Strathmore
UNIVERSITY

Determination of key parameters in Generalized Linear Models for claim reserving: General Insurance

Wasafisia Mang'eni Patrick - 070253

Submitted in partial fulfillment of the requirements for the Degree of Actuarial Science at Strathmore University

School of Finance and Applied Economics
Strathmore University
Nairobi, Kenya

[November, 2015]

DECLARATION

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

Name of Candidate: Wasafisia Mang'eni Patrick
Signature: .....
Date: 13/11/2015

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

Name of Supervisor: John Ocheche
Signature: .....
Date: 13/11/2015

School of Finance and Applied Economics
Strathmore University

ABSTRACT

Based on Generalized Linear Models, we are able to establish the key parameters to be used for determining the future claim frequency through the adoption of appropriate logical variable selection method given various variable interactions in the model. With ample calibration, we prove the robustness of the model by analyzing its predictability power through evaluation of the dispersion parameters. This facilitates the efficiency in claim reserving given the scarcity of resources and time to run the saturated models. We adopt the statistical based variable selection method for the model as opposed to experience selection. The general insurance industry could embrace the study to economically maintain and improve the certainty in their claim reserving. Relying on the sample analysis, general insurers and particularly motor vehicle insurers could estimate their future claim frequency at about 85% confidence interval basing on the model of the car and the age bracket of the policyholder.

ACKNOWLEDGEMENT

I would like to appreciate my immediate supervisor Mr. John Ocheche for giving me guidance and direction in the entire research. He questioned the logic of the study and brought in extensive practical experience. In addition, his contribution through weekly updates and critical evaluation of this research at each point has made it success.

My friends and classmates who have contributed in assessing my research topic, guiding me through difficult analysis and evaluation of my results are appreciated for their effort. Their moral support and company has made the research very successful.

My sincere gratitude goes to Kiva and Strathmore University for considering me to pursue this course and undertake this project. Their mentorship and basic support have made the entire journey of my undergraduate degree fruitful.

Table of Contents

GLOSSARY

The following is a summary of the abbreviations used

Abbreviation	Description
GLM	Generalized Linear Model
MLE	Maximum Likelihood Estimator
L	Likelihood
Chi	Chi-square
df	Degrees of freedom

1. INTRODUCTION

1.1 Background to the study

General insurance is non-life insurance which includes mainly automobile and homeowners policies. They provide payments depending on the loss from a particular event in return for payment of a premium over a specified time period. In general insurance, a claim arise as result of occurrence of an insured event such as property destruction by fire or car accident, (Taylor, 2000).

Unlike other industries, the insurance industry does not sell products as such, but promises. An insurance policy is a promise by the insurer to the policyholder to pay for future claims for an upfront received premium. The occurrence of the specified events and the amount of payment are both usually modelled as random variables because of high uncertainty involved.

As a result insurers do not know the upfront cost of their service, but rely on historical data analysis and judgment to derive a sustainable price for their offering. In General Insurance, most policies run for a period of 12 months. However, the claims payment process can take years or even decades due to delay process of claim verification, (England & R, 2002).

Therefore, an insurer need to set aside an amount as a reserve to meet such uncertain expenses in the future. The main goal of an actuarial reserving is concerned with the financial management of the financial security systems. These can be defined as mechanisms for reducing the adverse financial impact of random events that prevent the fulfilment of reasonable expectations set by the providers of insurance products.

Claim reserving is required to satisfy the future liability. This implies that an insurance company should put sufficient provisions from the premiums for the settlement of all the claims caused by insurance contracts. The greater challenge to general insurance is that claim payments and other parameters that affect claims are not generally normally distributed. This makes it complex and almost impossible in modeling the reserves using simple models that assume normal distribution of data.

(Jian Huang, 2012) Emphasizes on the need to use a model that incorporates various parameters that predict the clam amounts at any particular time. He proposes the use of a dynamic model such as GLM that considers both parametric and non-parametric data. While modeling a survival model, it was noted that nonparametric models may suffer from the curse of dimensionality due to difficulty in presenting such a model quantitatively. They also lack easy interpretation in parametric risk models.

Generalized Linear Models (GLM) were introduced by Nelder and Wedderburn (1972) and systematically summarized by McCullagh and Nelder (1989). They are considered by various scholars and researchers to be a powerful tool to analyze the relationship between a particular response variable and covariates, (Li Wang, 2011). Given a link function, the GLM expresses the relationship between the dependent and independent variables through a linear functional form. It is a flexible model that combines both parametric and nonparametric components.

GLMs are particularly defined to be a natural generalization of the familiar classical linear models because of the prescribed distribution family that should be used. The class of GLMs includes, as special cases: linear regression, analysis-of-variance models, log-linear models for the analysis of contingency tables and logit models for binary data in the form of proportions, (Renshaw, Haberman, & E., 1996). This class of stochastic models has been on high application in recent years to a wide range of problems involving mortality, multiple-state models, lapses, premium rating and reserving, (Beirlant, Antonio, & Jan, 2012).

In support of dynamic semiparametric model, (Beirlant, Antonio, & Jan, 2012) came to conclusion that GLM with semiparametric relative risk strike a good balance by allowing nonparametric risk for some covariates and parametric risk for others. The benefit of such models seem to be two-folds, (Beirlant, Antonio, & Jan, 2012). First, they have the merits of models with parametric risk. Including easy interpretation, easy estimation and easy inference. Second, their nonparametric part allows a flexible form for some continuous covariates whose patterns are unexplored and whose contribution cannot be assessed by simple parametric models.

In claim reserving, various models can be used such as Chain ladder Models which is used to estimate a reserve to be held over small range of development years, Bayesian Models for estimating posterior claims using prior claims with the probability likelihoods, Poisson Models used to estimate likelihood of future number of claims, Distributional Models which uses probability to estimate the future changes in claims, Bornhuetter–Ferguson (BF) Model which combines the loss ratios with chain ladder method by incorporating the most current information in the model to reduce variations and last but not least is Generalized Linear Model, (Merz, 2008). Each of the above models can be used to give an a proximate estimate depending on the type and nature of an insurance products, period over which an estimate need to be made, variables provided and accuracy required

(Beirlant, Antonio, & Jan, 2012) States that GLM incorporates various variables, both parametric and nonparametric that affect future cash outflows from perils insured. Such factors includes; one or more random variables which characterize the major dimensions of risks such as duration, size, number or delay, a set of well-defined states of nature together with a deterministic or stochastic changes between the states. The parametric factors may include; economic function, associated with the underlying variables and the states and transition events, which may also be deterministic or random. They are most often linked to uncontrollable external economic conditions such as market growth, inflation and currency risk. In addition, economic performance under the control of the profit margin, investment return and portfolio performance are also considered in determining the reserves.

1.2 Purpose of the study

The main purpose of this study was to analyze the key covariates for Generalized Linear Model in claim reserving by general insurer. This was through analysis of the deviance and interactions among variables that are used to predict claim amounts. This will help insurers to feed their linear models with an appropriate data. Knowing the exact information required in developing such a model will save time in model development, reduce random errors in the model and hence accuracy can be achieved.

