University of Nairobi

School of Mathematics

ISSN: 2410-1397

Master Dissertation in Mathematics

Modelling of Auto Insurance Claims Using Discrete Probability Distributions

Research Report in Mathematics, Number 25 , 2016

Vanis Kemunto Makori

December 2017

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

Master Dissertation in Mathematics

University of Nairobi
December 2017

Modelling of Auto Insurance Claims Using Discrete Probability Distributions

Research Report in Mathematics, Number 25 , 2016

Vanis Kemunto Makori

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

Prepared for	The Director Board Postgraduate Studies University of Nairobi
Monitored by	Director, School of Mathematics

Abstract

Within general insurance, pricing of premiums is always a challenging task. Frequency of claims plays a big part in pricing of premiums. Frequency of claims are determined by the attributes of a particular policy holder. Count regression analysis allows one to find out which characteristic of a policy holder plays a significant role in determining the frequency of claim and also in predicting the frequency of claims given the characteristics of a particular policy holder. The objective of this thesis is to find out which among the Poisson, $NB1$ and $NB2$ models is a better fit to the count data under consideration. The count data is from Kenendia insurance. The best model is chosen based on the log-likelihood method and the Akaike's Information Criteria (AIC).

Keywords Overdispersion, Count data, Poisson, Negative Binomial.

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

Vanis Kemunto Makori
Reg No. I56/81817/2015

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

Signature   Date

Dr Joseph Ivivi Mwaniki
School of Mathematics,
University of Nairobi,
Box 30197, 00100 Nairobi, Kenya.
E-mail: jimwaniki@uonbi.ac.ke

Dedication

This project is dedicated to the entire Makori's family especially my Dad Obiero Makori for their unconditional support throughout my studies God bless you abundantly

Contents

Abstract ii

Declaration and Approval iv

Dedication vi

Acknowledgments viii

1 Introduction 1

1.1 Background of Study 1
1.2 Problem Statement 2
1.3 Objectives 2
1.3.1 General Objectives 2
1.3.2 Specific Objectives 3
1.4 Significance of Study 3
1.5 Scope of Study 3
1.6 Organization 3
1.7 Literature Review 4

2 Methodology 6

2.1 Negative Binomial Distribution 8
2.2 Assessment of goodness of fit 13
2.2.1 Newton-Raphson numerical method for estimating β's 14
2.2.2 Using deviance to compare nested models 17
2.2.3 Deviance goodness of fit 17
2.2.4 Model selection strategy 17
2.2.5 Testing for over dispersion 19

3 Data Analysis 21

3.1 Data Description 21
3.2 The Models 23
3.2.1 The poisson model 23
3.3 Assessing goodness of fit of the models 25
3.4 Results and Analysis 26

4 Conclusion 29

4.1 Future Research 29

Bibliography 30

Bibliography 32

Acknowledgments

Special thanks to my supervisor Dr.Ivivi Mwaniki for his guidance and commitment to completion of this work. His valuable suggestions and review that made this work a success. To my friends especially my friend T.K.Langat and family thanks for your support. To my Head of Management Mr. Patrick Njoroge thanks for your encouragement and support. Above all, thanks to our Almighty God for seeing me this far and enabling me accomplish this work.

Vanis Kemunto Makori

Nairobi, 2017.

1 Introduction

1.1 Background of Study

The importance of insurance industry to the world economy cannot be understated. Insurance offers an economic re-mediation thus providing a way of minimizing financial loss resulting from unusual risks by pooling or spreading risk over a large number of entities. The main cash out-flow in every insurance industry is the claim payments which is wholly dependent on frequency of claims launched by the policy holder.

The frequency of claims is often referred to as claim count data. It is therefore important that insurance companies understand the nature and evolution of claims count data at any given point in time. The six main categories of general insurance are given as home and contents insurance, motor vehicle insurance, business insurance, mortgage insurance, workers compensation insurance and travel insurance. In this paper we focus on the general insurance.

The insurance process is simply summarized as a transfer of risk in exchange for a regular payment known as premium. To compute premiums actuaries have come up with several methods. These methods can be simplified in a general way to be the multiplication of the expected frequency of claims with the expected cost of claim. This implies that understanding the nature and evolution of count data is very vital.

General insurance plays a significant role in the economy of any country as it provides means of reducing financial risks. It has been estimated that over $750,000$ people are killed and tens of millions injured on the roads in third world countries each year. Thus, the most efficient and preferable means of guarding the risk is by insuring against it.

Actuaries from different insurance firms in different countries, in order to correctly price the insurance contracts always try to either come up with better models, or improve existing models so as to accurately estimate claims count data. The frequency of claims do vary from one policy holder to another. It depends on the characteristics of the insured. Thus finding a model that takes into account the characteristics of the insured is of prime importance. By finding an appropriate distribution pricing of products not only becomes easier but it also enables insurance companies to offer custom made insurance packages to their clients.

One of the widely used methods of modelling insurance count data is the classical method also known as the distributional approach method. Under this method, the assumption is that the frequency of claims follows a particular discrete distribution.

The Kenya Insurance Industry is governed by the Insurance Act (KIA) which was enacted in 1985. The Insurance Act of 2006 then established the Insurance Regulatory Authority (IRA), a body that ensures the effective administration, supervision, regulation and control of insurance and reinsurance business in Kenya. IRA describes the general insurance under which the motor insurance is classified as non-life insurance. The general insurance has a variety of products and these products vary between companies. Thus the policyholders are advised to be well conversant with their product disclosure statements before purchasing a given cover. Complaints regarding general insurance under which the auto insurance is categorized are considered three times higher to those of life insurance frequency of claims. Since the largest source of outflow of money in an insurance company is the claims payments, modelling of claims count data is very crucial. With good modelling of claims frequency, the amount of premium to be calculated will be transparent and this will be an advantage to both the insured and the insurer. According to [Haiss and Sümegi(2008)], non-life insurance is a the fastest growing areas for actuaries.

Section 4 of the Kenya insurance Act expressly highlights that no one shall use, cause to use or allow anyone to use a motor vehicle unless there is in force a policy of insurance or such a security in respect of third party risks.

1.2 Problem Statement

In pricing motor insurance policyholders who have accidents with a small size of loss are penalized in the same way with policyholders who have accidents with a big size of loss. To sort out this issue there is a need to develop a model that incorporates both the frequency and the severity components of a claim.

Actuaries have made significant strides in coming up with models that are suitable to model count data with different properties.

In this paper we seek to answer the question whether some of the discrete models proposed by various actuaries are good models for modelling count data. The models to be considered are the Poisson model, the negative binomial 1 model and the negative binomial 2 model.

1.3 Objectives

1.3.1 General Objectives

In this study we seek to model frequency of claims data by using both the poisson distributions and the negative binomial distribution and to show why the negative binomial distribution is a better model.

1.3.2 Specific Objectives

They include:

- To find out which explanatory variables have a significant effect on the frequency of claims.
- To compare both negative binomial model and the poisson model in modelling frequency of claim data.
- To show which is a better discrete distribution model to curb the property of overdispersion in frequency of claims.

1.4 Significance of Study

It is essential to have auto insurance since The potential costs resulting from the occurrence of an accident, whether a replacement or repair costs of the vehicle, other property or medical costs of the victims, are too huge to exercise the risk of being without sufficient coverage. This prompts individuals to take up risk covers. According to [Murat et al.(2002)Murat, Tonkin, and Jüttner] insurance companies especially those offering general insurance products trade under very competitive conditions .

1.5 Scope of Study

The scope of study is the non-life insurance sector basically the automobile insurance. The claims experience will consist of detailed information on the type of insurance claim, the risk factors and the corresponding claim amount.

1.6 Organization

The thesis is presented into four main chapters named as chapter one, chapter two, chapter three and chapter four. Chapter one is the introduction presenting the background to the study, then the statement of the problem, then the objective of the study, then the significance of the study, then the scope of the study and then the limitations of the study. Also in the first chapter we review the applicable literature comprising both the theoretical framework and the different perspectives of the study problems related to the auto insurance. Chapter two outlines the methodology to achieving the set objectives. Chapter three presents the data analysis, results obtained from the model and the accompanying discussions. In the final chapter, chapter four we give the summary of the findings and recommendations. The references are found in the bibliography section.

