# EVALUATION OF GENERALIZED LINEAR MODELS FOR MODELLING CLAIM FREQUENCIES IN VEHICLE INSURANCE

**Kinyanjui, Wanjiru; 091912**

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

**Strathmore Institute of Mathematical Sciences**
**Strathmore University**
**Nairobi, Kenya**

**November 2019**

## 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.



**Wanjiru Kinyanjui** [*Name of Candidate*]

..... [*Signature*]

28th Nov, 2019 ..... [*Date*]

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

**Dr. Linda Chaba** [*Name of Supervisor*]

..... [*Signature*]

28/11/19 ..... [*Date*]

Strathmore Institute of Mathematical Sciences

Strathmore University

Abstract

This study aims at evaluating the generalised linear models used for modelling vehicle insurance claim frequency. Modelling claim frequency, in turn, helps in the pricing and estimation of premiums. In this paper, claim frequencies will be modelled with respect to other risk factors present in the vehicle insurance data. This study makes use of data present in the CASdatasets that can be downloaded as an R package. The specific data used is brvehins1e that contains 393,071 observations. I further went ahead to randomly select 10,000 observations for computational purposes.

The four generalised linear models namely; Poisson, negative Binomial, zero inflated Poisson and zero inflated negative Binomial models were fitted to the data to evaluate how well they fit. Comparison of the models was done using the Akaike's Information Criteria, Bayesian Schwartz Information Criteria and as well as performing a Vuong Test. Significant variables in the model were determined using the p-values. The negative binomial model was determined as the better model when compared to the Poisson model. The zero inflated negative Binomial model was also seen to provide a better fit compared to the zero inflated Poisson model.

Table of Contents

List of equations



List of tables

# Chapter 1: Introduction

## 1.1 Background of study

Risk assessment has been a key issue in the insurance sector for a long time now. Particularly in vehicle insurance, there is reason to assess risk to help in pricing of premiums as well as to avoid making losses from selling policies that receive a high number of claims. Vehicle insurance offers protection in terms of finances as well as body injuries caused during accidents. Vehicle insurance will only protect the vehicle and an individual who has purchased a policy. According to (Denuit, 2007), actuarial science can be used in modelling claim counts which can be helpful in specifying the terms for the vehicle insurance.

The terms specified for vehicle insurance include premiums to be paid by a policyholder. Premium pricing has been discussed indepth by (Ramos, 2017) and it is clear that for premiums to be calculated, we require the frequency of claims best described as the claim count. Factors such as vehicle age, vehicle type and vehicle body are referred to as vehicle characteristics and they play a huge role in calculating the expected cost of future claims. Gender and age which compose the profile of an individual also play a role in pricing premiums. This has become vivid in the recent past as we see some insurance companies selling different policies for their lady policyholders. An example is the CIC LadyAuto. Based on a driver's past history, different insurance companies are seen to use the No Claim Discount (NCD) model to give a discount in premiums paid by the policyholders. To price a NCD sytem, the insured receives a discount in the absence of claims for a number of years (Kliger & Levikson, 2002)

Let the number of claims that may arise from vehicle insurance be denoted by N, which is a random variable, and $X_i$ represents the amount of the $i^{th}$ claim, $i=1,2...N$. We assume that the distribution of $N$ is a discrete distribution. This is because claims are discrete and non-negative in nature. Therfore, the number of claims can be modeled using discrete distributions

such as Binomial, Negative Binomial and Poisson distributions which are further discussed in the literature review.

In vehicle insurance, policyholders are divided in to different tarrif groups and for each group and we expect different values and numbers of claims. Since we expect different expected values for each group, an aspect of heteroskedasticity is brought up in our data. Heteroskedasticity is the aspect of variance not being constant across the observations. Therefore, this renders it impossible to use the ordinary least square (OLS) regression which is best referred to as the classical linear regression as OLS assumes homoskedasticity, which expects the errors in a a regression model to have constant variance conditional on the explanatory variables (Wooldridge, 2013). For instance, if a poisson distribution is used for the number of claims, the mean of the distribution is expected to be positive and following that mean depends linearly on the explanatory variables, it is not easy to guarantee its positivity. A logarithmic transformation is therefore used to guarantee that the mean will be positive leading us to a multiplicative model rather than a model with an additive effect on the mean (Anderson, et al., 2007). This then leads to the generalised linear model that has been used in actuarial work in the past and its work in vehicle insurance has been reviewed in details in the literature review.

In the past, actuaries have been relying mostly on one- way analysis for monitoring performance (Anderson, et al., 2007). One-way analysis only gives summarised statistics for each value of explanatory variables , not taking in to account the effect of other variables. One- way analysis can also be distorted by correlations between rating factors and also suffers from sequencing biasing. Two- way analysis also suffers similar drawbacks as those of the one-way analysis and hence need for multivariable methods that adjust for correlation and allow for investigation in to interaction terms (Anderson, et al., 2007).

Therefore, multivariable methods such as generalised linear models and minimum bias procedure can be used. One major disadvantage of minimum bias procedure is that once a solution has been calculated, no systematic test is given to find out the statistical significance of the tested variables (Anderson, et al., 2007). However, there is a connection that has been established between minimum bias procedures and generalised linear models that is well explained by (Mildenhall, 1999 ).

