UNIVERSITY OF GHANA

ADOPTING ZERO INFLATED MODELS FOR CLAIM COUNTS AND THE GAMMA REGRESSION MODEL FOR CLAIMS COST IN DETERMINING ACTUARIAL PREMIUMS

BY

FRED MAWULI AMENU

(10379715)

THIS THESIS IS SUBMITTED TO THE UNIVERSITY OF GHANA, LEGON IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE AWARD OF MPHIL IN ACTUARIAL SCIENCE DEGREE

April, 2022

DECLARATION

Candidate's declaration:

I Fred Mawuli Amenu hereby declare that this submission is my own work towards the award of the Master of Philosophy degree and that, to the best of my knowledge, it contains no material previously published by another person nor material which had been accepted for the award of any other degree of the university, except where due acknowledgement had been made in the text.

SIGNATURE: ..... DATE: 11-04-22 .....

FRED MAWULI AMENU

(10379715)

Supervisors' declaration:

We hereby certify that this thesis was prepared from the candidates own work and supervised in accordance with guidelines on supervision of thesis laid down by the university of Ghana.

SIGNATURE: ..... DATE: ....05/04/2022.....

PROF. KWABENA DOKU-AMPONSAH

(Principal Supervisor)

SIGNATURE: ..... DATE: 05/04/2022 .....

DR. PERPETUAL ANDAM BOIQUAYE

(Co-Supervisor)

ABSTRACT

Insurance is the exchange of risk by an insured person through the payment of premiums for financial protection and economic benefit. The problem is how premiums should be charged so as to keep the industry alive to perform this basic function of insurance. Because of the Bonus-Malus system, or Hunger for Bonus system (also called No Claim Discount), and deductibles, most claims are not reported by policyholders, causing the number of claims to be dominated by zeros, which leads to over-dispersion in the data. In modeling the claim frequency, the Zero-Inflated Poisson (ZIP) and Zero-Inflated Negative Binomial (ZINB) models were adopted. The Gamma regression model was used to fit the claims cost data. The claim frequency regression model that best fits the claim frequency with the Gamma model for the claims cost was combined in determining the actuarial premium. These models were numerically illustrated with data obtained from a major non-life insurance company in Ghana and French Motor Third-Party Liability data from `https://www.kaggle.com/datasets/karansarpal/fremtpl2-french-motor-tpl-insurance-claims`. The score test demonstrated the inability of the Poisson model to appropriately model the claims data due to the inflation of zeros in the data. The ZIP and ZINB were both found to be superior to their conventional equivalents based on the Vuong test statistics. The ZIP was chosen as an appropriate model for analyzing claim frequency data for both the French and Ghanaian data based on the values of the AIC and BIC. The risk factors that were found to influence claim frequency and claim cost were discovered to be different when both datasets were used. It is recommended that a separate analysis of claim frequency and claim cost be conducted with claim frequency receiving a high rating power.

DEDICATION

To the memory of my late father Benard Selome Komase

ACKNOWLEDGMENT

To God be the glory. He has done great things and will do even greater things in the future.

I am highly indebted to my Principal Supervisor, Prof. Doku Amponsah, for his suggestions and instructions given throughout this work. I am also grateful to my co-supervisor, Dr. Perpetual Andam Boiquaye, for her advice and assistance, which helped me in diverse ways.

All thanks to Dr. Dennis Arku for his open arms towards helping to provide the data needed for the completion of this work.

To all my lecturers in the Department of Statistics and Actuarial Science, University of Ghana: Dr. Richard Minkah, Dr. Lotsi Anani, Dr. Ezekiel Nortey, Dr. Godwin Debrah, Dr. Louis Aseidu, Mrs. Charlotte Chapman-Wardy, and Mr. K. Davies for their availability and inspiration, and for the impact they had on me during the coursework.

Thank you so much, my beloved Nana-Akua!! You have been a blessing to me through this difficult journey. To "baby" Nevaeh Marx for your youthful exuberance and "pressure" that ignites in me the strength to press on. To my colleagues, Samuella Adams, Philip Osei-Owusu, and Isaac Essel, for their love and kindness towards me through the thick and thin.

I cannot forget Mr. Newman Chiri for all the materials he has supplied me on research methods.

To Joshua Adagblenya, Twucent Ayeh, David Nador, Nelson Deh, my bosom friends, for their encouraging words.

Contents

List of Tables

List of Figures

List of Abbreviation

GLM Generalized Linear Model

ZIP Zero-Inflated Poisson

ZINB Zero-Inflated Negative Binomial

CDF Cumulative Distribution Function

PDF Probability Density Function

PMF Probability Mass Function

r.v Random Variable

EDF Exponential Dispersion Family

AIC Akaike Information Criteria

BIC Bayesian Information Criteria

OECD Organization for Economic Cooperation and Development)

NIC National Insurance Commission

ZIGP Zero-Inflated Generalized Poisson

GLMM Generalized Linear Mixed Models

NB Negative Binomial

EM Expectation Maximization

NSFG National Survey of Family Growth

OLS Ordinary Least Squares

MLE Maximum Likelihood Estimation

LR Likelihood Ratio Test

pmf Probability Mass Function

pdf Probability Density Function

SE Standard Error

Chapter 1

INTRODUCTION

In this chapter, the background and statement of the problem, the objectives of the study, research questions, the organisation of the study, and the significance of the study were laid out.

1.1 Background and Problem Statement

Globally, the insurance industry, both life and non-life has seen high level of growth in the past few years. Gross premium continued to rise in most countries in 2018 especially in the non-life sector (Global Insurance Market Trends [OECD], 2018). From 50 countries, the OECD (201) found out that, premium in all these 50 countries grew by 2.5% and 3.5% in life and non-life insurance sectors respectively. These statistics were not too far from that of Ghana in the previous year, as the 2018 National Insurance Commission(NIC) report shows that, despite the low penetration in the insurance industries, the premium rose from GHC2.4 billion in the previous year to GHC2.9 billion in 2018. That is an increase of 21 percent.

One basic function of insurance in general is to protect the insured individual when the event of interest happens. The insured individual trades his or her risk through a premium payment. (David, 2015). That is, premiums payment are exchanged for pecuniary protection which is seen in the claims paid to the insured in the occurrence of the event of interest. The problem is how will premium be charged so as to keep the industry alive to perform this fundamental role. The

subject of risk, incorporates premiums in to the surplus process. Risk theory, a synonym for non-life insurance mathematics, deals with the modelling of claims that arrive in an insurance business and gives advices on how much premium has to be charged in order to avoid bankruptcy (ruin) of the insurance company (Miskosch,2009).

Since insurance portfolios involves heterogeneity of risk, i.e. the risk of an insured is different from other person seeking the same insurance, Insurers therefore have the task to charge insurance tariff (premiums) that is in all cases fair and equitable to the insured. In order to achieve this, insurers group insureds in to various homogeneous risk groups. In that, individuals with the same risk characteristics are charged the same premium so as to remove or if possible eliminate the problem of adverse selection and moral hazards. That is, on one hand good risks, with low risk profiles could pay too much and eventually prefer to leave the company, on the other hand, bad risks may find a uniform tariff to be in their favour and eventually, prefer to stay with the company. This can undermine the solvency of the insurance company (Antonio & Valdez, 2012). Actuarial models are to be employed so as to help in the classification of insurance risk by building a tariff structure for the insurance company which incorporate all forms of risk.

Actuaries use statistical tools, mostly regression models in developing a rate making structure for the insurance industry. In modelling claim count, the standard Poisson regression model is normally used, however due lack of reporting of claims leading to no claims within a given insurance period, hence presence of many zeros in insurace portfolio, zero-inflated regressing models are applied. The most common of these zero inflated models is the zero-inflated poisson (ZIP) model. The ZIP is a mixture of a Poisson distribution and a zero point mass.(Wolny-Dominiak,2013). The availability of many zeros in the data will mostly lead to over-dispersion which leads to greater variability in the data.To account for latent factors, the Zero-inflated negative binomial(ZINB) is

used (Wolny-Dominiak,2013). These two zero-inflated models were compared to the classical poisson regression model which assume equality of means and variance. It was investigated to find out whether there was a departure from the classical Poisson model in modelling claims count for the data.

A regression model is a model of the average response variable given the covariate.i.e. a model of conditional expectation. Regression establishes a relationship between some covariate(predictors) and an outcome (response) variable but does not in anyway suggest cause-effect (causal) relationship. Mostly this relationship is expressed in the form of an equation that predicts a response variable (Denuit et al, 2007). In the case of claim count data, the response variable will be the expected number of claims reported by a particular policyholder. The predictors or explanatory variables will be the various risk factors which relates to the policyholder, vehicles types and types of contract. The claim count or frequency may be influenced in the autoinsurance product by age, gender, marital status, use of the car, geography, type of car (e.g sport car). Since claims data might not always follow a normal distribution (the use of ordinary linear regression model with constant variance), there is a need to use the generalized linear model (GLM).

The GLM is very useful in actuarial modelling where claim cost and claim counts follow asymmetric density that is non-Gaussian (David, 2013). The difference between the GLM and the ordinary linear regression model was looked at in section 3 (literature review). The Poisson, logistic and linear models are all special cases of the GLM (Denuit et al, 2007). The ZIP, ZINB and the classical Poisson regression techniques were therefore used in fitting the claims count. With the claims cost, explanatory variables may include the use of car,age, vehicle type, gender, etc. The gamma regression technique was used in modelling the claims cost. Finally, by the law of large numbers, pure premium is an expected cost of all claims that insurers report during coverage period and this is calculated by applying a regressing model that incorporates various available risk factors.

1.2 Research Objectives

The purpose of this research is to determine the risk factors in calculating actuarial premiums by combining the conditional expectations of claim count and claim amount. The objectives are as follows:

i To identify the risk factors that should be incorporated in the determination of actuarial premiums

ii To determine the appropriate claim count model (among Poisson, ZIP and ZINB) suitable for the analysis of the claim counts data.

iii To show how the actuarial premiums can be calculated by separately modelling claim counts and claims amount.

1.3 Research questions

i What are the risk factors that can be incorporated in the determination of premiums?

ii Which statistical model is suitable for the analysis of claim counts?

1.4 History and Contribution

1.4.1 GLM

The generalized linear model, mostly referred to in the statistical world by its acronym GLM dates back to the work of Nelder and Wedderburn(1972). In

their ground-breaking research, they showed how to use linearity to unify diverse statistical technique (McCullagh & Nelder, 1989). GLMs are ordinary linear regression models extended to a class of distribution called the exponential family of distribution. An ordinary linear regression model of Y_i as response variables and X_i as explanatory variables is defined as;
$Y_i = X^T\beta + \epsilon$ where Y_i follows a normal distribution. That is $Y_i \sim N(\mu_i, \sigma^2)$ and the expectation of Y, $\mathbb{E}[Y] = \mu = X^T\beta$. It is assumed that before this linearity will hold, Y_i are identically and independently distributed. This is actually the basis of most analyses in continuous data (Dobson & Barnett,2008). For Y_i, if $i = 1$, that is, one observation, this will produce a simple linear regression model. As noted by Dobson and Barnett (2008):

1. the response variable (Y_i) may have other distribution other than the normal distribution, they may be categorical other than continuous.

2. And the relationship between the response variable and the explanatory variables need not be of the simple linear regression, It may be a multiple regression (that is there may be several explanatory variables).

As said earlier, the random components (errors and response variable) and the systematic components ($u_i = \mu = X_i^T\beta$) are assumed to follow a normal distribution (Nelder & Wedderburn,1972). However, the so called 'nice' properties of the normal distribution are also shared by the exponential dispersion family of distributions. In order to estimate the parameters, there is a transformation of the systematic effect. That is, there is non-linear function (logarithmic function), $g(.)$ relating the mean parameter μ_i to the linear component $X_i^T\beta$. This is expressed as $g(\mu_i) = X_i^T\beta$. The function, $g(.)$ is normally called link function, since it is a link between the random component and the systematic component (McCullagh & Nelder, 1989). The discussion above can be modified to produce a three-part specification for the GLM model as stated in McCullagh and Nelder (1989)

1. The random component: the components of Y_i with $\mathbb{E}[Y] = \mu$

2. The systematic component: the covariates $x_1, x_2, ..., x_p$ produce a linear predictor, η given by, $\eta = X^T\beta$.

3. The link between the random and systematic components: $\eta = g(\mu_i)$.

1.4.2 Ghana's Insurance Industry in Perspective

In this section, a brief history about insurance in Ghana was given and this was put into perspective about the subject of interest, pricing of motor insurance policies. This section is sourced from NIC website.

1.4.2.1 The Legal Setting of Insurance in Ghana

The Insurace Act, 2006 which is the the seven hundred and twenty-fourth act of the parliament of the republic of Ghana. In the spirit of that Act, all insurance business are suppose to operate and be regulated. This Act provides regulations regarding the, establishment admininistration, regulation, supervision, and monitoring of the industry in Ghana. The Act gives these powers to statutory established institution called the National Insurance Commission (NIC). Among other important functions of the Commission, these are some of them as put out by Insurance Act (2006).

- The NIC has the sole power to license insurers and insurance intermediaries who transact insurance business in Ghana.
- In consultation with relevant bodies, approve and set standards to the conduct of insurance business and insurance intermiadiary business.
- approve, where appropriate, the rate of insurance premiums and commissions in respect of any class of insurance

- arbitrate insurance claims referred to the Commission by any party to an insurance contract
- supervise and approve transactions between insurers and their re-insurer

As stated by the NIC, the Act is line with the principles of International Association of Insurance Supervisors (IAIS). The commission in a nutshell must see to it that, the public is protected against financial loss caused by dishonesty, incompetence, malpractice or insolvency on the side of insurers or insurance intermidiaries

1.4.2.2 Types of Companies

As stated by the Middle East Insurance Review (2020), there are more than 130 insurance companies in good standing as at November 2020. The breakdown is as follows:

- 27 non-life companies
- 17 life companies
- 3 reinsurance companies
- 91 insurance broking companies

Important to note that National Insurance Commission [NIC], (2019) stated that, there are about 5 reinsurance brokers, 1 reinsurance contact office and over 7000 insurance agents

1.4.2.3 Classes of Business

There are two main types of insurance; Life insurance and non-life insurance. The various classes of insurance in Ghana, as stipulated in Ghana is as follows;

Non-Life Business

- Fire burglary, and property damage
- Accident
- Marine and Aviation
- Motor
- General Liability

Life Insurance Products

- Universal Life
- Funeral
- Whole Life Insurance
- Endowment Insurance
- Term life Insurance
- Group Life

Since the study focuses on non-life insurance business, a synopsis of this sector with emphasis on motor-insurance was given.

1.4.2.4 Motor Insurance

Just like there are various names for non-life insurance (such as casuality insurance, general insurance, property insurance) which describes insurance other than life, there are various names for Motor insurance; they include, car insurance, autoinsurance, motor vehicle insurance, etc. Motor Insurance in Ghana and all over

the world is divided into three major part: Third party, Third Party, Fire and Theft (TPFT), and Comprehensive. Motor Third Party Insurance is the largest lines of business in the non-life insurance sector, contributing to about 37% of non-life insurance premiums in both 2018 and 2019. The TPFT also contributed 23% to premium income for non-life insurance companies, comprehensive insurance and other types of insurance contributed the remaining proportion. Worthy to note that, the percentage contribution of the comprehensive insurance is very small. Ghana Oil and Gas Insurance Pool (GOGIP) contributed to 15% of the rest 40%, liability and engineering insurance contributed 6% each and financial loss insurance contributed to 5%, Others including, Marine and Aviation, Personal Accident other short-term products contributed 8% (NIC, 2019).

Motor insurance is a type of non-life insurance policy where a legitimate liability which may arise out of the use of any type of automobile is covered. Three main type of motor insurance are explained below;

1. **Third Party**

   This type of policy covers damage to other people (injury or death to a third party), or their vehicles, and property caused by the vehicle of the insured. It does not cover the insured or the insured's vehicle. It is compulsory or rather legal for all vehicles to be registered under this policy. If accidents are caused due to drunkeness, the driver being a minor, the "accident" being purposeful, and the vehicle being used for a purpose it is not insured for (for instance a private individual car registered as such is used for commercial purpose), and if the vehicle is being stolen (TPFT covers this)

2. **Third Pary, Fire and Theft- TPFT**

   Just like its names, the TPFT, entails cover for the loss of the insured vehicle either through theft or fire. Damages to your vehicle caused by collision, overturning, etc are not covered under this policy ("Motor Insurance >

Policies", 2021)).

3. **Comprehensive**

   The Comprehensive Motor Insurance policy includes all the two policies above and adds damage caused to your own vehicle due to accidental collision, or over-turning, fire, external explosion, self-ignition, or lightening, theft, burglary, housebreaking, malicious act, flood, storm, hurricane, volcanic eruption or earthquake ("Motor Insurance > Policies", 2021)

1.4.2.5 Charging Motor Insurance Premium in Ghana

The flat premium for Third Party Motor Insurance (and also TPFT and Comprehensive) will be 85% of the current basic premium for Private Individual and Private Corporate vehicles and 90% for all other classes of vehicles. In charging the final premium, the insurer will then add all other expenses including administrative charges, commissions, underwriting expenses, among others (which are normally called in actuarial literature as security or safety loadings)to the flat premium. In December 2018, the governement of Ghana cancelled the No Claim Discount (NCD) which sees to it that individual policyholders who did not register any claim throughout the coverage period will have their premiums reduced upon renewal depending on the class of motor insurance policy taken and how many years of no claim. However, the NCD was reinstituted in May 2020 by the NIC with modalities on how to implement it. The NCD system used in Ghana is similar to the Bonus Malus System (BMS) used in most European countries

1.5 Significance of Study

The results and findings of this research will contribute to existing literature on pricing of non-life insurance. It outlines various steps in calculating actuarial

premiums by modelling claim counts and claims cost separately. This will serve as a guide to insurance practitioners as they consider various exogenous risk factors in computing premiums.

1.6 Organization of Study

This research is organised in to 5 sections; The first section above looks at the background of non-life insurance and history of models used in determining the premiums for the insurance risk takers. Section two will see to the review of existing literature in non-life insurance pricing. Section 3 considers the methodology, and sections 4 and 5 look at results and analysis, and conclusions respectively. The appendices spell out some relevant codes written in R statistical software.

Chapter 2

LITERATURE REVIEW

A review of relevant literature on the subject and scope of the study is required in almost all academic studies. Previous work in the domain of actuarial modeling of claim counts and amounts, as well as the use of GLM in the pricing of pure premiums in the non-life insurance market, is reviewed in this study. The researcher reviewed literature on Poisson, ZIP, and ZINB regression models in particular. Also, some basic probability theories that are pertinent to the research were explained. In general, this chapter is divided into three sections: Section 1 covers the pricing of non-life insurance premiums. Section 2 looks at how the GLM is utilized in the non-life business to price pure premiums. The final portion delves into a number of past research conducted when there is a presence of zero-inflation in a dataset.

2.1 Pricing of Non-Life Insurance Premiums

A non-life insurance policy is an agreement between an insurance company and a customer (policyholder), which the insurer undertakes to indemnify the customer for some unforeseeable losses over a period, mostly one year for a tariff called premium Insurance is therefore an adventure of risk (an economic risk). It is largely a risk management endeavour. The policyholder, file for claims any time the insured event occurs. Conditioned on the fact that it is a risky but a business venture, insurers must set premiums that are fair, equitable and as well profitable.

As recognized earlier, premium determination is done through the law of large

numbers, where the losses incurred by few are pooled or spread across an entire homogeneous group to provide a substantially accurate prediction of future losses. The premium is generally set by the insurance company ahead of any claims. Based on this, it is not out of board for insurance industries to employ the predictive capacity of regression models in the calculation of insurance premiums. This is done by incorporating several risk factors as talked about earlier.

However, insurers must do this keeping in mind the competitiveness in the insurance market. A research by Accentuate Financial Services in 2017 about global distribution and marketing of insurance product reveals that from 32,175 insurance customers surveyed across 18 markets, the price was found to be the number one driver of customer loyalty. Stating that 52% of auto-insurance customers choosing it as their top lever. With this in mind, premium are set not only to cover the loses or risk of the policy holder but as well cover the expenses and administrative margins or loadings. Statistically, insurers should set premium in such a way that it involves the risk of the policyholder. Premiums must be based on expected average loss that is transferred from the policyholder to the insurer. Risk is the likelihood that something unpredictable could occur. In the context of this thesis, theft or damage of your motor vehicle or an injury to passengers. From a statistical point of view, risk is stated as τ as

$$\tau = \mathbb{E}(\frac{L}{\epsilon}) \tag{2.1}$$

where L is the loss and ϵ is the period for which the insurance is valid (exposure). On the basis of independence of frequency from the claims amount, it can be stated that

$$\tau = \mathbb{E}(\frac{L}{\epsilon}) = \mathbb{E}(\frac{L}{N}|N > 0) \times \mathbb{E}(\frac{N}{\epsilon}) = \mathbb{E}(S) \times \mathbb{E}(F) \tag{2.2}$$

Where N is the number of claims and S the severity of claim (size of claim) and F

the claim frequency. This suggests that premiums (incorporating all significant risk factors) should be calculated by multiplying the expected claims cost by expected frequency of claims. That is,

$$\text{risk (pure) premium} = \text{claim severity} \times \text{claim frequency} \tag{2.3}$$

.

2.2 Pricing with GLM

Under the History of GLM, it was stated that, the ordinary linear regression or slightly larger general linear models are not suitable for calculation of premiums due to:

1. The event of interst (claims cost and claim number) do not always follow a normal distribution. They are often skewed to the right (Ohlsson and Johansson,2010). And they are not always continuous.

2. In an ordinary linear model, the expectation of the random variable, the mean is a linear function of the covariates, while multiplicative models fit reasonably well to insurance data than the linear models.

Ohlsson and Johansson (2010) stated clearly how the GLM, which is a class of statistical methods generalizes the ordinary regression model into two aspects, each of which tackles the problem encountered with linear models.

1. Probability distribution: instead of assuming the normal distribution, the GLMs work with general class of distribution which contains a number of discrete and continuous distributions.

2. Model for the mean: In linear models, the mean is a linear function of

the covariates, In GLMs, some monotone transform of the mean is a linear function of the covariates.

The idea of GLM was introduced by Nelder and Wedderburn in 1972. In 1983, McCullagh and Nelder, take it up by applying the GLM to motor insurance data (Ohlsson & Johansson, 2010). However, not until the second half of the 90s that the use of GLMs by actuaries start spreading. Ohlsson and Johansson (2010) noted that, this was partly due to the deregulation of the insurance markets in many countries.

After the paper written by Nelder and Wedderburn (1972), some remarkable works were done by many actuaries, statistician and scientists by using GLM. David(2013) pointed out that many scientists and authors have succeeded to highlight or improve the assumptions imposed by the practical application of GLM in non-life insurance.

Among some of the reviewed literatures, are: Dobson and Barnett (2008) where the authors well-outlined the GLM and give some useful discussions on the application of GLM using R software. The book written by Ohlsson and Johansson (2010) is an eye opener on the theoretical setting of GLM, and its application to real world data. The authors stated some advantages of the use of GLM over other regression or predictive model in pricing non life insurance data. They suggested that:

1. GLM constitutes a general statistical theory which has well established techniques for estimating standard errors, constructing confidence intervals, testing model selection and other statistical features.

2. GLMs are used in many areas of statistics, so that one can draw from the developments both within and without of actuarial science.

3. and finally, there are standard software (R, SAS, etc) for fitting GLMs that

can easily be used in analyzing claim frequency and amount and hence premium calculation.

Sarul and Balaban (2013) took a look at premium pricing and risk assessment for claim amount based on generalized linear models. In this research the authors studied the importance for sustainable customer relationships by incorporating individual risk into premium calculation. Logarithimic gamma model is the best per the analysis of the customers that forms the data set and risks assessment was made by evaluating the coefficient of variation. The authors found out that 0.1% of the customers of the portfolio forms high risk group with regard to the change in the coefficient of variation according to ranges of claim amounts, i.e. maximum, minimum, and average.

Antonio and Valdez (2012) used Generalized Linear Model to model insurance premiums based on a priori risk. Since not all important risk factors may be observable, the authors used Generalized Linear Mixed Models (GLMMs) to incorporate posteriori risk (taking in to account the history of reported claims).The authors stated that the premium for an insurance contract is calculated after some claims history has been revealed by accounting for both the experience of the individual policyholder together with that of the whole portfolio to which the contract belongs. It reveals that, both types of rating in premium calculation is of importance because, the posterior or experience ratemaking is a way to penalize bad risk and reward good risk. They suggest that a posterior premium allows one to correct and adjust the previous priori premium, making the price discrimination even more fair and reasonable.