1.7 Literature Review

Non-life insurance especially the auto-insurance business is not only a fast growing area that needs actuarial input [Haiss and Sümegi(2008)], it also offers an interesting challenge since it manages a large number of scenarios involving different types of risks.

The major aim of an insurance company is to compute a premium that correctly covers the type of risk an insured is insured against. The price should factor in attributes of the customer because these attributes play a big role in classifying one's riskiness. The frequency of claims plays a vital role in calculation of premiums. Therefore, it is of prime importance to come up with models that correctly describes the evolution of frequency of claims also referred to as count data. Since no two persons have exactly similar attributes, this implies that to come up with a suitable model that correctly estimates the frequency of claims, models that incorporates risk factors should be considered. According to [Boucher et al.(2008)Boucher, Denuit, and Guillén], regression analysis of count data allows the identification of the explanatory variables and the estimation of expected number of claims conditional on the individual characteristic of policy holder.

The advent of Generalized Linear Models (GLMs) played a vital role in the development of count data models, [Cameron and Trivedi(1999)]. The Poisson regression model which is a GLM model was introduced in the paper [Nelder(1977)] and deeply looked into by Gourieroux and company in their paper [Gourieroux et al.(1984)Gourieroux, Monfort, and Trognon]. Within the context of general insurance, [McCullagh and Nelder(1989)] showed that a Poisson structure is realised a priori when using the GLM method to estimated the frequency of claims.

[Smyth and Jørgensen(2002)] puts forward the Poisson distribution as the model to be used in modelling frequency of claims. Although it statistically it is condusive and favourable , in the paper [Gourieroux and Jasiak(2004)] it is emphasized that the Poisson distribution has some shortcomings which limits its usage. The Poisson model has the equidispersion property which is the equality of mean and variance as shown to hold by [Cameron and Trivedi(1999)]. Absence of equidispersion is unobserved heterogeneity. In the analysis of count data the hidden heterogeneity leads to overdispersion.

A heterogenopus portfolio implies that all the insured have a constant but unequal underlying risk of having an accident. That is, the expected frequency of claims differs from an insured to an insured. Various research papers such as Hausman et.al(1984), Cameron and Trivedi(1990); Gurmu(1991); Charpenter and Denuit(2005);Hilbe(2014) suggest the use of negative binomial distribution as the better discrete distribution alternative. The standard negative binomial is obtained from a mixture of Poisson and gamma distributions. The alternative to the Poisson distribution that is the negative binomial distribution which is preferred when there is overdispersion. Negative binomial distribution is a discrete distribution used whose parameters are n and p with mean and variance as Klugman et.al(1998). According to Denuit.et.al(2007) the satisfactory alternative discrete distribution to the poisson distribution is the negative binomial distribution.

2 Methodology

Of interest to actuaries is the estimation of the number of claims in which the Poisson model is often used. Though, the poisson model has enough literature backing, it still has a limitation of equidispersion. Its popularity is seen from its various theoretical properties as defined in [Mikosch(2009)] in these ways:

- The process starts at zero $N(0) = 0$ almost surely.
- The process has independent increments for any $t_i, i = 0, ..., n$ and $n \geq 1$ such that $0 = t_0 < t_1 < \cdots < t_n$ the increments $N(t_i) - N(t_{i-1}), i = 1, ..., n$ are mutually independent.
- There exists a non-decreasing right continuous function $m : [0, \infty) \to [0, \infty)$ with $m(0) = 0$ such that the increments $N(t) - N(s)$ for $0 < s < t < \infty$ have a Poisson distribution, $Pois(m(t) - m(s))$ where m is denoted as the mean value function of $N(t)$.
- With probability 1, the sample paths of the process $N(t)$ are right-continuous for $t \geq 0$ and have limits from the left for $t > 0$. This implies that $N(t)$ has cadlag sample paths.

Within the actuarial literature context, the number of claims that occurs conditional on the characteristics of the policy holders follows approximately, the Poisson distribution. The poisson distribution is known to model the number of events which may occur in any of a large of number of trials but the probability of occurrence in any given trial is small. Given the discrete random variable k_i (observed number of claims in r^{th} intervals of period of time the interval from t_{r-1} to t_r), conditioned by the vector of explanatory variables X_i (which is the policy holder's characteristics), is assumed to be Poisson distributed, then the PDF of k_i is given by:

$$p(k_i|x_i) = \begin{cases} \frac{e^{-\lambda_i}\lambda_i^{k_i}}{k_i!}, & \text{for } k_i = 0, 1, 2, \ldots \text{and} \quad \lambda_i > 0 \\ \\ 0, & \text{otherwise} \end{cases} . \tag{1}$$

Thus, equation (1) can be described to give the probability that K_i takes the realisation $k_i (k_i \in \mathbb{N})$, conditional on the different attributes of customers.

Despite the poisson distribution being considered as a major reference point in the modeling of count data the property of equidispersion sets it back since it has its mean and variance equal.

$$\mathsf{E}(k_i|x_i) = \mathsf{var}(k_i|x_i) = \lambda_i$$

Under Generalised Linear Models (GLM) the link function relates the expected value of the explained variable to the linear predictor. From literature it is common knowledge that the link function associated to the Poisson distribution is the logarithmic function.i.e.

$$\ln(\lambda_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \tag{2}$$

Where the $\beta_j s$ are the regression coefficient which are to be approximated. Equation (2) implies that

$$\begin{aligned}\lambda_i &= e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}} \\ &= e^{\mathbf{x}_i^T \beta}\end{aligned}$$

The estimation of parameters is done using the maximum likelihood function, with the likelihood function of the poisson distribution defined as:

$$L(\beta) = \prod_{i=1}^{n} \frac{e^{-\lambda_i} \lambda_i^{k_i}}{k_i!} \tag{3}$$

$$= \prod_{i=1}^{n} \frac{e^{-e^{x_i^T \beta}} (e^{x_i^T \beta})^{k_i}}{k_i!} \tag{4}$$

Taking natural logarithms of both sides of equation (3) we get;

$$\ln L(\beta) = \sum_{i=1}^{n} [k_i \ln \lambda_i - \lambda_i - \ln k_i!] \tag{5}$$

$$= \sum_{i=1}^{n} [k_i x_i^T \beta - e^{x_i^T \beta} - \ln k_i!] \tag{6}$$

The maximum likelihood estimator (MLE) is the solutions of the equations obtained by differentiating the log-likelihood in terms of regression coefficients and solving for them by equating to zero. Although poisson is frequently used in modeling claim data the model is restrictive in the type of data being used. Count data with overdispersion characteristic renders the model unsuitable. [Denuit et al.(2007)Denuit, Maréchal, Pitrebois, and Walhin] states that overdispersion arises because in real life no two drivers are identical. Drivers have unique attributes such as swiftness of reflexes in case of impending dander, aggressiveness when driving, consumption of different types of drugs, etc. Since the insurer cannot observe these attributes overdispersion arrises. Unobserved heterogeneity leads to overdispersion. According to [Cameron and Trivedi(1990)] a regression based test for overdispersion can be conducted. To consider the possibility of presence of extra-Poisson variation(overdispersion), we extend the Poisson model given in (1) to include random effects. Let $y_1, y_2, \ldots, y_n$ be continuous, positive valued, independent and identically distributed random variables such that given x_i and y_i, k_i Poisson(y_i, λ_i). The assumption is that the first and second moments of $y_i^{\prime}s$ are finite. Without loss of generality, we take expectation and variance of the $y_i^{\prime}s$ to be 1 and α respectively. According to [Collings and Margolin(1985)],

$$\text{var}(k_i|x_i) = \lambda_i + \alpha\lambda_i^2$$

thus, the poisson model can be tested against the model with extra-poisson variation by testing the null hypothesis $H_0 : \alpha = 0$ against the alternative hypothesis $H_1 : \alpha > 0$. The hypothesis of no overdispersion is rejected, when after comparing the statistics calculated value with the theoretical one the test appears to be significant. Thus the other alternatives are considered.

2.1 Negative Binomial Distribution

Negative binomial distribution is the preferable alternative for the drawbacks identified in the poisson distribution. Traditionally negative binomial distribution is a mixture of both the poisson and gamma distribution. Where the number of accidents is Poisson distributed, but there is gamma- distributed unobserved individual heterogeneity reflecting the fact that the true mean is not perfectly observed. Following that negative binomial distribution theoretically has a greater variance than the mean we going to construct the general negative binomial probability density function and showcase some of its pproperties.

