Aggressiveness in driving is shown by speed, acceleration and braking while proficiency in car controlling is shown by cornering and turning. Suddenly changing speeds, speeding and tailgating are indicators of aggressive driving behavior. Therefore, monitoring driving

speed, acceleration and braking can help insurers determine aggressive drivers hence adjust their premiums accordingly. American Automobile Association conducted a study from 2003 through 2007 on aggressive driving where it was determined that speeding was the leading factor for fatal car crashes. Aggressive driving accounted for 56% of fatal car accidents in the 5-year period [21]. In addition, NHTSA (National Highways Traffic Safety Administration) reported in 2012 that fatal car crashes occur when driving speed is above 55 mph [21]. Besides, monitoring cornering behavior helps insurers predict the possibility of car rollover since it is a dangerous kind of accident and has very high fatality rates [15].

3 Methodology

3.1 Introduction

This chapter covers data modelling approaches, variable significance and other model considerations. Generalized linear models specifically gamma and Poisson models are discussed. The gamma has been used to model claim severity while Poisson models claim frequency.

3.2 Generalized Linear Models

Generalized linear models (GLMs) is a criterion of modelling the relationship between a dependent variable whose result we are interested in predicting and one or more independent variables [12]. In auto insurance ratemaking, the outcome variable can either be claim severity, claim frequency, loss ratio or pure premium while predictor variables are insured's characteristics or policy terms included in rating models by insurers [26]. In traditional rating models, the predictors include type of car, age of both car and the insured, use of the car among others.

GLMs have been applied by insurers in ratemaking, reserving and underwriting in general insurance for over 50 years. The basic ideas of GLMs were initiated by Nelder and Wedderburn in 1972 [19]. Most countries started to regulate their insurance markets in late $90s$ thereby made the application of generalized linear models become widespread. GLMs have a capacity to assess relationships in data that is not necessarily normally distributed hence suitable for modelling motor insurance claim amounts and frequency [8]. In auto insurance, claim amount and frequency are generally presumed independent and non-identical [15].

The formal expression of randomness of any particular risk's outcome denoted by y_i is as follows;

$$y_i \sim Exponential(\mu_i, \phi) \tag{1}$$

Where the *exponential* represents the exponential family distributions while μ_i and ϕ represent mean and dispersion of the distribution. All the members of the exponential family share this common trait of taking the two parameters [8]. Parameter ϕ is assumed to be the same for all entries while μ_i is entry-specific hence its unique to every risk[27].

The relationship between the predictors and the model prediction, μ_i, modelled by GLMs is;

$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \tag{2}$$

The transformation of μ_i denoted as $g(\mu_i)$ above is the link function determined by the user. The linear predictor is the right-hand side of (eqn.3) containing the intercept β_0. $x_{i1}, \cdots, x_{ip}$ are the predictor variables and $\beta_1, \beta_2, \cdots, \beta_p$ are the coefficients. A GLM software estimates the values for the coefficients and the intercept.

A GLM allows for a transformation of the mean as the link function hence provides model flexibility in relating μ_i to the predictors. More options in identifying a model that best suits the industry is a result of the flexibility brought by the link function [14]. For the insurance industry rating plans, the natural log is specified as the link function as it produces a rating structure that is multiplicative thus the equation;

$$\ln(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

$$\mu_i = exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip})$$

$$\mu_i = e^{\beta_0} \times e^{\beta_1 x_{i1}} \times e^{\beta_2 x_{i2}} \times \cdots \times e^{\beta_p x_{ip}} \tag{3}$$

Components of a GLM include a distribution for the data, a link function and linear predictors. Canonical link functions for exponential family distributions are as follows;

Table 1. Expontential distributions, their link functions and linear predictors

Distribution	Link	Linear predictor
Normal	Identity	$g(\mu) = \mu$
Poisson	Log	$g(\mu) = log(\mu)$
Gamma	Inverse	$g(\mu) = log\left(\frac{1}{\mu}\right)$
Binomial	Logit	$g(\mu) = log\left(\frac{\mu}{1-\mu}\right)$

3.3 Data analysis methods

3.3.1 Frequency model

Claim frequency is the number of claims per risk unit covered within a given duration, normally a year for auto insurance policies. A Poisson distribution is used in modelling counts of events that occur during a specified time interval hence commonly applied in actuarial studies as a claim counts distribution [20].

In this case, for driver i, the probability of random variable x_i taking value xi is given by the probability density:

$$Pr(X_i = x_i) = \frac{e^{-\mu_i}\mu_i^{x_i}}{x_i!} \tag{4}$$

Where μ_i is the average number of claims incurred within the cover period. Since the maximum likelihood estimator is the standard estimator for this model, the likelihood function becomes:

$$l(\mu) = \prod_{i=1}^{n} \frac{e^{-\mu_i}\mu_i^{x_i}}{x_i!}$$

And log likelihood function is described as:

$$LL(\mu) = \sum_{i=1}^{n} [\mu_i ln x_i - \mu_i - ln x_i!]$$

3.3.2 Severity model

The gamma distribution skewed to the right with a sharp peak and a zero lower bound which are similar characteristics exhibited by claim severity distributions thus the most commonly applied distribution for modelling claim severity [20].

For a driver i, individual claim costs $y_{i1}, y_{i2}, y_{i3}, \cdots, y_{in}$ are assumed to be independent and follow a gamma distribution with a probability density function described as follows;

$$Pr(Y_i = y_i) = \frac{1}{\beta_i^{\alpha_i}\Gamma\alpha_i}(y_i\alpha_i)^{\alpha_i} e^{(\frac{-y_i\alpha_i}{\beta_i})}, y_i > 0 \tag{5}$$

The aim is to estimate the shape parameter, α and the scale parameter, β which can used to estimate the future claim severity.