## 1.2 Problem Statement

Every insurance company's main aim is to make profit. In the recent past, however, most insurance companies have been making losses which can be attributed to poor pricing of premiums as is seen in (Gilneko & Mironova, 2017). According to (Garrido, Genest, & Schulz, 2016), during pricing of the pure premium, one considers the claim frequency, claim severity and a correction term to reflect the dependence in the two.

Claim frequency in the past has been calculated using both one way and two-way analysis. These have been having their disadvantages such as being distorted by correlation between rating factors and do not consider interdependencies between factors. Therefore, need arises to identify a suitable way of modelling claim frequency using a multivariable model such as generalized linear models which adjust for correlations and allow investigation into interdependencies. For instances where a Poisson generalised linear model has been used, issues have been raised concerning the variance in data (Valecký, 2016). A Poisson distribution assumes that the mean and variance is the same all along in data and there is need for further investigation on this.

## 1.3 Research Objectives

1. To review different generalized linear models available for modelling claim frequency.

2. To compare the performance of different generalised linear models with different distributions for claim frequency.
3. To determine factors that affect claim frequency.
4. To determine the characteristics of individuals purchasing vehicle insurance.

## 1.4 Research Questions

1. Which different generalized linear models are available for modelling claim frequency?
2. How do different generalised linear models compare when modelling claim frequency.
3. What are the factors affecting claim frequency?
4. What are the characteristics of individuals purchasing vehicle insurance?

## 1.5 Significance of research

This study is of benefit to actuaries, insurance companies that sell vehicle insurance policies as well as individuals who would be interested in using generalised linear models for predictive modelling.

The findings of this study will enable insurance companies to know what factors to consider when pricing premiums for vehicle insurance. It will help in allocation of costs when it comes to budgeting in the insurance companies.

The findings will also be of benefit to the society at large and the common man as it will give a clear picture of how to expect claim frequency to have an effect on premium pricing.

## 1.6 Scope of study

This study focused on motor vehicle insurance. It particularly looked at the claim frequency in vehicle insurance data. Moreover, it looked at the different models in which claim frequency can be modelled as well as the various factors that affect the claim frequencies.

# Chapter 2: Literature review

## 2.1 Overview of Generalized Linear Models

Since their establishment, generalized linear models have gained a lot of popularity (Nelder & Wedderburn, 1972). Generalized linear models are not only being used in the actuarial world but also in other fields. As elaborated by (Powers, Meyer, Roebuck, & Vaziri, 2005), GLM models are being used in the medicinal world to help in predictive modelling. (Famoye & Singh, 2006) state that the generalized Poisson model has been used to model dispersed count data. By default, insurance claims data is a form of count data. The generalized Poisson model hence is a good competitor to the negative binomial model when count data is dispersed. It is used to model count data that are affected by several known predictor variables. (Famoye & Singh, 2006) go ahead to propose the use of the zero-inflated generalized Poisson (ZIGP) regression model for their study which involved modelling domestic violence data with too many zeros. They go ahead to estimate the model parameters using the method of maximum likelihood. Moreover, GLM models have gained popularity in life insurance and are being used continuously. GLM models have been used in the construction of life tables too. The advantages of using GLM models has been dealt with in many papers but (Anderson, et al., 2007) goes ahead to give advantages such as GLM models allow for adjustment of correlation among variables as well as investigation in to interaction terms. GLMs extend the framework of linear regression models with normal distribution to the class of distributions from the exponential family (Kafkova & Krivankova, 2014)

GLM models require certain assumptions to hold for them to be well used. (Ewald & Wang, 2015) give the assumptions in place when using GLM models. The assumptions include:

1. Error terms can follow several distributions from the exponential family.

2. Variance needs not to be constant.
3. Covariates can be transformed so that their effect must not be additive.

GLM models allow for variance to adjust with mean.

As seen in (Ewald & Wang, 2015), generalised linear models are made of 3 components which include:

1. Random component which gives the distribution of the error term.
2. Systematic component that that gives a summation of all covariates that develop the predicted value
3. Link function which links the linear predictor to the expected value of the dataset.

$$E[Y_i] = g^{-1}(\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_n x_{i,n}) + e_i$$

*Equation 2.1: Expected value of Y which combines the link function and linear predictor under GLMs*

Where:

$g^{-1}$ or is the link function

$(\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_n x_{i,n})$ is the systematic component

$e_i$ is the random component

The set of probability density functions (PDF) of the exponential family is written in the form:

$$f(y) = exp\left[\frac{y\theta - b(\theta)}{a(\varphi)} + c(y, \varphi)\right]$$

*Equation 2.2: Probability density function of the exponential family*

Distributions from the exponential family are such as binomial, Poisson, negative binomial and gamma distributions. GLMs for count data prefer using the Poisson distribution with a link function of logarithmic. However,

this is not standard as one can change the distribution to suit the data available. A summary of the Poisson and negative binomial distribution, zero inflated Poisson and zero inflated negative binomial regression models which are most popular for count data is found in the methodology section.

The table below shows common link functions for various distributions:

| **Distribution** | **Link function** | $g(\mu)$ |
|---|---|---|
| Normal | Identity | $\mu$ |
| Poisson | Log | $ln(\mu)$ |
| Binomial | Logit | $ln\left(\frac{\mu}{1-\mu}\right)$ |
| Gamma | Inverse (reciprocal) | $\frac{1}{\mu}$ |
| Exponential | Log | $ln(\mu)$ |

*Table 2.1Canonical link functions for distributions*

Generalised linear models give a simple way to analyse the effect of different factors of interest on an observed effect.