According to the i^{th} individual, consider the number of claims k_i, given the parameter λ_i has a Poisson(λ_i) distribution. i.e.

$$p(k_i|x_i) = \begin{cases} \frac{e^{-\lambda}\lambda_i^k}{k_i!}, & \text{for } k_i = 0, 1, 2, \ldots \text{and} \quad \lambda > 0 \\ \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

λ_i denotes the different underlying risk of the i^{th} policyholder to have an accident. Suppose that λ_i follows a gamma (α_i, τ_i) distribution, with pdf of the form;

$$f_{\lambda_i}(\lambda_i; \alpha_i, \tau_i) = \begin{cases} \frac{\tau_i^{\alpha_i}}{\Gamma\alpha_i} e^{\tau_i\lambda_i}\lambda_i^{\alpha_i - 1}, & \text{for } \lambda_i \geq 0, \alpha_i \geq 0, \tau_i \geq 0 \\ \\ 0, & \text{otherwise} \end{cases} . \tag{8}$$

with mean $\mathsf{E}(\Lambda_i) = \frac{\alpha_i}{\tau_i}$ and variance $\mathsf{var}(\Lambda_i) = \frac{\alpha_i}{\tau_i^2}$ The unconditional distribution of the number of claims k_i will be:

$$\begin{aligned} p(k_i) &= \int_0^\infty p(k_i|x_i) f(\lambda_i) \mathrm{d}\lambda_i \\ &= \int_0^\infty \frac{e^{-\lambda_i}\lambda_i^{k_i}}{k_i!}\frac{\tau_i^{\alpha_i}}{\Gamma\alpha_i} e^{-\tau_i\lambda_i}\lambda_i^{\alpha_i-1}\mathrm{d}\lambda_i \\ &= \frac{\tau_i^{\alpha_i}}{\Gamma\alpha_i \quad k_i!}\int_0^\infty e^{-(1+\tau_i)\lambda_i}\lambda_i^{\alpha_i+k_i-1}\mathrm{d}\lambda_i \end{aligned}$$

By letting $y_i = (1+\tau_i)\lambda_i$ we have that $\lambda_i = \frac{y_i}{1+\tau_i}$ which implies that $\mathrm{d}\lambda_i = \frac{\mathrm{d}y_i}{1+\tau_i}$. Therefore;

$$\begin{aligned} p(k_i) &= \frac{\tau_i^{\alpha}}{\Gamma\alpha_i \quad k_i!}\int_0^\infty e^{-y_i}\left(\frac{y_i}{1+\tau_i}\right)^{\alpha_i+k_i-1}\left(\frac{1}{1+\tau_i}\right)\mathrm{d}y_i \\ &= \frac{\tau_i^{\alpha_i}}{\Gamma\alpha_i \quad k_i!}\left(\frac{1}{1+\tau_i}\right)^{\alpha_i+k_i}\int_0^\infty e^{-y_i}y_i^{\alpha_i+k_i-1}\mathrm{d}y_i \\ &= \frac{\tau_i^{\alpha_i}}{\Gamma\alpha_i \quad k_i!}\left(\frac{1}{1+\tau_i}\right)^{\alpha_i}\left(\frac{1}{1+\tau_i}\right)^{k_i}\Gamma(\alpha_i+k_i) \\ &= \frac{\Gamma(\alpha_i+k_i)}{\Gamma\alpha_i \quad k_i!}\left(\frac{\tau_i}{1+\tau_i}\right)^{\alpha_i}\left(\frac{1}{1+\tau_i}\right)^{k_i} \\ &= \frac{\Gamma(\alpha_i+k_i)}{\Gamma\alpha_i \quad \Gamma(k_i+1)}\left(\frac{\tau_i}{1+\tau_i}\right)^{\alpha_i}\left(\frac{1}{1+\tau_i}\right)^{k_i} \qquad ,\alpha_i,\tau_i > 0 \quad \text{and} \quad k_i = 0,1,2,\cdots \end{aligned}$$

which is a general probability density function of the Negative binomial with parameters α_i and τ_i

To find the mean and the variance of the negative binomial distribution, we use the probability generating function.

Re-writing the negative binomial probability generating function we have

$$\begin{aligned} p(k_i) &= \frac{\Gamma(\alpha_i+k_i)}{\Gamma\alpha_i \quad \Gamma(k_i+1)}\left(\frac{\tau_i}{1+\tau_i}\right)^{\alpha_i}\left(\frac{1}{1+\tau_i}\right)^{k_i} \\ &= \frac{(\alpha_i+k_i-1)!}{(\alpha_i-1)!k_i!}\left(\frac{\tau_i}{1+\tau_i}\right)^{\alpha_i}\left(\frac{1}{1+\tau_i}\right)^{k_i} \\ &= \binom{\alpha_i+k_i-1}{k_i}\left(\frac{\tau_i}{1+\tau_i}\right)^{\alpha_i}\left(\frac{1}{1+\tau_i}\right)^{k_i} \qquad ,\alpha_i,\tau_i > 0 \quad \text{and} \quad k_i = 0,1,2,\cdots \end{aligned}$$

If we let $\frac{\tau_i}{1+\tau_i} = p_i$ it implies that $\frac{1}{1+\tau_i} = 1 - p_i$ and the negative binomial distribution is now given by

$$p(k_i) = \binom{\alpha_i + k_i - 1}{k_i} p_i^{\alpha_i}(1-p_i)^{k_i} \quad , \alpha_i, 0 \le p_i \le 1 \quad \text{and} \quad k_i = 0, 1, 2, \cdots$$

From the definition of a probability generating function, we have

$$\begin{aligned} G_{k_i}(s) &= \sum_{k_i=0}^{\infty} p(k_i) s^{k_i} \\ &= \sum_{k_i=0}^{\infty} \binom{\alpha_i + k_i - 1}{k_i} p_i^{\alpha_i}(1-p_i)^{k_i} s^{k_i} \end{aligned}$$

which implies that

$$G_{k_i}(s) = p_i^{\alpha_i} \sum_{k_i=0}^{\infty} \binom{\alpha_i + k_i - 1}{k_i} \Big(s(1-p_i) \Big)^{k_i} \tag{9}$$

Using the Maclaurin expansion of $(1-x)^{-q}$ we have that

$$(1-x)^{-q} = 1 + (-q)(-x) + \frac{1}{2!}(-q)(-q-1)(-x)^2 + \frac{1}{3!}(-q)(-q-1)(-q-2)(-x)^3 + \dots \tag{10}$$

From equation (10) above we deduce that the coefficient of x^k is given by

$$\begin{aligned} \frac{1}{k!}(-1)^k(-q)(-q-1)(-q-2)\dots(-r-k+1) &= \frac{(q+k-1)(q+k-2)\dots(q+1)q}{k!} \\ &= \binom{k+q-1}{k} \end{aligned}$$

Therefore;

$$(1-x)^{-q} = \sum_{0}^{\infty} \binom{k+q-1}{k} x^k$$

If we let $x = s(1-p_i)$, equation (9) becomes

$$\begin{aligned} G_{k_i}(s) &= p_i^{\alpha_i} \sum_{k_i=0}^{\infty} \binom{\alpha_i + k_i - 1}{k_i} \Big(s(1-p_i) \Big)^{k_i} \\ &= p_i^{\alpha_i} \Big(1 - s(1-p_i) \Big)^{-\alpha_i} \\ &= \left(\frac{p_i}{1 - s(1-p_i)} \right)^{\alpha_i} \end{aligned}$$

It can easily be shown that by back substitution of p_i the probability generating function (PGF) of a negative binomial (α_i,τ_i) is given by

$$\begin{aligned} G_{k_i}(s) &= \left(\frac{\frac{\tau_i}{1+\tau_i}}{1-\left(\frac{1}{1+\tau_i}\right)s}\right)^{\alpha_i} \\ &= \left(\frac{\tau_i}{1+\tau_i-s}\right)^{\alpha_i} \end{aligned}$$

Using the probability generating function (PGF) both the mean and variance of the distribution can be easily computed as shown below.