Frees (2010) writes that the GLM has provided a unifying framework that not only encompasses many models but also provides a platform for new ones, including gamma regression for fat-tailed data and Tweedie distributions for two-part data. The issue with the Tweedie GLM is that it does not separately model claim counts and claims cost. However, in insurance modelling the most widely

used mixture model is the Tweedie distribution (Denuit et al, 2007). It includes distributions that are mixtures of discrete and continuous components–two parts model. The discrete part representing the claim counts and the continuous part representing the claims cost. The Tweedie is seen as a Poisson sum of gamma random variables known as aggregate loss in actuarial science. (Denuit et al, 2007). In this thesis the gamma model was used in modelling the claims cost. Later in the methodology, reasons for the choice of the gamma model were given to that effect.

2.3 Zero-Inflated Models for Claim Counts

Count data in health, manufacturing, insurance, economics has been widely modelled by Poisson regression.(Ismail & Zamani, 2013). But as indicated earlier, due to the dominance of zeros in most count data, which leads to overdispersion in the data, the Poisson regression will be inappropriate to model such type of data. If used to model a data with excess zeros, it may turn to underestimate the standard errors and overstate the significance of regression parameters (Ismail & Zamani, 2013). Cameron and Trivedi (2013) iterates that the default computed Poisson maximum likelihood t-statistics will be considerably overinflated, and this can lead to very erroneous and overly optimistic conclusions of statistical significance. It is therefore necessary to think of other regression models which will model the zero components of such a count data.

Diane Lambert (1992) was the first to apply zero inflated Poisson model on a count data. In his research he seeks to study soldering defects on print wiring board (Denuit et al, 2007). In the experiment, he found out that even the richest log-linear Poisson model which has a three way interaction, predicts poorly (Lambert, 1992).

Zero inflated models attempts to account for excess zeros. Zero inflated mod-

els estimate two equations, one from the count model and one for the excess zeroes.(Frees, 2009).

Yip and Yau (2005) motivated by the fact that accurate modelling of the claims count distribution is one of the essential steps in calculating policy rates, and zero inflation problem in a claim count distribution of a motor insurance data, study the use of the Poisson, the NB, ZIP, ZINB and ZIGP. The study proposed the use of zero inflated models which accommodates the extra zeros possibly caused by the unreported minor losses. They found out that, the NB, ZIP, ZINB fits the data reasonably well, but the ZIGP fits the data the most.

Bouchera and Denuit (2008) also used zero inflated models to model claims count distribution. The authors showed that the zero inflated model allow for more flexibility in the prediction of credibility premiums. In their paper, they showed the derivation of credibility formulae for some generalized zero-inflated models.

Sarul and Sahin (2015) applied zero inflated and hurdle models in general insurance in the modelling of claim frequency. They agree that, insurance data, due to deductibles and no claim discounts are dominated with many zeros than expected which is inconsistent with the Poisson process assumption of equality of mean and variance values. They concluded that zero inflated Poisson model is superior to the standard Poisson model and zero inflated negative binomial model is superior to the negative binomial model. The AIC results of the claim counts analysis showed that both the ZINB and the hurdle Negative binomial model fits the data the most.

Boucher *et al.* (2009) stated that the hunger for bonus phenomenon makes the insured to not report all of his accidents to save bonus on his next year's premium. In their paper they assume that the number of accidents is based on a Poisson distribution but the claim frequency is produced by a censored version of the Poisson distribution. They therefore extended the zero inflated Poisson models

to panel (longitudinal) data. That is, they adopted a multivariate zero inflated models and proposed an approximation of the accident distribution which can provide insight into the behaviour of insureds.

On how to use zero inflated models in modelling claim count data, Wolny-Dominiak (2013) proposed a 4-step procedure for modelling zero-inflation effect. He applied it on an insurance data by performing ZIP, ZINB and ZIGP. He found out that, the enlargement of the ZIP and ZINB models to ZIGP model could give better estimations. In other to choose the best model, some statistician sought to use Wald test and likelihood ratio tests in testing for overdispersion, however in their paper, Zamani and Ismail (2013), showed that the score test has an advantage over other tests. They claim the score test can be used to determine whether a more complex model is appropriate without fitting the more complex model. They also prove that the score test for testing ZIP regression model against ZINB is the same as the score test for testing ZIP against ZINB alternatives. They found that even though the ZIGP regression is a good competitor for ZINB regression model, in several cases, ZINB model may not provide converged values in the iterative technique of the fitting procedure and thus ZIGP model may be considered an appropriate alternative. Additionally, they revealed that when zero inflation and over dispersion exist in count data, ZIGP regression model behave similarly to ZINB regression model. consequentially, in their studies, the score test for testing ZIP against ZINB is equal to that of testing ZIP against ZIGP.

Ridout *et al.* (2001) also compared the score test to other goodness of fit test for testing zero-inflated Poisson regression models against ZINB. They revealed that the principle advantage of the score test in comparison with the likelihood ratio test is that the score test does not require the more complex model to be fitted.

Lambert (1992) gave an EM algorithm for fitting zero-inflated Poisson regression models which is easy to implement in any statistical package that includes facili-

ties for fitting weighted generalized linear models for Poisson and binomial data and has facilities for handling matrices. However, he stated that, the principle disadvantage of the score test in comparison with the likelihood ratio test is that the asymptotic distribution of the score statistic is often approached more slowly than that of the LR statistic. He believes the significant levels derived from the score statistic can be misleading, specially in small samples. From practical perspective, they suggest that if the uncorrected score test gives even a weak indication that the ZIP model is inappropriate, say at 10% significance level, a ZINB model should be fitted to the data.

Tennekoon (2017) on applying binomial-thinned zero inflated Poisson model on counting unreported abortions (by using NSFG data) used the binomial-thinned zero inflated Poisson model, where abortion rate is incorporated in to the Poisson model through an operator called binomial thinning operator. In the model, the actual number of intentional abortions of a woman was defined as a function of her exposure (t_i); the probability of being in the group of women that includes those who decide not to be pregnant, are infertile, or are strictly against abortion (ψ_i); and the expected number of intentional abortions during a unit period (λ_i). The model was used to predict the reporting probabilities of each individual which she believes can be used to correct the bias due to under reporting in any model in which the number of abortions is used as the dependent variables or as one of the covariates. The model estimates the average reporting rate in the NSFG during 2007-2013 as 35.3% with a standard error of 8.2%. A striking observation from her analysis was that, the proportion of women with zero abortions among the younger age groups was found to be smaller than the older age groups, and, the mean number of abortions of those younger women who would consider having abortion during a given period is smaller compared to older women. About the binomial thinned zero-inflated Poisson model, the researcher set out that, the model is mostly appropriate when the count data is suspected to be under-reported and if there is over-reporting of a count data, the

model could be modified or be used in addition to several other count models available.

In this research the zero inflated models were applied on a motor insurance data. As said before, a troublesome aspect of insurance data is the excess number of zeros, relative to a specified model (Frees, 2010). This might be due to the insureds being reluctant in reporting claims, fearing that a reported claim will result in higher future insurance premiums. Meaning there is a higher than expected number of zeros because of the non-reporting of claims. The zero inflated model represents the claims number as a mixture of a point mass at zero and another claims frequency distribution. Frees (2010) again pointed out that, the point mass can be interpreted as the tendency of non-reporting.

Chapter 3

METHODOLOGY

In this chapter, the theoretical concepts and frameworks underlying this research are outlined. By defining some statistical tools to enable us to achieve the purpose of this study in calculating actuarial premiums by modelling claim counts and claims amounts separately. Statistical concepts under probability theory, count models, and claim amounts are all looked at. The theoretical setting of these models were also set up. In line with the purpose of the study, the theoretical definitions of the model comparison approaches in the selection of the appropriate models for the claim count data were laid out. All of this is done to set the background for the data analysis section of this study.

3.1 Source of Data

Data was obtained from an open online database, `https://www.kaggle.com/datasets/karansarpal/fremtpl2-french-motor-tpl-insurance-claims`. on number of claims and amount of claims for autoinsurance. The claim counts and claims cost data have exogenous variables such as Driver's age, Car's age, Brand of car, among others. Another dataset was obtained from a major motor-insurance company in Ghana. The descriptions of these datasets are found in chapter four. Since in this study, various statistical distributions were used, a brief overview of the basic of probability theory is presented below.

3.2 Basics of Probability Theory

3.2.1 CDF and PDF

1. The function, F with the domain, $D_f = \mathbb{R}$ is a cumulative distribution function (CDF) if:

   (a). F is non-decreasing and right continuous; and

   (b). $\lim_{x \to -\infty} F(.) = 0$ and $\lim_{x \to +\infty} F(.) = 1$

2. The function f with $D_f = \mathbb{R}$ is called probability density function (PDF) if it is:

   (a). non-negative, $f(.) > 0$; and

   (b). integrable and $\int_{-\infty}^{+\infty} f(.)ds = 1$

3.2.2 Sigma Algebra and Probabillity Space

A probability measure $\mathbb{P}$ in $(\Omega, \mathbb{F})$ is a real-valued function, $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ such that:

1. $\mathbb{P}(\Omega) = 1$

2. $\mathbb{P}(A) \geq 0 \, , \forall A \in \mathcal{F}$

3. $$\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{n} \mathbb{P}(A_i), \forall A_i \in \mathcal{F} \text{ and} A_i \cap A_j, i \neq j$$

hence the space, $(\Omega, \mathbb{F}, \mathbb{P})$ is known as probability space.

3.2.3 A Random Variable and Alternative Definition of CDF

1. Let $\mathbb{P}$ in $(\Omega, \mathbb{F}, \mathbb{P})$ be a probability space. The function $X : \Omega \to \mathbb{R}$ is a random variable (r.v) if $\forall x \in \mathbb{R}$ it holds:

$$\{\omega : X(\omega) \leq x\} \in \mathcal{F}$$

2. Let X be a random variable. The function $F_X(x) = \mathbb{P}(X \leq x)$, $x \in \mathbb{R}$ is called a cumulative distribution function (CDF) of X

If a random variable takes at most a countable number of possible values, it is said to be a discrete random variable, if uncountable, it is called a continuous random variable.

3.2.4 Discrete and Continuous Probability Distribution

1. A random variable X is called discrete if it attains only countable different values $x_i, x_2, ...$The distribution function X is given by:

$$F_X(x) = \sum_{(x_i \leq X)} \mathbb{P}(X = x_i) \tag{3.1}$$

2. Random variable X is called continuous if there exist a density function f_x such that:

$$F_X(x) = \int_{-\infty}^{x} f_x(s)ds \tag{3.2}$$

3.2.5 Relationship Between CDF and PDF

Assume f is continuous on the interval $[a, b]$ and define

$$F(x) = \int_a^x f(s)ds \; ; x \in [a, b] \tag{3.3}$$

Then F is continuous on $[a, b]$ and $\frac{dF(x)}{dx} = f(x), \forall x \in (a, b)$.

Since the mean and variance of distributions used in the study will be of use, a general statistical definitions of them is given in the next section.

3.2.6 Mean, Variance and kth moments

3.2.6.1 Mean

1. Disrete r.v

   Let X be a discrete random variable with realisable values x_i and a density (technically, probability mass function–PMF), $f(x) = P(X = x)$, then the expectation of X is

   $$\mathbb{E}(X) = \sum_{i=1}^{\infty} x_i \mathtt{P}(X = x) \tag{3.4}$$

2. Continuous r.v Let X be a continuous random variable with density function, f_x, $\mathbb{E}[X]$ exists if and only if

   $$\int_{-\infty}^{\infty} \mid x \mid f_x(x)dx < \infty$$

   then

   $$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_x(x)dx \tag{3.5}$$

3.2.6.2 Second Moment

1. Discrete r.v

   Let X be a discrete random variable with realisable values x_i and a density (technically, probability mass function), $f(x) = P(X = x)$, then the second moment of X is defined as

$$\mathbb{E}[X^2] = \sum_{i=1}^{\infty} x_i^2 \mathtt{P}(X = x) \tag{3.6}$$

2. Continuous r.v

   Let X be a continuous random variable with density function, f_x, $\mathbb{E}[X^2]$ exists if and only if

$$\int_{-\infty}^{\infty} \mid x^2 \mid f_x(x)dx < \infty$$

   then

$$\mathbb{E}[X^2] = \int_{-\infty}^{\infty} x^2 f_x(x)dx \tag{3.7}$$

3.2.6.3 kth moment

1. Discrete r.v Let X be a discrete random variable with realisable values x_i and a PMF, $f(x) = P(X = x)$, then the kth moment of X is defined as:

$$\mathbb{E}[X^k] = \sum_{i=1}^{\infty} x_i^k \mathtt{P}(X = x) \; ; k = 1, 2, ... \tag{3.8}$$

2. Continuous r.v

   Let X be a continuous random variable with density function, f_x, $\mathbb{E}[X^k]$

exists if and only if

$$\int_{-\infty}^{\infty} \mid x^k \mid f_x(x)dx < \infty$$

then

$$\mathbb{E}[X^k] = \int_{-\infty}^{\infty} x^k f_x(x)dx,\ k = 1, 2, ... \tag{3.9}$$

3.2.6.4 Variance

The variance of a random variable can be defined in any of the following ways:

$$\text{var}(X) = \mathbb{E}[X - E(X)]^2 \tag{3.10}$$

$$\text{var}(X) = \mathbb{E}[X^2] - [\mathbb{E}(X)]^2 \tag{3.11}$$

3.3 Generalized Linear Model-GLM

In this section a brief theoretical setting, estimation of parameters using maximum likelihood estimation method and some useful properties of the GLM are illustrated.

3.3.1 Theoretical Setting of a GLM model

As noted by McCullagh and Nelder (1989), the generalized linear model is an extension of the classical model. For the classical model the expectation of the random variable (Yi), μ is seen as a linear combination of the explanatory variables (X_i). Also, in the ordinary linear model, it is assumed that the random

variable (Y_i) (and random errors) always follow a normal distribution. But as noted earlier, with real world data, that is always not the case.

The ordinary linear model is first of all defined and thereafter extended to GLM. By the ordinary linear regression model, for a response variable (Y_i) and explanatory variables (X_i), the relationship between Y_i is established by the model;

$$Y_i = \beta_0 + \sum_{j}^{p} \beta_j x_{ij} + \epsilon_i = X^T\beta + \epsilon;\ i \neq j \tag{3.12}$$

Where the index i is the number of observations. The βs are the unknown parameters (p) to be investigated and ϵ_i represents the error term. The mean of the random variable,

$$\mathbb{E}[Y] = \mu = \beta_0 + \sum_{j}^{p} \beta_j x_{ij} = X^T\beta;\ i \neq j \tag{3.13}$$

It is assumed that $Y_i \overset{iid}{\sim} N(\mu, \sigma^2)$. It can be seen from the above model that, the mean is linearly related to the explanatory variables. The challenges of the classical linear model is the normality assumption of the random response and error variable, and the constancy of the mean and variance (which is a strong assumption).

The generalized linear model extended the OLS model by assuming that the distribution of Y_i will belong to a family of exponential distributions. Also the mean will not be a linear combination of the explanatory variables but rather the mean is related to the explanatory variables through a scale transformation. The simple theoretical setting follows for the GLM from the equation of the ordinary linear model, where the mean will not be linearly related to the explanatory

variables, but a different linear predictor, η. That is

$$g(\mu) = \beta_0 + \sum_{j}^{p} \beta_j x_{ij} = \eta = X^T\beta;\ i \neq j \tag{3.14}$$

This is called the systematic component. The mean parameter will now be related to the explanatory variables through a scale transformation called the link function. This means the link function connects the random components to the systematic component.

3.3.2 The Three Components of GLM

1. The response variable/random component

   The response variable Y_i for $i = 1, 2, ..., N$ with means μ_i which are assumed to be independent but take the same distribution from the exponential family.

2. Linear Predictor/systematic component $\eta = X^T\beta$

   where X^T is an $N \times p$ matrix of explanatory variables and β denotes $p \times 1$ vector of unknown parameters. That is

$$X^T = \begin{bmatrix} x_{11} & x_{12} & ... & x_{1p} \\ x_{21} & x_{22} & ... & x_{2p} \\ . & . & ... & . \\ . & . & ... & . \\ . & . & ... & . \\ x_{N1} & x_{N2} & ... & x_{Np} \end{bmatrix}$$

   and

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \cdot \\ \cdot \\ \cdot \\ \beta_p \end{bmatrix}$$

3. Link function $g(\mu) = \eta = X^T\beta$

   As said earlier, the link function establishes the relationship between the mean and its linear predictors. Which implies

$$\mu = g^{-1}(\eta) = g^{-1}(X^T\beta) \tag{3.15}$$

3.3.3 Properties of The Link function

The link function must possess the following properties: It

1. must be monotonic
2. must be differentiable

   As noted by Dobson and Barnett (2008), these properties suggest that the link function, $g(\cdot)$ is flat or strictly increasing or decreasing with μ but it cannot be increasing for some values of μ and be decreasing for other values

3.3.4 The Distribution of the Response Variable

The distribution of the Y_i under the GLM is now analysed. Under the GLM, the random variable follows the exponential family of distribution. The theoretical setting of the exponential family of distribution and the various parameters of some common distributions are explained and then the definition of the GLM for

the various distributions that were used on the variables of interest (claim counts and claims cost) were briefly given.

3.3.5 The Theoretical Setting of Exponential Family of Distribution

Exponential family of distribution is a parametric set of probability distributions of a certain form based on some useful algebraic properties. It is also sometimes called exponential class of distribution or Koopman-Darmois family of distribution after the name B.O koopman and G. Darmois. The exponential is a parametric family of various distributions. The most common parametrisation of the probability density of the exponential family of distribution is

$$f_Y(y|\theta) = \exp[a(y)b(\theta) - c(\theta) + d(\theta)] \tag{3.16}$$

However the equivalent definition given in Nelder and Wilberburn (1972) which is termed in Lindsey (1997) as Exponential Dispersion Family was adopted in this study. Assume that each component of Y_i has a distribution in the exponential family of the form

$$f_Y(y;\theta,\phi) = \exp[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)] \tag{3.17}$$

for some specific functions of $a(.), b(.)$ and $c(.)$. If ϕ is known, or if $a(\phi) = \phi$ then this is an exponential family model with a canonical parameter θ or the distribution is said to be in a canonical or standard form (Dobson and Barnett, 2008). The $b(\theta)$ is termed the natural parameter of the distribution and any additional parameter to the parameter of interest, θ is referred to as the nuisance parameter. There are various distributions under the exponential family; we have normal, binomial, gamma, Poisson, etc. Some are discrete (e.g. binomial and Poisson) and others continuous (e.g. normal and gamma). The exponential form

of the normal, gamma, Poisson and binomial distributions are explained below to enhance one understanding of the parameters of the pdf of an exponential family of distributions

3.3.5.1 Normal Distribution as a Member of Exponential Family of Distribution

The probability density function of the Normal Distribution is:

$f_Y(y;\theta,\phi) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp[\frac{-(y-\mu)^2}{2\sigma^2}]$.

expanding the power of the exp, we have

$$\begin{aligned}
f_Y(y;\theta,\phi) =& \frac{1}{\sqrt{2\pi\sigma^2}}\exp[\frac{-(y^2-2y\mu+\mu^2}{2\sigma^2}] \\
=& \frac{1}{\sqrt{2\pi\sigma^2}}\exp[\frac{-y^2+2y\mu-\mu^2}{2\sigma^2}] \\
=& \frac{1}{\sqrt{2\pi\sigma^2}}\exp[\frac{2y\mu-\mu^2}{2\sigma^2}-\frac{y^2}{2\sigma^2}] \\
=& \exp\left(\log\left[\frac{1}{(2\pi\sigma^2)^{1/2}}\right]\right)\exp\left(\frac{2y\mu}{2\sigma^2}-\frac{\mu^2}{2\sigma^2}-\frac{y^2}{2\sigma^2}\right) \\
=& \exp\left(\log(2\pi\sigma^2)^{-1/2}+\frac{y\mu}{\sigma^2}-\frac{\mu^2}{2\sigma^2}-\frac{y^2}{2\sigma^2}\right) \\
=& \exp\left(\frac{y\mu-\mu^2/2}{\sigma^2}-\frac{1}{2}\left[\log(2\pi\sigma^2)+\frac{y^2}{\sigma^2}\right]\right) \\
=& \exp\left(\frac{y\mu-\mu^2/2}{\sigma^2}-\frac{1}{2}\left[\frac{y^2}{\sigma^2}+\log(2\pi\sigma^2)\right]\right) \qquad (3.18)
\end{aligned}$$

comparing equation 3.18 to the general density of the exponential family of distribution (equation 3.14), $\theta = \mu, \phi = \sigma^2$ and $a(\phi) = \phi, b(\theta) = \frac{\theta^2}{2}, c(y,\phi) = -\frac{1}{2}\left[\frac{y^2}{\sigma^2}+\log(2\pi\sigma^2)\right]$

3.3.5.2 Poisson Distribution as a Member of Exponential Family of Distribution

The probability density function for the Poisson distribution, Y is:

$f_Y(y;\theta,\phi) = \frac{\mu^y e^{-\mu}}{y!}$ where $y = 0, 1, 2, ...$

taking logarithm of the density function

$$\begin{aligned}\log f(y,\theta,\phi) &= \log\left[\frac{\mu^y e^{-\mu}}{y!}\right] \\ &= \log(\mu^y e^{-\mu}) - \log y! \\ &= \log \mu^y + \log e^{-\mu} - \log y! \\ &= y\log\mu - \mu - \log y!\end{aligned}$$

taking exponent of both sides

$$\begin{aligned}\exp(\log f(y,\theta)) &= \exp(y\log\mu - \mu - \log y!) \\ f(y,\theta,\phi) &= \exp(y\log\mu - \mu - \log y!) \qquad (3.19)\end{aligned}$$

comparing equation 3.19 to 3.17,

$\theta = \log u,\ \phi = 1$ and $a(\phi) = 1 = \phi,\ b(\theta) = \mu,\ c(y,\phi) = -\log y!$

3.3.5.3 Binomial distribution as a Member of Exponential Family of Distribution

Consider a series of binary events, called trials, each with only two possible outcomes, "success" or "failure". Let the random variable Y be the number of successes in n independent trials in which the probability of success, p, is the same in all trials. Then Y has the binomial distribution with probability density function,

$f(y,\theta,\phi) = \binom{n}{y} p^y (1-p)^{n-y}$; where $y = 0, 1, 2, ..., n$

but $\binom{n}{y} = \frac{n!}{y!(n-y)!}$, hence;

$f(y,\theta,\phi) = \frac{n!}{y!(n-y)!} p^y (1-p)^{n-y}$

taking logarithm of both sides

$$\begin{aligned}
\log f(y,\theta,\phi) &= \log\left[\binom{n}{y} p^y (1-p)^{n-y}\right] \\
&= \log\binom{n}{y} + \log p^y + \log(1-p)^{n-y} \\
&= \log\binom{n}{y} + \log p^y + \log[(1-p)^n (1-p)^{-y}] \\
&= \log\binom{n}{y} + y\log p + \log(1-p)^n - \log(1-p)^{-y} \\
&= \log\binom{n}{y} + y\log p + n\log(1-p) - y\log(1-p) \\
&= y\log p - y\log(1-p) + n\log(1-p) + \log\binom{n}{y} \\
&= y\left[\log p - \log(1-p)\right] + n\log(1-p) + \log\binom{n}{y} \\
&= y\log\left[\frac{p}{1-p}\right] + n\log(1-p) + \log\binom{n}{y}
\end{aligned}$$

taking exponent of both sides

$$\exp(\log f(y,\theta,\phi)) = \exp\left[y\log\left(\frac{p}{1-p}\right) + n\log(1-p) + \log\binom{n}{y}\right]$$

thus

$$f(y,\theta,\phi) = \exp\left[y\log\left(\frac{p}{1-p}\right) + n\log(1-p) + \log\binom{n}{y}\right] \qquad (3.20)$$

comparing equation 3.20 to the general density of the exponential family of distribution (3.17), we have,
$\theta = \log\left(\frac{p}{1-p}\right), \phi = 1, a(\phi) = \phi$ and $b(\theta) = n\log(1-p), c(y,\phi) = \log\binom{n}{y}$

3.3.5.4 Gamma Distribution as a Member of Exponential Family of Distribution

The exponential family of distribution for the gamma model is given here. There are different parametrisation of the gamma distribution but the one stated in Lindsey (1997) was adopted.