## 2.2 Empirical Literature Review

Insurance companies have been known to sell policies to individuals with the purpose of indemnifying the policyholder in case of a loss or accident. In vehicle insurance, policyholders will be paid once they make a claim. Claim frequency can hence be defined as the number of claims occurring or better the number of times a claim occurs. Many researchers have been able to identify how generalized linear models can be used in modelling claim frequency and a lot of literature exists based on this.

(Yao, 2013) gives a detailed discussion on the use of generalised linear models in the pricing of non-life insurance. The discussion revolves around six overlooked facts of using GLMs for pricing which include the fact that model predictions depend on the mixture of rating factors in the data as well as the fact that the link function used in the model could bias it and

significantly change the lower and upper bound of the prediction. A conclusion is made that where a model of high uncertainty is used, the premiums estimated by the GLM could be underestimated. This therefore requires that a model of high certainty should be used or a mark -up should be added to the estimated premium to cater for the risk of high uncertainty. Moreover, Yao recommends that when there is uncertainty in a model being used for pricing, this should be communicated to the business management to allow them to make informed decisions concerning the business being put out to the public.

(Kafkova & Krivankova, 2014) echo the fact that actuaries in insurance companies struggle with getting the best model for estimation of insurance premium. They go ahead to explain that estimation of insurance premium depends on various risk factors which include the vehicle characteristics as well as the profile of the driver. (Kafkova & Krivankova, 2014) therefore, try to investigate and predict the relation between annual claim frequency and various risk factors.

From a dataset of 57,410 vehicles, (Kafkova & Krivankova, 2014) use generalised linear models to predict annual claim frequency. In their study, they also explain why the standard linear regression is not the best for modelling the claim frequencies. The main disadvantage of the standard linear regression model is that it tends to assume that the observations are normally distributed. This, however, may not be the case with insurance data as it tends to be skewed. (Kafkova & Krivankova, 2014) hence go ahead to find the best distribution for their insurance claim data to be a Poisson distribution. They go ahead to use the log-link function to transform it to a generalised linear model. On running the model on R, they find out that the risk factors that should be best considered when modelling annual claim frequency include vehicle age, age band of the policyholder and area of residence of the policy holder. However, the results which include risk factors to consider, vary based on different datasets. The use of the Poisson

distribution is, however, slightly limiting depending on the data available. This is because the issue of overdispersion is not dealt with as a Poisson distribution assumes same variance across the observations in the data. Therefore, need arises to consider other distributions that can deal with overdispersion.

To address the issue of overdispersion in the insurance claim data, (Valecký, 2016) went ahead to model claim frequency as an extension of the works of (Kafkova & Krivankova, 2014). He went ahead to show that when modelling claim frequencies, it is important to consider overdispersion, non-linear systematic component and interacted rating factors. (Valecký, 2016) shows that one of the methods used to address non-linearity is the use of fractional polynomials. He also compared the Poisson model to the negative binomial model which was derived as a Poisson-Gamma mixed model. The data used is from a Czech insurer and was collected over the period 2004 to 2008. To compare the different models, the standardised Pearson's coefficient and deviance residuals are analysed. To conclude, the Poisson model was seen not to cater for overdispersion. The negative binomial model was then employed to show that considering heterogeneity in insurance policies yields a better fit model. (Valecký, 2016) hence concluded that when using fractional polynomials and interactions in the modelling, large datasets should be used to avoid the methods being questionable. The modelling techniques mentioned also tend to be computationally demanding.

Research has also been done specifically on claim amounts best referred to as claim severity. (Smyth & Jørgensen, 2014) did a study that aimed at reconsidering the issue of producing fair and accurate insurance tariffs based on aggregated insurance data that gave the number of claims and the total cost of the claims. They noticed that in order to model the cost of insurance claims, it was key to model the dispersion of the costs and model their mean too. Modelling the dispersion in turn helps in making sure the estimated tariffs are as precise as possible. The use of double generalised linear models,

therefore, helps to handle cases and situations where only the total cost of claims and not number of claims has been recorded. A Tweedie's Compound Poisson distribution can therefore be used. It provides a highly efficient method of analysing insurance claims data. However, its efficiency is slightly flawed as most terms are likely to be found to be significant in the fitted model compared to from other methods.

Some researchers, moreover, are seen to have an interest in modelling both the claim frequency and the claim severity. (Goldburd, Khare, & Tevet, 2016) wrote about the use of generalised linear models in insurance rating emphasising on the application of the theory. Moreover, they bring out the aspect of pricing a pure premium by looking at both the claim frequency and the claim severity. For both claim frequency (count of claims per exposure) and claim severity (amount of claim), (Goldburd, Khare, & Tevet, 2016) indicate that they can be predicted by using GLMs. Modelling both the claim frequencies and claim severity separately provides a lot of insight as it allows us to see which factors are frequency-driven versus the factors that are severity-driven. After selecting the target variable, a distribution that fits the data most closely out of the set of the possible distributions is chosen and the model can then be fit. However, (Goldburd, Khare, & Tevet, 2016) indicate that GLMs have the limitation of assigning full credibility to data used. When using GLMs for modelling both claim frequencies and severities, it would therefore be important to come up with necessary assumptions that help curb the above limitation.