1.3 Problem Statement

Given the nature and high risks involved in general insurance, high prudence need to be demonstrated in determining the rating factors to be used in claims reserving model. In his research on issues in claim reserving and credibility, (Beirlant, Antonio, & Jan, 2012) states that any particular model to be used in claims reserving need to incorporate sufficient key covariates.

Despite the use of short term forecast models that makes use of absolute past observed claims to determine the current reserves to be held, an alternative is needed to incorporate other factors that can affect future claims. In this perspective, GLM is considered to be flexible and dynamic due to use of the key covariates in the model. However, the covariates need to be appropriately selected.

In particular, this study gave most attention in selection process of key covariates to be used in a Generalized Linear Model for claim reserving by general insurers. Taking in view of other stochastic models that are used in determining future claims, GLM has been evaluated on its accuracy and flexibility according this research.

1.4 Research Objectives

1. To determine the key covariates to be used in Generalized Linear Model for claim reserving by general insurers.
2. To establish an effective variable selection procedure under Generalized Linear Model.
3. To establish the predictive power of Generalized Linear Model for claim reserving in general insurance.

1.5 Research Questions

1. What are key variables to be used in Generalized Linear Model for claim reserving by a general insurer?
2. Which effective variable selection procedure or method can be used in selecting the significant factors to be used in GLM?
3. What is the predictive power of GLM in General Insurance claim reserving?

2. LITERATURE REVIEW

2.1 Introduction

The objective of this chapter is to articulate the conceptual foundation of the research. It presents ideas from the previous researchers, scholars, analysts and authors. The review focuses on theoretical, empirical framework on which this research builds on. The current models and need for GLM by insurers is also reviewed. Then research gap in the reviewed literature and the conceptual framework.

2.2 The need for basis of parameters selection

(Gray & L., 2011) Notes that effort to identify and support credibility of the causal claims have received significant interest in the research community, particularly over the past few decades. The statistical procedures designed to strongly support causal claims for treatment and intervention when the response variable of interest is dichotomous are suggested. Seven key features of generalized linear model regression studies that should play a critical role in estimating a causal effect are identified. These includes elaboration of research design, model specification, clarification of the link function, limitations and challenges of sample size, interpretation of treatment effect through ratios, statistical tests and evaluation of model fit and lastly the potential of generalized liner model in modeling the causal claims.

In statistical modelling, it is highly advisable to use various statistical techniques in deriving the inferences. (Beirlant, Antonio, & Jan, 2012) Argues that a single statistical technique cannot give the inference for a sound decision especially in selecting the parameters to be used in GLMs. To achieve high credibility in determining the causality effects on the response factors, quality methodology design is more preferred to particular statistical technique. The legitimacy of any model to yield some credible results rely heavily on the adequacy and the quality of the sample, the research problem and the research design as supported by (Jian Huang, 2012).

2.3 Consideration of Generalized Linear Model in Claim reserving

GLM is a semiparametric model that captures both qualitative and quantitative parameters used in modeling frequency and severity of claims. Fox (2008) states that pure parametric models may be useful but they are not the only key determinants in claim reserving. Parametric models have proved to be capable of describing the structure of complex actuarial and social phenomena. In general insurance, claims may not be effectively modelled using pure parametric models. This is emphasized by (Reilly, 2012) who states that regardless of the elegance of a given parametric model, there are few single parametric model capable of solving a problem in claim reserving. This gives us an insight to move beyond pure parametric models to semiparametric model that can give more accurate results due to current complexity of insurance products.

The current financial products data that has required the need for semiparametric model is the one that has binary component especially in generating the rating factors. The description of binary outcomes and data can be best modelled by Generalized Linear Model. This has been mainly used in determining the rating factors in premium calculation. In determining these rating factors, various interactions among the covariates are used in the model. This leads us to the Rubin's Causal Model framework by Little and Rubin (2002) which describes the problem in estimation of causal effect with pure independence assumption among parameters to be used in modelling. However, in general insurance, we expect variables to be mostly dependent and correlated. (Guo and Fraser, 2010) highlights that essential to Robin's model is the assumption of independence among the parameters under the study which is known as stable unit treatment value assumption. This assumption implies that one parameter's effect to the response factor has no effect to the other parameter. Therefore, GLM may give universal result and inference despite the participant subjectivity because interactions among the parameters are factored in the model as it has been done in this research.

2.4 Empirical framework

There are two possible approaches suggested by (Jian Huang, 2012) for estimating the parametric component and nonparametric component in GLM. The first is an application of the marginal integration approach to the nonparametric component of the GLM. They treated the summand of additive terms as a nonparametric component, which is then estimated as a multivariate nonparametric function. This strategy may be insufficient when the number of additive terms is not small. The other method of variable selection proposed by Buja, Hastie and Tibshirani (1989) is a combination of local scoring and kernel-based back fitting. The inefficiency with this method is the fact that it requires a large number of mathematical equations to be solved as observed by [Yu, Park and Mammen (2008)]. In addition, penalized function for variable selection may be difficult to be introduced in this method.

Therefore, kernel-based back fitting together with marginal integration methods seem to be expensive and time consuming to compute due to large system of equations to be solved. Ruppert, Wand and Carroll (2013) and Wood (2014) studied penalized regression splines which shared most of the practical benefit of smoothing spline methods. The time and computational cost may be overcome by use of R/S plus packages in practice for convenient implementation. However, (Jian Huang, 2012) observed that no theoretical justifications are available for these procedures in additive case. In GLM, additivity need to be made possible by connecting mean response to the linear predictors.

To select significant variables in semiparametric models, (Li Wang, 2011) used variable selection procedures for parametric models used by Fan and Li's (2001) via non concave penalized quasi-likelihood. This is where a large number of variables may be collected and the ones that are not significant are eliminated in forming appropriate required model. It is emphasized that it is important to select significant variables for both parametric and nonparametric regression models. However, their models do not cover the GLM. Therefore, traditional variable selection procedures such as stepwise deletion and subset selection may be extended to the GLM although they may be computationally expensive.

(Pang Du, 2010) Carried out research in penalized variable selection procedure in Cox model with semiparametric risk and noted that traditional procedures such as Akaike information criterion (AIC) and Bayesian information criterion could be used but have some deficiencies. They lack stability in addition to failure to incorporate random errors derived from the initial stage of variable selection. Therefore, GLM need to adopt a process that can capture the stochastic errors in variable selection. It also need to be shown that the dependence of mean response over covariates have the linearity form assumption.

2.5 Literature Review Summary

The support of argument for GLM over other stochastic models has been provided to encompass: its dynamic nature and flexibility to incorporate both parametric and nonparametric data and the power of GLM to model binary data.

The traditional variable selection procedures reviewed incudes; Bayesian information criterion and Akaike information criterion to be used in case there is no stochastic error to be incorporated in the model, Lasso and adaptive Lasso variable selection mostly applied in Cox models, non-concave penalized likelihood approach proposed to be used in Cox with Proportional Hazard models, marginal integration for nonparametric components, Kernel-based back fitting and local scoring approach for parametric components and polynomial spline smoothing method.