Consider a gamma distribution with density;

$f(y,\theta,\phi) = (\frac{\nu}{\mu})^{\nu}\frac{y^{\nu-1}e^{-\frac{\nu y}{\mu}}}{\Gamma(\nu)}$

Taking the natural log of both sides.

$$\begin{aligned}
\log(f(y,\theta,\phi)) =& \log\left[(\frac{\nu}{\mu})^{\nu}\frac{y^{\nu-1}e^{-\frac{\nu y}{\mu}}}{\Gamma(\nu)}\right] \\
=& \log(\frac{\nu}{\mu})^{\nu} + \log\left[\frac{y^{\nu-1}e^{-\frac{\nu y}{\mu}}}{\Gamma(\nu)}\right] \\
=& \nu\log(\frac{\nu}{\mu}) + \log\left[y^{\nu-1}e^{-\frac{\nu y}{\mu}}\right] - \log\Gamma(\nu) \\
=& \nu\left[\log\nu - \log\mu\right] + \log y^{\nu-1} + \log\left[e^{-\frac{\nu y}{\mu}}\right] - \log\Gamma(\nu) \\
=& -\frac{\nu y}{\mu} - \nu\log\mu + (\nu-1)\log y + \nu\log\nu - \log\Gamma(\nu) \\
=& \left[-\frac{y}{\mu} - \log\mu\right]\nu + (\nu-1)\log y + \nu\log\nu - \log\Gamma(\nu) \\
\log(f(y,\theta,\phi)) =& \frac{[-y\mu^{-1} - \log\mu]}{\nu^{-1}} + (\nu-1)\log y + \nu\log\nu - \log\Gamma(\nu)
\end{aligned}$$

taking exponent of both sides

$$\exp\left[\log(f(y,\theta,\phi))\right] = \exp\left[\frac{(-y\mu^{-1} - \log\mu)}{\nu^{-1}} + (\nu-1)\log y + \nu\log\nu - \log\Gamma(\nu)\right]$$

thus

$$f(y,\theta,\phi) = \exp\left[\frac{(-y\mu^{-1} - \log\mu)}{\nu^{-1}} + (\nu-1)\log y + \nu\log\nu - \log\Gamma(\nu)\right] \tag{3.21}$$

comparing equation 3.21 to the general probability density function of exponential

family of distribution,

$\theta = \mu^{-1} = \frac{1}{\mu}, b(\theta) = -\log\mu$ and $\phi(\theta) = \nu^{-1} = \frac{1}{\nu}, c(y,\theta) = (\nu-1)\log y + \nu\log\nu - \log\Gamma(\nu)$

3.3.6 Relationship Between Mean And Variance of an Exponential Family of Distribution

Consider the general PDF of the exponential family of distribution;

$f_Y(y_i;\theta_i,\phi) = \exp\left[\frac{y_i\theta - b(\theta_i)}{a_i(\phi)} + c(y_i,\phi)\right]$

For a parametrized family of distribution like the exponential family, the likelihood function is described as;

$\mathcal{L}(\theta \mid y) = f(y \mid \theta)$.

That is the likelihood function is considered a function of θ given y. For the exponential family of distribution, the likelihood function is written as a function of θ given ϕ and y. That is:

$$\mathcal{L}(\theta \mid \phi, y) = f(y \mid \theta, \phi) \tag{3.22}$$

The log-likelihood function is the log of the likelihood function and can be written as:

$$\ell(\theta \mid \phi, y) = \log\left[\mathcal{L}(\theta \mid \phi, y)\right] = \log[f(y \mid \theta, \phi)] \tag{3.23}$$

The mean and variance of Y_i is derived by making use of the following relations from inference theory.

1. By noting that the first partial derivative of the log-likelihood function (called the score function) is equal to zero.That is

$$\mathbb{E}\left[\frac{\partial \ell(\theta_i \mid \phi, y)}{\partial \theta}\right] = 0 \tag{3.24}$$

and

2.

$$\mathbb{E}\left[\frac{\partial^2 \ell(\theta_i \mid \phi, y)}{\partial \theta^2}\right] + \mathbb{E}\left[\frac{\partial \ell(\theta_i \mid \phi, y)}{\partial \theta}\right]^2 = 0 \tag{3.25}$$

.

For the exponential family of distribution,

$$\ell(\theta \mid \phi, y) = \log[f_Y(y_i; \theta_i, \phi)] = \log\left[\exp\left(\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right)\right]$$

$$\ell(\theta \mid \phi, y) = \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \tag{3.26}$$

for simplicity, we write $\frac{\partial \ell(\theta_i|\phi,y)}{\partial \theta} = \frac{\partial \ell}{\partial \theta}$, Thus:

$$\begin{aligned}\frac{\partial \ell}{\partial \theta} &= \frac{\partial}{\partial \theta}\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right] \\ &= \frac{y - \frac{\partial b(\theta)}{\partial \theta}}{a(\phi)}\end{aligned} \tag{3.27}$$

Now

$$\begin{aligned}\frac{\partial^2 \ell}{\partial \theta^2} &= \frac{\partial}{\partial \theta}\left[\frac{\partial \ell}{\partial \theta}\right] = \frac{\partial}{\partial \theta}\left[\frac{y - \frac{\partial b(\theta_i)}{\partial \theta}}{a(\phi)}\right] \\ &= -\left[\frac{\frac{\partial^2 b(\theta)}{\partial \theta^2}}{a(\phi)}\right]\end{aligned}$$

hence

$$\begin{aligned}\mathbb{E}\left[\frac{\partial^2 \ell}{\partial \theta^2}\right] &= \mathbb{E}\left[-\frac{\frac{\partial^2 b(\theta)}{\partial \theta^2}}{a(\phi)}\right] \\ &= -\frac{\frac{\partial^2 b(\theta)}{\partial \theta^2}}{a(\phi)}\end{aligned} \tag{3.28}$$

knowing that $\mathbb{E}\left[\frac{\partial \ell}{\partial \theta}\right] = 0$.

This implies:

$$\begin{aligned}\mathbb{E}\left[\frac{y - \frac{\partial b(\theta)}{\partial \theta}}{a(\phi)}\right] &= 0 \\ \frac{\mathbb{E}[y] - \mathbb{E}\left[\frac{\partial b(\theta)}{\partial \theta}\right]}{a(\phi)} &= 0 \\ \mathbb{E}[y] - \mathbb{E}\left[\frac{\partial b(\theta)}{\partial \theta}\right] &= 0 \cdot a(\phi) \\ \mathbb{E}[y] - \mathbb{E}\left[\frac{\partial b(\theta)}{\partial \theta}\right] &= 0 \\ \mathbb{E}[y] = \mu &= \frac{\partial b(\theta)}{\partial \theta}\end{aligned} \tag{3.29}$$

We know that

$$\begin{aligned}\mathbb{E}\left[\frac{\partial^2 \ell}{\partial \theta^2}\right] + \mathbb{E}\left[\frac{\partial \ell}{\partial \theta}\right] &= 0 \\ \text{hence} \\ -\frac{\frac{\partial^2 (\theta)}{\partial \theta^2}}{a(\phi)} + \left[\frac{y - \frac{\partial b(\theta)}{\partial \theta}}{a(\phi)}\right]^2 &= 0\end{aligned} \tag{3.30}$$

from $\mu = \frac{\partial b(\theta)}{\partial \theta}$

$$-\frac{\frac{\partial^2(\theta)}{\partial\theta}}{a(\phi)} + \left[\frac{y-\mu}{a(\phi)}\right]^2 = 0 \tag{3.31}$$

we know from inferential statistics that, variance, $var(y) = [y-\mu]^2$

$$\begin{aligned}
-\frac{\frac{\partial^2(\theta)}{\partial\theta^2}}{a(\phi)} + \frac{var(y)}{a^2(\phi)} &= 0 \\
\frac{var(y)}{a^2(\phi)} &= \frac{\frac{\partial^2(\theta)}{\partial\theta^2}}{a(\phi)} \\
&= \frac{\frac{\partial^2(\theta)}{\partial\theta^2}}{a(\phi)} a^2(\phi) \\
var(y) &= \frac{\partial^2(\theta)}{\partial\theta^2} a(\phi) \qquad (3.32) \\
var(y) &= \mu^2 a(\phi) \qquad (3.33)
\end{aligned}$$

Hence, it is apparent that the variance of (Y_i) is a product of $\frac{\partial^2(\theta)}{\partial\theta^2}$ which depends on the canonical parameter and hence on the mean, and $a(\phi)$ which is independent of θ. The variance is therefore considered a function of the mean, μ and can be written as $V(\mu)$

Most at times $a(\phi)$ can be written as $\frac{\phi}{w}$ where ϕ, called the dispersion parameter is contant over the observations and w is a known prior weight that varies from observation to observation (McCullagh & Nelder, 1989).

3.3.7 The Joint Probability Density Function of GLMs

David (2013) specify various response variables that can be assumed by the random variable and suggested regression models that can be considered in fitting them. For a regression model with Y_i as the variable of interest and the X_i as the exogenous variables, the Y can assume any of the followings.

1. A binary variable that can have the value zero or one, the event of interest being the probability of risk occurrence for which it applies the binomial regression models (such as probit, logit, and log-log complementary models);

2. A count variable with values belonging to the set of natural numbers. In this thesis, claim frequency where the Poisson regression can be applied;

3. Real positive variable, with values belonging to the set of positive real numbers. In the context of this thesis, claims cost, where the Gamma regression model can be applied.

The marginal density of Y_i which are conditioned on the explanatory variables X_i is;
$f_Y(y_i;\theta_i,\phi) = \exp\left[\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i,\phi)\right]$
Since the Y_i are independent but not identically distributed, their joint probability density function is;

$$f_Y(y_1, y_2, ..., y_n; \theta_i, \phi) = \prod_{i=1}^{n} f(y_i \mid \theta_i, \phi) \tag{3.34}$$

$$\begin{aligned} f_Y(y \mid \theta, \phi) =& \prod_{i=1}^{n} \exp\left[\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i,\phi)\right] \\ =& f_Y(y \mid \theta, \phi) = \exp\left[\frac{\sum_{i=1}^{n} y_i\theta_i - \sum_{i=1}^{n} b(\theta_i)}{a_i(\phi)} + \sum_{i=1}^{n} c(y_i,\phi)\right] \end{aligned} \tag{3.35}$$

Attention is turned to the claim counts, also known in the world of actuarial science and insurance as claim frequency.

3.4 Claim Count Models

Event count defines the number of times an even occurs. For example, number of accidents (road, fire, earthquakes), number of passengers, number of claims

among others. This type of data is called count data or frequency data. Count data are normally modelled with a type of distribution called discrete distribution. They are mostly modelled by Poisson and negative binomial distribution. Count data regression methods, despite their relatively recent origin has build an impressive body of statistical research on univariate discrete distributions (Cameron & Trivedi, 2013).

3.4.1 Poisson Distribution

As noted by Cameron and Trivedi (2013), the Poisson distribution is the benchmark parametric model for count data. For a discrete random variable Y_i, which represents claim frequency, with intensity/rate parameter, μ, then the probability mass function(PMF), the probability that Y_i takes the value $y_i \in \mathbb{N}$ of Y, length of period, t set to unity is;

$$f(Y_i = y_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}$$

for the insured i where $y = 0, 1, 2, ...$

3.4.1.1 Derivation of The Poisson Distribution

Here we show that the Poisson Distribution can be derived from the Binomial Distribution.

The Poisson distribution was derived as a limiting case of the binomial by the french Mathematician Denis Poisson.

Proof:

If Y_i is a binomial random variable, then

$P(Y_i = y_i) = \binom{n}{y} p^y (1-p)^{n-y}$

We know that for $Y_i \sim bin(n, p)$,

$\mathbb{E}(Y) = np$ and for

$Y_i \sim Pois(\mu_i)$,

$\mathbb{E}(Y) = \mu_i$.

$\Rightarrow np = \mu_i$

$\Rightarrow p = \frac{\mu_i}{n}$

Thus

$$\begin{aligned} P(Y_i = y_i) &= \binom{n}{y} \left(\frac{\mu_i}{n}\right)^y \left(1 - \frac{\mu_i}{n}\right)^{n-y} \\ &= \binom{n}{y} \left(\frac{\mu_i}{n}\right)^y \left(1 - \frac{\mu_i}{n}\right)^{n} \left(1 - \frac{\mu_i}{n}\right)^{-y} \\ &= \frac{n!}{y!(n-y)!} \left(\frac{\mu_i}{n}\right)^y \left(1 - \frac{\mu_i}{n}\right)^{n} \left(1 - \frac{\mu_i}{n}\right)^{-y} \\ &= \frac{n!}{(n-y)!n^y} \frac{\mu_i^y}{y!} \left(1 - \frac{\mu_i}{n}\right)^{n} \left(1 - \frac{\mu_i}{n}\right)^{-y} \end{aligned} \tag{3.36}$$

Taking the limit of $P(Y_i = y_i)$, we end up taking limit of each components in equation 3.36. That is:

$$\begin{aligned} \lim_{n\to\infty} \frac{n!}{(n-y)!n^y} &= \lim_{n\to\infty} \frac{n(n-1)(n-2)...(n-x+1)}{n^y} \\ &= \lim_{n\to\infty} \frac{n(n-1)(n-2)...(n-x+1)}{n \cdot n \cdot n \cdot ... \cdot n} \\ &= \lim_{n\to\infty} \left[\frac{n}{n} \cdot \frac{n-1}{n} \cdot \frac{n-2}{n} \cdot ... \cdot \frac{n-x+1}{n}\right] \\ &= \lim_{n\to\infty} \left[1\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right) \cdot ... \cdot \left(1 - \frac{x-1}{n}\right)\right] \\ &\approx \lim_{n\to\infty} [1(1-0)(1-0) \cdot ... \cdot (1-0)] \\ \lim_{n\to\infty} \frac{n!}{(n-y)!n^y} &\approx 1 \end{aligned} \tag{3.37}$$

The limit of the second component (in equation 3.36) will not be touched. So we

take the limit of the third. That is;

$$\lim_{n\to\infty}\left(1-\frac{\mu_i}{n}\right)^n = e^{-\mu_i} \tag{3.38}$$

and finally

$$\lim_{n\to\infty}\left(1-\frac{\mu_i}{n}\right)^{-y} = 1 \tag{3.39}$$

substituting equations 3.37, 3.38 and 3.39 into equation 3.36, we have

$$\begin{aligned}\lim_{n\to\infty} P(Y_i = y_i) &= \frac{\mu_i^y}{y!}e^{-\mu_i} \\ \lim_{n\to\infty} bin(n,p) &= \frac{e^{-\mu_i}\mu_i^y}{y!}\end{aligned} \tag{3.40}$$

3.4.1.2 Theoretical Setting of The Poisson Regression model

For Y_i dependent variables which represent the claim frequency, the X_i is the vector of linearly independent regressors that are thought to determine Y_i. A regression model based on this distribution follows by conditional distribution of y_i on a k-dimensional vector of covariates $X_i = [x|i, ..., x]$ and the parameters β, through a continous function μ, such that the

$$\mathbb{E}(y_i \mid x_i) = \mu(x_i, \beta) \tag{3.41}$$

Having found out that through the log-linear transform of the density function of the Poisson,

$$\mu_i = e^{\sum_{j=1}^{p}\beta_j X_{ij}} \tag{3.42}$$

This parametrized or transformed mean parameter reveals clearly that the conditional mean is multiplicative and no more linear, as in the ordinary least squares

regression model. That is

$$\begin{aligned}\mu_i =& e^{\sum_{j=1}^{p}\beta_j X_{ij}} \\ =& e^{\beta_1 X_{i1}+\beta_2 X_{i2}+...+\beta_p X_{ip}}\end{aligned}$$

by the law of exponents

$$\mu_i = e^{\beta_1 X_{i1}} e^{\beta_2 X_{i2}} \cdot ... \cdot e^{\beta_p X_{ip}} \tag{3.43}$$

for i number of insured

3.4.1.3 Mean and Variance of the Poisson Regression Model

We now look at the 1st moment (the mean) and the the variance by using the general definition of mean and variance of exponential family of distribution. For the Poisson distribution, we already prove that, considering the general definition of exponential family of distributions;

$$\theta = \log \mu, a(\phi) = 1 = \phi, b(\theta) = \mu, c(y, \phi) = -\log y!$$

and we know that the mean of an exponential family of distribution is

$$\mathbb{E}[y_i] = \mu_i = \frac{\partial b(\theta_i)}{\partial \theta}$$

and the variance is

$$var(y_i) = \frac{\partial^2(\theta_i)}{\partial \theta^2} a_i(\phi)$$

3.4.1.4 Proof of The Mean of The Poisson Model

Since for the Poisson, $\theta = \log \mu$, this implies $\mu = e^{\theta}$. So;

$$\begin{aligned} \mathbb{E}[y_i] = \mu_i &= \frac{\partial}{\partial \theta} e^{\theta} \\ &= e^{\theta} \end{aligned}$$

$$\text{but } \theta = \eta = \sum_{j=1}^{p} \beta_j x_{ij}$$

hence

$$\mu_i = \exp(\sum_{j=1}^{p} \beta_j x_{ij})$$

$$\mu_i = \exp(X^T \beta) \tag{3.44}$$

3.4.1.5 Proof of The Variance of The Poisson Model

for the Poisson $a(\phi) = 1$, therefore the:

$$\begin{aligned} var(y_i) &= \frac{\partial^2(\theta_i)}{\partial \theta^2} a_i(\phi) \\ &= \left[\frac{\partial^2}{\partial \theta^2} e^{\theta} \right] a_i(\phi) \\ &= \frac{\partial}{\partial \theta} \left[\frac{\partial}{\partial \theta} e^{\theta} \right] \cdot 1 \\ &= \frac{\partial}{\partial \theta} \left[e^{\theta} \right] \\ &= e^{\theta} \end{aligned}$$

but $\theta = \eta = \sum_{j=1}^{p} \beta_j x_{ij}$, hence

$$var(y_i) = \exp(\sum_{j=1}^{p} \beta_j x_{ij}) \tag{3.45}$$

since variance is a function of μ we can write

$$var(y_i) = V(\mu_i) = \exp(\sum_{j=1}^{p} x_{ij}\beta_j)$$

$$V(\mu_i) = \exp(X^T\beta) \tag{3.46}$$

One can see that the mean of the Poisson regression model is not constant but changes due to changes in the regressors and also it can be seen that the means and variances of the claim frequency are equal. In statistical term, this assumption is a particular form of heteroskedasticity which is due to equidispersion. (Cameron & Trivedi, 2013)

3.4.1.6 Parameter Estimation For The Poisson Model

The standard estimator for the Poisson model is the maximum likelihood estimator (MLE). Given independent observations, the likelihood function is

$$\mathcal{L}(\beta) = \prod_{i=1}^{n} \frac{e^{-\mu_i}\mu^{y_i}}{y_i!}$$

$$= \prod_{i=1}^{n} \frac{e^{-\exp\left(\sum_{j=1}^{p}\beta_j x_{ij}\right)} \exp(\sum_{j=1}^{p}\beta_j x_{ij})^{y_i}}{y_i!}$$

we take the natural log, to find the log-likelihood function, $\ell(\theta)$

$$\ell(\theta) = \log\left[\prod_{i=1}^{n} \frac{e^{-\exp(\sum_{j=1}^{p}\beta_j x_{ij})} \exp(\sum_{j=1}^{p}\beta_j x_{ij})^{y_i}}{y_i!}\right]$$

$$= \sum_{i=1}^{n} \log\left[\frac{e^{-\exp(\sum_{j=1}^{p}\beta_j x_{ij})} \exp(\sum_{j=1}^{p}\beta_j x_{ij})^{y_i}}{y_i!}\right]$$

$$= \sum_{i=1}^{n}\left[\log[e^{-\exp(\sum_{j=1}^{p}\beta_j x_{ij})}] + \log[\exp(\sum_{j=1}^{p}\beta_j x_{ij})^{y_i}] - \log y_i!\right]$$

$$= \sum_{i=1}^{n}\left[-\exp(\sum_{j=1}^{p}\beta_j x_{ij}) + y_i\sum_{j=1}^{p}\beta_j x_{ij} - \log y_i!\right]$$

$$\ell(\theta) = \sum_{i=1}^{n}\left[y_i\sum_{j=1}^{p}\beta_j x_{ij} - \exp(\sum_{j=1}^{p}\beta_j x_{ij}) - \log y_i!\right] \tag{3.47}$$

$$\ell(\theta) = \sum_{i=1}^{n}\left[y_i X^T\beta - \exp(X^T\beta) - \log y_i!\right] \tag{3.48}$$

$$\text{since } \mu = \exp(\sum_{j=1}^{p}\beta_j x_{ij}), \text{this implies, } \log\mu = \sum_{j=1}^{p}\beta_j x_{ij}$$

hence, the log-likelihood function can be written as (in terms of the mean)

$$\ell(\theta) = \sum_{i=1}^{n}\left[y_i\log\mu - \mu - \log y_i!\right] \tag{3.49}$$

In order to obtain the maximum likelihood estimators, β_j in equation 3.48 we take the first two partial derivatives of the likelihood function and equate them to zero.

$$\begin{aligned}
\frac{\partial \ell(\beta)}{\partial \beta_j} &= \sum_{i=1}^{n} \left[y_i x_{ij} - x_{ij} e^{\sum_{j=1}^{p} \beta_j x_{ij}} \right] = 0 \\
&= \sum_{i=1}^{n} \left[y_i - e^{\sum_{j=1}^{p} \beta_j x_{ij}} \right] x_{ij} = 0 \\
&= \sum_{i=1}^{n} \left[y_i - e^{X^T \beta} \right] x_{ij} = 0 \qquad (3.50) \\
\frac{\partial \ell(\beta)}{\partial \beta_j \beta_k} &= \sum_{i=1}^{n} \left[-e^{(\sum_{j=1}^{p} \beta_j x_{ij})} x_{ij} x_{ik} \right] = 0 \\
&= \sum_{i=1}^{n} \left[-e^{X^T \beta} x_{ij} x_{ik} \right] = 0 \qquad (3.51)
\end{aligned}$$

equations 3.50 and 3.52 can be written in terms of the mean μ_i as

$$\frac{\partial \ell(\beta)}{\partial \beta} = \sum_{i=1}^{n} \left[y_i - \mu_i \right] x_{ij} = 0 \qquad (3.52)$$

$$\frac{\partial \ell(\beta)}{\partial \beta_j \beta_k} = -\sum_{i=1}^{n} \left[\mu x_{ij} x_{ik} \right] = 0 \qquad (3.53)$$

The MLE equations (3.39 and 3.41) suggest that, the unweighted residual is orthogonal to the regressors. However, the maximum likelihood equations are non-linear in the p unknowns β_j, and there is no analytical solution for $\hat{\beta}_j$. Iterative methods, mostly gradient methods such as Newton-Rahpson are used to compute the $\hat{\beta}_j$. (David, 2015).

3.4.2 Zero Inflated Poisson model-ZIP

As said earlier, researchers have turn to using ZIP to model count data inflated by excess zeros. The ZIP distribution is obtained by mixing a distribution degenerate at zero with the Poisson distribution. This allows for the explanatory variables

to be incorporated into both the zero process and the Poisson distribution.