When it comes to modelling claim frequency and severity, a detailed study is done by (Gilneko & Mironova, 2017). Their study was based on Russian Motor Own Damage Insurance (MOD) and the data analysed was provided by a leading insurance company in Russia. They recommend using a hurdle model for modelling claim frequencies and a GLM-Gamma distribution model for claim severities. This clearly shows that GLMs can not only used for modelling claim frequencies as is my objective in this study but also claim

severities which is basically the claim amounts. The hurdle model has been used for claim frequencies as it describes the link between regressors and the count dependent variable that has a large number of observations with zero values which is very important in vehicle insurance (Mullahy, 1986). This model basically consists of two components; the component for zero values (when a policyholder does not claim) and the component for positive count values which basically cover for claims that are occurring once or more times. However, the limitation when using the hurdle model is that insufficient data can lead to vagueness of results got. (Gilneko & Mironova, 2017) recommend that data used should include more factors such as vehicle's mileage to ensure clear results.

However, some researchers believe that when count data contains extra zeros, zero-inflated regression models should be used. (Yip & Yau, 2005) are seen to echo the fact that claims frequency data may tend not to follow the traditional Poisson distribution. If the data is zero-inflated, extra dispersion is seen to appear as the number of observed zeros in the data could differ from the number of expected zeros. Under the Poisson and negative binomial distribution assumptions. Moreover, Moreover, (Yip & Yau, 2005) go ahead to give different parametric zero-inflated count distributions that can accommodate the excess zeros from the insurance claim data. Such distributions include, but are not limited to, zero-inflated Poisson distribution and zero-inflated negative binomial distribution. After testing the goodness of fit using the Akaike's Information Criteria (AIC), Bayesian Information Criteria (BIC) and the Pearson $\chi^2$ statistic, the zero-inflated count models are seen to be suitable for data with many observed zero claims. This was verified using an automobile insurance claims data set that had extra zeros present.

(Rodriguez, 2013) has written a detailed paper comparing the different models available for modelling count data with overdispersion. He compares the extra-Poisson variation, negative binomial model, zero inflated models

and hurdle models. In his comparison, he notes that the Poisson distribution assumes equal variance and equal mean but more often than not, data is found to have its variance being greater than the mean. Need therefore arises to consider other models that can accommodate the extra variation. The negative binomial regression model comes in to play as it caters for overdispersion as well as unobserved heterogeneity that may be present in data. To get the moments such as mean and variance which are necessary for determining overdispersion, the law of iterated expectations can be used. Since the Poisson model is a special case of the negative binomial model when $\sigma^2 = 0$, the likelihood ratio test can then be used to compare the two models. (Rodriguez, 2013) went ahead to present zero-inflated models which include the zero-inflated Poisson model and the zero-inflated negative binomial model that can be used to model empirical data that shows a lot of zeros. The zero-inflated models tend to be very appealing to use but interpretation may not always be easy depending on the data being analysed. Hurdle models have also been identified as being useful in modelling data with excess zeros. A truncated Poisson model falls under this class and differs from the zero-inflated Poisson model as its classes are observed rather than latent, where one consists of observed zeros and the other of observed positive counts. Interpretation of results in these models is easier compared to zero-inflated models.

(Ismail & Zamani, 2013) have studied the generalized Poisson and negative binomial regression models through the mean-variance relationship. Moreover, they have compared the zero-inflated negative binomial and zero-inflated generalized Poisson regression models through their mean-variance relationship as well and suggested their application for both over-dispersed data as well as data with excess zeros. The generalized Poisson and negative binomial regression model are seen to be suitable for application while using over-dispersed or under-dispersed count data. For the zero-inflated models, (Ismail & Zamani, 2013) fitted the models to a claim count dataset that

differed from the one used to fit the generalized Poisson and negative binomial models.

(Ismail & Zamani, 2013) also went ahead to compare the two divisions separately using different tests. To test overdispersion in Poisson versus negative binomial regression models, the likelihood ratio test is performed that is one-sided. The Wald test is also done to test for overdispersion between the previous two models. The Vuong test and the information criteria are also brought into light when comparing the generalized Poisson and the negative binomial model. When comparing the zero-inflated models, they used both the likelihood ratio tests and the Wald test. They show that zero inflated models are best if the occurrences of count events depend on specific conditions or time.

(Shi, Feng, & Ivantsova, 2015) have done a detailed study on dependent frequency-severity modelling of insurance claims. In their study, they explore methods that can allow for correlation among frequency and severity components for micro-level insurance data. The hurdle modelling framework is therefore introduced under which two approaches are suggested to accommodate the dependency between frequency and severity of insurance claims. Based on conditional probability decomposition, the first approach treats the number of claims as a covariate in the regression model for average claim size. The second approach employs a copula approach to formulate the joint distribution of the number and size of claims. The two approaches helped in accommodating heavy tails, excess zeros as well as over-dispersion. However, it is noted that more work needs to be done in examining the frequency-severity dependency in a longitudinal context.

From the detailed literature review done above, it is seen that in the past, different methods and models have been used for modelling claim frequencies. There are different generalized linear models which exist for modelling insurance claim data which is a form of count data.

The Poisson model has been identified as being helpful for modelling count data that has mean which is equal to its variance. In the case of over dispersion, the negative binomial regression model has been identified to help solve the issue. Negative binomial regression models have been noted to accommodate over-dispersed data and in turn give valid results.