Therefore, there is a need for research to determine the appropriate Generalized Linear Model design for claim reserving to fill the gap.

3. METHODOLOGY

3.1 Introduction

This chapter addresses the methodology used to investigate the subject addressed. The rationale upon which the method of inquiry rests is highlighted together with hypothesis amongst the variables. Appropriate scientific models to test for the existence of hypothetical relationship are proposed herein.

This study was quantitative in nature hence it has dealt with mostly numerical values to arrive at the results.

3.2 Population and Sampling Method

3.2.1 Population

This study sought to implement the covariates selection for GLM in general insurance. In Kenya, there are forty nine insurance companies of which only seven focus only on life insurance. The remaining forty two carry out general insurance as their part of business.

3.2.2 Sample and Sampling Method

The universe of the study was the given forty-nine insurance companies in the Kenyan economy. The sample unit is a single general insurance company. The key focus is the list of general insurance. Non probabilistic sampling method was used in sample selection. Non-probability sampling was used because strongly established company was required to be selected as a good representative of a general insurance provider in the industry. The general insurance selected was Jubilee Insurance and motor vehicle insurance claims data for 2010-2014 was filtered out for the purpose of this study. Data from this general insurance provider was considered to be stable compared to young small general insurers in Kenya.

3.2.3 Data collection methods and procedures

Data on the variables of interest in this study includes the observed number of claims especially on motor vehicle insurance, the characteristics of both the asset insured and the policyholder, the premium rating factors used and other factors used in claim acceleration. This data was collected from primary sources kept by Jubilee Insurance.

3.3 Research Design

This gives an in-depth view of the underlying theoretical framework to the Generalized Linear Models, variable selection procedures and the reliability of adopted models for this study.

3.3.1 Fundamentals of Generalized Linear Models

In GLM, the distribution of data is assumed to be non-normal. This is particularly important in actuarial work where data very often do not have a normal distribution. In mortality, the Poisson distribution is used in modelling the force of mortality, μ_x while binomial distribution for the initial rate of mortality, q_x. On the other hand, in general insurance, the Poisson distribution is often used for modelling the claim frequency and gamma or lognormal distribution for the claim severity.

The aims of data analysis exercise in this research was to decide which variables or factors are important predictors for the risk being considered and then to quantify the relationship between the predictors and the risk in order to assess the appropriate claim reserve levels.

GLM relates a variable called a response variable which was the main interest of this research to be modelled, to variables or factors called predictors or covariates or independent variables which we have information about. It uses the sample data to define the relationship between the mean μ of the response variable and the covariates. In order to do this, we first define the distribution of the average number and size of claims to be of an exponential family which is used in GLMs. The covariates will be allowed to relate to the claim severity or claim frequency.

The exponential family as stated by (Jian Huang, 2012) includes the normal, Poisson, binomial, gamma and exponential distributions. The distribution of a random variable Y is said to belong to an exponential family if it has its density of the following form:

$$f_\gamma(y; \Theta, \varphi) = \exp\left[\frac{(y\Theta - b(\Theta))}{a(\varphi)} + c(y, \varphi)\right] \quad (1)$$

Where a, b and c are functions. The two parameters in the above density are Θ, which is called the natural parameter, is the one which is relevant to the model for relating the response (Y) to the covariates, and φ is known as the scale parameter or dispersion parameter. Because of the assumption of exponential family, Θ is a function of $\mu = E(Y)$ only.

3.3.2 Ingredients of Generalized Linear Model

3.3.2.1 A distribution for the data

It is assumed that the distribution of data falls under the category of exponential families. (Beirlant, Antonio, & J, 2007) Advocates in this case that gamma distribution could be selected to model the size of motor insurance claims, Poisson distribution to model the number of claims and normal distribution to model the a. However, the distribution of the data was defined on outset and it was observed that claim frequency follows Gaussian distribution.

3.3.2.2 A linear predictor

The covariates, also known as explanatory or predictor variables, enter the model through the linear predictor. This is where the parameters occur which have to be estimated.

The linear predictor η is a function of the covariates. In our case,

$$\eta = \beta_0 + \beta_1 w + \beta_2 x + \beta_3 y + \ldots\ldots \quad (2)$$

Where η was assumed to be claim frequency in this research with covariates; the type of car insured, age of policyholder, gender of the main driver and location of residence of the policyholder.

Looking deeper into the model, there are two types of covariates in GLM; variables and factors. A variable is a type of covariate whose real numerical value enters the linear predictor directly. Examples of variables in car insurance context are age, mileage or number of years in which driving license has been held. The other main type of covariate is a factor. It takes categorical value. It includes; sex of the policyholder or car type. This type of covariate was parameterized so that the linear predictor has a term a_1 for male and a_2 for female.

Therefore, a model which includes an age effect and an effect for sex of the policyholder could have a linear predictor combining both age variable and gender factor. More interactions were put in the model and solved using the appropriate statistical computer package.

3.3.2.3 Link function

We need to connect the mean response to the linear predictor, E(Y) = Linear predictor.

$$\mu = \eta = \beta_0 + \beta_1 w + \beta_2 x + \beta_3 y + \ldots\ldots \quad (3)$$

We therefore take some function of the mean response which is the link function that gives the relationship $g(\mu) = \eta$. Where g is the link function and η is the linear predictor. The link function is the missing link because in GLM, we are trying to determine a relationship between the mean of the response variable and the covariates.

By setting the link function $g(\mu) = \eta$, then, assuming that the link function is invertible, we can then make μ the subject of the formula.

$$\mu = g^{-1}(\eta) \quad (4)$$

Technically, it is necessary for the link function to be differentiable and invertible in order to fit a model.

Inverting the formula helps us have the relationship between the mean of the response variable and the covariate in an exponential form by using an appropriate canonical link function. For each exponential distribution family, the natural or canonical link function is defined by $g(\mu) = \Theta(\mu)$. Θ is the natural parameter for the exponential family form and that Θ is a function of the mean of the distribution μ.

Therefore, for any value η there is a unique value of μ to make prediction about the future. The canonical link functions for the possible models to be used in these case are:

Table 3.1: Prescribed link functions

Probability	Link name	Function
Normal	Identity	$g(\mu) = \mu$
Poisson	Log	$g(\mu) = \log \mu$
Binomial	Logit	$g(\mu) = \log [\frac{\mu}{1-\mu}]$
Gamma	Inverse (Reciprocal)	$g(\mu) = \frac{1}{\mu}$

Other link functions exist and can be quite complex for specific modelling purposes. As a basis of this research, the above four functions were considered to be sufficient as guide in analysis of this research. These link functions work well with each of the above distributions but it not obligatory that they be used in each case. However, the implications of the choice of the link function on the possible values for μ need to be considered. For example, if the data have Poisson distribution, then μ must be positive. If the log link function is used, then $\eta = \log \mu$ and $\mu = e^{\eta}$. Thus μ is guaranteed to be positive, whatever value such as positive or negative the linear predictor takes. The same is not true if identity link function is used. In this research, claim frequency assumed the normal distribution hence identity link function was implied to be used.