The mean is obtained from the following relationship.

$$\mathsf{E}(K_i) = G'(s)\big|_{s=1}$$

Since,

$$G'(s) = \frac{\alpha_i \tau_i^{\alpha_i}}{(1+\tau_i-s)^{(\alpha_i+1)}}$$

This implies that;

$$\begin{aligned} \mathsf{E}(K_i) &= G'(s)\big|_{s=1} \\ &= \frac{\alpha_i \tau_i^{\alpha_i}}{(1+\tau_i-1)^{(\alpha_i+1)}} \\ &= \frac{\alpha_i \tau_i^{\alpha_i}}{\tau_i^{(\alpha_i+1)}} \\ &= \frac{\alpha_i}{\tau_i} \end{aligned}$$

To compute the variance, we use the following relationship.

$$\mathsf{Var}(K_i) = \left[G''(s) + G'(s) - \big(G'(s)\big)^2\right]_{s=1}$$

Computing $G''(s)$ we have;

$$\begin{aligned} G''(s) &= \frac{d}{ds}G' \\ &= \frac{d}{ds}\left[\frac{\alpha_i \tau_i^{\alpha_i}}{(1+\tau_i-s)^{(\alpha_i+1)}}\right] \\ &= \frac{\alpha_i(\alpha_i+1)\tau_i^{\alpha_i}}{(1+\tau_i-s)^{(\alpha_i+2)}} \end{aligned}$$

Substituting for s in the above equation we get

$$\begin{aligned}
G''(1) &= \frac{\alpha_i(\alpha+1)\tau_i^{\alpha_i}}{(1+\tau_i-1)^{(\alpha_i+2)}} \\
&= \frac{\alpha_i(\alpha_i+1)}{\tau_i^2} \\
&= \left(\frac{\alpha_i}{\tau_i}\right)^2 + \frac{\alpha_i}{\tau_i^2}
\end{aligned}$$

Therefore variance is given by;

$$\begin{aligned}
\mathsf{Var}(K_i) &= \Big[G''(s) + G'(s) - \big(G'(s)\big)^2\Big]_{s=1} \\
&= \left(\frac{\alpha_i}{\tau_i}\right)^2 + \frac{\alpha_i}{\tau_i^2} + \frac{\alpha_i}{\tau_i} - \left(\frac{\alpha_i}{\tau_i}\right)^2 \\
&= \frac{\alpha_i}{\tau_i^2} + \frac{\alpha_i}{\tau_i} \\
&= \frac{\alpha_i}{\tau_i}\left(1 + \frac{1}{\tau_i}\right)
\end{aligned}$$

Comparing the variance and the mean obtained of the negative binomial distribution we see that the variance exceeds the mean. This is the overdispersion property. This suggests that negative binomial distribution can be used to model count data containing unobserved heterogeneity components. The term α_i plays the role of a dispersion factor and it is a constant.

To improve the Poisson model so that it can be robust and be used in modelling count data containing the property of overdispersion, [Boucher et al.(2008)Boucher, Denuit, and Guillén] argues that the more intuitive approach is the introduction of a random heterogeneity term θ_i of mean 1 and variance α_i in the mean parameter of the Poisson distribution. If the θ_i parameter follows a gamma distribution then it is known that this mixed model will result in a negative binomial distribution. [Greenwood and Yule(1920), Lawless(1987), Dionne and Vanasse(1989)] states that to ensure that the heterogeneity mean is equal to 1, both parameters of the gamma distribution are chosen to be equal to $\frac{1}{\theta_i}$. Thus, from this the mean and variance are given by $\mathsf{E}(k_i) = \lambda_i$ and $\mathsf{var}(k_i) = \lambda_i + \alpha_i\lambda_i^2$ respectively.

[Cameron and Trivedi(1986)] considers a more general class of negative binomial distribution (NBp) having the same mean λ_i, but a variance of the form $\lambda_i + \alpha\lambda_i^p$. To come up with a ditribution of this type we use a heterogeneity factor that has a Gamma distribution with whose mean is 1 and variance given by $\alpha\lambda^{p-2}$. Different values of p yields different forms of negative binomial distribution. If $p = 2$ the $NB2$ model coincides with the normal negative binomial distribution. To obtain the $NB1$ model from the general model p takes the value 1. The probability mass function of $NB1$ is given by:

$$f(k_i, \lambda_i, \alpha_i) = \frac{\Gamma(k_i + \frac{\lambda_i}{\alpha_i})}{\Gamma(k_i + 1)\Gamma(\frac{\lambda_i}{\alpha_i})}(1+\alpha_i)^{\frac{\lambda_i}{\alpha_i}}(1+\frac{1}{\alpha_i})^{-k_i} \tag{11}$$

The mean and the variance of $NB1$ model is thus given by $\mathsf{E}(k_i) = \lambda_i$ and $\mathsf{var}(k_i) = \lambda_i(1+\alpha_i)$ respectively. The log likelihood function is given below

$$\ln L(\alpha_i, \beta) = \sum_{i=1}^{n} \left\{ \left(\sum_{j=0}^{k_i - 1} (j + \frac{\lambda_i}{\alpha_i}) \right) - \ln k_i - (k_i + \frac{\lambda_i}{\alpha_i}) \ln(1+\alpha_i) + k_i \ln \alpha_i \right\} \tag{12}$$

The probability mass function of the $NB2$ model is similar to the general probability mass function of a negative binomial distribution. The mean and the variance of the $NB2$ is $\mathsf{E}(k_i) = \lambda_i$ and $\mathsf{var}(k_i) = \lambda_i(1+\alpha_i\lambda_i)$.The log-likelihood function corresponding to $NB2$ is as shown below;

$$\ln L(\alpha_i, \beta) = \sum_{j=1}^{n} \left\{ -\ln(k_i) + \sum_{j=1}^{k_i} \ln(\alpha_i k_i - j\alpha_i + 1) - (k_i + \frac{1}{\alpha_i}) \ln(1 + \alpha_i \lambda_i) + k_i \ln(k_i) \right\} \tag{13}$$

In the paper [Cameron and Trivedi(1999)] it is argued that there is less interest in the estimation of α_i, major attention is given to the estimation of β. Boucher and Guillen in their paper [Boucher et al.(2008)Boucher, Denuit, and Guillén] they argue that the process of estimating parameters is approximately the same for the three models.

To estimate the values of the $\beta' s$ we use numerical methods. The most widely used method is the Newton-Raphson technique. It is explained in the assessment of goodness of fit section below.

2.2 Assessment of goodness of fit

To assess the goodness of fit of various models residual deviance is used. In this study we shall in general terms refer to residual deviance as just deviance denoted by D. i.e. $D = 2(\ln(Ls) - \ln(Lm))$ where ln is the natural logarithm function, Ls is the maximized likelihood of the saturated model and Lm is the maximized likelihood under the fitted model. The smaller the value of the deviance implies a better fit. To estimate the value of the regression coefficients, the Newton-Raphson numerical method technique is used.

2.2.1 Newton-Raphson numerical method for estimating β's

Newton-Raphson method is among the most widely used methods in numerical analysis. It is used to estimate the roots of functions. The method is explained below.