3.4.2.1 Mixed probabilities for ZIP

As noted above, the ZIP model can be divided in to two groups, those events from the zero group, and those events that can be predicted by the Poisson group. The probability ω_i that observation i is in "always-0 group" is predicted by the characteristic function of observation i, which can be written as

$$\omega_i = F(Z_i, \alpha) \tag{3.54}$$

where Z_i is the vector of covariates and α_i, the vector of coefficients of logit or probit regression. The probability that, the observation i is in "not always-0 group" becomes $1 - \omega_i$. For observations in "not always-0 group", their positive count outcome can be predicted by standard Poisson model. This can be written as

$$P(y_i \mid x_i) = \frac{\mu_i^{y_i}}{y_i} e^{-\mu_i}$$

where μ_i is the conditional mean

The three probabilities that can be generated by the ZIP are;

1. zero counts in "always-0 group" (which are called structural zeros in Xia et al (2012))

$$P(Y_i = 0 \mid X_i, Z_i) = \omega_i \times 1 \tag{3.55}$$

since the individuals with zero count has probability of 1

2. Zero counts in "not always-0 group" (which are called sampling zeros in Xia et al (2012)).

$$\begin{aligned} P(Y_i = 0 \mid X_i, Z_i) =& (1 - \omega_i) \times \frac{\mu_i^0 e^{-\mu_i}}{0!} \\ =& (1 - \omega_i) \times e^{-\mu} \end{aligned} \tag{3.56}$$

3. and finally a non-zero count in "not always-0 group"

$$f(Yi = y_i) = (i - \omega_i)\frac{\mu_i^{y_i}}{y_i} e^{-\mu_i} \tag{3.57}$$

By combining equations 3.44, 3.45 and 3.46 above, the overall PMF of the ZIP distribution is

$$f(Yi = y_i) = \begin{cases} \omega_i + (1 - \omega_i)e^{-\mu_i}, & \text{for } y_i = 0 \\ (i - \omega_i)\frac{\mu_i^{y_i}}{y_i} e^{-\mu_i}, & \text{for } y_i > 0 \end{cases} \tag{3.58}$$

3.4.2.2 Theoretical Setting of The ZIP Model Under the Exponential Family of Distribution

We now turn to find the various parameters of the ZIP under the exponential family of distribution.

$$f(y_i \mid, \theta, \phi) = \begin{cases} \omega_i + (1 - \omega_i)e^{-\mu_i}, & \text{for } y_i = 0 \\ (i - \omega_i)\frac{\mu_i^{y_i}}{y_i} e^{-\mu_i}, & \text{for } y_i > 0 \end{cases} \tag{3.59}$$

for the count data part, let the density be denoted as f_R. Hence

$f_R(y_i \mid, \theta, \phi) = (i - \omega_i)\frac{\mu_i^{y_i}}{y_i} e^{-\mu_i}$, $y > 0$

we take a natural logarithm of both sides

$$\begin{aligned}\log f_R(y_i \mid, \theta, \phi) &= \log\left[(i-\omega_i)\frac{\mu_i^{y_i}}{y_i}e^{-\mu_i}\right] \\ &= \log(1-\omega_i) + \log\mu_i^{y_i} + \log e^{-\mu_i} - \log y! \\ &= y_i \log\mu_i + \log(1-\omega_i) - \mu_i - \log y_i!\end{aligned}$$

taking exponent of both sides

$$\exp\left[\log f_R(y_i \mid, \theta, \phi)\right] = \exp\left[y_i \log\mu_i + \log(1-\omega_i) - \mu_i - \log y_i!\right]$$

this becomes

$$f_R(y_i \mid, \theta, \phi) = \exp\left[y_i \log\mu_i - \mu_i + \log(1-\omega_i) - \log y_i!\right] \tag{3.60}$$

comparing this to the density of exponential family of distributions
$\theta = \log\mu_i, a(\phi) = 1, b(\theta) = \mu_i, c(y,\phi) = \log(1-\omega_i) - \log y_i!$. This means for Z_i and X_i, sets of covariates, Z_i (a vector of covariates that predict the zeros in the data) is linked to ω_i through the logit link (or logistic regression) and X_i is link to μ_i through the log transformation. That is

$$Z^T\alpha = \log\left[\frac{\omega_i}{1-\omega_i}\right] \tag{3.61}$$

and

$$X^T\beta = \log\mu_i \tag{3.62}$$

3.4.2.3 Mean and Variance of ZIP Model

We turn our attention to the 1st Moment and Variance of the ZIP model and compare this with the standard poisson regression

1. Mean of ZIP model

$$\mathbb{E}(y) = (1 - \omega_i)\mu_i \tag{3.63}$$

2. Variance of ZIP model

$$V(y) = (1 - \omega_i)\mu_i(1 + \mu^2\omega_i) \tag{3.64}$$

3.4.2.4 Proof of 1st Moment/Mean of ZIP

$$\begin{aligned}
\mathbb{E}(y \mid X, Z) &= \sum_{y=1}^{\infty} \left[y_i(1 - \omega_i)\frac{\mu_i^{y_i}}{y_i!}e^{-\mu_i} \right] \\
&= (1 - \omega_i)e^{-\mu_i} \sum_{y=1}^{\infty} y_i \frac{\mu_i^{y_i}}{y_i!} \\
&= (1 - \omega_i)e^{-\mu_i} \sum_{y=1}^{\infty} y_i \frac{\mu_i \mu_i^{y_i - 1}}{y_i(y_i - 1)!} \\
&= (1 - \omega_i)e^{-\mu_i}\mu_i \sum_{y=1}^{\infty} \frac{\mu_i^{y_i - 1}}{(y_i - 1)!} \\
&= (1 - \omega_i)e^{-\mu_i}\mu_i \sum_{y=0}^{\infty} \frac{\mu_i^{y_i}}{y_i!}
\end{aligned}$$

by Euler's formula $\sum_{y=0}^{\infty} \frac{\mu_i^{y_i}}{y_i!} = e^{\mu_i}$ hence

$$\begin{aligned}
\mathbb{E}(y \mid X, Z) &= (1 - \omega_i)e^{-\mu_i}\mu_i e^{\mu_i} \\
&= (1 - \omega_i)\mu_i e^{-\mu_i + \mu_i} \\
&= (1 - \omega_i)\mu_i \cdot 1 \\
\Rightarrow \mathbb{E}(y \mid X, Z) &= (1 - \omega_i)\mu_i
\end{aligned} \tag{3.65}$$

3.4.2.5 2nd Moment of ZIP With Proof

We obtain the 2nd moment by its definition

$$\begin{aligned}
\mathbb{E}(y^2 \mid X, Z) &= \sum_{y=1}^{\infty} \left[y_i^2 (1-\omega_i) \frac{\mu_i^{y_i}}{y_i!} e^{-\mu_i} \right] \\
&= (1-\omega_i) e^{-\mu_i} \sum_{y=1}^{\infty} y_i^2 \frac{\mu_i^{y_i}}{y_i!} \\
&= (1-\omega_i) e^{-\mu_i} \sum_{y=1}^{\infty} y_i^2 \frac{\mu_i \mu_i^{y_i-1}}{y_i (y_i-1)!} \\
&= (1-\omega_i) e^{-\mu_i} \mu_i \sum_{y=1}^{\infty} y_i \frac{\mu_i^{y_i-1}}{(y_i-1)!}
\end{aligned}$$

let y=y+1 which means y-1=y, then

$$\begin{aligned}
&= (1-\omega_i) e^{-\mu_i} \mu_i \sum_{y=0}^{\infty} (y_i+1) \frac{\mu_i^{y}}{(y_i)!} \\
&= (1-\omega_i) e^{-\mu_i} \mu_i \left[\sum_{y=0}^{\infty} y_i \frac{\mu_i^{y}}{y_i!} + \frac{\mu_i^{y}}{y_i!} \right] \\
&= (1-\omega_i) e^{-\mu_i} \mu_i \left[\sum_{y=0}^{\infty} y_i \frac{\mu_i^{y}}{y_i!} + \sum_{y=0}^{\infty} \frac{\mu_i^{y}}{y_i!} \right] \\
&= (1-\omega_i) e^{-\mu_i} \mu_i \left[\sum_{y=1}^{\infty} y_i \frac{\mu_i^{y}}{y_i!} + \sum_{y=0}^{\infty} \frac{\mu_i^{y}}{y_i!} \right] \\
&= (1-\omega_i) e^{-\mu_i} \mu_i \left[\sum_{y=0}^{\infty} y_i \frac{\mu_i \mu_i^{y-1}}{y_i (y_i-1)!} + \sum_{y=0}^{\infty} \frac{\mu_i^{y}}{y_i!} \right] \\
&= (1-\omega_i) e^{-\mu_i} \mu_i \left[\sum_{y=0}^{\infty} \frac{\mu_i \mu_i^{y-1}}{(y_i-1)!} + \sum_{y=0}^{\infty} \frac{\mu_i^{y}}{y_i!} \right] \\
&= (1-\omega_i) e^{-\mu_i} \mu_i \left[\mu_i e^{\mu_i} + e^{\mu_i} \right] \\
&= (1-\omega_i) (\mu_i^2 e^{\mu_i} e^{-\mu_i} + \mu_i e^{\mu_i} e^{-\mu_i}) \\
&= (1-\omega_i) (\mu_i^2 e^{-\mu_i+\mu_i} + \mu_i e^{-\mu_i+\mu_i}) \\
&= (1-\omega_i) (\mu_i^2 \cdot 1 + \mu_i \cdot 1) \\
\mathbb{E}(y^2 \mid X, Z) &= (1-\omega_i) (\mu_i^2 + \mu_i) \qquad (3.66)
\end{aligned}$$

Now the variance by inference theory is (using equations 3.65 and 3.66)

$$\begin{aligned} var(y_i) &= \mathbb{E}(y_i) - [\mathbb{E}(y_i)]^2 \\ &= (1-\omega_i)(\mu^2+\mu) - [(1-\omega_i)\mu]^2 \\ &= \mu^2 + \mu - \omega\mu^2 - \omega\mu - [1 - 2\omega_i + \omega_i^2]\mu^2 \\ &= \mu^2 + \mu - \omega\mu^2 - \omega\mu - \mu^2 - 2\omega\mu_i^2 + \omega_i^2\mu^2 \\ &= \mu - \omega\mu + \omega\mu_i^2 - \omega_i^2\mu^2 \\ &= \mu(1-\omega_i) + \mu_i^2\omega(1-\omega_i) \\ &= (1-\omega_i)(\mu + \mu_i^2\omega) \\ var(y) &= (1-\omega_i)\mu(1+\mu\omega) \end{aligned} \tag{3.67}$$

One can see clearly that the variance of the ZIP model is greater than its mean.

That is, $(1-\omega_i)\mu(1+\mu\omega) > (1-\omega_i)\mu$

Which can be rewritten as

$\mathbb{E}(y_i)(1+\mu\omega) > \mathbb{E}(y_i)$

Hence the ZIP has made provision for over-dispersion. It can also be seen from the definition that as ω_i approaches zero, the ZIP model will be approximately a Poisson distribution. We can illustrate this from the ratio of the variance to the mean of ZIP model.

$$\begin{aligned} \frac{var(y_i)}{\mathbb{E}(y_i)} &= \frac{(1-\omega_i)\mu(1+\mu\omega)}{(1-\omega_i)\mu} \\ &= (1+\mu\omega) \end{aligned} \tag{3.68}$$

making μ the subject from the mean of the ZIP model, equation 3.65

$$\mathbb{E}(y \mid X, Z) = (i - \omega_i)\mu$$
$$\mu = \frac{\mathbb{E}(y_i)}{1 - \omega_i} \tag{3.69}$$

and placing it in equation 3.57, it produces

$$= 1 + [\frac{\mathbb{E}(y_i)}{1 - \omega_i}]\omega_i \tag{3.70}$$

and the ratio of variance to mean finally becomes

$$\frac{var(y)}{\mathbb{E}(y_i)} = 1 + [\frac{\omega_i}{1 - \omega_i}]\mathbb{E}(y_i) \tag{3.71}$$

It can be demonstrated that, as ω_i approaches zero,

$\frac{var(y)}{\mathbb{E}(y_i)} = 1$, suggesting that the "structural zeros" decrease to zero, which implies equality of means and variance

3.4.2.6 Parameter Estimation of The ZIP model

For a ZIP model, the density is

$$f(Y_i = y_i) = \begin{cases} \omega_i + (1 - \omega_i)e^{-\mu}, & \text{for } y_i = 0 \\ (1 - \omega_i)\frac{\mu_i^{y_i}}{y_i}e^{-\mu}, & \text{for } y_i > 0 \end{cases}$$

We adopt the loglikelihood function as posited in Penman and Brager (2021) The log-likelihood of the ZIP model is divided in to three categories, ℓ_1, ℓ_2 and ℓ_3. Hence, for the ZIP model, the log-likelihood function is

$$\ell = \ell_1 + \ell - \ell_3 \tag{3.72}$$

Where

$$\ell_1 = \sum_{(i:y_i=0)} \log[\lambda_i + e^{-\mu}] \tag{3.73}$$

Equation 3.73 explains the log-likelihood function when the event (in this case claim counts) are zeros

$$\ell_2 = \sum_{(i:y_i>0)} [y_i \log \mu_i - \mu_i - \log y!] \tag{3.74}$$

Equation 3.74 suggests that the event count can be modelled by the standard Poisson

$$\ell_3 = \sum_{i=1} \log(1 + \lambda_i) \tag{3.75}$$

Equation 3.75 suggests that the event count is 1

hence, the log-likelihood function of the ZIP model is:

$$\ell = \sum_{(i:y_i=0)} \log[\lambda_i + e^{-\mu}] + \sum_{(i:y_i>0)} [y_i \log \mu_i - \mu_i - \log y!] - \sum_{i=1} \log(1 + \lambda_i) \tag{3.76}$$

In order to incorporate the explanatory variables, we note, ω_i has a logit link with the explanatory variables from the zero process and the μ has log-link from the count data process.

$$\begin{aligned} Z^T\alpha &= \log\left[\frac{\omega_i}{1-\omega_i}\right] \\ e^{Z^T\alpha} &= \frac{\omega_i}{1-\omega_i} = \lambda_i \end{aligned} \tag{3.77}$$

and

$$\begin{aligned} X^T\beta &= \log\mu \\ e^{X^T\beta} &= \mu \end{aligned}$$

hence the log-likelihood function for the ZIP model can be rewritten as:

$$\ell = \sum_{(i:y_i=0)} \log[\frac{\omega_i}{1-\omega_i} + e^{-\mu}] + \sum_{(i:y_i>0)} [y_i \log \mu_i - \mu_i - \log y!] - \sum_{i=1} \log[1 + \frac{\omega_i}{1-\omega_i}]$$

$$\ell = -n \log[1 + \frac{\omega_i}{1-\omega_i}] + \sum_{(i:y_i=0)} \log[\frac{\omega_i}{1-\omega_i} + e^{-\mu}] + \sum_{(i:y_i>0)} [y_i \log \mu_i - \mu_i - \log y!]$$

To finally incorporates the explanatory in order to estimate the parameters, It is worth noting to recall;

$$\begin{aligned} \lambda_i &= e^{Z^T\alpha} = \frac{\omega_i}{1-\omega_i} \\ \lambda_i &= \frac{e^{Z^T\alpha}}{1+e^{Z^T\alpha}} \end{aligned} \tag{3.78}$$

Hence, in terms of the the Z, X

$$\ell = -n \log[1 + \frac{e^{Z^T\alpha}}{1+e^{Z^T\alpha}}] + \sum_{(i:y_i=0)} \log[\frac{e^{Z^T\alpha}}{1+e^{Z^T\alpha}} + e^{\exp(X^T\beta)}] + \sum_{(i:y_i>0)} \Big[y_i \log e_i^{X^T\beta} - e_i^{X^T\beta} - \log y!\Big]$$

$$= -n \log[1 + \frac{e^{Z^T\alpha}}{1+e^{Z^T\alpha}}] + \sum_{(i:y_i=0)} \log[\frac{e^{Z^T\alpha}}{1+e^{Z^T\alpha}} + e^{\exp(X^T\beta)}] + \sum_{(i:y_i>0)} \Big[y_i X^T\beta_i - e_i^{X^T\beta} - \log y!\Big] \tag{3.79}$$

The parameters (α and β) will be obtained using R software.

Even though the negative binomial regression model is not of ultimate interest in modelling the claim count data, we need to consider it first before its corresponding zero inflated model for easy understanding

3.4.3 Negative binomial model

We say a variable Y has a negative binomial distribution, write $NB(\mu, \gamma)$, where μ is the mean and $\gamma > 0$ is the dispersion parameter, if

1. there is a latent variable, $\Theta \sim \Gamma(\gamma, \gamma)$
2. conditionally, given, $\Theta, Y \sim Poi(\Theta\mu)$.

1. (**1st Definition**)

$$f(Y = y) = \frac{\Gamma(\gamma + y)}{\Gamma(\gamma)y!} \left(\frac{\gamma}{\gamma + \mu}\right)^{\gamma} \left(\frac{\mu}{\gamma + \mu}\right)^{y} \tag{3.80}$$

assume $p = \frac{\mu}{\gamma+\mu} \in (0, 1)$ and $y \in \mathbb{N}_0$, then the pmf of the NB can be re-parametrised as

2. (**2nd Definition**)

$$\begin{aligned} f(Y = y) &= \binom{y + \gamma - 1}{y} (1 - p)^{\gamma} p^{y} \\ &= \binom{y + r - 1}{r - 1} p^{r} (1 - p)^{y} \end{aligned} \tag{3.81}$$

where $r = \gamma$ and $p = \frac{\gamma}{\gamma+\mu}$

3.4.3.1 Negative binomial as a Member of The Exponential Family of Distributions

To show that the negative binomial is a member of the exponential family of distribution, we use the parametrization given in Dobson and Barnett (2008).

$$f(y;\theta,\phi) = \binom{y+r-1}{r-1} p^r (1-p)^y$$

take natural log of both sides

$$\begin{aligned}
\log\left[f(y;\theta,\phi)\right] &= \log\left[\binom{y+r-1}{r-1} p^r (1-p)^y\right] \\
&= \log\binom{y+r-1}{r-1} + \log p^r + \log(1-p)^y \\
&= \log(1-p)^y + \log p^r + \log\binom{y+r-1}{r-1} \\
&= y\log(1-p) + r\log p + \log\binom{y+r-1}{r-1}
\end{aligned}$$

taking exponent of both sides

$$\begin{aligned}
\exp\left[\log\left[f(y;\theta,\phi)\right]\right] &= \exp\left[y\log(1-p) + r\log p + \log\binom{y+r-1}{r-1}\right] \\
f(y;\theta,\phi) &= \exp\left[y\log(1-p) + r\log p + \log\binom{y+r-1}{r-1}\right] \qquad (3.82)
\end{aligned}$$

comparing equation 3.82 to the general form of exponential distributions

$\theta = \log(1-p), a(\phi) = 1 = \phi, b(\theta) = -r\log p, c(y,\phi) = \log\binom{y+r-1}{r-1}$

3.4.3.2 Mean And Variance of NB (Proof From The Exponential Form)

Recall; the mean of an exponential family of distribution is

$$\mathbb{E}[y_i] = \frac{\partial b(\theta)}{\partial \theta}$$

We know from equation 3.82 that, $b(\theta) = -r\log p$ but in order to find the first derivative of this we need to write $b(\theta)$ in terms of θ

$\theta = \log(1-p)$, hence

$$e^{\theta} = 1 - p$$

$$p = 1 - e^{\theta}$$

$$\log p = log(1 - e^{\theta})$$

hence

$$\begin{aligned} b(\theta) &= -r\log(1-e^{\theta}) \\ \mathbb{E}[y_i] &= \frac{\partial[-r\log(1-e^{\theta})]}{\partial\theta} \\ &= \frac{-r(-e^{\theta})}{1-e^{\theta}} \\ &= \frac{re^{\theta}}{1-e^{\theta}} \end{aligned}$$

in terms of p

$$\mathbb{E}[y_i] = \mu = \frac{r(1-p)}{p} \tag{3.83}$$

Recall the variance of exponential family of distribution as:

$$var(y_i) = \frac{\partial^2 b(\theta_i)}{\partial\theta^2} a_i(\phi)$$

Hence, the Variance of the NB becomes;

$$
\begin{aligned}
var(y_i) &= \frac{\partial^2[-r\log(1-e^{\theta})]}{\partial\theta^2}a_i(\phi) \\
&= \frac{\partial}{\partial\theta}\left[\frac{\partial[-r\log(1-e^{\theta})]}{\partial\theta}\right]\cdot 1 \\
&= \frac{\partial}{\partial\theta}\left[\frac{re^{\theta}}{1-e^{\theta}}\right]\cdot 1 \\
&= \frac{\partial}{\partial\theta}\left[\frac{re^{\theta}}{1-e^{\theta}}\right] \\
&= \frac{\frac{\partial}{\partial\theta}re^{\theta}-\frac{\partial}{\partial\theta}(1-e^{\theta})}{(1-e^{\theta})^2} \\
&= \frac{(1-e^{\theta})\frac{\partial}{\partial\theta}re^{\theta}-re^{\theta}\frac{\partial}{\partial\theta}(1-e^{\theta})}{(1-e^{\theta})^2} \\
&= \frac{(1-e^{\theta})re^{\theta}-re^{\theta}(-e^{\theta})}{(1-e^{\theta})^2} \\
&= \frac{(re^{\theta}-re^{2\theta}+re^{2\theta}}{(1-e^{\theta})^2} \\
&= \frac{re^{\theta}}{(1-e^{\theta})^2} \\
var(y_i) &= \frac{r(1-p)}{p^2}
\end{aligned}
\tag{3.84}
$$

we know that $\mu = \frac{r(1-p)}{p}$ and $p = \frac{\gamma}{\gamma+\mu}$. Hence

$$
\begin{aligned}
var(y_i) = \frac{r(1-p)}{p^2} &= \frac{r(1-p)}{p}\frac{1}{p} \\
&= \mu\left[\frac{1}{\frac{\gamma}{\gamma+\mu}}\right] \\
&= \mu\left[\frac{\gamma+\mu}{\gamma}\right] \\
&= \mu\left[\frac{\gamma}{\gamma}+\frac{\mu}{\gamma}\right] \\
var(y_i) &= \mu\left[1+\frac{\mu}{\gamma}\right]
\end{aligned}
\tag{3.85}
$$

3.4.3.3 The Regression Form of The NB

The PMF of negative binomial is as previously defined

$$\begin{aligned} f(Y=y) &= \frac{\Gamma(\gamma+y)}{\Gamma(\gamma)y!}\left(\frac{\gamma}{\gamma+\mu}\right)^{\gamma}\left(\frac{\mu}{\gamma+\mu}\right)^{y} \\ &= \frac{\Gamma(y+r)p^{r}(1-p)^{y}}{\Gamma(y+1)\Gamma(r)} \end{aligned}$$

we incorporate the regressors in to the model after we obtained the log-likelihood equation. $\mathcal{L}$ and ℓ denotes the likelihood function and the log-likelihood function respectively.

$$\begin{aligned} \mathcal{L} &= \prod_{1=i}^{n} f(Y_i = y_i \mid X_i) \\ &= \prod_{1=i}^{n}\left[\frac{\Gamma(y+r)p^{r}(1-p)^{y}}{\Gamma(y+1)\Gamma(r)}\right] \\ \log\mathcal{L} = \ell &= \log\prod_{1=i}^{n}\left[\frac{\Gamma(y+r)p^{r}(1-p)^{y}}{\Gamma(y+1)\Gamma(r)}\right] \\ &= \sum_{1=i}^{n}\Big[\log(\Gamma(y+r))\Big] + \sum_{i=1}^{n} r\log p + \sum_{i=1}^{n} y\log(1-p) - \sum_{i=1}^{n}\log\Gamma(y+1) - \sum_{i=1}^{n}\log\Gamma(r) \\ &= \sum_{1=i}^{n}\Big[\log(\Gamma(y+r))\Big] + nr\log p + \sum_{i=1}^{n} y\log(1-p) - \sum_{i=1}^{n}\log\Gamma(y+1) - n\log\Gamma(r) \end{aligned} \tag{3.86}$$

including γ and μ

we know; $\log p = \log(1-e^{\theta}), \log(1-p) = \theta \; and \; \theta = r$

$$= \sum_{1=i}^{n}\Big[\log(\Gamma(y+\theta))\Big] + n\theta\log(1-e^{\theta}) + \sum_{i=1}^{n} y\theta - \sum_{i=1}^{n}\log\Gamma(y+1) - n\log\Gamma(\theta) \tag{3.87}$$

Where $\theta = X^{T}\beta = \log\left[\frac{\mu}{\mu+\gamma}\right]$ through the canonical link and hence $e^{X^{T}\beta} = \left[\frac{\mu}{\mu+\gamma}\right]$, normally written in term of only the mean as $\mu = e^{X^{T}\beta}$ since the partial derivatives of the parameters do not have closed form, the parameters are obtained

through MLE algorithms in R software.

3.4.4 Zero Inflated Negative Binomial Model

If one replaces the Poisson distribution in the ZIP model, the negative binomial is obtained. Hence, the general form of the ZINB is as defined in the next section

3.4.4.1 Theoretical Setting of ZINB

$$f(y_i \mid, \theta, \phi) = \begin{cases} \omega_i + (1-\omega_i)\frac{\gamma}{\gamma+\mu}, & \text{for } y_i = 0 \\ (1-\omega_i)\frac{\Gamma(\gamma+y)}{\Gamma(\gamma)y!}\left(\frac{\gamma}{\gamma+\mu}\right)^{\gamma}\left(\frac{\mu}{\gamma+\mu}\right)^{y}, & \text{for } y_i > 0 \end{cases} \tag{3.88}$$

3.4.4.2 Mean and Variance of ZINB

The mean of ZINB as posited by Xia *et al.* (2012), is

$$E(y) = \mu(1-\omega) \tag{3.89}$$

and the variance is

$$var(y) = \mu(1-\omega)(1+\mu[\omega+\alpha]) \tag{3.90}$$

As suggested by SAS Global Forum (2008), it is more straight forward to estimate $\alpha = 1/\gamma$ instead of γ.

From the definition of the ZINB, ω is written as a function of $Z^T\pi$, where Z is the vector of covariates that predict zero inflation in data and π is the vector of coefficients of the zero-inflated process

The function that relates the vector $Z^T\pi$ to the probability, ω is called zero inflated link function and can be specified as either logit or probit or other models of binary outcomes.