3.3.3 Deviance of model fitting

The parameters in this study were estimated using maximum likelihood estimation. The log-likelihood function $l(y; \Theta, \varphi) = \log (f_Y(y; \Theta, \varphi))$, depends on the parameters in the linear predictor through a link function. The maximum likelihood estimates can be obtained by maximizing this function with respect to the parameter in the linear predictor.

3.3.4 Model selection

The likelihood of the possible models designed was compared with the likelihood under the saturated model to assess the adequacy of the model. For comparisons, the key assumption was that the saturated model used the same distribution and link function as the selected model.

If the likelihood of the design model is equal or almost equal to the likelihood under saturated model, then the model is concluded to be a good model. On the other hand, the log of the likelihood ratio could be used in model selection. The natural log of the likelihood ratio statistic is;

$$\text{Log}\frac{Ls}{Lm} = ls - lm \tag{5}$$

Where Ls = log Ls and Lm = log Lm

Given a sample size n, the log-likelihood is;

$$L(y; \Theta, \varphi) = \sum_{i=0}^{n} \log f_Y(y; \theta_i, \varphi_i)$$

$$= -\frac{n}{2}\log 2\prod\sigma^2 - \sum_{i=1}^{n} \frac{(y_i-\theta_i)^2}{2\sigma^2} \tag{6}$$

Then the scaled deviance is used to make an inference about the model. In GLM, scaled deviance is defined as twice the difference between the log-likelihood of the current model and the saturated model. The scaled deviance is the absolute value of the difference in the log-likelihood. Since the log-likelihood of the saturated model will always be greater than or equal to that for the current model, the scaled deviance S can be expressed as;

$$S = 2(ls - lm) \tag{7}$$

The deviance for a range of models was observed for decision making. From point of view of the model fit, the smaller the deviance, the better the model.

3.3.5 Residuals analysis and assessment of the model fit

Upon selection of the possible model using the deviances, it was then checked by looking at the residuals and the significance of the parameters. The residuals are always based on the difference between the observed responses y_i and the fitted responses $\tilde{u}$. These deviance residuals were plotted and their distribution observed.

The deviance table is suitable in identifying the model that is suitable for the data under consideration. The selected model should be examined by looking at residuals plot and significance of the parameters. The assumption of GLM require that the residuals should show no patterns. The presence of pattern implies that something has been missed in the relationship between the predictor and the response. In case of such pattern, other model specification were to be tried out. In addition, plots of the residuals against the variables and factors should be examined for patterns and other outliers.

3.3.6 Goodness of fit

Statistical tests were used for the acceptability of a particular model once fitted. These included Pearson's Chi-square test and Likelihood ratio test. An empirical comparison was made between the statistical parameters of the model and standard parameters from probability table.

4. RESULTS AND DATA ANALYSIS

4.1 Preliminary Analysis

The data collected and considered in this analysis is from Jubilee Insurance. It contains the risk factors categorized into homogenous classes for various claims. The variables provided are; gender of the policyholder, class of car, the age of policyholder and the district where policyholder lived. For analysis purpose, gender is classified as either male or female. Car groups were classified as either small, medium, semi-large and large. Age of policyholders was categorized into four cohorts; class 1 (18-29) years, class 2 (30-42) years, class 3 (43-65) years and class 4 (Above 65) years. The districts where the policyholder lived were classified into two, Nairobi and major cities or other towns other than the cities.

The main justification of these categorization was to simplify the analysis and incorporate these variables into GLM appropriately. The grouping of age is based on the variation of risk characteristics of individuals across different ages.

Basing on the data collected, the claim frequency is the main focus of analysis in answering the research questions and hence meet the objectives of this research. The number of claims between 2010 and 2014 is provided together with the number of policies written that resulted to the claims. With this information, the average number of claims was modelled.

4.2 Determination of key parameters

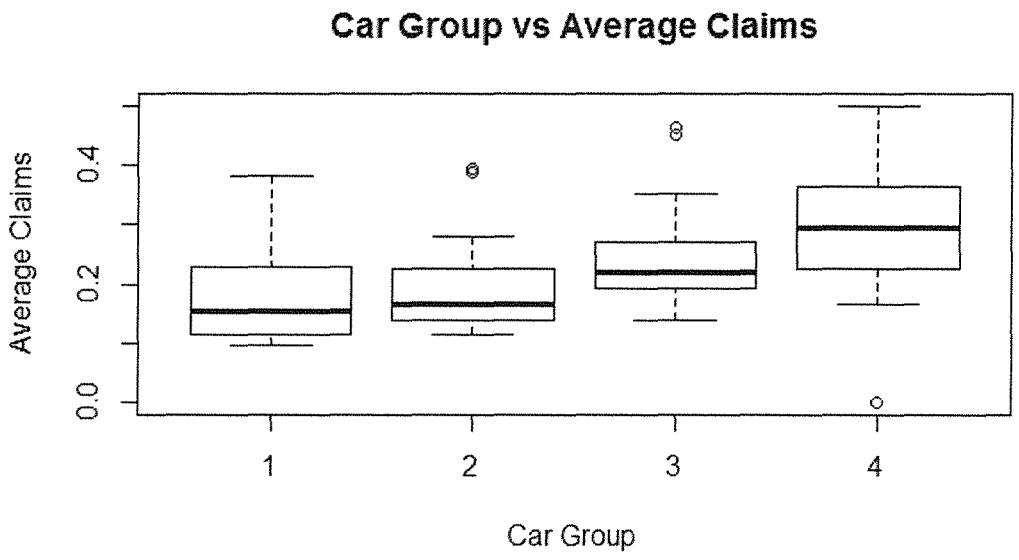
Given the age and gender of the policyholder, the district lived by the policyholder and category of the car, we need to draw the box plot to establish the relationship between these parameters and the response factor which is the average number of claims. The main focus is the average number of claims computed as number of claims divide by the number of policies.

4.2.1 Box plot Analysis

The following box plots were plotted to show the relationship and effect of claim frequency with various variables.