Let $l(g)$ be a function whose root we want to estimate. Suppose g_r is the true root that is unknown and g_0 is the first initial estimate. Expanding $l(g)$ about g_0 by taylor series we get;

$$l(g) = l(g_0) + (g - g_0)l'(g_0) + \frac{1}{2!}(g - g_0)^2 l''(g_0) + \cdots$$

Equating the first two terms to zero and solvin for g we get

$$l(g) \approx l(g_0) + (g - g_0)l'(g_0) = 0 \tag{14}$$

$$\implies g = g_1 = g_0 - \frac{l(g_0)}{l'(g_0)} \tag{15}$$

g_1 is then used in place of g_0 to compute g_2 the new approximation of g_r. The process is repeated over and over until the newly approximated value is not significantly different from the previous approximation. Generalizing, we get the following iteration formula;

$$g_j = g_{j-1} - \frac{l(g_{j-1})}{l'(g_{j-1})} \quad , j = 1, 2, \ldots \tag{16}$$

The iteration is said to converge if

$$\lim_{j \to \infty} g_j \to g_r.$$

The Newton-Raphson iteration function given in (16) is generalized to solve a system of linear equations with multiple variables by using matrices and vectors. The equation to solve this type of system of equations is given by

$$\mathbf{w}^{(\mathbf{j})} = \mathbf{w}^{(\mathbf{j-1})} - J^{-1}(\mathbf{w}^{(\mathbf{j-1})})G(\mathbf{w}^{(\mathbf{j-1})}) \quad j = 1, 2, \ldots \tag{17}$$

where ;

$\mathbf{w}$ is a $n \times 1$ vector given by $\mathbf{w}^T = (w_1, w_2, \ldots, w_n)$, $w_i \in \mathbb{R}, i = 1, 2, \ldots, n$, G is a vector function such that $G : \mathbb{R}^n \to \mathbb{R}^n$, that is

$$G(w_1, w_2, \ldots, w_n) = \begin{bmatrix} g_1(w_1, w_2, \ldots, w_n) \\ g_2(w_1, w_2, \ldots, w_n) \\ \vdots \\ g_n(w_1, w_2, \ldots, w_n) \end{bmatrix}$$

where g_i maps $\mathbb{R}^n$ to $\mathbb{R}$ for $i = 1,2,\ldots,n$.

$J(\mathbf{w})$ is a Jacobian matrix. The inverse on the Jacobian matrix is the one in use in (17). This inverse is given by

$$J^{-1}(\mathbf{w}) = \begin{bmatrix} \frac{\partial g_1(\mathbf{w})}{\partial w_1} & \frac{\partial g_1(\mathbf{w})}{\partial w_2} & \cdots & \frac{\partial g_1(\mathbf{w})}{\partial w_n} \\ \frac{\partial g_2(\mathbf{w})}{\partial w_1} & \frac{\partial g_2(v)}{\partial w_2} & \cdots & \frac{\partial g_2(\mathbf{w})}{\partial w_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_n(\mathbf{w})}{\partial w_1} & \frac{\partial g_n(\mathbf{w})}{\partial w_2} & \cdots & \frac{\partial g_n(\mathbf{w})}{\partial w_n} \end{bmatrix}^{-1}$$

To use the Newton-Raphson method, we follow the steps outlined below

Step 1: Obtain the initial vector $\mathbf{w}^{(0)}$ given by $\mathbf{w}^{(0)T} = (w_1^{(0)}, w_2^{(0)}, \ldots, w_n^{(0)})$.

Step 2: Calculate the Jacobian matrix $J(\mathbf{w})$ and the vector function $G(\mathbf{w})$ when $\mathbf{w} = \mathbf{w}^{(0)}$ to get $J(\mathbf{w}^{(0)})$ and $G(\mathbf{w}^{(0)})$ respectively.

Step 3: By Gaussian elimination solve for $\mathbf{e}^{(\mathbf{0})}$ in the linear system $J(\mathbf{w}^{(0)})\mathbf{e}^{(\mathbf{0})} = -G(\mathbf{w}^{(0)})$ to get $\mathbf{e}^{(\mathbf{0})} = -J^{-1}(\mathbf{w}^{(0)})G(\mathbf{w}^{(0)})$.

step 4: solve for $\mathbf{w}^{(\mathbf{1})} = \mathbf{w}^{(0)} + \mathbf{e}^{(\mathbf{0})} = \mathbf{w}^{(0)} - J^{-1}(\mathbf{w}^{(0)})G(\mathbf{w}^{(0)})$.

step 5: $\mathbf{w}^{(\mathbf{1})}$ is then used to compute $\mathbf{w}^{(\mathbf{2})}$ which is then used to compute $\mathbf{w}^{(\mathbf{3})}$ and so on. The iteration is stopped after say j iterations if the difference between $\mathbf{w}^{(\mathbf{j-1})}$ and $\mathbf{w}^{(\mathbf{j})}$ is negligible. $\mathbf{w}^{(\mathbf{j})}$ is the estimate of the roots of the system of equations.

To apply the Newton-Raphson method in the estimation of the coefficients of the explanatory variables we need to define the following terms appropriately to fit equation (17).

Let $G(\beta)$ be a $q \times 1$ vector function of the first derivatives of the log-likelihood. It is called the efficient scores function and is given by;

$$G(\beta) = \left(\frac{\partial logL(\beta)}{\partial \beta_1}, \ldots, \frac{\partial logL(\beta)}{\partial \beta_q}\right).$$

The maximum likelihood estimate, $\hat{\beta}$ is obtained by solving $G(\hat{\beta}) = 0$.

The Jacobian matrix $J(\beta)$ is a $q \times q$ matrix of negative second order derivatives of the natural logarithm of the likelihood, such that the entry in the k^{th} row and l^{th} column is given by

$$J(\beta)_{kl} = -\frac{\partial^2 logL(\beta)}{\partial\beta_k \partial\beta_l}.$$

$J(\beta)$ is called the observed information matrix. The modification of equation (17) to be used to estimate the β's is thus given by;

$$\beta^{(j)} = \beta^{(j-1)} - J^{-1}(\beta^{(j-1)})G(\beta^{(j-1)})$$

Suppose that after k iterations it is observed that there is no significant change in the log-likelihood function, this will mean that the iteration has converged and the estimates of the β's is given by the entries of vector $\beta^{(k)}$.

To calculate the variance covariance matrix C we compute the negative inverse of the observed information matrix at $\hat{\beta}$, the estimate of β obtained by the maximum likelihood method. This is given by

$$C = -J^{-1}(\bar{\beta}) = -\begin{bmatrix} \left.\frac{\partial^2 logL(\beta)}{\partial\beta_1^2}\right|_{\beta=\hat{\beta}} & \left.\frac{\partial^2 logL(\beta)}{\partial\beta_1\partial\beta_2}\right|_{\beta=\hat{\beta}} & \cdots & \left.\frac{\partial^2 logL(\beta)}{\partial\beta_1\partial\beta_q}\right|_{\beta=\hat{\beta}} \\ \left.\frac{\partial^2 logL(\beta)}{\partial\beta_2\partial\beta_1}\right|_{\beta=\hat{\beta}} & \left.\frac{\partial^2 logL(\beta)}{\partial\beta_2^2}\right|_{\beta=\hat{\beta}} & \cdots & \left.\frac{\partial^2 logL(\beta)}{\partial\beta_2\partial\beta_q}\right|_{\beta=\hat{\beta}} \\ \vdots & \vdots & \cdots & \vdots \\ \left.\frac{\partial^2 logL(\beta)}{\partial\beta_q\partial\beta_1}\right|_{\beta=\hat{\beta}} & \left.\frac{\partial^2 logL(\beta)}{\partial\beta_q\partial\beta_2}\right|_{\beta=\hat{\beta}} & \cdots & \left.\frac{\partial^2 logL(\beta)}{\partial\beta_q^2}\right|_{\beta=\hat{\beta}} \end{bmatrix}^{-1}$$

The main diagonal entries of this matrix gives the variance of the maximum likelihood estimates $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_q$ whose square-root is the standard error of the corresponding estimates.

The estimates represented by vector $\hat{\beta}$ are asymptotically unbiased. The $(1-\alpha)100\%$ confidence interval of the β's is given by the formula $[\hat{\beta}_j \pm z_{\alpha/2} se(\hat{\beta}_j)]$ for $j = 1, 2, \ldots, q$. Where $se(\beta_j)$ is the standard error of β_j for $j = 1, 2, \ldots, q$ and $z_{\alpha/2}$ gives the upper $\alpha/2$-point of the standard normal distribution. In the event that the computed confidence interval of a particular parameter β contains zero, then there is a chance that the true value of parameter β takes the value zero in the presence of the other parameters. In this regard, it will be prudent to conduct a hypothesis test to test the significance of the parameter to the model in the presence of the other parameters.

To test the null hypothesis that $\beta_j = 0, j = 1, 2, \ldots, q$ in the presence of all the other terms , we use the test statistic $W^2 = \left(\frac{\hat{\beta}_j}{se(\beta_j)}\right)^2$. This statistic is then compared to the percentage points of a chi-squared distribution on one degree of freedom. If the test statistic is big we will reject the null hypothesis and concluded that the covariate whose coefficient is β_j is not significant in the presence of the other covariates in the model. Its effect will have be analysed by fitting it alone.