In order to see the statistical capability of the ZINB, we find the ratio of the variance to the mean of the ZINB

$$\frac{var(y)}{\mathbb{E}y} = 1 + \mu(\omega + \alpha) \tag{3.91}$$

replacing μ as $\dfrac{\mathbb{E}y}{1-\omega}$

$$= 1 + \frac{\mathbb{E}y}{1-\omega}(\omega + \alpha)$$

$$\frac{var(y)}{\mathbb{E}y} = 1 + \left[\frac{\omega + \alpha}{1-\omega}\right](\mathbb{E}y) \tag{3.92}$$

There are two major things to take away from the mean, the variance and the ratio of the variance to mean. These are

1. for ZINB, the variance is greater that the mean. i.e. $var(y) > \mathbb{E}(y)$ which implies $\mu(1-\omega)(1+\mu[\omega+\alpha]) > \mu(1-\omega)$. This demonstrates that the ZINB model has the capability to model overdispersion present in the data

2. since the coefficient of the mean, $\frac{\omega+\alpha}{1-\omega}$ in the ratio of the variance to the mean has both the dispersion parameter, α and the zero inflated parameter ω, it demonstrates that the ZINB incorporates both heterogeneity in data and overdispersion. This explains the reason why using the NB to model the claim count was ignored.

   In this thesis, the emphasis is also to compare zero inflated models to one of the mostly used distributions of count data (e.g. claim count), which is the Poisson Model

3.4.4.3 Parameter Estimation of The ZINB

For parameter estimation as said previously, The re-parametrised form of the ZINB was used, which will take $\gamma = \frac{1}{\alpha}$, since it is easier to estimate the α. Therefore

$$f(y_i \mid, \theta, \phi) = \begin{cases} \omega_i + (1-\omega_i)g(y_i = 0), & \text{for } y_i = 0 \\ (i-\omega_i)g(y_i), & \text{for } y_i > 0 \end{cases} \tag{3.93}$$

where

$$g(y_i = 0) = \left(\frac{1}{1+\alpha\mu)}\right)^{\alpha^{-1}} \tag{3.94}$$

$$g(y) = \frac{\Gamma(y+\alpha^{-1})}{\Gamma(y+1)\Gamma(\alpha^{-1})}\left(\frac{1}{1+\alpha\mu)}\right)^{\alpha^{-1}}\left(\frac{\alpha\mu}{1+\alpha\mu}\right)^{y} \tag{3.95}$$

As stated earlier, the function that relates the vector $Z^T\pi$ to the probability, ω is called zero inflated link function and can be specified as either logit or probit or other models of binary outcomes. That is

$$\begin{aligned} Z^T\pi &= \log(\frac{\omega}{1-\omega}) \\ \Rightarrow \omega &= \frac{e^{Z^T\pi}}{1+e^{Z^T\pi}} \\ \text{where } e^{Z^T\pi} &= \lambda_i \\ \Rightarrow \omega &= \frac{\lambda}{1+\lambda} \end{aligned} \tag{3.96}$$

and the $X^T\beta$ is linked to the μ by the log function

$\log\mu = X^T\beta$

Hence the log-likelihood function, ℓ as posited by Penman and Brager (2021) is

$$\ell = \ell_1 + \ell_2 + \ell - \ell_4 \tag{3.97}$$

$$\ell_1 = \sum_{(i:y_i=0)} \log[\lambda(1+\alpha\mu)^{-\alpha^{-1}}] \tag{3.98}$$

$$\ell_2 = \sum_{(i:y_i>0)} \sum_{i=0}^{1} \log(1+\alpha^{-1}) \tag{3.99}$$

$$\ell_3 = \sum_{(i:y_i>0)} \left[-\log y! - (y+\alpha^{-1})\log(1+\alpha\mu) + y\log\alpha + y\log\mu\right] \tag{3.100}$$

$$\ell_4 = \sum_{i=0} \log(1+\lambda) \tag{3.101}$$

3.5 Model selection for fitting claim counts

As indicated by Xia *et al.* (2012), that though ZIP and ZINB address structural zeros, it is difficult to tell whether they are the appropriate choice for the data at hand. Goodness of fit criteria was therefore applied to help in selecting the best fit model. First, the score test was conducted to check for zero inflation in the data, secondly the Vuong Test was conducted to select the best between a zero inflated model and its ordinary counterpart. In this case Poisson vs ZIP model, and NB vs ZINB. And finally goodness of fit statistics and criteria were used to assess the predictive performance of the three models considered (Ordinary Poisson, ZIP model and ZINB).

3.5.1 The Score Test

The score test is a statistical test introduced by Van den Boerk (1995) which tests for zero inflation in the data. Even though, the score test is used to test for the fit of the zeros it also accounts for the means of the fitted Poisson distribution. As noted by Van den Boerk (1995), this is one of the importance of the score test

because it compares the number of zeros to the mean of the observations. Unlike other test statistics, the score test does not require the fitting of the inflated models but just a Poisson (since it is the distribution under the null hypothesis). Wolny-Dominiak (2013) stated that, in testing whether there are many observed zeros, the score test assumes that the ω_i is the same for all observation (e.g. all policies). We test against the null hypothesis that the probability of having zeros in the data is zero. That is

$$H_0 : \omega = 0 \tag{3.102}$$

The score test statistics is defined as

$$S(\hat{\beta}) = \frac{\left[\sum_{i=1}^{n} \frac{1-e^{\mu_i}}{e^{\mu_i}}\right]^2}{\left[\sum_{i=1}^{n} \frac{1-e^{\mu_i}}{e^{\mu_i}}\right]^2 - n\bar{y}} \tag{3.103}$$

where $\bar{y}$ is the average of the claims count, and it is known that the $S(\hat{\beta})$ follows an asymptotic $\mathcal{X}^2$ distribution with 1 degree of freedom. As noted earlier, the presence of excess zero leads to overdispersion since there will be great variability in claim data as a result of the zeros. Hence rejection of the null hypothesis implies presence of overdispersion which suggest a non-suitability of the Poisson regression for fitting the data.

3.5.2 The Vuong Test

Posited to by Wolny-Dominiak (2013), the Vuong test compares two models based on Kullback-Leibler information criteria as a measure of the distance between these models (in this case Poisson vs ZIP, and NB vs ZINB).

If two competing claims count distributions have densities $f_1(x, \beta_1)$ and $f_2(x, \beta_2)$, then the null hypothesis is:

if the distribution are proven to be nested;

$$H_0 : LR \equiv \overset{\max}{\beta}\{\mathbb{E}[\log f_1(x, \beta_1) - \log f_2(x, \beta_2)]\} = 0 \tag{3.104}$$

If the two distributions are non-nested then, the H_0 statistics will become;

$$H_0 : LR \equiv \sqrt{n}\sum_{i=1}^{n}[\log f_1(x, \beta_1) - \log f_2(x, \beta_2)] \tag{3.105}$$

Or Vuong test statistic (V) can be calculated as

$$V = \frac{m_i\sqrt{n}}{sd(m)} \tag{3.106}$$

where, $sd(m), n$ are the standard deviation of m and sample size respectively, and

$$m_i = \log\left[\frac{f_1(x, \beta_1)}{f_2(x, \beta_2)}\right] \tag{3.107}$$

V follows an asymptotically standard normal distribution and the test is directional (Xia et al ,2012). Large values interprets f_1 as the best model fit and larger negative values suggest f_2 is the favoured model, and a value close to zero signifies that the two models are not different(they are equivalent) or do fits the data well.

In this study the Vuong Test for non-nested models was used. It should be noted that, there are literatures written to the use of Vuong Test for Zero Inflated Models and its misuse. (Vuoung, 1989; Wilson, 2015)

3.5.3 Goodness of Fit Tests

The classical statistical tests that are based on the likelihood approach that are normally used in model specification and selection are Wald, likelihood ratio, and Lagrange multiplier. (Cameron and Trivedi, 2013). In this research, other

goodness of fit tests like AIC and BIC was used in testing the performance of the claim counts models.

3.5.3.1 Akaike Information Criteria (AIC)

The Akaike Information Criteria is named after the Japanese statistician Hirotugu Akaike. The AIC is used to measure the information loss when using statistical model to represent a process that generate a data. The smaller the AIC, the less the information loss. It is defined as

$$AIC = 2k - 2\log \mathcal{L}. \tag{3.108}$$

where $\mathcal{L}$ denotes the maximum value of the likelihood function and k, the number of estimated paramaters in the model. Where the term $2k$ is the penalty for the complexity of the model (Frees, 2010). This means, fit cannot be improved upon by introducing additional parameters. AIC is mostly used for comparing non-nested models (Xia *et al.*, 2012)

3.5.3.2 Bayesian Information Criterion (BIC)

This information criterion was developed by Schwarz in 1978 as an alternative to AIC, using Bayesian methods (Frees, 2010)

$$BIC = \log k - 2\log \mathcal{L} \tag{3.109}$$

What BIC does is that, it gives more weight to the number of parameters. This means it will suggest a more parsimonious model than AIC. However, as noted by Xia *et al.* (2012), this can lead to over simplified models. In essence, The AIC and BIC is used to penalize over-fitting.

3.6 The Gamma Regression Model For Claims Cost

David (2013) pointed out that, claim amount is more difficult to predict than claim frequency. This is because, there are no distributions for positive real values. However, many researchers consider the gamma model in fitting claims cost. Other model assessed in modelling claim amounts are inverse Gaussian, log-normal, exponential, Weibull, etc.

The density of the gamma distribution can be written as;
We recall the previously defined.

$$f(c_i) = \left(\frac{\nu}{\mu_i}\right)^{\nu} \frac{c_i^{\nu-1} e^{-\frac{\nu c_i}{\mu_i}}}{\Gamma(\nu)},\ c_i > 0$$

Where c_i denotes the cost of claims for insured i assumed to be independently distributed.

3.6.1 Mean and Variance of The Gamma Distribution

3.6.1.1 Mean

$$\begin{aligned}
\mathbb{E}(C) &= \int_0^\infty \left(\frac{\nu}{\mu_i}\right)^\nu \frac{c_i^{\nu-1} e^{-\frac{\nu c_i}{\mu_i}}}{\Gamma(\nu)} c_i \\
&= \frac{1}{\Gamma(\nu)} \left(\frac{\nu}{\mu_i}\right)^\nu \int_0^\infty c_i c_i^{\nu-1} e^{-\frac{\nu c_i}{\mu_i}} \\
&= \frac{1}{\Gamma(\nu)} \left(\frac{\nu}{\mu_i}\right)^\nu \int_0^\infty c_i^{\nu-1+1} e^{-\frac{\nu c_i}{\mu_i}} \\
&= \frac{1}{\Gamma(\nu)} \left(\frac{\nu}{\mu_i}\right)^\nu \int_0^\infty c_i^{\nu} e^{-\frac{\nu c_i}{\mu_i}} \\
&= \frac{1}{\Gamma(\nu)} \left(\frac{\nu}{\mu_i}\right)^\nu \int_0^\infty c_i^{\nu} e^{-\frac{\nu c_i}{\mu_i}} \frac{\Gamma(\nu+1)}{\Gamma(\nu+1)} \frac{\left(\frac{\nu}{\mu}\right)^{\nu+1}}{\left(\frac{\nu}{\mu}\right)^{\nu+1}} \\
&= \frac{\Gamma(\nu+1)}{\Gamma(\nu)} \left(\frac{\nu}{\mu_i}\right)^\nu \frac{1}{\left(\frac{\nu}{\mu}\right)^{\nu+1}} \int_0^\infty \frac{\left(\frac{\nu}{\mu}\right)^{\nu+1}}{\Gamma(\nu+1)} c_i^{\nu} e^{-\frac{\nu c_i}{\mu_i}}
\end{aligned}$$

but since the gamma distribution, denoted $\Gamma(\nu, \frac{\nu}{\mu})$ is a legitimate pdf then

$$\int_0^\infty \left(\frac{\nu}{\mu_i}\right)^\nu \frac{c_i^{\nu-1} e^{-\frac{\nu c_i}{\mu_i}}}{\Gamma(\nu)} dx = 1 \tag{3.110}$$

hence

$$\int_0^\infty \frac{\left(\frac{\nu}{\mu}\right)^{\nu+1}}{\Gamma(\nu+1)} c_i^{\nu} e^{-\frac{\nu c_i}{\mu_i}} dx = 1 \tag{3.111}$$

since it is a gamma distribution of $\Gamma(\nu+1, \frac{\nu}{\mu})$, then;

$$\begin{aligned}
\mathbb{E}(C) &= \frac{\Gamma(\nu+1)}{\Gamma(\nu)} \left(\frac{\nu}{\mu_i}\right)^{\nu} \frac{1}{\left(\frac{\nu}{\mu}\right)^{\nu+1}} \cdot 1 \\
&= \frac{\Gamma(\nu+1)}{\Gamma(\nu)} \left(\frac{\nu}{\mu_i}\right)^{\nu} \left(\frac{\nu}{\mu}\right)^{-\nu-1} \\
&= \frac{\Gamma(\nu+1)}{\Gamma(\nu)} \left(\frac{\nu}{\mu_i}\right)^{\nu-\nu} \left(\frac{\nu}{\mu}\right)^{1} \\
&= \frac{\Gamma(\nu+1)}{\Gamma(\nu)} \left(\frac{\nu}{\mu_i}\right)^{0} \left(\frac{\nu}{\mu}\right)^{-1} \\
&= \frac{\Gamma(\nu+1)}{\Gamma(\nu)} \cdot 1 \left(\frac{\mu}{\nu}\right) \\
&= \frac{(\nu+1-1)!}{(\nu-1)!} \left(\frac{\mu}{\nu}\right) \\
&= \frac{\nu(\nu-1)!}{(\nu-1)!} \left(\frac{\mu}{\nu}\right) \\
\mathbb{E}(C) &= \mu
\end{aligned} \tag{3.112}$$

3.6.1.2 Variance

Before we find the variance, we first find the second moment

$$\begin{aligned}
\mathbb{E}(C^2) &= \int_0^{\infty} c_i^2 \left(\frac{\nu}{\mu_i}\right)^{\nu} \frac{c_i^{\nu-1} e^{-\frac{\nu c_i}{\mu_i}}}{\Gamma(\nu)} \\
&= \frac{1}{\Gamma(\nu)} \left(\frac{\nu}{\mu_i}\right)^{\nu} \int_0^{\infty} c_i^{\nu-1+2} e^{-\frac{\nu c_i}{\mu_i}} \\
&= \frac{1}{\Gamma(\nu)} \left(\frac{\nu}{\mu_i}\right)^{\nu} \int_0^{\infty} c_i^{\nu+1} e^{-\frac{\nu c_i}{\mu_i}} \\
&= \frac{1}{\Gamma(\nu)} \left(\frac{\nu}{\mu_i}\right)^{\nu} \int_0^{\infty} c_i^{\nu+1} e^{-\frac{\nu c_i}{\mu_i}} \\
&= \frac{1}{\Gamma(\nu)} \left(\frac{\nu}{\mu_i}\right)^{\nu} \int_0^{\infty} c_i^{\nu+1} e^{-\frac{\nu c_i}{\mu_i}} \frac{\Gamma(\nu+2)}{\Gamma(\nu+2)} \frac{\left(\frac{\nu}{\mu}\right)^{\nu+2}}{\left(\frac{\nu}{\mu}\right)^{\nu+2}} \\
&= \frac{\Gamma(\nu+2)}{\Gamma(\nu)} \left(\frac{\nu}{\mu_i}\right)^{\nu} \frac{1}{\left(\frac{\nu}{\mu}\right)^{\nu+2}} \int_0^{\infty} \frac{\left(\frac{\nu}{\mu}\right)^{\nu+2}}{\Gamma(\nu+2)} c_i^{\nu} e^{-\frac{\nu c_i}{\mu_i}}
\end{aligned}$$

but since the gamma distribution, denoted $\Gamma(\nu, \frac{\nu}{\mu})$ is a density then

$$\int_0^\infty \left(\frac{\nu}{\mu_i}\right)^\nu \frac{c_i^{\nu-1} e^{-\frac{\nu c_i}{\mu_i}}}{\Gamma(\nu)} dx = 1$$

hence

$$\int_0^\infty \frac{\left(\frac{\nu}{\mu}\right)^{\nu+2}}{\Gamma(\nu+2)} c_i^{\nu+1} e^{-\frac{\nu c_i}{\mu_i}} dx = 1 \tag{3.113}$$

since it is a gamma distribution of $\Gamma(\nu+2, \frac{\nu}{\mu})$, hence

$$\begin{aligned}
\mathbb{E}(C^2) &= \frac{\Gamma(\nu+2)}{\Gamma(\nu)} \left(\frac{\nu}{\mu_i}\right)^\nu \frac{1}{\left(\frac{\nu}{\mu}\right)^{\nu+2}} \cdot 1 \\
&= \frac{\Gamma(\nu+2)}{\Gamma(\nu)} \left(\frac{\nu}{\mu_i}\right)^\nu \left(\frac{\nu}{\mu}\right)^{-\nu-2} \\
&= \frac{\Gamma(\nu+2)}{\Gamma(\nu)} \left(\frac{\nu}{\mu_i}\right)^{\nu-\nu} \left(\frac{\nu}{\mu}\right)^{-2} \\
&= \frac{\Gamma(\nu+2)}{\Gamma(\nu)} \left(\frac{\nu}{\mu_i}\right)^{0} \left(\frac{\nu}{\mu}\right)^{-2} \\
&= \frac{\Gamma(\nu+2)}{\Gamma(\nu)} \cdot 1 \left(\frac{\mu}{\nu}\right)^2 \\
&= \frac{(\nu+2-1)!}{(\nu-1)!} \left(\frac{\mu}{\nu}\right)^2 \\
&= \frac{(\nu+1)!}{(\nu-1)!} \left(\frac{\mu}{\nu}\right)^2 \\
&= \frac{(\nu+1)(\nu+1-1)!}{(\nu-1)!} \left(\frac{\mu}{\nu}\right)^2 \\
&= \frac{(\nu+1)(\nu)!}{(\nu-1)!} \left(\frac{\mu}{\nu}\right) \\
&= \frac{(\nu+1)(\nu)(\nu-1)!}{(\nu-1)!} \left(\frac{\mu^2}{\nu^2}\right) \\
\mathbb{E}(C^2) &= \mu^2 \frac{(\nu+1)}{\nu}
\end{aligned} \tag{3.114}$$

Now we find the variance of the gamma distribution

$var(c_i) = \mathbb{E}(C^2) - [\mathbb{E}(C)]^2$

$$\frac{\mu^2(\nu+1)}{\nu} - \mu^2$$
$$\frac{\mu^2(\nu+1) - \nu\mu^2}{\nu}$$
$$\frac{\nu\mu^2 + \mu^2 - \nu\mu^2}{\nu}$$
$$var(c_i) = \frac{\mu^2}{\nu} \tag{3.115}$$

3.6.2 Maximum likelihood Estimate of The Gamma Model

The likelihood function of the gamma distribution is obtained by taking the product of the marginal densities.

$$\mathcal{L} = \prod_{i=1}^{n} \left[(\frac{\nu}{\mu})^{\nu} \frac{y^{\nu-1} e^{-\frac{\nu y}{\mu}}}{\Gamma(\nu)} \right]$$

we take the log of the above to find the log-likelihood function

$$\begin{aligned}
\ell = \log \mathcal{L} &= \log \prod_{i=1}^{n} \left[(\frac{\nu}{\mu})^{\nu} \frac{y^{\nu-1} e^{-\frac{\nu y}{\mu}}}{\Gamma(\nu)} \right] \\
&= \sum_{i=1}^{n} \left[\log \{ (\frac{\nu}{\mu})^{\nu} \frac{y^{\nu-1} e^{-\frac{\nu y}{\mu}}}{\Gamma(\nu)} \} \right] \\
&= \sum_{i=1}^{n} \left[\log (\frac{\nu}{\mu})^{\nu} + \log \left(\frac{y^{\nu-1} e^{-\frac{\nu y}{\mu}}}{\Gamma(\nu)} \right) \right] \\
&= \sum_{i=1}^{n} \nu \log (\frac{\nu}{\mu}) + \sum_{i=1}^{n} \log \left[y^{\nu-1} e^{-\frac{\nu y}{\mu}} \right] - \sum_{i=1}^{n} \log \Gamma(\nu) \\
&= \sum_{i=1}^{n} \nu \left[\log \nu - \log \mu \right] + \sum_{i=1}^{n} \log y^{\nu-1} + \sum_{i=1}^{n} \log \left[e^{-\frac{\nu y}{\mu}} \right] - \sum_{i=1}^{n} \log \Gamma(\nu) \\
&= -\sum_{i=1}^{n} \frac{\nu y}{\mu} - \sum_{i=1}^{n} \nu \log \mu + \sum_{i=1}^{n} (\nu - 1) \log y + \sum_{i=1}^{n} \nu \log \nu - \sum_{i=1}^{n} \log \Gamma(\nu) \\
&= -\sum_{i=1}^{n} \frac{\nu y}{\mu} - n\nu \log \mu + \sum_{i=1}^{n} (\nu - 1) \log y + n\nu \log \nu - n \log \Gamma(\nu)
\end{aligned} \tag{3.116}$$

$\theta = \mu^{-1} = \frac{1}{\mu}, b(\theta) = -\log \mu$ and $a(\phi) = \nu^{-1} = \frac{1}{\nu}, c(y, \theta) = (\nu - 1) \log y + \nu \log \nu - \log \Gamma(\nu)$

3.6.3 Finding the Mean of Gamma Model by Using The Definition of EDF

$$\theta = \frac{1}{\mu} \tag{3.117}$$

but

$$b(\theta) = -\log\mu = -\log(\frac{1}{\theta}) = \log(\theta^{-1})^{-1} = \log(\theta^{1})$$

$$b(\theta) = \log\theta \tag{3.118}$$

$$\mathbb{E}(C) = \frac{\partial b(\theta)}{\partial\theta} = \frac{\partial\log\theta}{\partial\theta} = \frac{1}{\theta} = \frac{1}{1/\mu}$$

$$\mathbb{E}(C) = \mu$$

$$\ell = -\sum_{i=1}^{n}\frac{\nu y}{\exp(X^T\beta)} - n\nu\log X^T\beta + \sum_{i=1}^{n}(\nu-1)\log y + n\nu\log\nu - n\log\Gamma(\nu) \tag{3.119}$$

We incorporate the explanatory variables by using the log link function, $\log\mu = X^T\beta$ defined parameters in exponential family of distribution.

3.6.4 The Link Function of The Gamma Model

From the above defined parameters, the link function for the gamma model is

$$\theta = \mu^{-1}$$

$$X^T\beta = \mu^{-1}$$

Making the Mean the subject,

$$\mu = \theta^{-1}$$

$$\mu = (X^T\beta)^{-1} \tag{3.120}$$

Due to the limitation of the inverse link, in that it becomes undefined when η is zero and μ could assume negative values, most actuarial literature uses, a log link function as seen under the Poisson model. In other words for μ to be positive, then $\eta = X^T\beta$ must be negative, and this places a restriction on the β

3.6.5 Finding the Variance of Gamma Model Using The Definition of EDF

Recall the Variance of EDF and together with equation 3.108;

$$\begin{aligned} var(c) &= \frac{\partial^2 b(\theta)}{\partial\theta^2} a(\phi) \\ &= \frac{\partial^2 \log\theta}{\partial\theta^2}\nu^{-1} \\ &\frac{1}{\theta^2}\cdot\frac{1}{\nu} \end{aligned}$$

from equation 3.107

$$\begin{aligned} &= \frac{1}{1/\mu^2}\cdot\frac{1}{\nu} \\ var(c) &= \frac{\mu^2}{\nu} \end{aligned}$$

3.6.6 Measuring The Goodness of Fit For The Gamma Model

In GLM analysis, one of the key goodness of fit measures is the Deviance. The deviance measures the variation between the full or saturated model and the model under consideration. If this discrepancy measure is between the full model and the fitted model, we have what we call the residual deviance (in most statistical softwares, like R Statistical Software) and when it is between the full model and the null model it is called the null deviance. The full model which is described as a perfect model, is obtained when the number of parameters to estimate is

the same as the number of observations (Jong & Heller, 2008). The fitted model also called the proposed model is set up to explain that, the data points can be estimated by an intercept and a given number of parameters. The null model assumes the absence of all predictors and explains the data points with only a single parameter. In comparing the fit between the count models, the analysis focused on other goodness of fit measures than the deviance. But since it was assumed that the claims amount follows a gamma distribution, the null and the residual deviances are compared to make a decision on the fit of the claims amount model. The formal definition the null deviance and the residual deviance (casually called deviance) for the Gamma Model are given below and proofs of the Deviance are shown in the appendix A.

3.6.6.1 Null Deviance

The null deviance is twice the log-likelihood ratio statistic for testing the null model as against the full model. Here the null model is denoted as n and the full model is denoted as f. ℓ denotes log-likelihood function. Hence, ℓ_f and ℓ_n denotes the log-likelihood function for the full model and the null model respectively. Hence the deviance is formally defined as

$$D(y;\hat{\mu}) = 2(\ell_f(y;y) - \ell_n(\hat{\mu};y)) \tag{3.121}$$

where the deviance $D(y;\hat{\mu})$ is the deviance.