4.2.1.1 Car Group and Claim Frequency

Table 4.1



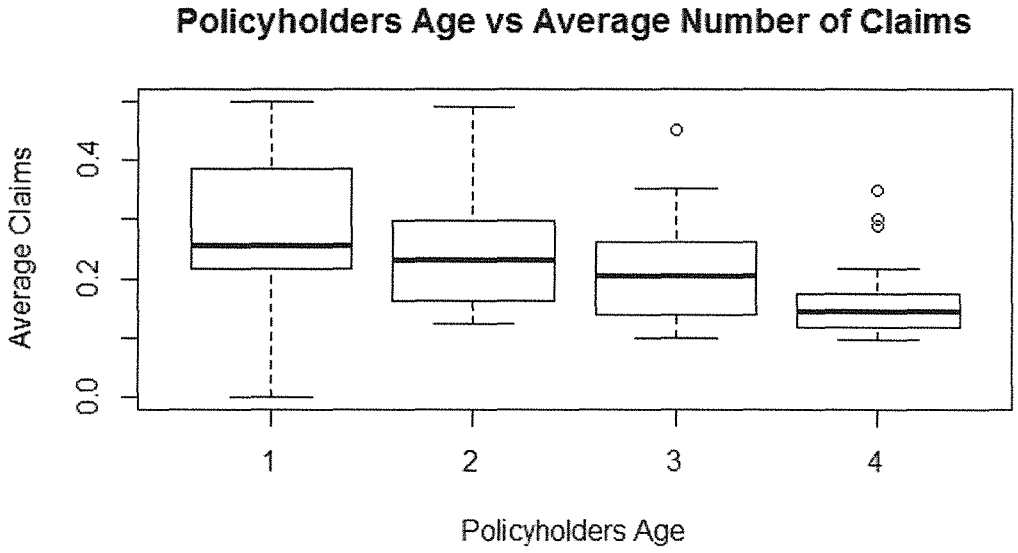
The average number of claims varies across the motor vehicle classes. The claim frequency increases with car size. According to these results, smaller sized cars have a smaller probability of claiming compared to large sized cars. The probability of making a claim increases as the car size moves from group 1 to group 4. Therefore, group 4 which contains large sized cars have a higher probability of causing an accident and hence claim being launched followed by group 3, 2 then 1. However, most of small cars and semi-large will have more than the average number of claims with few making extremely less than average number of claims. On the other hand, most of large cars will have less than average number of claims with only a few making extremely large number of claims.

Small and semi-large vehicles have a very small deviation in the average number of claims. Therefore, claim frequency for the two classes of cars could me estimated with high accuracy in the model being achieved. This is not the case for large vehicles which were observed to have a very high variation in the average number of claims. This class proves to have a very high standard deviation in average number of claims. Therefore, there could be a high risk in modeling the average number of claims for car group 4. This might pose a high risk on reserving as the amount that could be reserved for this class might not match the future experience.

In relation to generalized linear model, car group is a key parameter in estimating the average number of claims by general insurers. This is based on the findings above where the average number of claims tend to vary significantly with the group of car insured.

4.2.1.2 Age of the policyholder and Claim Frequency

Table 4.2



Claim frequency varies with the age of the policyholder in motor vehicle insurance. The age group between 18 years old and 29 years old was observed to have the highest average number of claims than all other groups. Policyholders aged 30-42 and 43-65 have a relatively lover probability of making a claim and hence lower risk than their junior

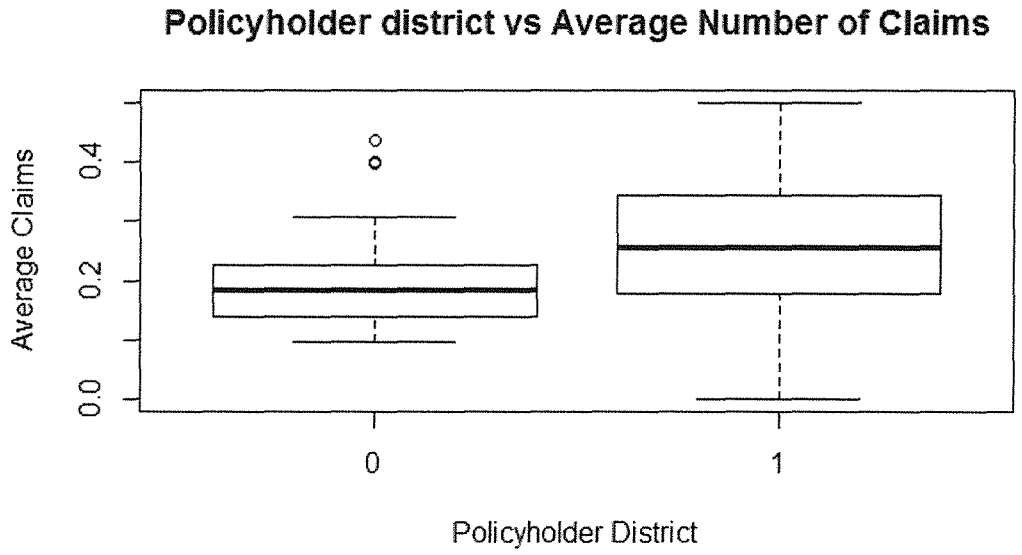
counterparts. Persons aged 65 and above have minimum risk in motor vehicle insurance basing on these statistics. However, the policyholders aged 43-65 have a slightly higher risk of causing an accident and hence making a claim as compared to those aged 30-42 and those aged above 65.

The results indicate that there was a huge difference in the average number of claims for policyholders aged 18-29 and 43-65. However, the average number of claims tend to be smooth with small deviation in the number of claims made. Therefore, claim frequency for policyholders aged 65 and above could be modelled with high certainty.

Policyholder's age is therefore considered one of key variable to be considered in determining the reserves to be held by general insurers to pay for the expected future claims. Depending on the main age bracket of the majority policyholders, the insurer is able to estimate the average number of claims expected and the level of certainty to be put on the projections.

4.2.1.3 Policyholder's district and claim frequency

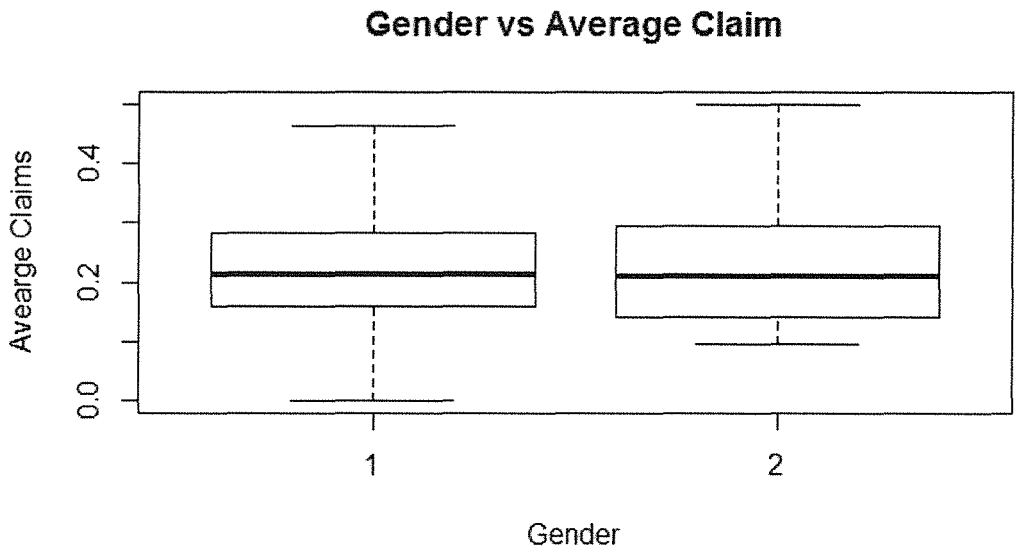
Table 4.3



The average number of claims does not vary with the district where the policyholder lived. However, the majority of policyholder were observed to reside in Nairobi and other major cities. Quite a few of policyholders live in other town in the country. This observation reflects the national wide statistics on penetration of general insurance and the main target market for the products. Most of policyholders are based mainly in the cities with a few in other small towns and upcountry.