The maximum likelihood estimator is the standard estimator for this model and it is defined as;

$$l(\beta) = \prod_{i=1}^{n} \frac{1}{\beta_i^{\alpha_i}\Gamma\alpha_i}(y_i\alpha_i)^{\alpha_i} e^{(\frac{-y_i\alpha_i}{\beta_i})}$$

With mean $E[y_i] = \mu_i$ and variance $Var(y_i) = \frac{\mu_i^2}{v}$, thus the log likelihood function can be expressed as follows;

$$l(\beta) = \prod_{i|z_i>0} \prod_{k=1}^{n} (\frac{1}{\Gamma v}(\frac{v y_{ik}}{\mu_i})^v exp(-\frac{v y_{ik}}{\mu_i})\frac{1}{y_{ik}})$$

The log likelihood function is differentiated with respect to scale parameter β to get the partial derivative as follows;

$$\frac{\partial LL(\beta|y)}{\partial \beta_j} = \frac{\partial}{\partial \beta_j} \sum_{i|z>0} \sum_{k=1}^{n} (-vln\mu_i - \frac{vc_{ik}}{\mu_i}) = 0$$

For driver i, the estimated claim cost is defined as $\hat{y}_i = \hat{\mu}_i = exp(r_i\hat{\beta})$

The maximum likelihood estimates of $\hat{\beta}$ is the solution of;

$$\sum_{i|z_i>0} (z_i - \frac{y_i}{\hat{y}_i})r_i = 0$$

3.3.3 Goodness of fit

To find a model that fits the data adequately, we consider residuals. Smaller residuals indicate that the model is better. A "Full model" is one with the smallest residuals whereas a "Null model" is one with the largest residuals [13]. Somewhere between the null and full models lie the optimal model. Scaled deviance or simply deviance, D is used to check residuals. The scaled deviance is described as double the difference of the log-likelihood of the saturated model and the model under consideration which is the current model. In a saturated model, observed values are equal fitted values since the model has as many parameters as number of observations[6]. The scaled deviance D is used in this study with a chi-square test ${\chi^2}_{(\alpha;n-p)}$, where α is the significance level and p parameters. If $D > {\chi^2}_{(\alpha;n-p)}$, the null hypothesis is rejected concluding that the model is good in respect to residuals; rejected null hypothesis H_0: residuals deviance D is significantly large hence model is not good.

3.3.4 Other model considerations

When modelling rating plans for insurance, is it required that there are constant elements in the plan and changing elements dependent on individual covers. Rating algorithms usually have a base value of the loss amount that varies depending on policyholder characteristics and it is arrived at by a non-GLM-based analysis. In that case, the base value

included in the GLM is not assigned an estimated coefficient but still remains part of the insurance rating plan [14]. The base value is considered a predictor with a coefficient of 1 in the model to ensure the other estimated coefficients in the model give optimal results in its presence.

4 Data Analysis and discussion

4.1 Introduction

This chapter covers data and its, preparation, analysis and discussion. Telematics and claims data from a Kenyan insurer of policy year 2019 containing 523 exposure units is used in this study. The two data sets used are telematics data and claims data.

4.2 Analysis, presentation and discussion

The telematics data contained variables of driver ID, speed, distance driven and time of day. Variables contained in claims data include driver ID, claim date and claim amount. The two datasets were grouped into speed bands, distance bands and day or night driving. Policies of cars driven for less than 200km were excluded from the analysis. Policies missing any of the variables in telematics data and claims data were removed from the analysis. Policies in the claims data that were absent in telematics data were omitted from the analysis. Duplicates were removed from both telematics data and claims data.

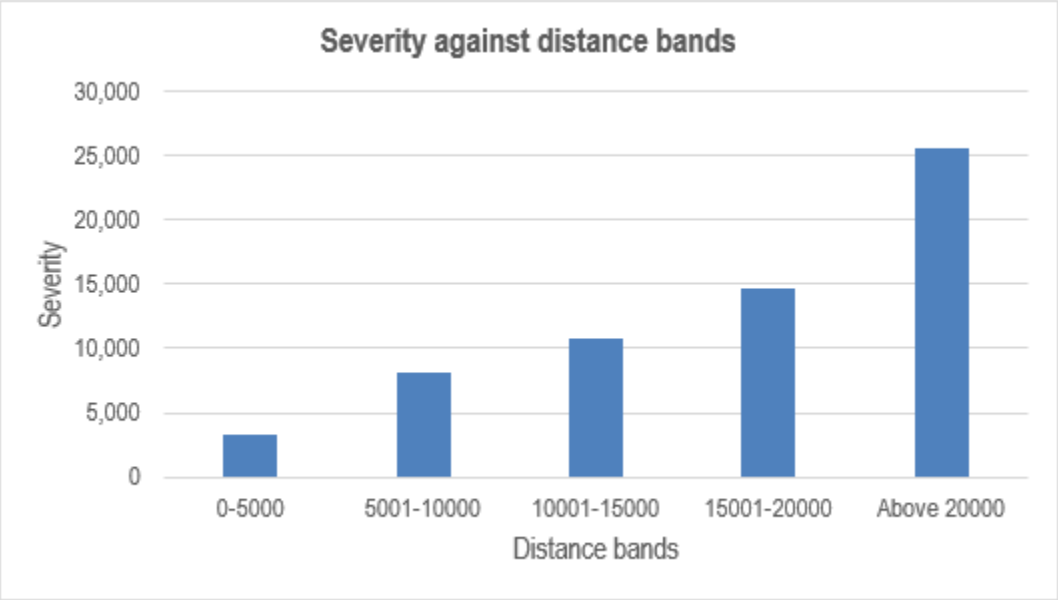
Assumptions made include; the data is correct apart from the corrected errors during data preparation, and the claims trend remain similar. Data preparation, grouping and graphs fitting was done in Microsoft Excel. R software was used in fitting the models applying the GLM package and Summaries.