3.6.6.2 The Deviance-The Residual Deviance

The residual deviance is formally defined as twice the log-likelihood ratio statistic for testing the fitted model as against the full model. Here the fitted model is denoted as m and the full model is denoted as f. ℓ denotes log-likelihood function. Hence, ℓ_f and ℓ_m denotes the log-likelihood function for the full model and the

fitted model respectively. Hence the deviance is formally defined as

$$D(y;\hat{\mu}) = 2(\ell_f(y;y) - \ell_m(\hat{\mu};y)) \tag{3.122}$$

where the deviance $D(y;\hat{\mu})$ is the "deviance" or residual deviance. The log-likelihood estimate of the fitted model, ℓ_m is usually expressed in terms of the mean parameter. For the Gamma model, the deviance is obtained as:

$$D(y;\hat{\mu}) = 2\sum_{i=1}^{n}\left[-\log(\frac{y_i}{\hat{\mu}_i}) + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}\right] \tag{3.123}$$

3.6.6.3 The Scaled Deviance

When the residual deviance is divided by the dispersion parameter, a deviance called the scaled deviance is obtained. Scaled deviance is usually denoted by:

$$D^*(y;\hat{\mu}) = D(y;\hat{\mu})/\phi \tag{3.124}$$

In SAS statistical software, deviance is stated to mean scaled deviance. This may be due to the fact that, when assessing the fitness of the models whose dispersion parameter is not 1 (i.e. $\phi \neq 1$) as in cases like the Poisson, there is a need to scale the residual deviance with the (scale) factor ϕ. The scaling is important for in the Gamma model since its dispersion parameter is not 1. The scale deviance for the Gamma model is given below:

$$D^*(y;\hat{\mu}) = 2\nu\sum_{i=1}^{n}\left[-\log(\frac{y_i}{\hat{\mu}}) + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}\right] \tag{3.125}$$

since $\phi = \nu^{-1}$

3.6.6.4 Assessing The Use of Deviance As a Goodness of Fit Test

The deviance is assumed to be approximately $\mathcal{X}_{n-p}$. That is a chi-squared distribution with $n-p$ degree of freedom. (Jong & Heller, 2008). A large value of deviance suggests that, the model lacks fit, a smaller value of deviance suggests the fitted model fit the data well. If a model fit the data well, it means the log-likelihood of the fitted model is close to the log-likelihood of the full model. But as noted by Jong and Heller (2008), the log-likelihood of the fitted model could not be greater than that of the full model.

One can also divide the deviance by the degree of freedom, $n-p$, and if the result is larger than one, it means the model lacks fit.

One can also compare the null deviance and the residual deviance calculated for the fitted regression model. If the residual deviance and its degree of freedom is lower than that of the null deviance and its degree of freedom, it suggests that, adding explanatory variables to the model improved the model.

McCullagh and Nelder (1989) warned against the use of deviance as a measure of goodness of fit, stating that one cannot rely on it as an absolute measure of goodness of it but can be used to compare two nested models. They emphasized this by saying that, the deviance in itself does not need to be assumed to follow a chi-squared distribution, but rather the difference between two deviances (of two models) is quitely approximated by the chi-squared distribution

3.7 Pure Premium Model

As noted by Ohlsson and Johansson (2010), the standard GLM tariff analysis is to do separate analyses for claim frequency and claims cost, and the premium is calculated by multiplying the results. They suggested the importance of separating

the frequency and claim analysis based on the fact that:

1. claim frequency is usually much more stable than claims cost and often much of the power of rating factors is related to claim frequency and the factors can be affected with greater accuracy.

2. a separate analysis gives more insight into how a rating factor affects the pure premium. Hence the pure premium model models the cost of claims and the frequency of claims incorporating the risk factors of the policyholders.

The expected annual cost of claims, (S) through the collective risk model is:

$$\begin{aligned}
\mathbb{E}(S) &= \mathbb{E}\left[\sum_{i=1}^{Y}[C_i]\right] \\
&= \mathbb{E}[\mathbb{E}(S|Y)] \\
&= \mathbb{E}\left[\mathbb{E}\left[\sum_{i=1}^{Y}[C_i|Y]\right]\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{Y}\mathbb{E}[C_i|Y]\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{Y}\mathbb{E}[C_i]\right] \\
&= \mathbb{E}[Y\mathbb{E}[C_i]] \\
&= \mathbb{E}[Y_1] \times \mathbb{E}[C_1]
\end{aligned}$$

using the link function

$$= g_y^{-1}(\eta_y) \times g_y^{-1}(\eta_c) \tag{3.126}$$

Where $g_y^{-1}(\eta_y)$ is the inverse of the link function from the claim count model and $g_y^{-1}(\eta_c)$ is the inverse of the link function from the claims cost model

Chapter 4

ANALYSIS AND DISCUSSION

4.1 Introduction

In this study, model selection methods were used to determine the distribution that best fit the claim count. The model that best fits the claim count will be used to estimate the expected claim count and hence be incorporated into the premium model defined previously in section 3.7. The expected claim amount is estimated by the Gamma model. In this chapter, discussion and analysis of the results and findings of the study are presented. The models were applied to the European data obtained from kaggle.com and the insurance data obtained from a major insurance company in Ghana.

Part I

Analysis and Results Obtained Using the French Insurance Data

4.2 Description and composition of data

The data used in this study was obtained from an open online database, www.kaggle.com. The direct link to this data will be provided in the appendix. Two datasets were obtained: on claim counts and claims amount. The two datasets freMTPLfreq and freMTPLsev with risk factors are collected for 413,169 motor third-party liability policies (observed mostly on one year). freMTPLfreq contains the risk features and claim counts, whilst freMTPLsev contains claims amount. Both tables can be linked together via the corresponding policy ID. The claim counts data is made up of 10 variables, 2 of which, that is policy ID and exposure are not used in the analysis. The variables in the dataset can be seen in the table 4.1. The claimNb, which is the dependent variables describes the number of claims recorded and or paid within the coverage period of one year. The CarAge variable describes the number of years the vehicle was in use, the DriverAge reveals the age of the insured driver, the Power variable is a categorical variables that describes the capacity of the car. We have the Brand variable of the car divided into seven groups, Gas variable describing the fuel type of the car, Region variable describes the area of residence of the policyholder, and Density variable describing the inhabitants in the city that the driver of the car lives in, measured in number of inhabitants per square kilometres. The uniform continuous variable Exposure tells as the period of exposure for a policy in years. Table 4.1 summarizes the lists of variables in the dataset

4.2.1 Distribution of Claim Counts

figure 4.1 and table 4.2 show the distribution of one of the target variables, claim counts. The number of claims submitted by the policyholder during the coverage period ranges from 0 to 4 claims. And it can be seen from the claim counts data that, approximately 96.3% of the policyholders did not report any claims

Table 4.1: List of Variables in our dataset

Variable	Type	Description	Values
Brand	Character	The brand of the car, divided into 7 groups	A - Renault Nissan and Citroen B - Volkswagen, Audi, Skoda and Seat, C - Opel, General Motors and Ford, D - Fiat, E - Mercedes, Chrysler and BMW, F - Japanese (except Nissan) and Korean brands, G - Other
CarAge	Numeric	The vehicle age, in years	From 0 to 100 years
ClaimAmount	Numeric	The cost of the claim	From 5 to 182,467
ClaimNb	Numeric	The number of claims during the exposure period	From 0 to 4 claims
Density	Numeric	The density of inhabitants, in number of inhabitants per km^2	From 2 to 27,000 sq. km
DriverAge	Numeric	The driver's age, in years (in France, people can drive a car from the age of 18)	From 18 to 99 years
Gas	Character	The car's fuel type	Diesel or Regular (petrol)
Power	Character	The power of the car (split into ordinal categories)	From d to h
Region	Character	The policy region in France (based on the 1970-2015 region classification). 6 different Regions	Centre, Ile-de-France, Bretagne, Pays-de-la-Loire, Aquitaine, Nord-Pas-de-Calais, Other

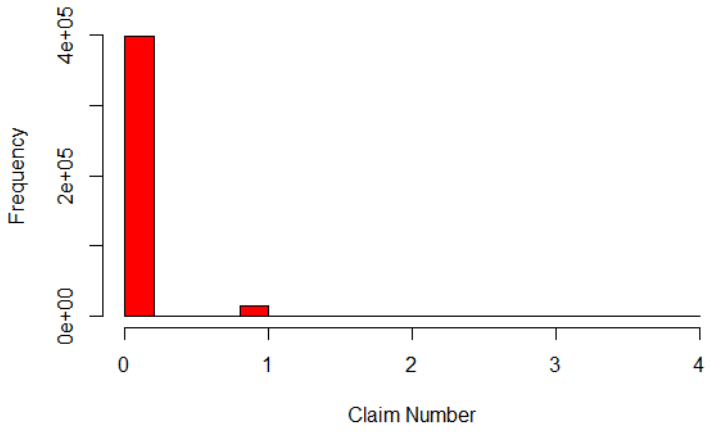

Figure 4.1: Bar Chart Showing The Distribution of Claim Count

throughout the coverage period. followed by 3.5% of them making one claim and an insignificant number of the policyholders submitted 2, 3 and 4 claims. This is an indication of excess zeros in the data. Since the claim count is non-negative as show in table 4.2 and figure 4.2, the Poisson distribution will be one of the first distribution to consider fitting to the data.

Table 4.2: Distribution of Claim Count

Claim Number	Frequency	Percent
0	397,779	96.3
1	14,633	3.5
2	726	0.2
3	28	0.0
4	3	0.0
Total	413,169	100.0

4.3 Limitation of Data

Data was obtained from www.kaggle.com. Even though this data is real data from a French Motor third-party liability policy, it covered only a year.

4.4 Claim Count Distribution fit

One can see from figure 4.1 that, the count data does not follow a normal distribution. Even though, sometimes Ordinary Least Squares (OLS) regression can be used in modelling claim counts, the assumption of normality of this model, makes it unfit for the data. Sometimes, the OLS can be log transformed, but due to the zeros in the data, data can be loss during the log transformation leading to bias in estimates. Again, due to the presence of excess zeros in the data and the purpose of this research, the standard Poisson model was compared to the zero-inflated models of ZIP and ZINB.

4.4.1 The Poisson Regression Model

Since the outcome variable, claim count occurred in a given period of time, it is ideal to fit the Poisson distribution first to the data. The Poisson regression model was fit to the freMTPLfreq data at a significant level of 5% ($\alpha = 0.05$). The estimates of the regressors, the standard errors and z-statistics as well as the p-values are all obtained from the Poisson regression model. From table 4.3, it can be seen that the variable Power is significant for e, f, g, i, j and k. It implies that this factors influence the occurrence of claim at $\alpha = 0.05$. Age of car and Driver's age are also found to be significant. What this means is that, the car's age and the driver's age both influence the number of claims submitted by a policyholder. For brands; Japanese (except Nissan) or Korean brand of automobile is the only variable among the brands of car that influences the number of claims submitted in a coverage period. The rest of the brands were found to be statistically insignificant at $\alpha = 0.05$. The regular fuel type was found to be statistically significant. This implies that, the regular fuel type a car uses influences the number of claims the policyholder can submit within an insured period. Apart from the Nord-Pas-de-Calais region, the rest of the

variables under Region were statistically significant. That is, the policy region in France (based on the 1970-2015 region classification) are seen to influence the number of claims during the exposure period. The density of inhabitants, in the city that the driver of the car lives in, (in number of inhabitants per km^2) was also found to be a contributing factor in the occurrence of claims within an exposure period since it is statistically significant at the alpha risk level of 0.05. Due to the dominance of zeros in the data as seen in figure 4.1 and table 4.2, it is of statistical importance to fit zero inflated regression models since the presence of this zeros can lead to overdispersion in the data. We therefore fit the Zero Inflated Poisson (ZIP) model and Zero Inflated Negative Binomial (ZINB) model to the data and later on use comparison and selection tests to select the model that best fits the data among the standard Poisson count model, ZIP, and ZINB. As noted in the Methodology section, zero inflated models account for both structural zeros and sampling zeros. It is noteworthy that, from the coefficient of the intercept, it can be interpreted that, if all variables are zero or absent, the Poisson model is still significant at $\alpha = 0.05$.

4.4.2 Zero Inflated Poisson Model

Where the standard Poisson model becomes incapacitated in dealing with excess zeros in a given data, the Zero Inflated Poisson count model is adopted to deal with the excess zeros and its effect of overdispersion in the data. We applied therefore, ZIP model to the freMTPLfreq data at a level of significance of $\alpha = 5\%$. From the count model in table 4.4, it can be seen that, the age of the car and the driver's age are both significant. Among the brand or vehicle makeup, Japanese or Korean Brand, Mercedes, Chrysler or BMW Brand are both significant. The regular fuel type of a car also influences the occurrence of claims in an exposure period since it is significant at $\alpha = 0.05$. Among the regions of policyholders, Haute-Normandie, lle-de-France, Limousin are all found to be significant. It

Table 4.3: Poisson Model Showing Regressors, Estimates, Std. Error, z-value and Pr(>|z|) at $\alpha = 0.05$

Regressors	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.1749	0.0582	-54.56	2×10^{-16}*
Powere	0.0849	0.0280	3.03	0.0024*
Powerf	0.0961	0.0273	3.52	0.0004*
Powerg	0.0427	0.0272	1.57	0.1166
Powerh	0.0613	0.0388	1.58	0.1145
Poweri	0.1328	0.0430	3.09	0.0020*
Powerj	0.1511	0.0437	3.46	0.0005*
Powerk	0.1563	0.0560	2.79	0.0052*
Powerl	0.0767	0.0832	0.92	0.3568
Powerm	0.0924	0.1192	0.77	0.4384
Powern	0.1250	0.1364	0.92	0.3593
Powero	0.1594	0.1383	1.15	0.2491
CarAge	-0.0068	0.0015	-4.37	1.23×10^{-5}*
DriverAge	-0.0016	0.0006	-2.90	0.0037*
BrandJapanese (except Nissan) or Korean	-0.5557	0.0453	-12.26	2×10^{-16}*
BrandMercedes, Chrysler or BMW	-0.0320	0.0532	-0.60	0.5468
BrandOpel, General Motors or Ford	0.0691	0.0448	1.54	0.1232
Brandother	-0.0458	0.0622	-0.74	0.4621
BrandRenault, Nissan or Citroen	-0.0412	0.0392	-1.05	0.2932
BrandVolkswagen, Audi, Skoda or Seat	0.0171	0.0459	0.37	0.7102
GasRegular	-0.0640	0.0172	-3.72	0.0002*
RegionBasse-Normandie	0.1639	0.0563	2.91	0.0036*
RegionBretagne	0.1921	0.0388	4.95	7.32×10^{-7}*
RegionCentre	0.0902	0.0337	2.68	0.0075*
RegionHaute-Normandie	-0.3386	0.0742	-4.57	4.98×10^{-6}*
RegionIle-de-France	0.0968	0.0401	2.41	0.0158*
RegionLimousin	0.2905	0.0777	3.74	0.0002*
RegionNord-Pas-de-Calais	0.0070	0.0449	0.15	0.8769
RegionPays-de-la-Loire	0.1312	0.0399	3.29	0.0010*
RegionPoitou-Charentes	0.1724	0.0471	3.66	0.0002*
Density	1.442×10^{-5}	1.951×10^{-13}	7.39	1.47×10^{-6}*
*** means significant at** $\alpha = 0.05$				

is important to also note that, the intercept is also significant for the count model. For the zero-inflation model, the age of the car, the driver's age are found to be significant at $\alpha = 0.05$. Again,for the brand, Japanese or Korean Brand, Mercedes, Chrysler or BMW Brand are both significant. For the Zero-inflated model, the regions, Bretagne, Centre, Normandie, and Poiton-Charentes are significant. All the variables significant for the zero-inflated model are said to be predictors of zeros in the data.

4.4.3 Zero Inflated Negative Binomial Model

The ZINB model is noted for being able to model overdispersion present in a data due to excess zero and heterogeneity. Zero Inflated Negative Binomial was therefore fitted to the FreMTPLfreq data at a level of significance of $\alpha = 0.05$. In Table 4.5, the Regressors, the Estimates, the standard error (Std. Error), z-values and p-values are all shown. From table 4.5, it can be seen that, the age of the car, driver's age are both statistically significant. This implies that the car's age and the age of the driver, influence the frequency of claims in an exposure period. The vehicle brands; Japanese or Korean, and Mercedes, Chrysler or BMW are also found to influence the occurrence of the number of claims in a coverage period. The fuel type, regular (petrol) used also influences the number of claims in the insured period. It was found out to be statistically significant. Among the regions of the insureds; Haute-Normandie, lle-de-France, and Limousin were found to be statistically significant. For the zero-inflation model, the age of the car, the driver's age, Japanese or Korean brand, Mercedes, Crysler or BMW brand, and Opel, General Motors or Ford were all found to be predictors of the zeros in the data. Bretagne, Centre, Haute-Normandie and Poiton-Charentes regions are also predictors of zeros in the data. It is worth noting that, when the count model and zero-inflation model have their predictors evaluated at zero, the model will still be statistically significant. This is due to the fact that, the intercept of both

Table 4.4: Zero-inflated Poisson Model Showing Regressors, Estimates, Std. Error, z-value and Pr(>|z|) For Both The Count Model and Zero-Inflated Model at $\alpha = 0.05$

Count model coefficients (poisson with log link)				
Regressors	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.868109	0.133942	-13.947	2×10^{-16}*
CarAge	-0.065546	0.003688	-17.774	2×10^{-16}*
DriverAge	-0.004307	0.001031	-4.177	2.96×10^{-5}*
BrandJapanese (except Nissan) or Korean	0.456026	0.115197	3.959	7.54×10^{-5}*
BrandMercedes, Chrysler or BMW	0.318523	0.118541	2.687	0.00721*
BrandOpel, General Motors or Ford	0.084090	0.102261	0.822	0.41090
Brandother	-0.167910	0.135064	-1.243	0.21380
BrandRenault, Nissan or Citroen	0.044038	0.090307	0.488	0.62580
BrandVolkswagen, Audi, Skoda or Seat	0.080222	0.105854	0.758	0.44854
GasRegular	-0.067026	0.033199	-2.019	0.04350*
RegionBasse-Normandie	0.221484	0.130344	1.699	0.08928
RegionBretagne	-0.035745	0.088306	-0.405	0.68563
RegionCentre	-0.135680	0.078717	-1.724	0.08477
RegionHaute-Normandie	-0.770898	0.157400	-4.898	9.70×10^{-7} *
RegionIle-de-France	0.180533	0.084491	2.137	0.03262*
RegionLimousin	0.361385	0.164551	2.196	0.02808*
RegionNord-Pas-de-Calais	0.042480	0.106856	0.398	0.69097
RegionPays-de-la-Loire	0.068799	0.090012	0.764	0.44467
RegionPoitou-Charentes	-0.024304	0.100705	-0.241	0.80929
Zero-inflation model coefficients (binomial with logit link)				
(Intercept)	1.296994	0.263255	4.927	8.36×10^{-7}*
CarAge	-0.158473	0.010726	-14.774	2×10^{-16}*
DriverAge	-0.007269	0.002054	-3.538	0.000403*
BrandJapanese (except Nissan) or Korean	1.315905	0.226221	5.817	6.00e-09*
BrandMercedes, Chrysler or BMW	0.663561	0.251682	2.637	0.008376*
BrandOpel, General Motors or Ford	-0.000243	0.239621	-0.001	0.999191
Brandother	-0.535602	0.398350	-1.345	0.178770
BrandRenault, Nissan or Citroen	0.160756	0.211036	0.762	0.446213
BrandVolkswagen, Audi, Skoda or Seat	0.073500	0.242347	0.303	0.761674
GasRegular	-0.008224	0.066472	-0.124	0.901538
RegionBasse-Normandie	0.148185	0.223844	0.662	0.507971
RegionBretagne	-0.464865	0.166814	-2.787	0.005324*
RegionCentre	-0.485064	0.143599	-3.378	0.000730*
RegionHaute-Normandie	-0.955917	0.346955	-2.755	0.005866*
RegionIle-de-France	-0.090488	0.139304	-0.650	0.515968
RegionLimousin	0.168979	0.274950	0.615	0.538832
RegionNord-Pas-de-Calais	0.049722	0.179968	0.276	0.782333
RegionPays-de-la-Loire	-0.089355	0.162125	-0.551	0.581529
RegionPoitou-Charentes	-0.397159	0.194737	-2.039	0.041403*
*** means significant at** $\alpha = 0.05$				

models are statistically significant at $\alpha = 0.05$

4.5 Model Comparison for Count Models

4.5.1 Score Test

The Score Test, tests for zero inflation in the data. From table 4.6, it can be concluded that, the data is inflated with zeros since the null hypothesis of not having zeros in the data (i.e. the probability of having zero, $\omega = 0$) is rejected at a significance level of $\alpha = 0.05$. This confirms, as seen from the observed zeros that, there is occurrence of zero-inflation in the data and hence overdispersion due to the high variability of zero claims. This suggests that, the ordinary Poisson model is not appropriate fit for the data since this can lead to underestimation of standard errors and regression parameters (Wolny-Dominiak, 2013). The result can be seen in table 4.6

4.5.2 The Vuong Test

The Vuong non-nested test is used to compare the predicted probabilities of two non-nested models. In this case, the Vuong test was used to compare the zero inflated Poisson model to the non-zero inflated Poisson (standard Poisson) model. Eventhough the Negative Binomial is not of interest here, it was shown that, the zero-inflated negative binomial turned out to be superior to the ordinary negative binomial. Same for Zero-Inflated Poisson, being superior to the standard Poisson model. This is tested under the null hypothesis that the two models are indistinguishable and the test statistic is asymptotically distributed standard normal. i.e. $V \sim N(0, 1)$. The Vuong Test results are shown for Poisson vs ZIP and NB vs ZINB in table 4.7.

Table 4.5: Zero-inflated Negative Model Showing Regressors, Estimates, Std. Error, z-value and Pr(>|z|) For Both The Count Model and Zero-Inflated Model at $\alpha = 0.05$

Count model coefficients (Negbin with log link)				
Regressors	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.868227	0.133941	-13.948	2×10^{-16}*
CarAge	-0.065545	0.003688	-17.775	2×10^{-16}*
DriverAge	-0.004307	0.001031	-4.177	2.95×10^{-5}*
BrandJapanese (except Nissan) or Korean	0.456089	0.115197	3.959	7.52×10^{-5}*
BrandMercedes, Chrysler or BMW	0.318589	0.118537	2.688	0.00719*
BrandOpel, General Motors or Ford	0.084117	0.102259	0.823	0.41074
Brandother	-0.167813	0.135056	-1.243	0.21404
BrandRenault, Nissan or Citroen	0.044083	0.090306	0.488	0.62544
BrandVolkswagen, Audi, Skoda or Seat	0.080293	0.105851	0.759	0.44813
GasRegular	-0.067024	0.033197	-2.019	0.04349*
RegionBasse-Normandie	0.221542	0.130336	1.700	0.08917
RegionBretagne	-0.035701	0.088300	-0.404	0.68598
RegionCentre	-0.135642	0.078711	-1.723	0.08484
RegionHaute-Normandie	-0.770758	0.157391	-4.897	9.73×10^{-7}*
RegionIle-de-France	0.180502	0.084485	2.136	0.03264*
RegionLimousin	0.361364	0.164540	2.196	0.02808*
RegionNord-Pas-de-Calais	0.042425	0.106849	0.397	0.69133
RegionPays-de-la-Loire	0.068817	0.090006	0.765	0.44452
RegionPoitou-Charentes	-0.024256	0.100699	-0.241	0.80965
Log(theta)	10.935177			
Zero-inflation model coefficients (binomial with logit link)				
(Intercept)	1.2967303	0.2632852	4.925	8.43×10^{-7}*
CarAge	-0.1584852	0.0107263	-14.775	2×10^{-16}*
DriverAge	-0.0072697	0.0020544	-3.539	0.000402*
BrandJapanese (except Nissan) or Korean	1.3161513	0.2262549	5.817	5.99×10^{-9}*
BrandMercedes, Chrysler or BMW	0.6638740	0.2517029	2.638	0.008351*
BrandOpel, General Motors or Ford	-0.0001128	0.2396516	0.000	0.999624
Brandother	-0.5352438	0.3983307	-1.344	0.179040
BrandRenault, Nissan or Citroen	0.1609387	0.2110666	0.763	0.445761
BrandVolkswagen, Audi, Skoda or Seat	0.0737443	0.2423743	0.304	0.760931
GasRegular	-0.0081975	0.0664724	-0.123	0.901852
RegionBasse-Normandie	0.1483699	0.2238359	0.663	0.507426
RegionBretagne	-0.4647977	0.1668124	-2.786	0.005331*
RegionCentre	-0.4849913	0.1435967	-3.377	0.000732*
RegionHaute-Normandie	-0.9555450	0.3469106	-2.754	0.005879*
RegionIle-de-France	-0.0905313	0.1393025	-0.650	0.515764
RegionLimousin	0.1690170	0.2749440	0.615	0.538731
RegionNord-Pas-de-Calais	0.0495982	0.1799685	0.276	0.782860
RegionPays-de-la-Loire	-0.0893112	0.1621235	-0.551	0.581713
RegionPoitou-Charentes	-0.3970740	0.1947344	-2.039	0.041445*
*** means significant at $\alpha = 0.05$**				

Table 4.6: Score Test for Zero Inflation

Chi-square	df	p-value
770.51221	1	$< 2.22 \times 10^{-16}$

Table 4.7: The Vuong Test Showing The z-values, Models and p-values

	Vuong z-statistic	Models	p-value
Vuong Test For Poisson vs ZIP			
Raw	-11.44031	ZIP > Poisson	2.22×10^{-16}
AIC-corrected	-11.24373	ZIP > Poisson	2.22×10^{-16}
BIC-corrected	-10.16925	ZIP > Poisson	2.22×10^{-16}
Vuong Test For NB vs ZINB			
Raw	-5.854620	ZINB > NB	2.3905×10^{-9}
AIC-corrected	-5.583748	ZINB > NB	1.1769×10^{-8}
BIC-corrected	-4.103216	ZINB > NB	2.0372×10^{-5}

4.5.3 AIC and BIC

We use the AIC and BIC to test which of the model (Poisson, ZIP and ZINB) performed better in fitting the claim counts data. The AIC and BIC methods require the maximization of the log-likelihood. The smaller the AIC and BIC, the better the fit. Even though the AIC can be calculated for models not fit by maximum likelihood, it is normally used for comparing models fitted by maximum likelihood to the same data. The AIC and BIC for the count models was computed; Poisson, ZIP and ZINB as seen in table 4.8. It was found out that, the model that fit the data the most is Zero Inflated Poisson Model. This is because, its AIC or BIC was the smallest among the three candidate models. One should take note of the fact that, the ZINB was closest to the ZIP. The closest in the fit of both ZIP and ZINB model can be confirmed in their predicting abilities of zeros in the data. The *predict* function in R was used to predict the zeros generated by Poisson, ZIP and ZINB models to see which of them comes close to the 96.27513% observed zeros in the data. Table 4.9 summarizes this results.