4.2.1.4 Policyholder's gender and claim frequency

Table 4.4



The analysis indicated above for gender impact to claim severity is a clear indication that the average number of claims does not vary with gender. However, the number of women who took out insurance was relatively higher than the number of men who took out insurance. This results were in conjunction with (Reilly, 2012) who indicates that women are more risk averse in nature compared to men. Therefore, more women in realty tend to take up an insurance cover as compared to men.

Basing on the results, this cannot be a key variable GLM in claim reserving for motor vehicle insurance because gender does not have any impact on the average number of claims. This information could only be useful in marketing the motor vehicle insurance products because it guides the main target market.

4.2.2 Analysis of other variables

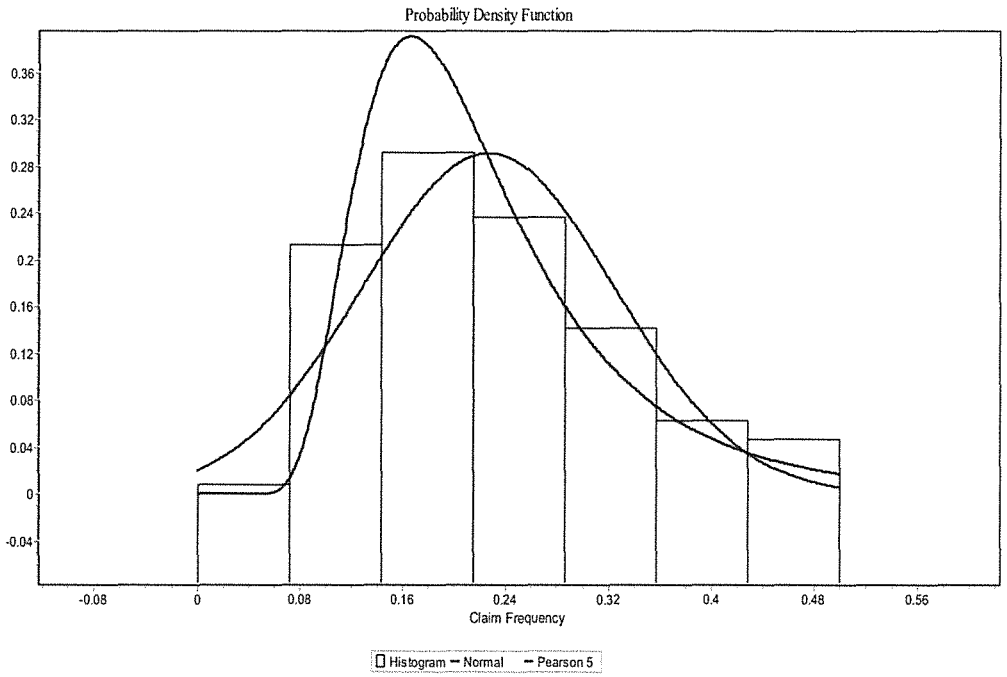
Some variables were not considered in this analysis because of various reasons. The occupation and alcoholic status of the policyholder were not considered in this analysis. The impact they have on claim frequency is minimal and proving such characteristics tend to outweigh the benefit of their consideration.

In addition, different insurers do use different classification of such policyholder characteristic. This makes it difficult to differentiate their respective definition of who is considered to be alcoholic. To avoid this ambiguity and make the model more parsimonious, such character traits are disregarded.

4.3 Model Selection

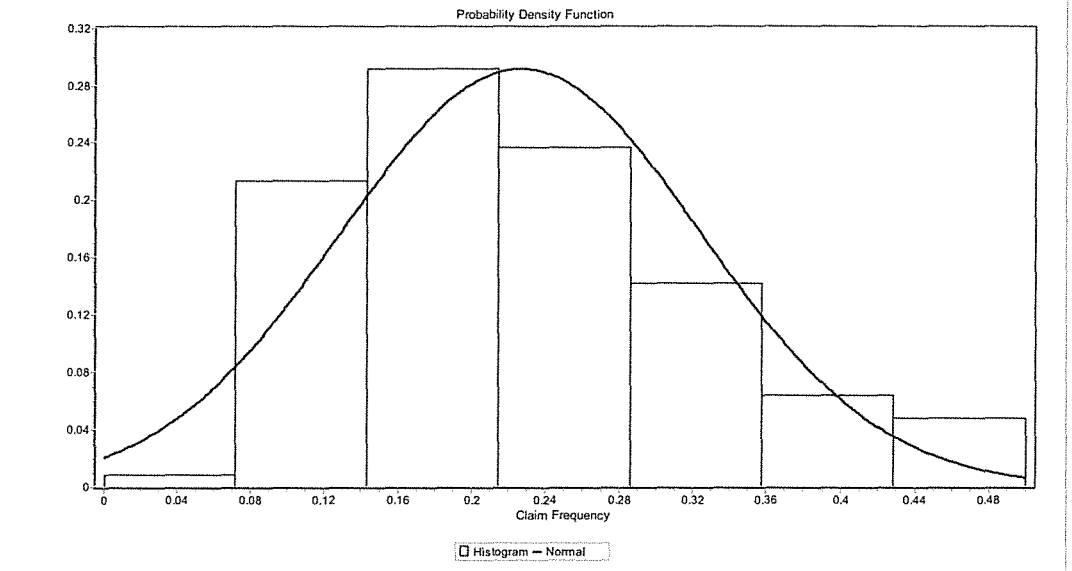
4.3.1 Distribution of data

Table 4.3 a): Distribution of Claim Frequency



- Pearson 5 as an original best fitted distribution.
- Normal distribution as an estimate fitted distribution.

Table 4.3 b): Fitted distribution of Claim Frequency



The graphs above shows the fitted distribution of claim frequency on the probability density function. Using the data provided, the best probability distribution fit was Pearson 5 as indicated in graph 4.3 a) which is always approximated to normal distribution for computation purpose.

The normal distribution model for the average number of claims was considered to be optimal among the exponential family. The following reasons justify the selection of Gaussian distribution:

1. The law of central limit theorem states that a given probability distribution tends to follow normal distribution if the sample size is large enough such that n>20. The analysis was based on a sample of 127 items hence it is large enough to assume normal distribution.
2. Although (Pang Du, 2010) states that the number of claims in general insurance follows Poisson distribution, the average number of claims does not follow the same principles. Poisson distribution is discrete distribution whereas the average number of claims follow a continuous probability distribution. The average number of claims indicated to be on a continuous state space hence the preferred exponential distribution family that can be relative to Poisson is normal distribution.

3. Gaussian distribution is one of the exponential family distribution that is used to model the average number of claims using GLM in general insurance, (England & R, 2002). This could therefore be used to particularly modelling claim reserves in GLM for motor vehicle insurance.

4.3.2 Residual deviance analysis

The relationship between different variables are combined together and their respective impact on the average number of claims recorded by observing the changes in the degree of freedom and residual deviances. The objective of combining the parameters of the model together while recording their impact on the response variable is to estimate a parsimonious saturated model. Interactions are added until the model is saturated if saturation can be achieved.