However, researchers tend to agree that when there is preponderance of zeros in count data, other models need to be examined. This then leads to zero-inflated models that are seen to accommodate both overdispersion and extra zeros in data. Many zero-inflated models exist but our attention is particularly drawn to the zero-inflated Poisson model and the zero-inflated negative binomial model. No model has so far been concluded to be the best for modelling claim frequencies. This is because the goodness of fit of a model will depend on the data that has been fit to it.

# Chapter 3: Methodology

## 3.1 Summary of the GLM models

### 3.1.1 Poisson distribution

The following are the characteristics of a Poisson distribution; events are known to occur randomly and singly, it is purely discrete and λ which is a single parameter, defines the distribution and proportion to the frame of measurement.

The probability density function (pdf) of a Poisson distribution can be written as follows:

$$P\,(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

*Equation 3.1: Probability density function of a Poisson distribution*

for $x = 0,1,2 \ldots$

Where:

$e$ is the base of natural logarithms

$\lambda$ is the mean number of "successes"

$x$ is the number of "successes" in question

### 3.1.2 Negative Binomial Distribution

According to (Cook, 2009) the probability distribution function of a negative binomial distribution can be summarised as follows:

The distribution counts the number of the trial at which the $r^{th}$ success occurs. This is denoted as:

**Negative Binomial- Type 1**

The probability density function of the distribution can be denoted as:

$$P(X = x|p,r) = \binom{x-1}{r-1} p^r (1-p)^{x-r}$$

*Equation 3.2: Probability density function of a negative binomial distribution-Type 1*

for integer $x \geq r$. Here, $0 < p < 1$ and $r$ is a positive integer number.

The mean in this case is: $\lambda = \frac{r}{p}$ and the variance is $Var(x) = \frac{r(1-p)}{p^2}$

**Negative Binomial-Type 2**

This is the most commonly used type of the negative binomial distribution. The probability density function of the distribution is denoted as:

$$P(X = x) = \frac{\Gamma(r+x)}{\Gamma(x+1)\Gamma(r)} p^r (1-p)^x$$

*Equation 3.3: Probability density function of a negative binomial distribution-Type 2*

for integer $x \geq r$. Here, $0 < p < 1$ and $r$ is a positive integer number.

In this case, the mean is; $\lambda = \frac{r(1-p)}{p}$ and the variance remains the same as that of the negative binomial-type 1; $Var(x) = \frac{r(1-p)}{p^2}$

3.1.3 Zero Inflated Regression Models

These models are often used to account for the predominance of excess zeros frequently observed in count data. There are many zero inflated models that exist but for this study, we will look at the zero inflated Poisson model and the zero inflated negative binomial model.

1. **Zero Inflated Poisson Model (ZIP)**

In the recent past, the model has gained popularity among researchers handling purely zero inflated count data. The model assumes that with probability $p$, the only possible observation is zero and with a probability $1-p$, a Poisson random variable $(\lambda)$ is observed. The probability density function can then be defined as:

$$P(Y_i = y_i) = \begin{cases} p_i + (1-p_i)e^{-\lambda_i}, & y_i = 0 \\ (1-p_i)\dfrac{\lambda_i^{y_i}}{y_i!}e^{-\lambda_i}, & y_i \geq 0 \end{cases}$$

*Equation 3.4:Probability density function of a ZIP distribution*

Where $0 \leq p_i < 1$ and $\lambda_i > 0$, with mean, $E(Y_i) = (1-p_i)\lambda_i$ and variance, $Var(Y_i) = (1-p_i)\lambda_i(1+p_{i\lambda_i})$.

When $p_i = 0$, the zero inflated Poisson model reduces to Poisson regression. The covariates can then be incorporated by using a log link for $\lambda_i$ and a logit link function for $p_i$.

2. **Zero Inflated Negative Binomial Model (ZINB)**

This model has been used to handle both zero inflated data and over-dispersed data. The ZINB type 1(ZINB-1) can be obtained by mixing a distribution degenerate at zero with the NB type 1(NB-1) distribution. This can be done by allowing the incorporation of explanatory variables in both the zero process and the NB-1. The same can be done to obtain the zero inflated negative binomial type 2 (ZINB-2).

3.2 Description of data

We used the brvehins1 dataset that is a subset of CASdatasets. This dataset is publicly available and can be accessed on the internet (CASdatasets 1.0-6.zip). The dataset brvehins1 of 1,965,355 vehicle insurance policies has been split in a random manner in to five sub-datasets of 393,071 policies. Each record in the dataset includes risk features, claim amount and claim history for the year 2011. Due to the size of data, we randomly selected the brvehins1e dataset which is one of the five sub-datasets. We further went

ahead to randomly select 10,000 entries to be used for analysis for computational purposes. For both modelling and data analysis, R software was used to run codes.

3.3 Fitting models

This entailed estimation of parameters that minimize the differences between the data values and the predicted values. Estimation of parameters was done using the maximum likelihood estimation (MLE) method. The four models described above were fitted to the 10,000 observations of the brvehins1 dataset.

3.4 Model comparison

The performance of the fitted models was compared based on how well they fit the data. The comparisons were done based on the Vuong test and both the Akaike's and Bayesian Schwartz information criteria.