4.3 Data Presentation

4.3.1 Severity

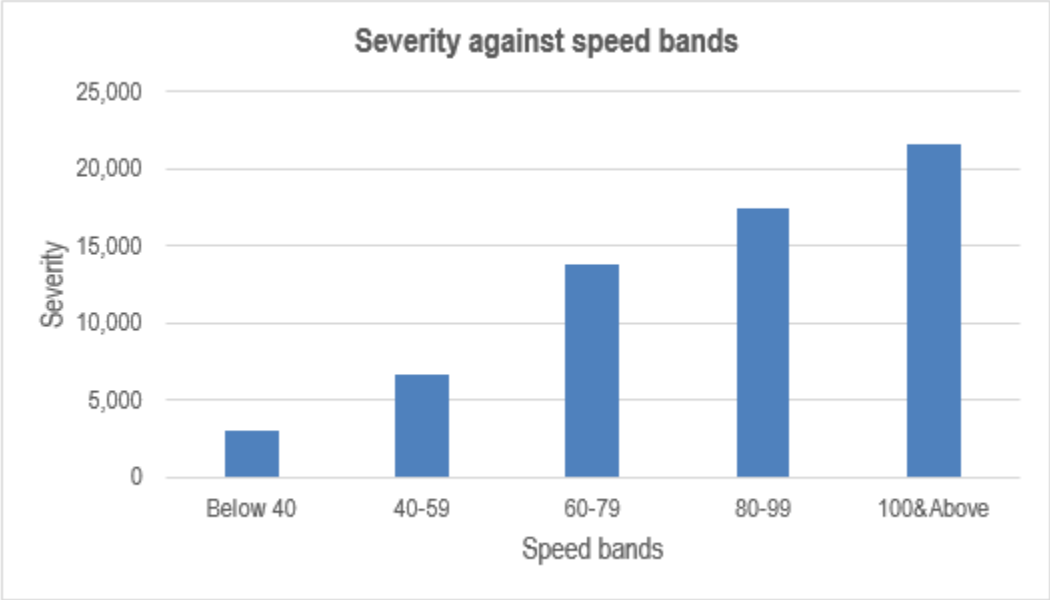
This section looks at how severity varies with distance, speed and time of day. Bar graphs created in excel have been used to present the data. Distance and speed were grouped into five bands each while time was grouped into day or night. The three variables are plotted on the x-axis.

Figure 1. A graph of severity against distance



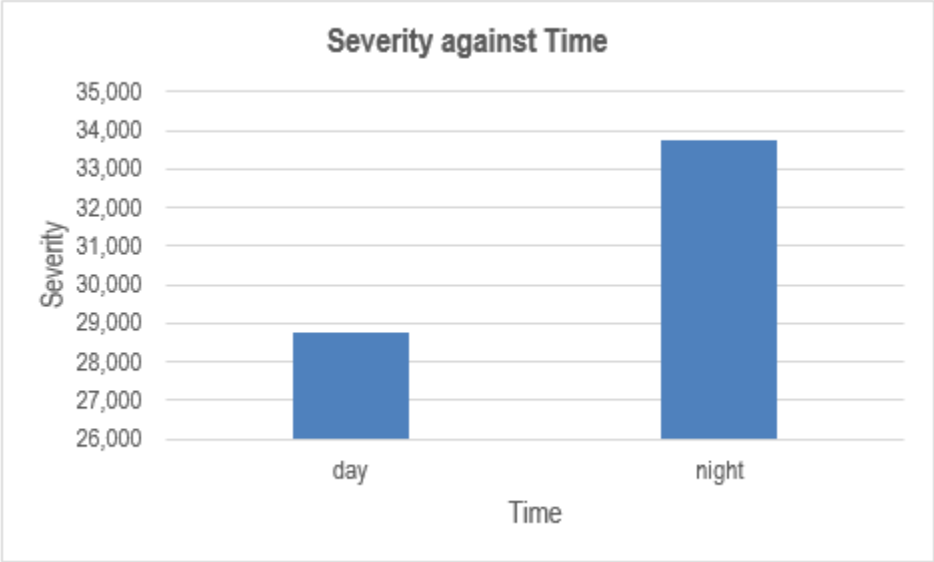
We observe that severity increases with distance. Above 20000 distance band has the highest severity while distance band 0-5000 has the least severity. Generally, cars driven for longer distances are more exposed to accident risk hence the high severity.

Figure 2. A graph of severity against speed



We observe that severity increases with speed. Speed band of below 40 has the lowest severity while 100&Above band has the highest severity. The two speed bands below 60 combined have a lower severity than any of the other individual higher bands.