An optimal model to be considered should be the one that imitates the saturated model with sufficient key parameters that are significant to the model.

4.3.2.1 Degree of freedom and residual deviance analysis

The interactions among the parameters from the obtained variables yielded below results and used in selection of the optimal model. Because saturation could not be achieved using the Pearson and Residual indices, an alternative method was used in model selection. If saturation cannot be achieved as was in this case, the best model was selected basing on rule of relationship between difference of degrees of freedom and difference in the deviance.

The rule: The best model should be selected if twice the difference of degrees of freedom starts being greater than the difference in the deviance.

Table 4.7

The table below was generated to give as a summary of the deviance and degrees of freedom.

Model X – Saturated model that contains all provide covariates

pa – policyholder age

cg – car group

pd – policyholder district

gn – policyholder gender

Table: Degree of freedom and residual deviance analysis

			Differences		
Formula	Deviance	d.o.f	deviance	d.o.f	2(p-q)
Model X 1	1.073E-28	0			
			-47.7430	-124	-248
Model1 pa	47.743	124			
			0.9990	0	0
Model2 cg	46.744	124			
			24.8160	1	2
Model3 cg+pa	21.928	123			
			10.8680	1	2
Model4 cg+pa+pd	11.060	122			
			0.3460	1	2
Model5 cg+pa+pd+cg.pa	10.714	121			
			0.6220	1	2
Model6 cg+pa+pd+cg.pa+cg.pd	10.092	120			
			1.3845	1	2
Model7 cg+pa+pd+cg.pa+cg.pd+pa.pd	8.7075	119			
			0.1601	1	2
Model8 cg+pa+pd+cg.pa+cg.pd+pa.pd+cg.pa.pd	8.5474	118			

The residual deviance decreases as more variable interactions are put in the model. It was observed that policyholders' age and the group of car have almost an equal residual deviance while district in which the policyholder lived and gender had a minimal impact on the change of the residual deviance. This indicated that the main key variables to be considered in the GLMs are policyholders' age and the car group.

Policyholders' age has the highest residual deviance, 47.743. Therefore, this is the most sensible variable to the claim frequency and hence attracts the highest consideration in claim reserving using the generalized linear model. The immediate next key parameter to be considered in modelling the reserves should relate to car group with the residual deviance of 46.744.

4.3.2.2 Inference to degree of freedom and residual deviance analysis

Basing on the degrees of freedom and residuals analysis, the following observations were made:

1. The Pearson and the deviance residuals were generally identical.
2. Deviance residuals are almost symmetrically distributed and have approximately normal distribution.
3. Residual plots of the models show no patterns.

These observations are in accordance with findings made by (Beirlant, Antonio, & J, 2007). In GLM, the three conditions are key conditions in selecting the parameters of the model. In addition, the conditions are used in establishing if the analysis meet the criteria underlying GLMs.

The most parsimonious model is model 3 which has a combination of policyholders' age and the cat group. This model was selected on the basis that this was the point where double the difference of degrees of freedom are greater than the difference of residual deviance. Therefore, it was concluded that car group and policyholders' age are the main parameter to be into the generalized linear model for estimating any future average number of claims.

The table below summarizes the residual deviance analysis for the model selected.

Table 4.8

```
Analysis of Deviance Table

Model 1: ac ~ cg
Model 2: ac ~ cg + pa
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       124     46.744
2       123     21.928  1   24.816 < 2.2e-16 ***
```

4.3.2.3 Summary analysis of model 3 deviance residual and coefficients
Table 4.9

```
Deviance Residuals:
     Min       1Q   Median       3Q      Max
-0.62477 -0.20116  0.05062  0.38570  1.14119

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.219597   0.015362  14.295  < 2e-16 ***
cg           0.034544   0.004023   8.586 3.32e-14 ***
pa          -0.041757   0.003539 -11.798  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.1782793)

    Null deviance: 59.939  on 125  degrees of freedom
Residual deviance: 21.928  on 123  degrees of freedom
```

This model gave dispersion parameter of 17.8 % compared to the target dispersion parameter of 0%. Therefore, the confidence of the model lies around 94.2% (100-17.8) in projecting the claim frequency using GLM for motor vehicle insurance using policyholders' age and car group as the key variables and parameters.

5. DISCUSSION, RECOMMENDATIONS AND CONCLUSIONS

5.1 Introduction

The nature of general insurance business mainly involves high risks and uncertainties, (Merz, 2008); hence the need for design of appropriate variable selection methodology in modeling the claim frequency. While establishing the claim reserving problem using generalized linear models for general insurance business, this analysis narrowed down to establish the procedure of selecting the key variables and the reliance that could be put on the model using motor vehicle insurance data. The analysis extended to determine the significance and impact of each variable to the claim frequency as the response factor. The analysis of these variable in particular to GLMs is key tool in risk measurement in general insurance, (Li Wang, 2011).

This paper sought to establish the impact of policyholder characteristics as main variable in GLM to claim frequency as a response factor. The analysis informs the Kenyan general insurers on the key variable to be considered in reserving for motor vehicle insurance claims with the strength to be put on each variable.

From literature review, it was evident that usage of GLM in estimating claim frequency for claim reserving purpose is a key tool for claim management in motor vehicle insurance by general insurers in Kenya. Appropriate projection of claim frequency and claim severity enables development of reliable reserving policies for the company hedging it against various risks. GLM was used as an appropriate tool of using the policyholder characteristics to quantify their respective impact to the claim frequency. The implied link function is identity link function for the normal distribution used to model the claim frequency.

The preceding chapter establishes the GLMs approach to model the claim frequency and determination of the key variables for the model. The plot box were used to establish the key variables to be used in the generalized linear model by analyzing their significance and relationship with the average number of claims for motor vehicle insurance. Null and residual deviance together with chi-square analysis were used to evaluate the impact of aggregating the provided variables to the claim frequency. In addition, the quantile and

residual plots we used to evaluate the dispersion of the estimates from the saturated model results.

This analysis is in line with the current concern for claim reserving as risk management strategy in the Kenyan general insurance industry. However, the key questions arising from these findings are; what are the key variables to be used in the GLM for modeling claim frequency? What is the prescribed methodology in selecting these variables? To what extend is the GLM prediction reliable? This chapter addresses above questions, evaluate the main constraints relating to this research then winds up with the recommendations and concluding remarks.

5.2 Discussion

Table 4.1, 4.2, 4.3 and 4.4 provides the detail of the relationship between the single variables and the response factor. This enable an establishment of the deterministic view of variable relationship with the dependent parameter of GLM. The results were obtained from the simple plot of the box plot using the car insurance data provided. It was established that car group and policyholders' age were the key variables that affect the average number of claims. On the other hand, the district lived by the policyholder and gender had insignificant impact to the claim frequency. The results are in accordance with the conclusions made by (Jewell, 1980) and a similar theory backed up by (Kunkler, 2004). However, although gender in this case did not show any effect to the average number of claims, (Pang Du, 2010) established a strong relationship to the average number of claims in Cox model for mortality analysis.