**a) <u>Vuong Test</u>**

First established by (Vuong, 1989), the test is proposed for non-nested models. The test simply states that under the null hypothesis that two non-nested models fit equally, that is, that the expected value of their log-likelihood ratio equals zero, then under the null hypothesis, $H_0$, the asymptotic distribution of the log-likelihood ratio statistic (LR) is normal. Under the Vuong test;

$H_0$: *competing models provide an equally good fit of the data*

$H_a$: *one model provides a better fit of the data*

When comparing two models with pdf $p_1(.)$ and $p_2(.)$, the Vuong statistic can be defined as (Ismail & Zamani, 2013):

$$V = \frac{\bar{m}\sqrt{n}}{s\,d(m)}$$

*Equation 3 5: Vuong Statistic*

Where $\bar{m}$ is the mean of $m_i$, $s\,d(m)$ is the standard deviation of $m_i$, $n$ is the sample size and $m_i = \ln\left(\frac{p_{1i}(y_i)}{p_{2i}(y_i)}\right)$. The significance level to be used is at the

discretion of the researcher. However, in this study, I will use 0.05 significance level.

**b) Information Criteria**

As we fitted the models, the aim was to minimise the loss of information in the models. The information criteria show how much information is lost in a model. Information criteria always help in creating a balance between the accuracy of fitting data and how complex a model is.

**i) Akaike Information Criteria (AIC)**

The AIC equation is given as, $AIC = -2l + 2k$ where $l$ is the logarithm of the likelihood function of the proposed model and $k$ is the number of model parameters. The model with the lowest AIC is always preferred. However, it is good to note that the AIC will be computed easily on R without necessarily imputing the equation.

**ii) Bayesian Schwartz Information Criteria (BIC)**

Similarly, the BIC shows how much information is lost in a model. The lower the BIC, the better the model. BIC penalizes a model with a larger number of parameters and larger sample size. It is defined as: $BIC = -2l + k\ln(n)$ where $l$ is the logarithm of the likelihood function of the proposed model and $k$ is the number of model parameters and $n$ is the sample size.

3.5 Determining factors that affect claim frequency

After finding the best fit model, we determined the factors affecting claim frequency based on that specific model. To do so, we will use the p-values.

**a) P-value**

Under the null and alternative hypotheses, the p-value will be used to decide whether to reject or fail to reject the null hypothesis. The p-value is the probability of obtaining a test statistic that is at least as extreme as the actual calculated value, if the null hypothesis is true.

In this study, the cut-off for the p-value was 0.05, such that if the calculated p-value was less than 0.05, the null hypothesis was rejected.

Chapter 4: Results and discussions

The results highlighted in this section are categorized into two. The first section gives the descriptive statistics of the data being analysed whereas the second section goes ahead to fit the models and give the suitable model for the data.

4.1: Descriptive statistics

The table below shows the variables that will be used in fitting the of the GLM models as well as their characteristics.

| **Notation** | **Frequency (N)** | **Percentage (%)** |
|---|---|---|
| **Gender** | | |
| Male | 5,113 | 51.13% |
| Female | 3,924 | 39.24% |
| Corporate | 963 | 9.63% |
| **Total** | **10,000** | 100% |
| **DrivAge - Age of insured** | | |
| 18-25 | 798 | 7.98% |
| 26-35 | 2,057 | 20.57% |
| 36-45 | 2,546 | 25.46% |
| 46- 55 | 2,316 | 23.16% |
| above 55 years | 2,283 | 22.83% |
| **Total** | **10,000** | **100%** |
| **State** | | |
| Sao Paulo | 2362 | 23.62% |
| Minas Gerais | 1065 | 10.65% |
| Parana | 1012 | |

| | | |
|---|---|---|
| Santa Catarina | 899 | 8.99% |
| Rio Grande do Sul | 874 | 8.74% |
| Rio de Janeiro | 746 | 7.46% |
| Others | 3042 | 30.42% |
| **Total** | **10,000** | **100%** |
| Yeargroup | | |
| 1950 – 1959 | 1 | 0.01% |
| 1960 – 1969 | 1 | 0.01% |
| 1970 – 1979 | 14 | 0.14% |
| 1980 – 1989 | 92 | 0.92% |
| 1990 – 1999 | 1297 | 12.97% |
| 2000 – 2009 | 6949 | 69.49% |
| 2010 and above | 1646 | 16.46% |
| **Total** | **10,000** | **100%** |
| Veh | | |
| Fiat | 1525 | 15.25% |
| Renault | 649 | 6.49% |
| GM Chevrolet | 1657 | 16.57 |
| Volkswagen | 1606 | 16.06% |
| Others | 4563 | 45.63% |

| | | |
|---|---|---|
| Total | **10,000** | **100%** |
| **Median number of claims:** | 1 | 1st Quartile: 0<br>3rd quartile: 1 |

*Table 4.2: Descriptive statistics of variables*

4.2.1 Model fitting and comparison

As described in the methodology, I went ahead to fit four different generalized linear models on R. The models fitted were Poisson, Negative Binomial as well as both zero inflated Poisson and zero inflated negative binomial models respectively.