Figure 3. A graph of severity against time

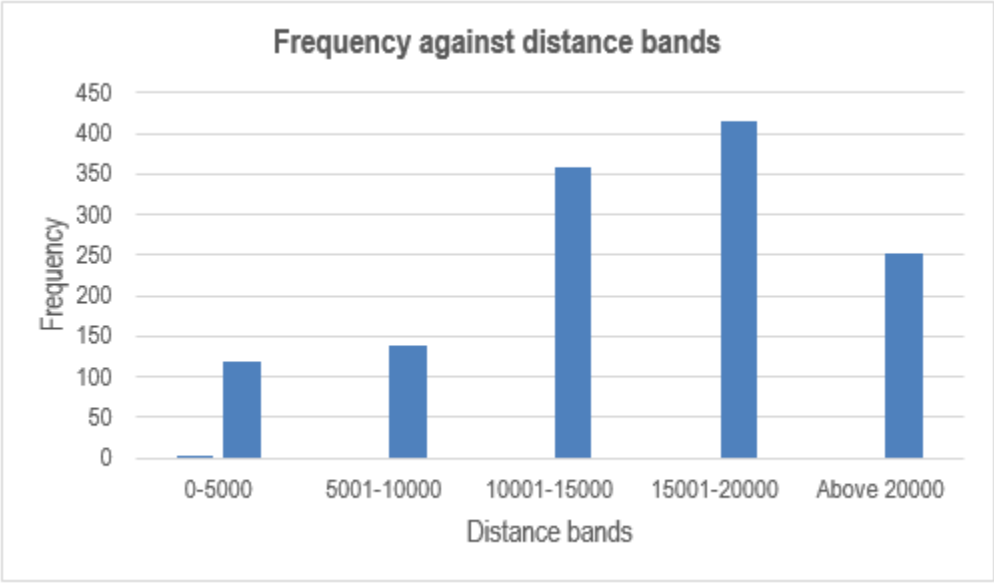


We observe that night driving has higher severity than day time driving. Night driving poses a higher risk to drivers as compared to daytime.

4.3.2 Frequency

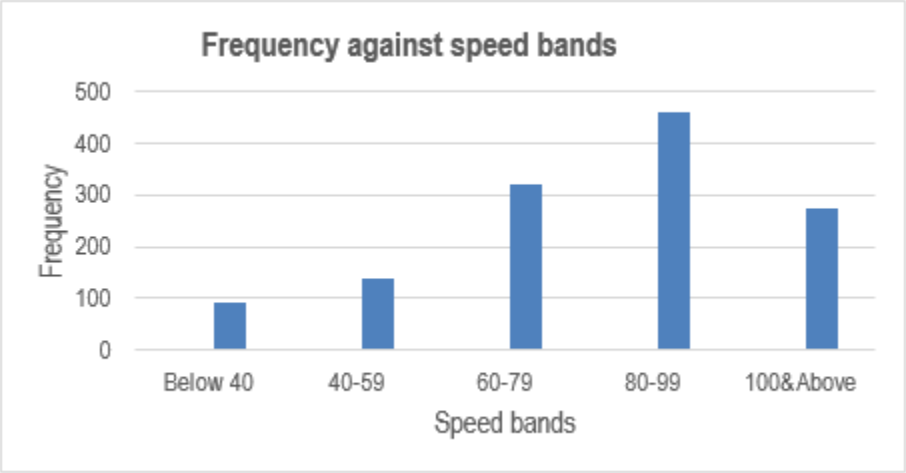
This section looks at how frequency varies with distance, speed and time of day. Bar graphs created in excel have been used to present the data. Distance and speed were grouped into five bands each while time was grouped into day or night. The three variables were plotted on the x-axis.

Figure 4. A graph of frequency against distance



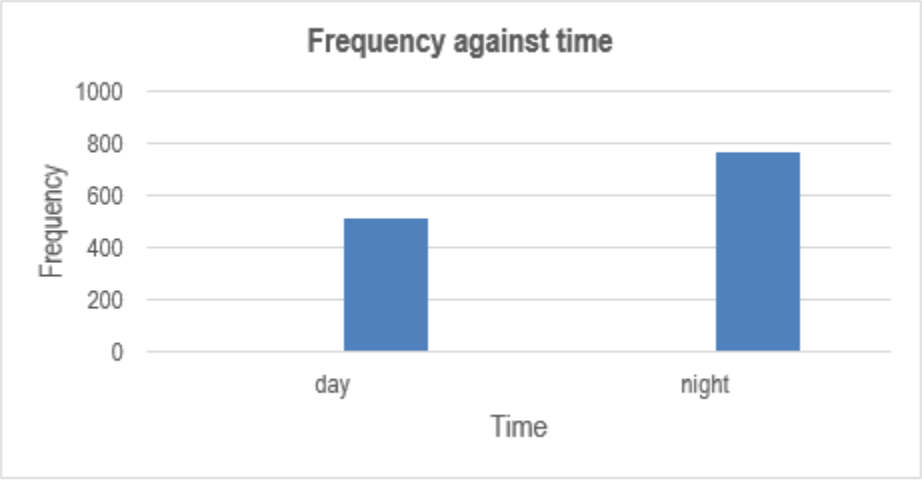
We observe that frequency increases with distance. Frequency is highest for the $15001 - 20000$ distance band followed closely by the $10001 - 15000$ band while Above 20000 has the third highest frequency.

Figure 5. A graph of frequency against speed



Generally, frequency increases with speed. Frequency is highest at $80 - 99$ speed band followed by $60 - 79$. The two speed bands below 60 when combined give the least frequency compared to the individual speed bands above 60.

Figure 6. A graph of frequency against time



Night time driving has higher frequency than day driving.

4.4 Model fitting

Models fitting was done in R software using the glm() package and Summaries. Severity and frequency were the respective response variables in the first and second models below. Estimated parameters of the severity and frequency models are shown below.

Figure 7. Parameter results for severity model

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               11.83214    0.18331  64.548  < 2e-16 ***
distanceband10001-15000    0.26704    0.17095   1.562  0.11888
distanceband15001-20000    0.56546    0.17338   3.261  0.00118 **
distanceband5001-10000     0.73356    0.19518   3.758  0.00019 ***
distancebandAbove 20000    1.79927    0.18952   9.494  < 2e-16 ***
speedband40-59            -0.94865    0.19132  -4.958 9.62e-07 ***
speedband60-79            -1.44837    0.15083  -9.602  < 2e-16 ***
speedband80-99            -1.03472    0.15224  -6.797 2.93e-11 ***
speedbandBelow 40         -1.43830    0.20183  -7.126 3.44e-12 ***
TimeNight                 -0.06352    0.09596  -0.662  0.50829
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From figure 7 above, all the speed and distance bands are significant except distance band $10001 - 15000$. Night time is also not significant. The significant variables indicate that there is correlation with severity. Coefficient estimates for distance bands are positive values indicating that distance has a positive correlation with severity whereas speed band estimates are negative values hence are negatively correlated with severity.