Use of deviance analysis in GLM was applied to conduct statistical impact of the combination of the variables to the response factor. This research adopted (Qian, Garbor, & R, 1996) method of variable selection by focusing on the change in the null deviance and residual deviance. The response factor as a component of GLM was specified to be the average number of claims. The deviance and degrees of freedom were observed to decrease as more interactions are incorporated in the model, (Qian, Garbor, & R, 1996). The objective was to establish which variable interactions yield a parsimonious model that could be used for reserving purpose for motor vehicle insurance. Using these statistical measures for variable selection, the appropriate interactions that yield an appropriate model

were the policyholders' age the car group with the dispersion factor of 17.8%. However, if a more conservative confidence interval is required by the general insurer such as 90% confidence interval, this model will be deemed redundant.

Table 4.5 and 4.6 shows the results of the plotted residuals deviance and residuals Pearson for the saturated model. This analysis was used as a data verification tool to establish any error in data and the model that was being mimicked. The results indicated the normal expected trend and desired probability distribution of errors as per the GLM assumptions. The saturated model gave accurate results with nil dispersion parameter hence it was deduced that quality data with no seen anomalies was collected. The model that was designed and selected reflects the saturated model with dispersion parameter of 17.8% which was tolerable, (Antonio, J, & T, 2005).

This research established self-explanatory and efficient methodology of variable selection by use of both statistical and empirical approach. The box plot generated gave an initial analysis of the relationships for single variables while the deviance analysis gave the final section with appropriate quantification of the impact of variable interaction in the generalized linear model. Such simplistic approach was derived by (Kaplan, 1958) and applied by (Pang Du, 2010) in Cox regression model in life insurance which could be extended to be applied in general insurance. However, (Doray, 1994) and (Gray & L., 2011) uses more complicated stochastic approaches.

5.3 Conclusions

This research was based on the data collected from an established general insurer in Kenya with a diversified risk portfolio. The key variable selection methodology adopted can be used as a guide by other insurers in the industry for statistical claim modelling. The use of this methodology saves time and replication of efforts in the industry is minimized. In addition, ease of computation with less computer processing power being required is an added advantage for this method. Various computer soft wares such as R and Easy Fit are easily integrated in the process therefore enhancing data analysis.

The analysis revealed that the main key variables to be used by a general insurer for fitting generalized linear models for clam frequency modelling are the age of the main driver or the policyholder and the model of the motor vehicle. In addition, the distribution of claim frequency for motor vehicle insurance was found to be Gaussian distribution with a mean of 0.2266 and standard deviation of 0.09787. However, the mean and standard error only apply to the particular company whose data was used.

The predictability power of GLM was seen to be about 82.2%. This confidence interval may seem to be quite satisfactory given the stochastic nature of general insurance business. This degree of assurance about the future experience of the general insurance claims can be used for appropriate reserving and planning purpose. However, the results might be slightly different among different general insurers and appropriate caution have to be observed while generalizing the results in the general insurance industry.

GLM with semiparametric relative risk scores a good balance in risk measurement by allowing nonparametric risk for some covariates and parametric risk for others, (Beirlant, Antonio, & Jan, 2012). This had two good benefits. First, they had the merits of models with parametric risk. Including easy interpretation, easy estimation in addition to easy inference. Second, their nonparametric component allowed a flexible form for some continuous covariates whose patterns are unexplored and whose contribution cannot be assessed by simple parametric models.

5.4 Limitation to the research

This research was faced with some constraints relating to data. There was inadequate data for some companies in the sample selected which limited our choice in the selection of the particular company to be used. Therefore, companies with insufficient data to facilitate this research were left out.

Secondly, the frequency of the data available was too low in that, even for the aggregated data, only annual data was available. Therefore, all variables had to be presented in an annual format. This was mostly disadvantageous for incorporating lapses in the policies as it was difficult to incorporate their impact in claim reserving. This general aggregation of data for the whole year without indicating the particular dates of transactions could not facilitate monthly or quarterly analysis. Annual analysis of data was the only possible solution.

The strict assumptions of GLM such as limiting the distribution of data to the five classes of exponential family limited the freedom in carrying out this research. In addition, the assumption that only the available variables are the only one that could mainly affect the claim frequency could not be conclusive for all the years. Some factors that are not quantified could significantly affect claim frequency at any particular time.

References

Antonio, K., J, B., & T, H. (2005). Discussion of "A Bayesian Generalized Linear Models for the Bornhuetter- Ferguson Method of Claims Reserving.". *North American Actuarial Journal, 9(3)*, 143-145.

Beirlant, Antonio, & J. (2007). *Insurance: Mathematics and Economics, 40(1) Actuarial Statistics With Generalized Linear Mixed Models*, 58-77.

Beirlant, Antonio, K., & Jan. (2012). Issues In Claims Reserving and Credibility. *The Journal of Risk And Insurance*, 643-646.

Doray, L. (1994). Reserve Under a Loglinear Location-Scale Regression Model. *Casuality Actuarial Science Forum*, 609-652.

England, P., & R, J. V. (2002). Stochastic Claim Reserving in General Insurance. *British Actuarial Journal, 8*, 443-544.

Gray, A. O., & L., D. (2011). Cause and Event: Supporting causal claims through logistic models. *Journal of Educational Psychology*, 245-261.

Jewell, W. S. (1980). Models in Insurance: Paradigms, puzzles, communications and revolutions. *Trans. Intnational Congress Actuaries, 5*, 87-141.

Jian Huang, J. L. (2012). Variable Seection in Nonparametric Additive Models. *The Annals of Statistics*, 228-242.

Kaplan, E. L. (1958). Nonparametric estimation from incomplete Observations. *J Am. Statist. Ass, 53*, 457-481.

Kunkler, M. (2004). Modelling Zeros in Stochastic Reserving Models. *Insurance: Mathematics and Economics, 34(1)*, 23-35.

Li Wang, X. L. (2011). Estimation and variable selection for Generalized Additive Partial Linear Models. *Journal of Mathematical Statistics*, 1827-1851.

McCullagh, P. (2007). Generalised Linear Models. *Journal Economic Statistics*, 47-68.

Merz, M. V. (2008). *Stochastic Claims Reserving In Insurance*. West Sussex, Enland: Wiley & Sons Ltd.

Pang Du, S. M. (2010). Penalized Variable Selection Procedure for Cox Models with semiparametric relative risk. *Journal of Mathematical Statistics*, 2092-2117.

Pourhamadi, M. (2000). Generalised Linear Models for Multivariate Normal Covariance Matrix. *Journal of Political Science, 82 (2)*, 425-435.

Qian, G., Garbor, G., & R, P. G. (1996). Generalised Linear Model Selection by the Predictive Least Quasi-deviance criterion. *Journal of Political Science, 8*, 41-54.

Reilly, F. K. (2012). *Investment Analysis and Portfolio Management*. South-Western College Publisher.

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

Stein, A, K., Froot, & C, J. (1998). Risk Management, Capital Budgeting and Capital Structure Policy for Financial Institutions: An Intergrated approach. *Journal of Financial Economics, 47 (1)*, 55-82.

The Actuarial Education Company. (2014). *CT6.* Oxford.