2.2.2 Using deviance to compare nested models

Since several values of explanatory variables are recorded. The modeling process dictates that we determine the explanatory variables by means of their statistical significance that should be included in the model or left out. To achieve this, there is need to come up with a method to assess each explanatory variable's contribution to the model.

Suppose we need to compare the goodness of fit of two models A and B. Suppose further model A is nested within model B. That is model A has u covariates and model B has $u+v$ covariates. To assess whether the additional v covariates improves the model or not, we need to do the log-likelihood ratio test which tests the null hypothesis that the additional v covariates takes the value zero in the presence of u covariates that is to say the simpler model is a better model or in other words the additional covariates are statistically insignificant in the presence of the other u covariates.

Using results from likelihood theory, the bigger the sample size implies that the difference in value of the deviances follows a chi-squared distribution under the null hypothesis that the additional covariates are not statistically significant. The difference in the number of parameters between the two model gives the degree of freedom of the Chi-square distribution. In this case it will be v.

2.2.3 Deviance goodness of fit

Deviance measures the closeness of the predicted values to the observed. Deviance can be used to test for the models goodness of fit. Though we anticipate the predicted values to be close to those observed in reality they will not be a perfect match even if the model has been specified correctly. To employ use of deviance in the assessment of goodness of fit, under the assumption that the model is well specified, we need to find out the expected variation in the observed frequency of claims around those predicted by the model, under the Poisson assumption. As we have seen above from the likelihood theory under the assumption that the model is well specified the difference in deviance between the model being proposed and the saturated model follows approximately a Chi-square distribution whereby the degrees of freedom is obtained by finding the difference between the number of parameters in the two models. The saturated model has n parameters since it employs a parameter for each frequency of claim observed. Suppose the model being proposed has q parameters, therefore the residual deviance will be tested by use of a Chi-squared distribution with $n-q$ degrees of freedom.

2.2.4 Model selection strategy

The first step in model selection involves identifying the set of explanatory variables that have a significant effect on the frequency of claims.

It is from this set that we will obtain the combination of covariates to be added in the model. If interaction is to be included in the model by the hierarchic principle [Nelder(1977)] the main effects have to be included as well. If the number of covariates is not too large we can fit all possible models and compare the values of their respective deviance statistic and then pick the model(s) that give the least value of these statistic. However in most cases the number of explanatory variables is huge thus fitting all the models is computationally costly. Several procedures have been developed to assist in selecting variables to be included in the model. They include forward selection procedure, backward selection procedure and step wise procedure. Under forward selection procedure, the null model is fitted first, then variables are added one by one. At each step, variables that lead to the largest decrease in the value of deviance on their addition are the ones included in the model. The process stops when the next variable to be added does not reduce the value of the deviance by more than a predetermined amount called the stopping rule [Collett(1994)]. Under backward selection, all variables are first fitted then the variables are eliminated one by one. At each step, variables that lead to the smallest increase in the deviance are the ones to be removed. The procedure goes on and on until the next variable to be eliminated leads to an increase in the deviance that exceeds the amount determined by the stopping rule. The step wise procedure is a combination of both the forward and the backward procedures. In this procedure, the variables already included will still be subjected to removal test at later steps. So after a variable has been added, the procedure checks if any of the variables already in the model can be omitted.

The above model selection procedures is simplified in the following steps

Step 1: Determine which variables significantly reduce the value of the deviance statistic. This is achieved by fitting models that contain the explanatory variables on their own then comparing them with the null model.

Step 2: All the variables that were viewed as significant in step 1 are fitted together then the variables are omitted one at a time just as is the case in the backward procedure. The ones that do not significantly decrease the value of the deviance are omitted.

Step 3: The variables that were left out in step 1 are added in this step one by one as was the case in the forward procedure. The ones that significantly reduce the size of the deviance are retained.

Step 4: The variables that finally make to the model in step 3 are tested once more to ensure that each will significantly increase the value of the deviance if they are removed. Also the variables left out are checked to ensure that none can significantly reduce the value of the deviance if included in the model.

To successfully carry out the above procedures the choice of the significant level is made flexible.

The standard measure of goodness of fit that is used to assess the adequacy of various models is the likelihood ratio that follows a $\chi^2_{\alpha,p}$ distribution(chi-square), level of significance $\alpha = 0.05$ and with p degrees of freedom, where p represents the number of explanatory variables included in the regression model as discussed in [Denuit and Lang(2004)]. This test is derived by finding the difference between the deviance of the regression model without explanatory variables and the deviance of the model with independent variables. With the deviance being twice the difference between the maximum log-likelihood and the log-likelihood of the fitted model. When the log-likelihood ratio is higher than the statistical theoretical value it means that the suggested regression model fits a well analyzed data. The standard method used to compare the two distribution is the likelihood ratio with the given expression:

$$LR = -2(LL_p - LL_{NB}),$$

where LL_p and LL_{NB} are the log-likelihood values under the Poisson and negative binomial models respectively. The resultant statistic has a chi-square distribution with a d.f equal to one. If calculated value is higher than the theoretical value then the NB models are chosen against Poisson,more-so the convient method used to chose between the $NB1$ and the $NB2$ is the log-likelihood function. The $NB2$ model is preferred to $NB1$ as it has higher log- likelihood, [Cameron and Trivedi(1999)].

2.2.5 Testing for over dispersion

In fitting a Poisson regression model to count data one of the assumptions is that the mean is equal to the variance a property called equidispersion. If this property is violated such that the conditional variance is greater than the conditional mean of the count data then the predictions made by the Poisson regression model will be inaccurate and will lead to wrong premium computations which will adversely affect the insurance business.

Therefore, after fitting a Poisson regression model we immediately test for overdispersion to check whether we are fitting the correct model. A quick test is to compute the ratio of the residual deviance generated by the model to the given degrees of freedom. If this ratio exceeds 1 it implies that there is overdispersion and Poisson regression model is not the appropriate model. However, what happens when the ratio slightly exceeds 1 do we just throw away the Poisson regression model? What value of the ratio is deemed significant?

To answer the above questions we use the test designed in [Cameron and Trivedi(1990)] which tests for overdispersion. The idea is very simple. Under a Poisson regression model the mean and the variance are equal. i.e. Suppose Y is the response variable in our case is the random variable representing the frequency of claims. If λ is the Poisson parameter then $\mathsf{E}(Y) = \mathsf{var}(Y) = \lambda$. The equality of mean and variance assumption is tested under the null hypothesis against the alternative hypothesis where $\mathsf{var}(Y) = \lambda + kg(\lambda)$. The constant $k < 0$ implies underdispersion and $k > 0$ implies overdispersion. Funtion $g(\cdot)$ is often either a linear or a quadratic monotonic function. The equivalence of the above test is a test where we test $H_0 : K = 0$ against $H_1 : K \neq 0$. The t statistic which asymptotically follows a normal distribution if the null hypothesis holds is used as the test statistic.

3 Data Analysis

3.1 Data Description

Analyzing this we will use data from Kenendia insurance. From the data, the frequency of claims is represented by the number of open complains. There are 12 explanatory variables. They include ;

Customer lifetime value which is the financial value the insurance company gets from a lifetime relationship with a particular customer. It gives the net present value of all the future cash flows attributed to the policyholder during the entire relationship with the insurance company. It gives the difference of all the projected premiums the customer will pay and the cost the company will incur with respect to a particular customer. This variable is a quantitative variable and it varies from customer to customer.

Coverage, this gives the level of coverage. It is a qualitative variable with three levels namely, basic, extended and premium.

Level of education, this qualitative variable captures the highest level of education f the policy holder. It has four levels namely, high school and below, college level, bachelors degree level and masters degree level.

Employment status, this categorical variable captures the nature of employment of the customer. The categories are employed, unemployed, medical leave and disabled.

Gender as an explanatory variable is divided into two categories male and female.

Income, this explanatory variable captures the amount a particular customer earns per month.

Marital status, this is a categorical variable with three categories, married, single and divorced.

Months since policy inception, this gives the number of months that have elapsed since taking of cover. It captures the number of months since policy inception to the day of data collection.

Policy type, as a categorical variables it captures the different types of policies available to the customers. They include; corporate auto, personal auto and special auto.

Total claim amount, this quantitative variable captures the total amount of money the policy holder can claim in the event of an accident.