The model with the closest predicted zeros to the observed zero is ZINB. This is however, the same as ZIP if the predicted values are corrected to 3 significant figures (i.e. both model will predict 96.3% of zeros, same as the observed zeros). It is worth noting that, the Ordinary Poisson did not do a good job in predicting

Table 4.8: AIC and BIC Values For Poisson, ZIP and ZINB

Models	df	AIC	BIC
Poisson	31	137665.8	138004.7
ZIP	38	136865.1*	137280.5*
ZINB	39	136867.1	137293.4
*** means the best model**			

Table 4.9: Observed Zeros, Predicted Zeros for the Poisson, ZIP and ZINB models

Observed(%)	Poisson (%)	ZIP(%)	ZINB(%)
96.27513	96.16257	96.2551	96.27465*
*** means the closest to the % observed zeros**			

the zeros in the data. Since the Zero Inflated Poisson model was found to be the best performer among the candidate count models. The ZIP will be used together with the claims amount model in calculating the actuarial premiums of the insurer.

4.6 Distribution and Fitting of Claims Amount

4.6.1 Distribution of Claim Amount

From the claim frequency data, it can be seen that claims rarely occur but from the claims amount distribution seen in figure 4.2, one can see that, the claim amount is skewed to the right. This means that, there are large claims when claims occured. As stated in Arku et al (2020), claims do not occur "frequently" but when they do, the severity is very high. From table 4.10, one can conclude that the claims amount data is asymmetrically distributed. The positive skewness value of 13.65 demonstrates that the data is right skewed and therefore have a right tail. The high kurtosis value of 215.77 indicates that the data have a longer tail. The other summary statistics that demonstrate the asymmetry and non-nomality of the data is the distance between the mean and the median. These confirm the choice of the Gamma distribution, hence the Gamma model in fitting the data. The histogram in figure 4.2 illustrates the same reasoning of how righly

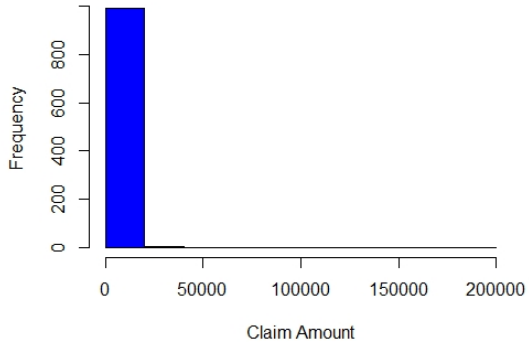

Figure 4.2: Bar Chart Showing The Distribution of Claims Amount

skewed the claim amount data is.

Table 4.10: Summary Statistics of The Claims Amount

mean	Std.Dev	Median	Minimum	Maximum	Range	Skew	Kurtosis	SE
2121.18	9122.15	1114.5	5	182467	182462	13.65	215.77	288.47

4.6.2 Fitting the Gamma Regression Model

After, the ZIP model was chosen to be the best fit for modelling the insurance claim frequency, attention is now turned to estimating the claims cost using the Gamma model. As noted by Fathia and Purwono (2019), the Gamma model is normally used for large insurance claims data. The Gamma model was fitted to the claims amount data and obtained the results in table 4.11. Although, the traditional link function as proven under the methodology for Gamma regression model is the inverse of the mean, we used the log link function due to the limitation of the inverse link function in taking 0 values and assuming negative values, and hence restricting the values of the regression coefficient. From the table, it can be seen that, the insurance claims amount is influenced by only the variables under Power. Power f, h and k all were found to influence the claims cost. The positive effect of these factors on claim amount is revealing because the power of a vehicle or the size of a vehicle can determine its speed and hence if an accident occurs, the gravity will be very high. These factors are different from the factors that influence the claim frequency under the ZIP model. This confirms the earlier

assertion of separately modelling claim frequency and claim amount that:

much of the power of rating factors is related to claim frequency (in this case 7 risk factors for claim frequency as against 3 for claims amount), and the factors can be affected with greater accuracy.

4.6.3 Assessing The Fit of The Gamma Model

By analyzing the null deviance and the residual deviance obtained from fitting the Gamma Model, one can conclude that the Gamma model provides an adequate fit for the claims amount data. The null deviance and its degree of freedom are all greater than the residual deviance of the Gamma model by 363.7 and 30 respectively.This explains that, the fitted model (Gamma model) is an improvement in predicting the response variable (Claims amount) without predictors. An assessment of the p-value (of 1) for the scaled deviance also emphasizes that, the Gamma model has a significant support in modelling the claims cost. This is because the null hypothesis of the gamma model providing a good fit could not be rejected at a significant level of $\alpha = 0.05$. This means the Gamma model shows no lack of fit (see table 4.11)

4.7 Premium Calculation Model

The pure premium is calculated by taking a product of the estimated claim frequency and estimated claims cost. This follows the assumption that, the claim frequency and claims cost are influenced by different risk factors. This can be seen from the log link part of the ZIP model (for claim frequency) and the Gamma model (for claims cost). Assuming, the estimated claim frequency is $\mathbb{E}(Y)$ and estimated claims cost is $\mathbb{E}(C)$, then the Pure premium, $\mathbb{E}(S)$, can be calculated

Table 4.11: The Gamma Model Showing Regressors, Estimates, Std. Error, z-value and Pr(>|z|) at $\alpha = 0.05$

Regressors	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.0862	0.7434	8.19	8.4×10^{-16}*
Powere	0.2641	0.2769	0.95	0.3404
Powerf	0.6688	0.2643	2.53	0.0116*
Powerg	0.1813	0.2605	0.70	0.4866
Powerh	0.8574	0.3992	2.15	0.0320*
Poweri	-0.1834	0.4318	-0.42	0.6712
Powerj	0.7726	0.4273	1.81	0.0709
Powerk	2.1023	0.6892	3.05	0.0023*
Powerl	0.9622	0.7963	1.21	0.2272
Powerm	-0.1040	1.0764	-0.10	0.9231
Powern	-0.9630	1.4480	-0.67	0.5062
Powero	0.0380	1.7617	0.02	0.9828
CarAge	-0.0327	0.0171	-1.92	0.0556
DriverAge	0.0084	0.0052	1.62	0.1056
BrandJapanese (except Nissan) or Korean	0.0716	0.4702	0.15	0.8790
BrandMercedes, Chrysler or BMW	0.5681	0.5306	1.07	0.2846
BrandOpel, General Motors or Ford	0.3220	0.4082	0.79	0.4304
Brandother	0.1569	0.5732	0.27	0.7843
BrandRenault, Nissan or Citroen	0.2001	0.3474	0.58	0.5647
BrandVolkswagen, Audi, Skoda or Seat	-0.2104	0.4126	-0.51	0.6102
GasRegular	0.3327	0.1716	1.94	0.0529
RegionBasse-Normandie	-0.0364	0.7185	-0.05	0.9596
RegionBretagne	0.6552	0.6242	1.05	0.2942
RegionCentre	0.7179	0.6000	1.20	0.2318
RegionHaute-Normandie	0.3224	1.0963	0.29	0.7687
RegionIle-de-France	1.0340	0.6483	1.59	0.1111
RegionLimousin	0.1903	0.7959	0.24	0.8111
RegionNord-Pas-de-Calais	-0.3359	1.1574	-0.29	0.7717
RegionPays-de-la-Loire	0.3753	0.6351	0.59	0.5546
RegionPoitou-Charentes	0.2486	0.6665	0.37	0.7093
Density	-0.3813×10^{-5}	1.991×10^{-5}	-1.91	0.0558

Results of Deviance

Model	Deviance	df	p-value
Null	1803.6	999	-
Fitted	1439.9	969	-
Scaled Deviance	243.3128*	999	1

* means significant at $\alpha = 0.05$

Table 4.12: Computation of The Pure Premium

Regressors	Estimate (ZIP)	Estimate (Gamma)	Actuarial Premium
Intercept	-1.868109	6.0862	4.2181
CarAge	-0.065546	-	-0.0655
DriverAge	-0.004307	-	-0.0043
BrandMercedes, Chrysler or BMW	0.318523	-	0.3185
GasRegular	-0.067026	-	-0.0670
RegionHaute-Normandie	-0.770898	-	-0.7709
RegionIlle-de-France	0.180533	-	-0.1805
RegionLimousin	0.361385	-	0.3614
Powerf	-	0.2641	0.2641
Powerh	-	0.6688	0.6688
Powerk	-	0.8574	0.8574

through the log link as;

$$\begin{aligned}\mathbb{E}(S) &= \mathbb{E}(Y) \times \mathbb{E}(C) \\ &= \exp(X_y^T \beta_y) \times \exp(X_c^T \beta_c) \\ &= \exp\{X_y^T \beta_y + X_c^T \beta_c\} \end{aligned} \tag{4.1}$$

Where X_y and β_y are vectors of risk factors and coefficients respectively from the claim frequency model (ZIP model), and X_c and β_c are vectors of risk factors and coefficients from the claims cost model (gamma model).

Empirically, the pure premium is obtained by putting together the log part of the ZIP model (in table 4.4)and the gamma model (in table 4.11) using only the significant variables. The results of the premium model is illustrated in table 4.12

4.8 Findings

The claim count, was dominated by a preponderance 96.3% of zeros. These observed zeros was confirmed by the Vuong test where, the zero-inflated Poisson and the zero-inflated negative binomial models were superior to their ordinary models. That is the two-part models (ZIP and ZINB) did better than their one-part models (Poisson and Negative). This was confirmed by the score test

where the null hypothesis of the probability of not having excess zeros in the data was rejected at $\alpha = 0.05$ significant level. This was also confirmed by the large dispersion parameter ($\theta = 56116.1$) of ZINB model. This confirms the assertion by David (2014) that, the Poisson and NB models are not able to distinguish between the insured that report no claim due to no accident or event occurring or insureds that do not report any claim due to the fear of being punished by the bonus-malus system. She stated also that the zero values and those strictly positive cannot be modelled by the same process. Although the ZIP and the ZINB through the statistical analyses were found to have a close AIC and BIC, the ZIP have the urge over both the ordinary Poisson and the ZINB. The ZIP model was therefore the preferred model for the claim frequency data. Based on the ZIP model, the car's age, the driver's age, the Japanese or Korean brand, the Mercedes, Chrysler or BMW brand were all found as significant risk factors in modelling the claim count. The regions of Haute-Normandie, Limousin Ill-de-France Bretagne, Centre and Poiton-Charentes were also found to influence the frequency of claims. The regular type of fuel was also found to significantly influence the occurrence of claims. The analysis of the significant factors of the ZIP indicated that, some of the risk factors associated with the count model is different from those from the binary model. Bretagne, Centre and Poiton-Charentes were the different factors found to predict the zeros apart from those that influence the count model. This agrees with David (2014) that the count model and binary models could be explained by different risk factors. From the findings, it is worth-noting that claim frequency decreases with both car's age and driver's age. That means, any increase in these variables will cause a decrease in the number of claims reported. The Gamma regression model was used in analysing the claims size. From the analysis, it was demonstrated that, three variables under the variable power were found as having significant influence on claims amount. These variables were all different from those that affects the claim count. Also, the contributing factors of claim frequency are more than that of

claim amount. This result concord with what Ohlsson and Johansson (2010) put forward that, it is of importance to model these two separately since often, much power of rating factors are associated to claim frequency.

Part II

Analysis And Results Obtained Using The Ghanaian Insurance Data

4.9 Description And Composition of Data

The data used in this section was obtained from a major non-life insurance company in Ghana. It covered the period 2013-2016. The two dependent variables: claim counts and claims amount are related to a motor insurance policy. Due to the dichotomous nature of the data obtained from this instiution, (that is the claim frequency was recorded as either zero or one) we regrouped the data based on the vehicle age. After this regrouping, a total of 3394 observations were used for the analysis. The variables used include Insurance Type (Comprehensive, Third Party), Usage Type (Private Individual, Private Corporate, Carriage, Others), Vehicle Age, Renewal.

4.10 Distribution of Claim Counts

In table 4.13 and figure 4.3 , the distribution of the claim frequency is presented. The number of claims registered by an insured ranges from 0 to 7. About 63.2% of the policyholders did not registered any claim. 34.1% of them registered 1 claim and 2.3% registered 2 claims. 0.27% registered 3 claims. Table 4.13 shows more details. The 63.2% of no claim means there exist excess zeros in the data.

Table 4.13: Distribution of Claim Count

Claim Number	Frequency	Percent
0	2146	63.2
1	1159	34.1
2	78	2.3
3	9	0.3
4	1	0.0
7	1	0.0
Total	3394	100.0

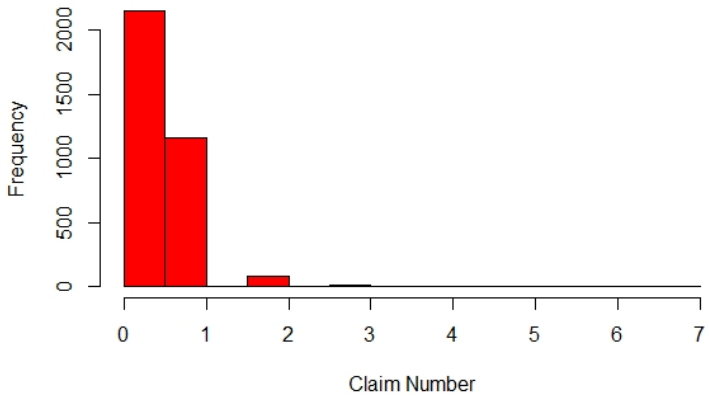

Figure 4.3: A Histogram Showing The Distribution of Claim Count

4.11 Regression Models for the Claim Count

The Poisson Model, the ZIP model and the ZINB were fitted to the Ghanaian claim count data as fitted to the French claim count data. All assumptions of these models remains the same as used in Part 1

4.11.1 The Poisson Model

Looking at the histogram, one concludes that, the data follows a Poisson distribution. The poisson regression model was therefore fitted to the data as a starting point. The result of the Poisson model can be seen in table 4.14.

Table 4.14: Poisson Regression Model at a Significant level of 5%

Regressors	Estimate	Std. Error	z value	$\Pr(>\|z\|)$
(Intercept)	0.168395	0.067654	2.489	0.01281*
Insurance.TypeTHIRD PARTY	-0.307405	0.068414	-4.493	7.01×10^{-6}*
Usage.TypeOthers	-0.242885	0.074356	-3.267	0.00109*
Usage.TypePrivate Cars (Corporate)	-0.059085	0.072549	-0.814	0.41541
Usage.TypePrivate Cars (Individual)	-0.225619	0.094621	-2.384	0.01711*
Veh.Age	-0.008118	0.003478	-2.334	0.01958*
Renewal.mean	-4.155646	0.401213	-10.358	2×10^{-16}*
* means significant at $\alpha = 0.05$				

The results of the Poisson regression model in table 4.14 reveal that, all the variables are significant in determining the frequency of claims except the private

corporate cars. It should be noted that, all these variables have a decreasing effect on claim frequency since the coefficient of these variables are all negative. For instance when the age of the vehicle insured increases, the claim frequency decreases.

4.11.2 The ZIP Model

As stated in Part 1 in the analysis of the ZIP model, the poisson model becomes mostly incapacitated in dealing with excess zeros in the data. The ZIP model is adopted to deal with this deficiency of the Poisson model. Table 4.15 shows the result of the ZIP model. With the data used here, one can see that, the ZIP model

Table 4.15: ZIP Regression Model at a Significant level of 5%

Regressors	Estimate	Std. Error	z value	$\Pr(>\|z\|)$
Count model coefficients (poisson with log link)				
(Intercept)	-0.0036137	0.0739733	-0.049	0.961038
Insurance.TypeTHIRD PARTY	-0.2238017	0.0679809	-3.292	0.000994*
Usage.TypeOthers	-0.2017794	0.0744391	-2.711	0.006715*
Usage.TypePrivate Cars (Corporate)	0.0807769	0.0760476	1.062	0.288150
Usage.TypePrivate Cars (Individual)	-0.1058548	0.0962730	-1.100	0.271538
Veh.Age	0.0009267	0.0037075	0.250	0.802619
Renewal.mean	-3.8285636	0.4047842	-9.458	2×10^{-16}*
Zero-inflation model coefficients (binomial with logit link)				
(Intercept)	-272.766	130.514	-2.090	0.0366*
Insurance.TypeTHIRD PARTY	47.881	22.968	2.085	0.0371*
Usage.TypeOthers	84.111	42.868	1.962	0.0498*
Usage.TypePrivate Cars (Corporate)	141.170	68.217	2.069	0.0385*
Usage.TypePrivate Cars (Individual)	109.261	52.718	2.073	0.0382*
Veh.Age	4.536	2.192	2.069	0.0385*
Renewal.mean	44.735	22.413	1.996	0.0459*
* means significant at $\alpha = 0.05$				

shows that, the third party insurance type, the other usage of vehicles (apart from corporate and individuals), and the the number of renewals were found to significantly determine the occurence of claims. All these three significant variables also have a decreasing effect on the occurence of claims. Worth noting is that, all the variables are seen to be predictors of the zeros in the data.

4.11.3 The ZINB Model

Where there are overdispersions in the data, the Poisson and the ZIP model become weak. The ZINB model performs better in this circumstance most at times. The ZINB was fitted to the data and the output was analysed. The

Table 4.16: ZINB Regression Model at a Significant level of 5%

Regressors	Estimate	Std. Error	z value	$Pr(>\|z\|)$
Count model coefficients (negbin with log link)				
(Intercept)	-0.007026	0.075516	-0.093	0.925872
Insurance.TypeTHIRD PARTY	-0.242161	0.069435	-3.488	0.000487*
Usage.TypeOthers	-0.173768	0.075200	-2.311	0.020847*
Usage.TypePrivate Cars (Corporate)	0.097258	0.077200	1.260	0.207732
Usage.TypePrivate Cars (Individual)	-0.127994	0.096076	-1.332	0.182787
Veh.Age	0.001303	0.003848	0.338	0.735007
Renewal.mean	0.179324	0.203814	0.880	0.378946
Log(theta)	5.750806	0.791291	7.268	3.66×10^{-13}*
Zero-inflation model coefficients (binomial with logit link)				
(Intercept)	-30.8750	9.2260	-3.347	0.000818*
Insurance.TypeTHIRD PARTY	1.9550	1.0301	1.898	0.057726.
Usage.TypeOthers	9.5246	3.7421	2.545	0.010921*
Usage.TypePrivate Cars (Corporate)	17.9766	6.1149	2.940	0.003284*
Usage.TypePrivate Cars (Individual)	9.1122	3.2345	2.817	0.004844*
Veh.Age	0.5909	0.1803	3.277	0.001049*
Renewal.mean	31.1774	8.3873	3.717	0.000201*
* means significant at $\alpha = 0.05$				

output in table 4.16 shows from the ZINB model that, third party insurance type, other usage of cars are the only variables significant in determining the probability of occurence of claims. All the other variables apart from the third party insurance type were all found to predict the excess zeros in the data.

Table 4.17: Score Test For Zero Inflation at $\alpha = 5\%$

Chi-square	df	p-value
125.28262	1	2.22×10^{-16}

The score test result suggests that, there is a zero-inflation in the data since the p-value of less than 2.22×10^{-16} leads to the rejection of the null hypothesis of no zero-inflation. This is an indication that, a zero-inflated model could be considered apart from the standard Poisson model

4.12 Model Comparison For The Counts Data

4.12.1 The Score Test

Following the outline as used in Part I, we first of all determine whether or not there is presence of excess zeros in the data. The score test tests the null hypothesis that the probability of having zero in the data is zero. Table 4.17 shows the result of the score test.

4.12.2 The Vuong Test

The Vuong Test was also used here to compare each zero-inflated model against its ordinary model. The Vuong test indicates how a two-part model does against its standard model. That is, the ZIP was compared to the Poisson and also the ZINB was compared to the negative binomial. The result of the Vuong test is detailed in table 4.18

Table 4.18: The Vuong Test Showing The z-values, Models and p-values at α=5%

	Vuong z-statistic	Models	p-value
Vuong Test For Poisson vs ZIP			
Raw	-7.786128	ZIP> Poisson	3.4547×10^{-15}
AIC-corrected	-6.372696	ZIP>Poisson	9.2867×10^{-11}
BIC-corrected	-2.040695	ZIP >Poisson	0.020641
Vuong Test For NB vs ZINB			
Raw	-13.18320	ZINB >NB	2.22×10^{-16}
AIC-corrected	-13.04049	ZINB >NB	2.22×10^{-16}
BIC-corrected	-12.60309	ZINB >NB	2.22×10^{-16}

The Vuong test results in table 4.18 show that, each of the zero-inflated models were superior to their ordinary models. That is, The ZIP did better than the standard Poisson. Also the ZINB did better than the ordinary negative binomial. This conclusion is valid because the p-values were all less than the significant level of 5% suggesting that the null hypothesis of "no difference between a zero

inflated model and its ordinary form" is rejected.

4.12.3 AIC and BIC

The AIC and the BIC values for the three main models were obtained and compared to evaluate which model best fit the claim counts data. All assumptions of the AIC and BIC as stated in Part I still hold. The details are shown in table 4.19

Table 4.19: AIC And BIC Values for Poisson, ZIP and ZINB

Models	df	AIC	BIC
Poisson	7	3215.8	3258.7
ZIP	14	3152.6*	3238.5*
ZINB	15	3265.7	3357.7
* means the best model			

Both AIC and BIC support the Zero-inflated Poisson model over the Poisson and the Zero-inflated negative binomial models since the ZIP has the smallest BIC and AIC values (which suggests smallest information loss). Eventhough the ZIP model was selected at the expense of the other two models it did poorly in predicting the zeros in the data. The ZINB did well in predicting the excess zeros in the data. Table 4.20 shows how each of these three models does in predicting the excess zeros.

Table 4.20: Observed and Predicted Zeros for Poisson, ZIP, and ZINB

Observed (%)	Poisson(%)	ZIP(%)	ZINB(%)
63.2	73.8	74.0	69.6*
* means the closest to the % of observed zeros			

The results from table 4.19 shows that, the ZINB did better than the other two models in predicting the zeros in the data. However, this is not encouraging because, the gap between the observed zeros and what is predicted by the ZINB model is not too close (a difference of 6.4%). This raises a question mark on not only the ZINB models but the models and the data used in the research. We shall talk about this in the ovarall conclusion and suggest the way forward.

Next, gamma model was fitted to the amount of claims paid.