The estimated parameters of the frequency model are shown below.

Figure 8. Parameter results for frequency model

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               1.042463   0.107106   9.733  < 2e-16 ***
distanceband10001-15000   0.188736   0.107806   1.751 0.079998 .
distanceband15001-20000   0.406591   0.106772   3.808 0.000140 ***
distanceband5001-10000    0.077856   0.126978   0.613 0.539781
distancebandAbove 20000   0.461939   0.112759   4.097 4.19e-05 ***
speedband40-59           -0.511360   0.105624  -4.841 1.29e-06 ***
speedband60-79           -0.789678   0.083661  -9.439  < 2e-16 ***
speedband80-99           -0.294603   0.077232  -3.815 0.000136 ***
speedbandBelow 40        -0.610905   0.123629  -4.941 7.76e-07 ***
TimeNight                 0.009637   0.057143   0.169 0.866073
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From above figure 8, all the speed and distance bands are significant except distance bands $5001 - 10000$ and $10001 - 15000$. Night time is also not significant. The significant

variables indicate that there is correlation with frequency. The positive estimates for distance bands indicate a positive correlation between distance and frequency whereas the negative estimates for speed bands indicate a negative correlation with frequency.

4.5 Tables

This section contains tables of claim severity, claim frequency and pure premium. The contingency tables were created in Microsoft excel. Distance bands are presented in rows while columns present speed bands.

Table 2. Claim severity

Distance	Below 40	40-59	60-79	80-99	100&Above
0-5000	64	245	429	1,186	1,340
5001-10000	621	668	1,742	1,934	3,106
10001-15000	726	1,481	2,425	2,586	3,543
15001-20000	873	1,566	2,796	4,160	5,355
Above 20000	718	2,622	6,406	7,607	8,306

Values in '000s

Table 2 above contains aggregated claim amounts grouped into distance and speed bands. The amounts are increasing with distance and speed. The *Above*20000 distance band and speed band 100&*Above* has the highest claim amount.

Table 3. Claim frequency

Distance	Below 40	40-59	60-79	80-99	100&Above
0-5000	14	7	15	36	46
5001-10000	25	7	46	25	36
10001-15000	23	38	113	140	44
15001-20000	18	66	105	145	82
Above 20000	10	22	41	113	66

Table 3 above contains claim frequencies grouped into distance and speed bands. Generally, the frequencies are increasing with distance and speed except for the *below*40 speed band. The $15001 - 20000$ distance band and speed band $80 - 99$ has the highest frequency.

Table 4. Pure premium

Distance	Below 40	40-59	60-79	80-99	100&Above
0-5000	103	461	857	2,267	4,066
5001-10000	823	2,532	2,955	5,317	10,913
10001-15000	1,727	3,410	5,748	7,725	11,785
15001-20000	2,141	5,499	7,462	15,420	28,112
Above 20000	3,088	5,763	9,718	36,869	47,819

Values in '000s

Table 4 above contains calculated pure premiums grouped into distance and speed bands. The amounts are increasing with distance and speed. The *Above*20000 distance band and speed band 100&*Above* has the highest pure premiums.

4.6 Evaluation of models

Scaled deviance was used to test the models' goodness of fit as follows; the severity model has a residual deviance of 450.28 on 522 degrees of freedom. The chi-square distribution returns a p-value of 1. Consequently, the null hypothesis is rejected at 5% level of significance; Ho: residual deviance is not significantly large and the model is good as far as the residuals are concerned.

The frequency model has a residual deviance of 442.96 on 522 degrees of freedom. From the chi-square distribution, the p-value is 1. Consequently, the null hypothesis is rejected at 5% level of significance; Ho: residuals deviance is not significantly large and the model is good as far as the residuals are concerned.

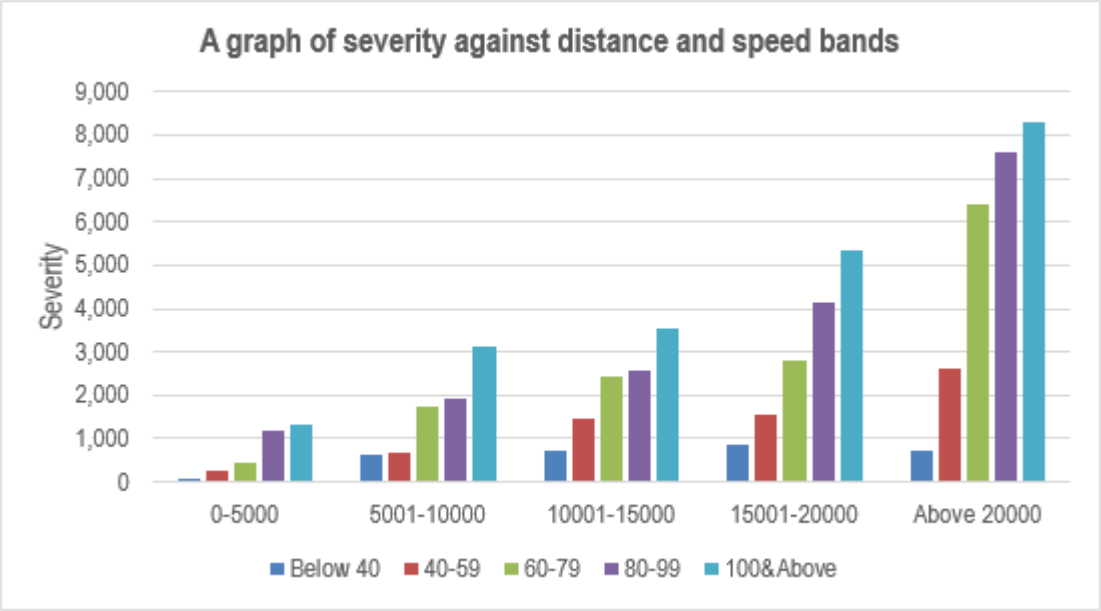
4.7 Modelling results

Results of the modelled severity and frequency are presented in this section. Microsoft excel was used to plot the graphs. Stacked bar graphs were used to present the results. Distance and speed bands were both plotted on the same x-axis.