Vehicle class captures the different classes of vehicles being covered. There are six classes, two-door car, four-door car, suv, luxury suv, luxury car and sports car.

Vehicle size, this explanatory variable captures the physical size of the vehicle. It is a categorical variable with three categories namely, medsize, small size and large size

The table below gives a the frequency distribution of the observed claim frequency

Table 1. Table showing observed claim frequency

Frequency of claims	Number of policyholders
0	225
1	39
2	18
3	12
4	3
5	3
6 and above	0

Analyzing the distribution of claims frequency we observe that the maximum number of complains by a customer is 5. On close scrutiny we observe that throughout the period of analysis 225 out of 300 policyholders did not launch any claim complain. This translates to a 75% of the total policyholders under investigation. 39(13%) of policyholders declared a single claim, 18(6%) of policyholders declared two claims, 12(4%) of customers declared to the insurer three claims and 3(1%) of policyholders informed the insurance company about the occurrence of four accidents and a similar percentage of policyholders declared the occurrence of five accidents.

The above distribution of claim frequency can easily be summarized in a histogram as shown below.

From the histogram above we observe that the frequency of claims is highly skewed to the right and the number of zero counts is big. Though the conditional distribution of the number of claims frequency given the covariates could be different from the marginal

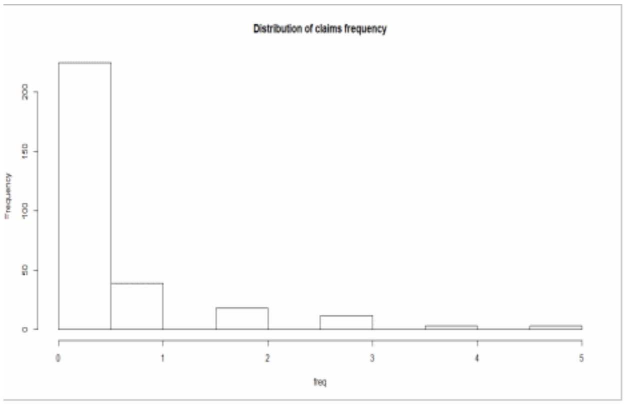

Figure 1. Histogram showing distribution of claims frequency

distribution, the big departure from symmetry spells doom for least-squares regression methods.

Table 2. Table showing the summary of count data

Variable	Mean	Median	Standard deviation	Variance	Minimum	Maximum
Frequency of claims	0.46	0.00	0.9651	0.9315	0	5

With a mean of 0.46 and a variance of 0.9315, it implies that the variance exceeds the mean. This together with the earlier observation of skewness to the right it indicates presence of heterogeneity in the data. Which implies that the policyholders have different and unique attributes which play apart in their proneness to accidents. On the positive side a huge number of 0 claims implies that majority of the policy holders are low risk hence good for business.

3.2 The Models

3.2.1 The poisson model

The table below give the results obtained after fitting the Poisson regression model. To check for significance of the explanatory variables on determining the frequency of claims we first fitted a model containing all the explanatory variables then fitted other models without one explanatory variable. Since the models containing all but one explanatory variables are nested within the model containing the all the explanatory variable, to assess the significance of the explanatory variable not fitted we employ the backward procedure technique discussed in the previous section above. To test the null hypothesis that a particular explanatory variable is not significant in the presence of the other explanatory variables we use the likelihood ratio statistic. The likelihood ratio statistic is obtained by subtracting the value of deviance from the model with all explanatory variables from the one with all but the explanatory variable in question. This likelihood ratio statistic has a chi-square distribution with degrees of freedom equivalent to the difference in the number of parameters in the two models.

Table 3. Table showing the summary of different poisson models

Covariates	Deviance	Degrees of freedom	LRS	χ^2 tabulated at 10% DoF
All Variables	386.51	273	-	-
All except Customer lifetime value	386.68	274	0.17	1.642(1)
All except Coverage	387.99	275	1.48	3.219(2)
All except Coverage	387.99	275	1.48	3.219(2)
All except Education	390.61	277	4.10	5.989(4)
All except employment status	395.08	277	8.57	$5.989(4)^*$
All except Gender	386.51	274	0.001	1.642(1)
All except Income	387.01	274	0.50	1.642(1)
All except Marital status	388.74	275	2.23	3.219(2)
All except Months since policy inception	387.34	274	0.83	1.642(1)
All except Policy type	386.89	275	0.38	3.219(2)
All except Total claim amount	386.74	274	0.23	1.642(1)
All except Vehicle class	396.80	278	10.29	$7.289(5)^*$
All except vehicle size	390.89	275	4.38	$3.219(2)^*$

From the table above, we see that at 10% level of significance all but three explanatory variables are not significant. The significant explanatory variables are employment status, vehicle class and vehicle size. This implies that these three covariates have an effect on the frequency of claims. They contribute to the process of explaining the variation in the frequency of claims among different policyholders.We therefore go ahead and fit a Poisson regression model using only the three explanatory variable. The table below gives a summary of the resulting coefficients and their standard errors.

Table 4. Table showing the summary of coefficients and their standard errors

Coefficients	Estimate	Standard error
Intercept	-1.53420	0.56313
EmploymentStatus-Employed	0.45749	0.46403
EmploymentStatus-Medical Leave	-0.39103	0.73139
EmploymentStatus-Retired	1.29511	0.54150
EmploymentStatus-Unemployed	0.75189	0.47540
Vehicle.Class-Luxury Car	0.85091	0.59316
Vehicle.Class-Luxury SUV	-14.64683	635.44420
Vehicle.Class-Sports Car	-1.42298	0.71830
Vehicle.Class-SUV	-0.07216	0.23892
Vehicle.Class-Two-Door Car	0.21082	0.20961
Vehicle.Size-Medsize	0.34647	0.35197
Vehicle.Size-Small	-0.07062	0.39003

Nevertheless, if the equality of the mean and variance assumption of the Poisson model is not met the model won't accurately give us the information on the relationship between the explanatory variables and the response variable which is the frequency of claims. Therefore, it is prudent for us to test for overdispersion. The method proposed by Cameron and Trivedi is used.

Table 5. Table showing the summary of Cameron and Trimedi test

Function	Parameter,k	Degree of freedom	Parameter Estimate	Standard Error	t-Value	Pr>
$g(\lambda) = \lambda_i$	k_0	1	0.0253	0.00467	2.65	< 0.0
$g(\lambda) = \lambda_i^2$	k_1	1	0.2356	0.04322	4.34	< 0.0

The table above gives the value of the test statistic which enable us test the null hypothesis that there is no overdispersion against an alternative of overdispersion. Using 0.05 as the level of significance we reject the null hypothesis . This implies there exists overdispersion in the count data. We therefore fit the $NB1$ and $NB2$ models which takes into account the presence of overdispersion. After fitting the two models we need to assess the best model.

3.3 Assessing goodness of fit of the models

To assess the goodness of fit of the models we use the Akaike's Information Criteria(AIC). In addition to the log-likelihood, the AIC adds term that penalizes depending on the number of explanatory variables. The addition of this penalizing term makes AIC a better criterion since it balances the goodness-of-fit with respect to inclusion of additional explanatory variables.

To calculate the value of AIC we use the following formula;

$$AIC = -2\ln L + 2p$$

Where ln is the natural logarithm function and p is the number of explanatory variables and other interaction terms included in the model.

A model with a smaller AIC terms implies a better fit.

After fitting the two models we get the following values of AIC

Table 6. Table showing values of AIC

Criterion	Poisson	NB1	NB2
AIC	602.21	593.18	542.43

From the table we observe that the AIC corresponding to $NB2$ model is smaller than the one corresponding to $NB1$ and Poisson models. This implies that $NB2$ model is a better model and gives the best fit.

3.4 Results and Analysis

The table below gives the estimate of values of the coefficients and their standard errors under the negative binomial $NB2$ regression model.