Comparing the first two models using the Vuong Test and the Akaike's and Bayesian information Criteria, the negative binomial model stood out to be superior. This is because the model has a lower AIC and BIC. The Vuong statistic is also -10.2101, which is greater than -1.96, showing that the negative binomial model provides a good fit of the data. The results of the fit of the model and tests taken were as summarised below:

| **Model type** | **AIC** | **BIC** | **Vuong test statistic** |
|---|---|---|---|
| Poisson | 36,790 | 37,550 | -10.2101 |
| Negative Binomial | 15,890 | 16,649 | -10.2101 |

*Table 4.2 1: Results of the AIC,BIC and Vuong after fitting the Poisson and negative binomial GLMs*

When it comes to fitting the zero inflated models, the variable that is considered is only the policyholder's age and the number of claims. This is because the variable of policyholder age is numeric in nature hence the component of the zero inflation can be analyzed from the data. A Vuong test was also carried out and both Akaike's and Bayesian information criteria used to compare the zero inflated Poisson and zero inflated negative binomial model whose results are as follows:

| Model type | AIC | BIC | Vuong test statistic |
|---|---|---|---|
| Zero Inflated Poisson | 34,353.8 | 34,405 | -7.32 |
| Zero Inflated Negative Binomial | 17083.52 | 17134.72 | -7.32 |

*Table 4.2 2: Results of the AIC,BIC and Vuong after fitting the zero-inflated Poisson and zero-inflated negative binomial models*

Moreover, I fitted the four models, considering vehicle age as the only variable to allow for equal comparison. The results showed that the zero inflated negative binomial model provided the best fit based on the 3 tests. The results are as shown in the table below:

| Model type | AIC | BIC | Vuong test statistic |
|---|---|---|---|
| Zero Inflated Poisson | 34,353.8 | 34,405 | - |
| Zero Inflated Negative Binomial | 17083.52 | 17134.72 | -1.40317 |
| Poisson model | 46880 | 46911.94 | - |
| Negative Binomial model | 17180 | 17222.27 | -1.40317 |

*Table 4.2 3 Results of the AIC,BIC and Vuong after fitting the four models*

4.2.2 Determining factors that are significant in modelling claim frequency

Since the negative binomial model was found to be the model of best fit when fitted with all variables, I went ahead to check which of the fitted variables was significant using various methods as stated in the methodology section.

4.2.2.1 P-Value

After running an analysis of variance (ANOVA) test on the saturated model, the p-values of the factors gender, state of residence, driver's age, the model of the vehicle and the vehicle year are lower than 0.05 which shows a confidence level of over 95%. This implies that the factors are statistically significant and should be included in the model. The results of the ANOVA test are as shown in the table below:

| **Variable** | **P-Value** | **Significance** |
|---|---|---|
| Gender | < 0.00000000000000022 | *** |
| State | < 0.00000000000000022 | *** |
| Vehicle's year of manufacture | < 0.00000000000000022 | *** |
| Driver's age | 0.007669 | ** |
| Vehicle type | < 0.00000000000000022 | *** |
| Significance codes: 0 '***' | 0.001 '**' 0.01 '*' 0.05 '.' 0.1 | " 1 |

*Table 4.2 4: P-values of showing the significant factors at 0.05*

The best model therefore to be used is the Negative Binomial Generalized Linear Model. The factors to be considered when fitting the model from the data are the policyholder's gender, age, state of residence, vehicle type and the year of manufacture of the vehicle.

# Chapter 5 Conclusion and recommendations

## 5.1 Conclusion

Insurance data is not normally distributed hence the need for finding another method of modelling it as opposed to the usual linear regression method. Generalised linear models have proven to be of importance when it comes to modelling non- normal data as they allow you to specify which distribution you want to assume from the exponential family. When it comes to modelling claim counts, the negative-binomial distribution model is seen to be the best. From this specific study, it is seen that the risk factors; driver's age, vehicle year, gender, vehicle type and state are of importance when it comes to modelling the claim frequency of vehicle insurance data. However, given that our data also included corporate policies sold, the risk factors may vary depending on the given type of data. Moreover, while fitting GLM models, it is good to bear in mind that your model should not be complex. This will aid in understanding the model in an easy and simple manner.

## 5.2 Limitations of the study

During analysis of the data, the zero inflated models failed to pick up various variables when running the code. This is still an area that I am still conducting further research on.

## 5.3 Recommendations

Generalised linear models in vehicle insurance would be a good starting point for students trying to further their knowledge on the topic of generalised linear models. This is because it helps you to apply your coursework into real life data. Moreover, this could be an area of interest to researchers as well as insurance companies that sell vehicle insurance policies.

## 5.4 Areas of further study

As we have seen, generalised linear models are a good way of modelling claim frequency of vehicle insurance data. It would also be better if the models are fit to more than one data set and comparison made to ensure that

the best model is identified. There is a need to further explore how else generalised linear models can be use in both life and non- life insurance. One could also seek to research on how to use generalised linear models in micro-insurance which is an upcoming sector in micro-insurance.

References

Anderson, D., Feldblum, S., Modlin, C., Schirmacher, D., Schirmacher, E., & Thandi, N. (2007, February). A Practitioner's guide to Generalized Linear Models. Towers Watson.

Cook, J. D. (2009, October 28). Retrieved from https://www.johndcook.com/negative_binomial.pdf

CT6-PN-16 Course Notes. (2016). The Actuarial Education Company.

Ewald, M., & Wang, Q. (2015, June). Predictive modeling: A modeller's introspection. Schaumburg, Illinois, United States of America: Society of Actuaries.

Famoye, F., & Singh, K. P. (2006). Zero-inflated Generalized Poisson regression model with an application to domestic violence data. *Journal of Data Science*(4), 117-130.

Garrido, J., Genest, C., & Schulz, J. (2016). Generalized linear models for dependent frequency and severity of insurance claims. *Insurance: Mathematics and Economics, 70*, 205-215.