4.7.1 Severity

The figure below shows how severity varies with distance and speed bands.

Figure 9. A graph of severity against distance and speed bands.

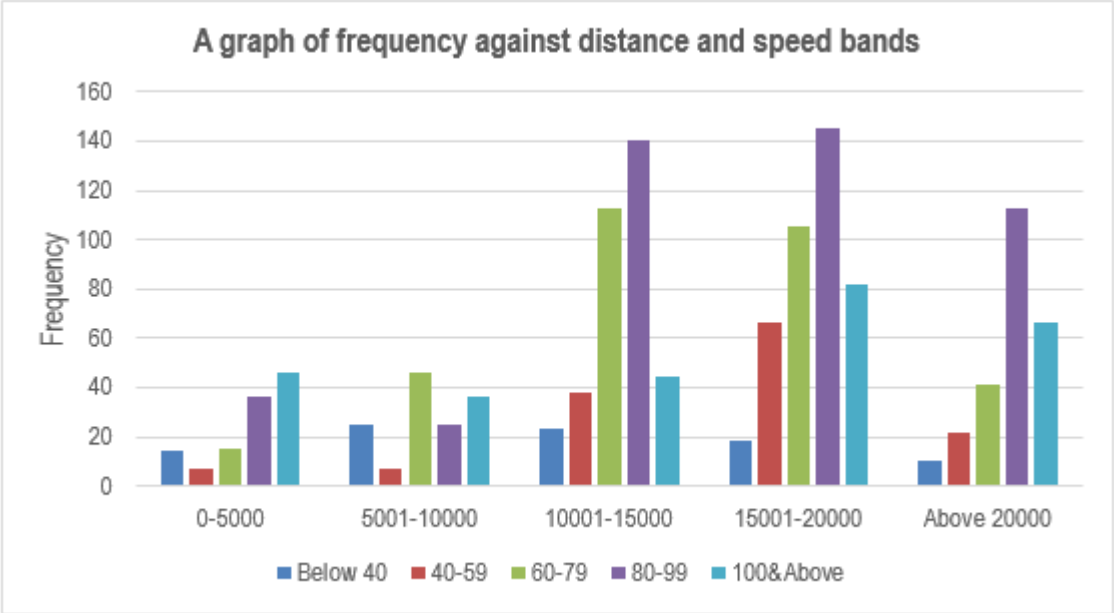


From figure 9 above, we note that severity increases with distance and speed. Speed band 100&Above has the highest severity in all distance bands while speed below 40 has the least severity in all the distance bands. Severity for the two speed bands below 60 is increasing at a slower rate unlike the other 3 speed bands.

4.7.2 Frequency

The figure below shows how frequency varies with distance and speed bands

Figure 10. A graph of frequency against distance and speed bands.



Generally, frequency increases with distance and speed. Frequency is highest at speed band $80-99$ and distance band $15001-20000$. Speed band $80-99$ has the highest frequency followed by $60-79$ band in 3 out of the 5 distance bands.

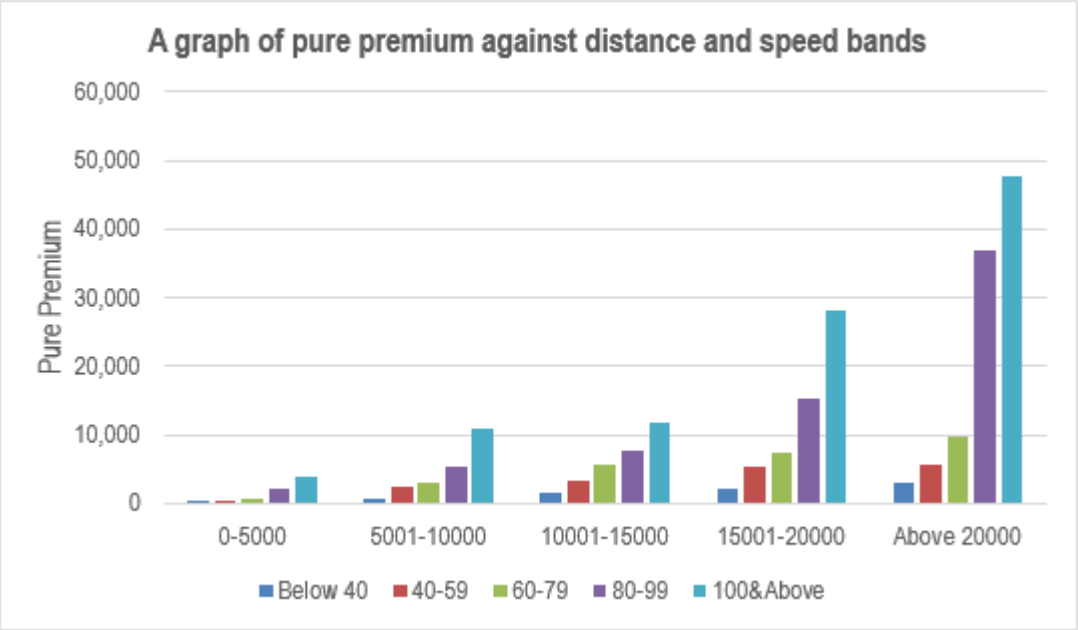
4.8 Pure premium results

The premium is calculated by multiplying severity with frequency and has been done in excel. The values for severity and frequency were estimated using the estimated parameters of severity model and frequency model respectively.

$$ExpectedPurePremium = claimfrequency \times claimseverity$$

The figure below shows how pure premium varies with distance and speed bands.