Table 7. Table showing values of esimates of coefficients under $NB2$ model

Coefficients	Estimate	Standard error
Intercept	-1.601	0.7592
EmploymentStatus-Employed	0.5188	0.6132
EmploymentStatus-Medical Leave	-0.3135	0.9139
EmploymentStatus-Retired	1.330	0.8169
EmploymentStatus-Unemployed	0.8049	0.6377
Vehicle.Class-Luxury Car	0.9038	0.59316
Vehicle.Class-Luxury SUV	-28.62	69680
Vehicle.Class-Sports Car	-1.410	0.8487
Vehicle.Class-SUV	0.06717	0.3495
Vehicle.Class-Two-Door Car	0.2403	0.3193
Vehicle.Size-Medsize	0.3243	0.4970
Vehicle.Size-Small	-0.09029	0.5434

From 7 above it can be observed that the estimated values of the coefficients is not very different from the one obtained under the poisson model. Though the standard errors of the estimated coefficients are slightly higher than the output from the fitted poisson model, it does not affect the significance of the estimated parameter.

By observing the regression coefficient we can establish the profile of policy holders who pose the highest risk to the insurance company. These type of policyholders are the ones who have the highest chance of launching more claims. A foreknowledge of these type of class of policy holders will enable the insurance company to differentiate the amount of premiums they charge their clients. In this regards, customers are charged premiums depending on their risk groups.

To compute premiums, insurance companies uses different components. Estimated frequency of claims is a significant component in the computation of premiums. The estimated frequency of claims for a new customer is computed by matching the characteristics of the customer in question to one category of the policyholders. To obtain the estimated frequency of policyholders in one of these categories the link function is used. As described in the methodology, for the negative binomial model, the link function is the logarithmic function.

Taking into consideration the estimates of parameters for the $NB2$ model, we can easily compute the estimated value of frequency of claims for the policyholders considered the most risky in the calculation shown below.

$$\lambda_{riskiest \quad group} = e^{-1.601+0.8040+0.9038+0.3243+150*0.0081} = 5.1914 \tag{18}$$

Which gives the expected frequency of claims from customers who have a similar characteristic to those who fall in the riskiest category for the insurer.

4 Conclusion

Correct computation of premiums allows an insurance company to meet its payment of claims obligation when they arise, to meet the daily operational costs and of course to make a profit. We have seen that to accurately price a premium, accurate estimation of frequency of claims is required. In this paper, we considered discrete distributions namely; the Poisson, the $NB1$ and the $NB2$ distributions to model count data. The models incorporated risk factors which play a crucial role in explaining the risk being insured. Upon testing the assumption of equality and mean of the Poisson model, the test technique employed in this paper reached the conclusion of existence of over dispersion within the insurance portfolio being analysed. In an attempt to find a better model that will take care of the over dispersion property the negative binomial models were fitted. Upon fitting the negative binomial models to the data, the results obtained showed that indeed the negative binomial models correct the shortcoming of the Poisson model since they give a better fit to the data. Comparing further $NB1$ model to $NB2$ model, we found out that $NB2$ model gives a better fit. Thus, based on the results obtained in this paper we can conclusively suggest that $NB2$ model is a better model to use to estimate and predict frequecy of claims.

4.1 Future Research

To look at other models that incorporate the problem of zero counts. To mention a few of these models we have zero-truncated models and Hurdle model.

To research on machine learning algorithms and techniques that give better fit.

Bibliography

[Allain and Brenac(2001)] E. Allain and T. Brenac. Modèles linéaires généralisés appliqués à l'étude des nombres d'accidents sur des sites routiers: Le modèle de poisson et ses extensions. *Recherche-Transports-Sécurité*, 72:3–16, 2001.

[Boucher et al.(2008)Boucher, Denuit, and Guillén] J.-P. Boucher, M. Denuit, and M. Guillén. Models of insurance claim counts with time dependence based on generalization of poisson and negative binomial distributions. *Variance*, 2(1):135–162, 2008.

[Cameron and Trivedi(1999)] A. Cameron and P. Trivedi. Essentials of count data regression.[online] http://www. econ. ucdavis. edu/faculty/cameron/research/cte01prepri nt. pdf, 1999.

[Cameron and Trivedi(1986)] A. C. Cameron and P. K. Trivedi. Econometric models based on count data. comparisons and applications of some estimators and tests. *Journal of applied econometrics*, 1(1):29–53, 1986.

[Cameron and Trivedi(1990)] A. C. Cameron and P. K. Trivedi. Regression-based tests for overdispersion in the poisson model. *Journal of econometrics*, 46(3):347–364, 1990.

[Collett(1994)] D. Collett. Modelling survival data. In *Modelling Survival Data in Medical Research*, pages 53–106. Springer, 1994.

[Collings and Margolin(1985)] B. J. Collings and B. H. Margolin. Testing goodness of fit for the poisson assumption when observations are not identically distributed. *Journal of the American Statistical Association*, 80(390):411–418, 1985.

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

[Denuit et al.(2007)Denuit, Maréchal, Pitrebois, and Walhin] M. Denuit, X. Maréchal, S. Pitrebois, and J.-F. Walhin. *Actuarial modelling of claim counts: Risk classification, credibility and bonus-malus systems*. John Wiley & Sons, 2007.

[Dionne and Vanasse(1989)] G. Dionne and C. Vanasse. A generalization of automobile insurance rating models: the negative binomial distribution with a regression component. *Astin Bulletin*, 19(2):199–212, 1989.

[Gourieroux and Jasiak(2004)] C. Gourieroux and J. Jasiak. Heterogeneous inar (1) model with application to car insurance. *Insurance: Mathematics and Economics*, 34(2): 177–192, 2004.

[Gourieroux et al.(1984)Gourieroux, Monfort, and Trognon] C. Gourieroux, A. Monfort, and A. Trognon. Pseudo maximum likelihood methods: Theory. *Econometrica: Journal of the Econometric Society*, pages 681–700, 1984.

[Greenwood and Yule(1920)] M. Greenwood and G. U. Yule. An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *Journal of the Royal statistical society*, 83(2):255–279, 1920.

[Haiss and Sümegi(2008)] P. Haiss and K. Sümegi. The relationship between insurance and economic growth in europe: a theoretical and empirical analysis. *Empirica*, 35 (4):405–431, 2008.

[Hilbe(2011)] J. M. Hilbe. Modeling count data. In *International Encyclopedia of Statistical Science*, pages 836–839. Springer, 2011.

[Holla(1967)] M. Holla. On a poisson-inverse gaussian distribution. *Metrika*, 11(1):115–121, 1967.

[Klugman et al.(2012)Klugman, Panjer, and Willmot] S. A. Klugman, H. H. Panjer, and G. E. Willmot. *Loss models: from data to decisions*, volume 715. John Wiley & Sons, 2012.

[Lawless(1987)] J. F. Lawless. Negative binomial and mixed poisson regression. *Canadian Journal of Statistics*, 15(3):209–225, 1987.

[McCullagh(1989)] P. McCullagh. An outline of generalized linear models. *Generalized linear models*, 1989.

[McCullagh and Nelder(1989)] P. McCullagh and J. Nelder. Generalised linear modelling. *Chapman and Hall: New York*, 1989.

[Mikosch(2009)] T. Mikosch. *Non-life insurance mathematics: an introduction with the Poisson process*. Springer Science & Business Media, 2009.

[Murat et al.(2002)Murat, Tonkin, and Jüttner] G. Murat, R. S. Tonkin, and D. J. Jüttner. Competition in the general insurance industry. *Zeitschrift für die gesamte Versicherungswissenschaft*, 91(3):453–481, 2002.

[Nelder(1977)] J. Nelder. A reformulation of linear models. *Journal of the Royal Statistical Society. Series A (General)*, pages 48–77, 1977.

[Seal(1983)] H. L. Seal. The poisson process: its failure in risk theory. *Insurance: Mathematics and Economics*, 2(4):287–288, 1983.

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

[Vasechko et al.(2009)Vasechko, Grun-Réhomme, and Benlagha] O. Vasechko, M. Grun-Réhomme, and N. Benlagha. Modélisation de la fréquence des sinistres en assurance automobile. *Bulletin Français d'Actuariat*, 9(18):41–63, 2009.

[Willmot(1987)] G. E. Willmot. The poisson-inverse gaussian distribution as an alternative to the negative binomial. *Scandinavian Actuarial Journal*, 1987(3-4):113–127, 1987.

[Yip and Yau(2005)] K. C. Yip and K. K. Yau. On modeling claim frequency data in general insurance with extra zeros. *Insurance: Mathematics and Economics*, 36(2):153–163, 2005.