Gilneko, E. V., & Mironova, E. A. (2017). Modern claim frequency and claim severity models: An application to the Russian motor own damage insurance market. *Cogent Economics & Finance*, 1.

Goldburd, M., Khare, A., & Tevet, D. (2016). *Generalised Linear Models for indurance rating.* Arlington: Casualty Actuarial Society.

Ismail, N., & Zamani, H. (2013). Estimation of claim count data using Negative Binomial, Generalized Poisson, Zero-Inflated negative binomial and Zero-Inflated generalized Poisson regression models. *Functional forms of negative binomial, generalized Poisson, zero-inflated negative binomial and zero-inflated generalized Poisson regression models*. Arlington County, Virginia, United States: Casualty Actuarial Society.

Kafkova, S., & Krivankova, L. (2014). Generalised linear models in vehicle insurance. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, 383-388.

Kliger, D., & Levikson, B. (2002). Pricing no claim discount systems. *Insurance: Mathematics and Economics*, 191-204.

Mildenhall, S. (1999 ). A systematic relationship between minimum bias and generalised linear models. *Proceedings of the Casualty Actuarial Society.*

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics, 33*(3), 341-365.

Nelder, J. A., & Wedderburn, R. W. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society*, 370-382.

Powers, C. A., Meyer, C. M., Roebuck, M. C., & Vaziri, B. (2005). Predictive modelling of total healthcare costs using pharmacy claims. *Medical care*, 1065-1072.

Ramos, P. L. (2017). Premium calculation in insurance activity . *Journal of Statitistics and Management Systems*, 39-65.

Rodriguez, G. (2013). Models for count data with overdispersion.

Shi, P., Feng, X., & Ivantsova, A. (2015). Dependent frequency-severity modeling of insurance claims. *Insurance; Mathematics and Economics.*

Smyth, G. K., & Jørgensen, B. (2014). Fitting tweedie's compound Poisson model to insurance claims data: Dispersion modelling. *The Journal of the International Actuarial Association*, 143-157.

Valecký, J. (2016). Modelling claim frequency in vehicle insurance. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, 683-684.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica, 57*, 307-333.

Wooldridge, J. M. (2013). *Introductory Econometrics* (5th ed.). South-Western, Cengage Learning.

Yao, J. (2013). *Generalized Linear Models for non-life pricing - Overlooked facts and implications.* London: Institute and Faculty of Actuaries.

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

Appendix

Codes run in R

```
library(CASdatasets)

library(dplyr)

library(MASS)

library(ggplot2)

library(boot)

library(pscl)


data("brvehins1e")

missing_values<-is.na(brvehins1e)

#counts the number of missing variables


new_data<-na.omit(brvehins1e)

#creates new data after omitting zero fields


View(new_data)

#allows us to view the new data


new_data$sum<-
new_data$ClaimNbRob+new_data$ClaimNbPartColl+new_data$ClaimNbT
otColl+new_data$ClaimNbFire+new_data$ClaimNbOther

#creates a new variable for total claims
```

```
summary(new_data)
#gives the summary statistics of the variables

library(ggplot2)
ggplot(new_data,aes(sum))+ geom_histogram()+scale_x_log10()
#gives a plot of frequency of claims

sdata=new_data[sample(nrow(new_data), 10000),]
#randomly selects 10,000 entries that will be used for computational
purposes

sdata$veh=substr(sdata$VehGroup,start = 1,stop = 4)
#allows us to group the vehicle type eg: Nissan, Citroen, based on the first
four characters
summary(as.factor(sdata$veh))

sdata$yeargroup=substr(sdata$VehYear,start = 1,stop = 3)
#allows us to group vehicles' years of manufacture
summary(as.factor(sdata$yeargroup))

View(sdata)
attach(sdata)
summary(sdata)
```

```
ggplot(sdata,aes(sum))+ geom_histogram()+scale_x_log10()

#gives a plot of frequency of claims from the data sdata


poissonmod1 <-
glm(sum~Gender+State+DrivAge+as.factor(yeargroup)+veh,family =
poisson(link = log),data=sdata)

#allows us to fit a glm model of the Poisson family and log-link function

poissonmod2 <- glm(sum~DrivAge,family = poisson(link = log),data=sdata)

poissonmod2


nbmod1 <- glm.nb(sum~Gender+State+DrivAge+as.factor(VehYear)+
veh,data=sdata)

#allows us to fit a glm model of the negative binomial family

nbmod1

nbmod2 <- glm.nb(sum~DrivAge,data=sdata)

nbmod2


AIC(poissonmod1,nbmod1)

BIC(poissonmod1,nbmod1, nbmod2, poissonmod2)

vuong(poissonmod1,nbmod1)

#returns the vuong test statistic
```

```
ZIP1 <- zeroinfl(sum~ DrivAge | DrivAge, dist="poisson", link = "logit", data
= sdata)

#fits a zero inflated poisson model

ZIP1


ZINB1 <- zeroinfl(sum~ DrivAge | DrivAge, dist="negbin", data = sdata)

#fits a zero inflated negative binomial model

ZINB1

AIC(ZIP1,ZINB1)

BIC(ZIP1,ZINB1)

vuong(ZIP1,ZINB1)

#returns the vuong test statistic for the zero-inflated models

anova(nbmod1)

#returns an analysis of variance for the model and gives the p-values
```