Figure 11. A graph of pure premium against distance and speed bands



From figure 10 above, we note that pure premium increases with distance and speed. Speed bands 100&Above and $80-99$ have the highest and second highest pure premiums respectively in all the distance bands. Distance bands $15001-20000$ and Above 20000 together with the two speed bands above 80 are the only ones with pure premium above $15,000$.

5 Conclusion

5.1 Introduction

This chapter covers summary of findings, conclusions drawn, recommendations and areas of further research. The conclusions drawn are based on data analysis, discussions and results from the previous chapter.

5.2 Summary of the findings

Based on 5% level of significance and the p-value, severity and frequency models have passed the test of goodness of fit. We observe that severity and frequency increase with distance. Distance band of Above 20000 has the highest severity and frequency while distance band $0-5000$ has the least severity and frequency. Generally, cars driven for longer distances are more exposed to accident risk on the road hence the high severity and frequency.

Frequency and severity increase with speed. Below 40 speed band has the least frequency and severity while 100&Above speed band has the highest severity and $80-99$ speed band has the highest frequency. A high speed is more likely to lead to a high claim frequency and severity as it enhances accident risk. From researches on leading causes of road accidents, it has been shown that driving above set speed limits is among the leading causes of road crashes.

We observe that night driving has higher frequency and severity than day time driving. Night driving poses a higher accident risk to drivers due to a number of unfavorable conditions such as poor road visibility compared to daytime, and high theft/crime cases which lead to a claim.

We note that the pure premium increases with distance and speed. Since pure premium is a product of severity and frequency, it also increases in a similar manner as both frequency and severity. Similarly, night driving has a higher pure premium.

5.3 Conclusion

From the above findings, both severity and frequency vary similarly with distance, speed and time of day. Consequently, the pure premium varies in a similar manner as frequency and severity. Driving for longer distances and/or at higher speeds is associated with

higher severity and frequency. The pure premium therefore increases with distance and speed.

Generally, cars driven for long distances are more exposed to accident risk on the road hence the high severity and frequency. On the other hand, a high speed is more likely to lead to a high claim frequency and severity as it enhances accident risk. As a result, pure premium rates for driving at high speeds and long distances should be relatively higher as compared to low speeds and short distances.

Since driving time was not significant, it should be categorized into smaller time frames instead of day or night so as to uncover periods when accident risk is high. This will enable insurers adjust premium rates according to risk exposure at different time periods.

These findings are representative of a part of auto policies covered and do not represent a generalized trend. However, the findings are useful to auto insurers since they are aimed at better risk quantification thus better pricing of auto policies.

5.4 Recommendation

This study recommends that the insurance industry adopts telematics technology in auto insurance premium pricing so as to accurately match insured risks to premium rates and fairly charge drivers as per their exposure to risk.

5.5 Areas of further research

This study was based on private auto insurance policies. A similar study can be conducted for commercial auto insurance policies. Also, the other exponential distributions instead of the Poisson and gamma distributions can be applied together with other predictor variables such as cornering and braking to evaluate their effect on the pure premium.

Appendix

R syntaxes

Syntax for severity model

```
severitydata <-
  read.csv("C:\\Users\\masese.victor\\Documents\\data1.csv",header=TRUE)
#importing severitydata to R
severitydata #calling the severitydata
attach(severitydata) #attaching severitydata column titles in R workspace
head(severitydata) #viewing the first few rows of severitydata
severitymodel <-
  glm(severity ~ distance.band + speed.band + Time, family = Gamma(link = "log"),
      data = severitydata) #running the severity model
severitymodel #calling the severity model
summary(severitymodel) #obtaining the summary of the severity model
y<-predict(severitymodel,type="response")#using model parameters to predict severity
y #calling the predicted severity values
severitydata$modelledseverity<-y #attaching predicted severity values to the dataset
pchisq(450.28, df=522, lower.tail = FALSE) #obtaining p-value using chi-square
write.csv(severitydata, "C:\\Users\\masese.victor\\Documents\\sdata.csv")
#exporting modelledseverity data to excel
```

Syntax for frequency model

```
frequencydata <-
  read.csv("C:\\Users\\masese.victor\\Documents\\data2.csv",header=TRUE)
#importing frequencydata to R
frequencydata #calling the frequencydata
attach(frequencydata) #attaching frequencydata column titles in R workspace
head(frequencydata) #viewing the first few rows of frequencydata
frequencymodel <-
  glm(frequency ~ distanceband + speedband + Time, family = poisson(link = "log"),
      data = frequencydata) #running the frequency model
frequencymodel #calling the frequency model
summary(frequencymodel) #obtaining the summary of the frequency model
y<-predict(frequencymodel,type="response")#using model parameters to predict freque
y #calling the predicted frequency values
frequencydata$modelledfrequency<-y#attaching predicted frequency values to the data
pchisq(442.96, df=522, lower.tail = FALSE) #obtaining the p-value using chi-square
write.csv(frequencydata, "C:\\Users\\masese.victor\\Documents\\fdata.csv")
#exporting modelledfrequency data to excel
```

Bibliography

[1] Antonio, K., and Beirlant, J. (2007). Actuarial statistics with generalized linear mixed models. *Insurance: Mathematics and Economics,40(1),* 58–76.

[2] Arumugam, S., and Bhargavi, R. (2019). A survey on driving behavior analysis in usage based insurance using big data. *Journal of Big Data,* 6(1).

[3] Bordoff, J., and Noel, P. (2010). Pay-as-You-Drive Auto Insurance. *Issues of the Day: 100 Commentaries on Climate, Energy, the Environment, Transportation, and Public Health Policy, 150.*

[4] Bruneteau, F., Hallauer, T., Noël, M., and Tusa, S. (2013). *Usage-based Insurance, Global Study, Catch up with the Telematics Revolution. October,* 1–104.

[5] Cerchiara, R. R., Edwards, M., and Gambini, A. (2008, October). Generalized linear models in life insurance: decrements and risk factor analysis under Solvency II. In *18th international AFIR colloquium.*

[6] David, M. (2015). Auto Insurance Premium Calculation Using Generalized Linear Models. *Procedia Economics and Finance,20(15),* 147–156.

[7] de Azevedo, F. C., Oliveira, T. A., and Oliveira, A. (2016). Modeling non-life insurance price for risk without historical information. *Revstat Statistical Journal,14(2),* 171–192.

[8] De Jong, P. and Heller, G. Z. (2008). Generalized linear models for insurance data. *C*ambridge Books.

[9] Deloitte. (2020). *Insurance Outlook Report 2019/2020 East Africa. September 2019.*

[10] Denuit, M., and Lang, S. (2004). Non-life rate-making with Bayesian GAMs. *Insurance: Mathematics and Economics,* 35(3), 627-647.

[11] Desyllas, P., and Sako, M. (2013). Profiting from business model innovation: Evidence from Pay-As-You-Drive auto insurance. *Research Policy,* 42(1), 101-116.

[12] Goldburd, M., Khare, A. and Tevet, D. (2016). *Generalized linear models for insurance rating.* Casualty Actuarial Society, CAS Monographs Series, (5).

[13] Graybill, F. A. (1976). *Theory and application of the linear model* (Vol. 183). North Scituate, MA: Duxbury press.

[14] Haberman, S. and Renshaw, A. E. (1996). Generalized linear models and actuarial science. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 45(4), 407-436.

[15] Husnjak, S., Peraković, D., Forenbacher, I., Mumdziev, M. (2015). Telematics system in usage based motor insurance. *Procedia Engineering, 100*(January), 816–825.

[16] kurylowicz lukasz. (2016). Usage-Based Insurance: the concept and study of available analyses. *Insurance Review,4,* 127–142.

[17] Kwofie, C., Yormekpe, Di. D., Mensah, S. O., and Botchway, P. (2018). Choosing between Alternative Motor Insurance Policies: A Discrete Choice Experiment. *International Journal of Mathematics and Mathematical Sciences,* 2018.

[18] Ma, Y. L., Zhu, X., Hu, X., and Chiu, Y. C. (2018). The use of context-sensitive insurance telematics data in auto insurance rate making. *Transportation Research Part A: Policy and Practice,* 113, 243-258.

[19] McCullagh, P. (1984). Generalized linear models. *European Journal of Operational Research, 16*(3), 285–292.

[20] Ohlsson, E., and Johansson, B. (2010). *Non-life insurance pricing with generalized linear models* (Vol. 174). Berlin: Springer.

[21] Paefgen, J., Fleisch, E., Ackermann, L., Staake, T., Best, J. and Egli, K. (2013). *Telematics strategy for automobile insurers. I-Lab Whitepaper,* 1–31.

[22] Petrini, L. (2017). Non life pricing. *empirical comparison of classical GLM with tree based Gradient Boosted Models Innovative approach to pure premium estimation. June,* 1–12.

[23] Segovia-Vargas, M. J., Camacho-Miñano, M. del M., and Pascual-Ezama, D. (2015). Risk factor selection in automobile insurance policies: a way to improve the bottom line of insurance companies. *Revista Brasileira de Gestao de Negocios,* 17(57), 1228–1245.

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

[25] Supervisor Mario W, ochbauer V, Christoph Buser AXA Winterthur, C.-S. (2016). *Predicting Claims Frequencies using Tree-Based Models. July.*

[26] Tbook, T. E. X. (n.d.). *Non-Life Insurance Pricing with Generalized.*

[27] Tevet, D. (n.d.). *CAS MONOGRAPH SERIES GENERALIZED LINEAR MODELS FOR INSURANCE RATING Second Edition (Issue 5).*

[28] Tselentis, D. I., Yannis, G., and Vlahogianni, E. I. (2016). Innovative Insurance Schemes: Pay as/how You Drive. *Transportation Research Procedia, 14,* 362–371.

[29] Verbelen, R., Antonio, K., and Claeskens, G. (2018). Unravelling the predictive power of telematics data in car insurance pricing. *Journal of the Royal Statistical Society. Series C: Applied Statistics, 67(5),* 1275–1304.

[30] Wriggins, J. B. (2010). Automobile Injuries as Injuries with Remedies: Driving, Insurance, Torts, and Changing the Choice Architecture of Auto Insurance Pricing. *Loy. LAL Rev.,* 44, 69.

[31] Wuthrich, M. V., and Buser, C. (2019). Data analytics for non-life insurance pricing. *Swiss Finance Institute Research Paper,* (16-68).

[32] Zantema, J., Van Ameisfort, D. H., Bliemer, M. C. J., and Bovy, P. H. L. (2008). Pay-as-you-drive strategies: Case study of safety and accessibility effects. *Transportation Research Record, 2078,* 8–16.

[33] Zhang, H., Xu, L., Cheng, X., Chen, W., and Zhao, X. (2017, September). Big data research on driving behavior model and auto insurance pricing factors based on UBI. In *International Conference On Signal And Information Processing, Networking And Computers* (pp. 404-411). Springer, Singapore.