4.13 Fitting the Gamma Model

We now fit the gamma model to the claims cost in other to estimate the average claims cost to be used in the calculation of the actuarial premium. The claims amount consist of claims paid out of 3394 policyholders. 1248 claims were paid. As stated in Part I of the fitting of the gamma model, we adopt the log link function instead of the actual inverse function due to its limitation in the modelling of the claim amount. Table 4.21 shows the results of the gammma model.

Table 4.21: The Results of The Gamma Model at $\alpha = 5\%$

Regressors	Estimate	Std. Error	t-value	Pr($>\|t\|$)
(Intercept)	10.248766	0.121838	84.118	2×10^{-16}*
Insurance.TypeTHIRD PARTY	-0.540702	0.115469	-4.683	3.14×10^{-6}*
Usage.TypeOthers	0.330450	0.123404	2.678	0.00751*
Usage.TypePrivate Cars (Corporate)	-0.509121	0.126110	-4.037	5.74×10^{-5}*
Usage.TypePrivate Cars (Individual)	-0.849905	0.155144	-5.478	5.20×10^{-8}*
Veh.Age	-0.049517	0.006303	-7.856	8.55×10^{-15}*
Renewal.mean	-0.559053	0.372841	-1.499	0.13401
AIC = 25625 Residual deviance = 2168.3 scaled deviance = 873.0908*				
* means significant at $\alpha = 0.05$				

The results of the gamma model suggests that for the motor insurance data, all the variables are significant in determining claims cost except the number of times renewals are made to a policy. The results also demonstrate that, apart from other usages of vehicle, the rest of the variables have a decreasing effect on claims amount. That is if the number of those using vehicles for other purposes increases, the claim amount will also increase. Five(5) significant risk factors were identified as determinants of claims cost. The scaled deviance which is obtained by dividing the residual deviance (also called deviance) by the dispersion parameter (in this case 2.483476) was found to be significant. This implies the gamma model fit the data well. This can be confirmed by the histogram in figure 4.4 and the positive

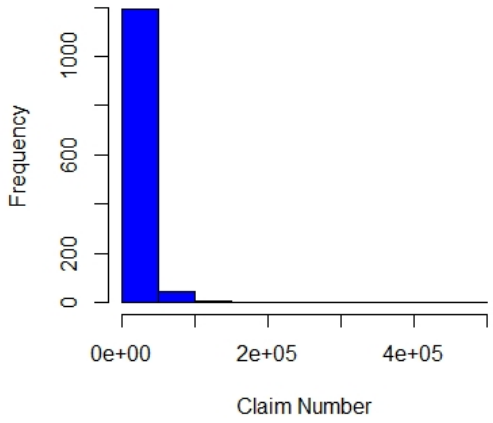

Figure 4.4: Bar Chart Showing The Distribution of Claims Amount

Table 4.22: Summary Statistics For Claims Amount

mean	sd	median	min	max	range	skew	kurtosis	se
12628.06	27770.42	4644.88	100	483672.6	483572.6	9.46	138.08	786.1

skewness and kurtosis in table 4.22 which show that, the claim amount data is non-gaussian. And the gamma distribution is a possible choice for such a data as explained in Part 1

4.14 The Premium Model

We adopt the same premium model used in Part I in this part as well. The significant factors in the claim counts model (in table 4.15) and the claims cost model (in table 4.21) were combined through the log link function to calculate the premiums. Table 4.23 shows the actuarial premium calculation.

4.15 Findings

4.15.1 The Claim Counts

There were excess zeros of 62.3% in the claims data. The presence of the excess zeros were confirmed by the score test. The score test reveals that, at a significant

Table 4.23: Computation of The Actuarial Premium

Regressors	Estimate (ZIP)	Estimate (Gamma)	Actuarial Premium
Intercept	-	10.248766	10.2488
Insurance.TypeTHIRD PARTY	-0.2238017	-0.540702	-0.7645
Usage.TypeOthers	-0.2017794	0.330450	0.1287
Usage.TypePrivate Cars (Corporate)	-	0.509121	-0.5091
Usage.TypePrivate Cars (Individual)	-	0.849905	-0.8499
Veh.Age	-	-0.049517	-0.0495
Renewal.mean	-3.8285636	-	-3.8286

level of 0.05, there null hypothesis of not having excess zeros present in the data was rejected indicating that there was a predominance of zeros in the data. The Vuong Test also proves that, the zero inflated models did better than their counterpart which suggest that there are excess zeros in the data. The AIC and BIC values give support to the ZIP model at the detriment of the other two models (Poisson and ZINB). However, the amount of zeros predicted by the selected model (ZIP) was farther away from the observed zeros than what was predicted by the ZINB model. Reasons for this contrary results will be given in section 5. Bassed on the AIC and BIC values, the ZIP model was selected and used for the calculation of the acturial premiums. It was found out that, the Third party insurance type, the other usage type of vehicles, and the number of times a policy is renewed were significant risk factors in determining the claim frequency. All these significant factors have a decreasing effect on the claim frequency. The zeros present in the data were found to be contributed to by all the factors (including, third party policy, other usage types, corporate cars, individual cars, vehicle age, and number of renewals).

4.15.2 Claims cost

The claims cost was assumed to follow a gamma distribution and this was proven right by the significance of the scaled deviance of the gamma model. The results

of the Gamma model show that, the third party policy, the other usage type, the corporate cars, individual cars, and vehicle age are all critical risk factors in determining the average amount of claims paid in an insurance period. It is should be noted that, out these risk factors only other usage type has a positive effect on the amount of claims. That is, if the number of policyholders using vehicles for other purposes increases, the amount of claim paid will also increase.

4.15.3 The Actuarial Premium

Combining the frequency of claims and the cost of claims, it is seen that the contributing risk factors of claim frequency were less for this data than that of claim amount. The number of times a policy is renewed affected claim frequency but did not affect claim amount. Similarly, the, vehicle age and private corporate and individual cars affected claims amount but not claim frequency. This confirms the reason for separate analysis of the claim frequency and claim amount since they could have different risk factors. However the number of risk factors associated with claim amount was more than the claim frequency, which is contrary to what is stated in Ohlsson and Johansson (2010) that most at times more rating is given to the claim frequency. Upon modelling the actuarial premiums, these risk factor will become a default homogenous groups where new policyholders will be registered and their premiums charged based on this already defined clases.

Chapter 5

SUMMARY, CONCLUSION AND RECOMMENDATION

5.1 Summary of Major Findings

In this section, the results and findings obtained using the French insurance data and the local data are summarized

5.1.1 The French Motor Liability Data

There was zero-inflation in the claim frequency data causing overdispersing and therefore incapacitating the ordinary Poisson model in modelling the data. The 96.3% of excess zeros present in the claim frequency data was confirmed by the Vuong test statistics, the score test and the dispersion parameter ($\theta = 56116.1$) of the ZINB model. In identifying the risk factors of claim frequency, the standard Poisson model, the Zero-Inflated Poisson model and the Zero-Inflated Negative Binomial model were considered. After which model specification and comparison methods were used in selecting the suitable models for the claim count data. The score test demonstrated the inability of the Poisson model in appropriately modelling the data due to the inflation of zeros in the data. The zero inflated models of Poisson and Negative Binomial were all found to be superior to their ordinary models as demonstrated by the Vuong test statistics which suggested that the claim frequency data contains excess zeros. To choose the better model among the three models under consideration (Poisson, ZIP and ZINB), the AIC

and the BIC results all agrees that, the ZIP model is suitable for analysing the claim frequency data since these values are the smallest for the ZIP model. The appropriate model in ZIP showed that about 11 factors are responsible for the claim count. By fitting the gamma model to the claims amount, It was realised that, the factors that influence the claims amount were different from that of the claim frequency.

5.1.2 Ghanaian Non-life Insurance Data

The Ghanaian non-life insurance data was also inflated with 63.2% of zeros. This presence of zeros weakens the ability of the Poisson model in modelling such a data. The Score test proves that, there was an excess zeros and the Vuong test also demonstrated the inability of the standard poisson and the negative binomial model in modelling the frequency of the insurance data. However, their inflated models, respectively, ZIP, and ZINB were found to perform better than their ordinary counterparts. In choosing among the three main candidates models, the AIC and the BIC criteria was used. It was found out that the AIC and the BIC values all gave support to the ZIP model over the ordinary Poisson model and the ZINB. However it should be noted that, the selected ZIP model did poorly in predicting the zeros in the data. The ZINB model did better in that regard. The ZIP model was selected as the best model in modelling the claims data. Out of the risk variables, three of them influence claims frequency. They include, the third party policy, the other usage of cars and the number of times a policy is renewed. That is three factors were found as determinants of claim frequency one of which is different from the factors affecting the claims cost. The results of the Gamma model show that, five(5) factors were found as determinants of claims cost three of which are different from the risk factors associated with the claim frequency.

5.2 limitation of Data and Comparison of Results

Modelling of claim frequency and claim amount is a significant aspect of actuarial science and insurance. Claim count modelling and claim amount modelling are important in determining the various risk factors that can be considered in computing an actuarial premium of an insurance portfolio. In this thesis, two sets of data were used; one from a French Motor ThirdParty Liability Policy (FMTL) and the other from a major non-life Insurance Company in Ghana (GMID). Even though the number of observations and the settings of these two datasets are different results were compared carefully in the following manner and some salient conclusion were drawn from it.

5.2.1 Observed Zeros and Predicted Zeros

The excess zeros present in the French ThirdParty Liability Policy is about 96% and about 63% of the Ghanaian data was dominated by excess zeros. It should however be noted that the number of observations in the FMTL is about 413,169 while that from the GMID is about 3,394. The best model selected in modelling both the FMTL and the GMID is the Zero-inflated Poisson Model. In the prediction of the zeros in the FMTL, both the ZIP and the ZINB did extremely well, such that by rounding to one decimal place they both predicted the zeros accurately. Contrary with the Ghanaian Motor Insurance Data (GMID), the selected ZIP model came second in predicting the zeros in the data. What is striking to note is, none of these models got so close to the 63% of the observed zeros. Only the ZINB estimated 69% of excess zeros which is still farther from the observed zeros. This sends the first red flag, suggesting that the Ghanaian data might not support the models (conclusion on this was made after the comparison)

5.2.2 The Vuong Test

The Vuong test to validate whether or not the ordinary models and their two-part counterpart are indistinguishable for both the FMTL and the GMID were 'similar'. That is, for the FMTL, the zero-inflated models, ZIP and ZINB did better than their ordinary models Poisson and negative binomial respectively. This also holds true for the GMID.

5.2.3 The Score Test

The score test, validates whether or not, there exist a presence of zeros in the data or not. In both FMTL and GMID, the results of the score test show that there was excess zeros. That is the probability of not having zeros in the the two datasets were different from zeros. Inferentially, the Poisson model will be weak in modelling the count data

5.2.4 AIC and BIC Values

The AIC and BIC values of both the FMTL and GMID all gave support to the ZIP model. However, for the FMTL, the support was not substantial for the AIC value. The AIC values of the ZIP was smaller than the ZINB by 2 units. This confirms the reason why, these two models (ZIP and ZINB) did marvelously well in predicting the excess zeros. As a matter of fact, in considering the predicted zeros to 2 decimal places or more, the ZINB predicts better than ZIP. With the GMID, there was a substantial support for the ZIP model over both ZINB and the standard Poisson model. However, under this model, the Poisson came second after the ZIP. This could be due to the fact that, the excess zeros were not too high as compared to the FMTL data. (see next section for more details)

5.2.5 Stand Error of Estimates

The standard error of the estimate measure how far the predicted values are from the actual values (or visually, the regression line). One thing encouraging for the FMTL data is that the standard errors of all three models are very small wich suggest a good predictive capacity of these models. Shockingly for the chosen model (the ZIP) for the GMID, the standard error for the zero-inflation is quite large relative to the regression coefficient. That is most of it are almost half of their respective coefficients. This again is another red flag for the ZIP model for the Ghanaian data. However, for the GMID, the standard error for the standard Poisson is the smallest, followed by the ZINB. This is revealing because, one can see from the predicted zeros that, the ZINB did better than the ZIP, no wonder, the standard error of the zero-inflation part of the ZINB was far better (smaller) than the ZIP.

5.2.6 Zero Inflation, Overdispersion and Sample Size

When the overdispersion test proposed by Cameron and Trivedi (1990) was conducted on both the French Insurance data and the Ghanaian Insurance data, it was found out that, the French that shows overdispersion under the Poisson distribution but the Ghanaian data shows underdispersion. Table 5.1 illustrates that. The alpha value in the overdispersion test shows whether there is overdispersion or not. If there is overdispersion, the alpha value will be greater than 0 (meaning the variance exceeds the mean), and if the alpha value falls below zero, it suggests there is underdispersion (the variance is less than the mean). The alpha value of the French data under the Poisson distribution was positive (showing overdispersion) while the alpha value under the Poisson distribution for the Ghanaian data was negative (showing underdispersion). Worth noting however is, the overdispersion in the French data is very significant at 5% significance

level while the underdispersion in the Ghanaian data was statistically significant. The mean, the variance and the variance ratio of the claim counts data for the french data show overdispersion in the data. It can be clearly seen that, the variance of 0.04164 is greater than the mean of 0.3986. The opposite was true for the Ghanaian data, having its variance (0.3176) less than the mean (0.3986) displaying underdispersion. The underdispersion in the Ghanaian data could be

Table 5.1: Test of Overdispersion

z-value	p-value	alpha value
	Ghanaian Data	
-14.224	1	-0.2999461
	French Data	
16.06	$< 2.2 \times 10^{-16}$*	0.0674068
	* means significant at $\alpha = 0.05$	
	Mean, Variance and Mean-Variance Ratio	
Mean	Variance	Ratio
	French Data	
0.3916	0.04164	1.06382
	Ghanaian Data	
0.3986	0.3176	0.7967

a reason why the standard errors of the ZIP and the ZINB models were relatively large compared to their regression coefficients. The (insignificant) underdispersion might be the reason why the Poisson model has a smaller standard error than the other two models (ZIP and ZINB)relative to their regression coefficients.

5.2.7 General Conclusion

The inadequacies of using the Poisson model to a count data when zeros dominates the data was seen in both the French data and the Ghanaian data. Though the ZIP model was taken for both the French data and the Ghanaian data using BIC, care should be taken, just like any other model when using these models. This was seen when the models were applied to the Ghanaian data that, even though both the AIC and the BIC values fully supported the ZIP model, the standard errors of the zero-inflated part was larger than its counterparts (the

count part) modelled by the Poisson model. It can be concluded that, the choice of a model can be affected by the data.

Even though, both the Ghanaian data and the French data show zero-inflation, the Ghanaian data was under-dispersed. This could be a reason again, where although the ZIP is good for modelling zero-inflated data, one should check for overdispersion. Cambell (2021) found out the following which could be used in explaining the contradiction of goodness of fit criteria for the count data for the Ghanaian data.

- small sample size based on score tests and AIC, and BIC can lead to substantial inflation (type 1 error)
- sufficiently large sample size based on score tests is not problematic
- ignoring overdispersion and zero-inflation can lead to invalid inference

Hence zero-inflation might not always lead to overdispersion and there is a need therefore for both to be checked including the standard errors when considering appropriate models for the count data. The lack of fit based on the standard error for the ZIP might be an explanation to the underdispersion in the Ghanaian data though it is zero-inflated.

5.3 Contribution and Recommendation

Based on the purpose and findings of this study, it was demonstrated how actuarial modelling of claim count can be done with real data and how the dominance of zeros in the data can affect the distribution of the data and hence showed ways and methods to evaluate the models that were used. Since these models were applied to a real-world data, the results in this research can be used together with other literature regarding the issue of modelling claim counts. The

insurance practitioner can use the methods and results in determining risk factors that contribute to the occurrence of claim frequency. The insurance company is provided with the benefit of knowing that, the risk factors associated with claim frequency and claims cost are mostly different and therefore the need to consider modelling them separately. In the determination of the actuarial premium, the results show, in consonance with other researches that, greater rating power should be giving to claim frequency than claims cost. In conclusion the actuarial and econometric models used here provides a robust statistical information to the insurer in ensuring that, there is a harmony between the premium paid by an insured and the associated observable risks. If this methods and results are studied by the insurer and applied duly, it will lead to maintaining a profitable institution since an accurate modelling of claim frequency means an accurate premiums charged which can save the insurance company from bankruptcy. The following recommendations are made to the insurance practitioner and interested researchers.

i Claim count data should be analysed accurately so as to determine the appropriate model to consider in modelling it especially when the data is inflated with zeros.

ii Zero-inflation and overdispersion should both be checked when adopting zero-inflated models since Zero-inflated models make provision for overdispersion.

iii The insurance practitioner should consider modelling the claim count separately from the claims cost in the calculation of premiums since these two are affected mostly by different risk factors

iv To the researcher, a further research could look at competing models for the claims cost instead of assuming it follows a gamma distribution.

v A comparative analysis could be done between the GLM model used in calculation of premiums and the Tweedie GLM. As noted by Quijano and Garrido

(2014) when parsimony is a subject of attention, then the Tweedie GLM should be considered. This suggests Tweedie GLM is a simpler model and does not separately estimate the mean of the claim frequency and claims cost

vi Another point worth considering for further research is to include merit rate making features like previous accidents, conviction records, and other experience of the individual policyholder than risk factors common to a homogeneous group. further work should be done on posterior rate making or experience rate making as called in Antonio and Valdez (2012).

References

Adeti, F. (2016). Modelling Count Outcomes from Dental Caries in Adults: A Comparison of Competing Statistical Models (M.PHIL). Kwame Nkrumah University of Science and Technology.

Antonio, K., & Valdez, E. A. (2012). Statistical concepts of a priori and a posteriori risk classification in insurance. AStA Advances in Statistical Analysis, 96(2), 187-224.

Arku, D., Doku-Amponsah, K., & Howard, N. (2020). A Markov-modulated tree-based gradient boosting model for auto-insurance risk premium pricing. Risk And Decision Analysis, 8(1-2), 1-13. https://doi.org/10.3233/rda-180050

Bailey, R. A., & Simon, L. J. (1960). Two Studies in Automobile Insurance Ratemaking. ASTIN Bulletin, 1(4), 192-217. https://doi.org/10.1017/s0515036100009569

Bohning, D., Dietz, E., Schlattmann, P., Mendonca, L., & Kirchner, U. (2000). Corrigendum: The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. Journal of the Royal Statistical Society: Series A (Statistics in Society), 163(1), 121-122. https://doi.org/10.1111/1467-985x.00160

Boland, P. (2007). Statistical and probabilistic methods in actuarial science. Chapman & Hall/CRC.

Boucher, J. P., & Denuit, M. (2008). Credibility premiums for the zero-inflated Poisson model and new hunger for bonus interpretation. Insurance: Mathematics and Economics, 42(2), 727-735. https://doi.org/10.1016/j.insmatheco.2007.08.003

Boucher, J. P., Denuit, M., & Guillen, M. (2009). Number of Accidents or Num-

ber of Claims? An Approach with Zero-Inflated Poisson Models for Panel Data. Journal of Risk and Insurance, 76(4), 821-846. https://doi.org/10.1111/j.1539-6975.2009.01321.x

Burnham, K., & Anderson, D. (2004). Multimodel Inference. Sociological Methods & Research, 33(2), 261-304. https://doi.org/10.1177/0049124104268644

Cameron, C. A., & Trivedi, P. K. (2013). Regression Analysis of Count Data (Econometric Society Monographs) (2nd ed.). Cambridge University Press.

Campbell, H. (2021). The consequences of checking for zero-inflation and overdispersion in the analysis of count data. Methods In Ecology And Evolution, 12(4), 665-680. https://doi.org/10.1111/2041-210x.13559

David, M. (2014). Modeling The Frequency Of Claims In Auto Insurance With Application To A French Case. Review of Economic and Business Studies, (13), 69-85.

David, M. (2015). Auto Insurance Premium Calculation Using Generalized Linear Models. Procedia Economics and Finance, 20, 147-156. https://doi.org/10.1016/s2212-5671(15)00059-3

Denuit, M., Marechal, X., Pitrebois, S., & Walhin, J. F. (2007). Actuarial Modelling of Claim Counts. Wiley.

Desmarais, B., & Harden, J. (2013). Testing for Zero Inflation in Count Models: Bias Correction for the Vuong Test. The Stata Journal: Promoting Communications On Statistics And Stata, 13(4), 810-835. https://doi.org/10.1177/1536867x13013004

Dionne, G., & Vanasse, C. (1992).Automobile insurance ratemaking in the presence of asymmetrical information. Journal of Applied Econometrics, 7(2), 149-165. https://doi.org/10.1002/jae.3950070204

Dobson, A. J., & Barnett, A. G. (2018). An Introduction to Generalized Lin-

ear Models (Chapman & Hall/CRC Texts in Statistical Science) (4th ed.). Chapman and Hall/CRC.

Fathia, F., & Purwono, Y. (2019). Pure Premium Estimation Towards Zero Inflated Claim Data of Accident Insurance Through The Generalized Linear Model. International Journal Of Scientific & Engineering Research, 10(1).

Frees, E. W. (2009). Regression Modeling with Actuarial and Financial Applications (International Series on Actuarial Science) (1st ed.). Cambridge University Press.

Insurance. Oxford Business Group. (2021). Retrieved 4 July 2021, from https://www.oxfordbus 2020/insurance.

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. Casualty Actuarial E Forum. Published.

Jansakul, N., & Hinde, J. (2002). Score Tests for Zero-Inflated Poisson Models. Computational Statistics & Data Analysis, 40(1), 75-96. https://doi.org/10.1016/s0167-9473(01)00104-9

Lambert, D. (1992). Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. Technometrics, 34(1), 1. https://doi.org/10.2307/1269547

Lindsey, J. K. (1997). Applying Generalized Linear Models (Springer Texts in Statistics) (Corrected ed.). Springer.

Mikosch, T. (2009). Non-life insurance mathematics. Second. Universitext.

McCullagh, P., & Nelder, J. A. (1989). Generalized Linear Models (Chapman & Hall/CRC Monographs on Statistics and Applied Probability) (2nd ed.). Chapman and Hall/CRC.

National Insurance Commission. (2019). https://nicgh.org/wp-content/uploads/2020/10/2019-NIC-Annual-Report.pdf

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized Linear Models. Journal of the Royal Statistical Society. Series A (General), 135(3), 370. https://doi.org/10.2307/2344614

Ohlsson, E., & Johansson, B. (2010). Non-Life Insurance Pricing with Generalized Linear Models (EAA Series) (1st ed. 2010, Corr. 3rd printing 2014 ed.). Springer.

OECD. (2018). Global Insurance Market Trends 2019.

Penman, N., & Brager, G. (2021, January 29). NCSS Statistical Software Documentation — NCSS Software Help. NCSS. https://www.ncss.com/software/ncss/ncss-documentation/

Pinquet, J. (1997). Allowance for Cost of Claims in Bonus-Malus Systems. ASTIN Bulletin, 27(1), 33-57. https://doi.org/10.2143/ast.27.1.542066

Quijano Xacur, O. A., & Garrido, J. (2015). Generalised linear models for aggregate claims: to Tweedie or not? European Actuarial Journal, 5(1), 181-202. https://doi.org/10.1007/s13385-015-0108-5

Raftery, A. (1999). Bayes Factors and BIC. Sociological Methods & Research, 27(3), 411-427. https://doi.org/10.1177/0049124199027003005

Ridout, M., Hinde, J., & Demetrio, C. G. B. (2001). A Score Test for Testing a Zero-Inflated Poisson Regression Model Against Zero-Inflated Negative Binomial Alternatives. Biometrics, 57(1), 219-223. https://doi.org/10.1111/j.0006-341x.2001.00219.x

Sarul, L. S., & Sahin, S. (2015). An application of claim frequency data using zero inflated and hurdle models in general insurance. Journal of Business

Economics and Finance, 4(4).

Motor Insurance > Policies. Sic-gh.com. (2021). Retrieved 4 July 2021, from https://sic-gh.com/subcat_select.cfm?prodcatID=2&tblNewsCatID=16.

van den Broek, J. (1995). A Score Test for Zero Inflation in a Poisson Distribution. Biometrics, 51(2), 738. https://doi.org/10.2307/2532959

Verico, P. (2002). Bonus-Malus Systems: Lack of Transparency and Adequacy Measure. ASTIN Bulletin, 32(2), 315-318. https://doi.org/10.2143/ast.32.2.1032

Vuong, Q. H. (1989). Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. Econometrica, 57(2), 307. https://doi.org/10.2307/1912557

Wilson, P. (2015). The misuse of the Vuong test for non-nested models to test for zero-inflation. Economics Letters, 127, 51-53. https://doi.org/10.1016/j.econlet.2014.12.029

Wolny-Dominiak, A. (2013). Zero-inflated claim count modeling and testing–a case study. Ekonometria, (39), 144-151.

Xia, Y., Morrison-Beedy, D., Ma, J., Feng, C., Cross, W., & Tu, X. (2012). Modeling Count Outcomes from HIV Risk Reduction Interventions: A Comparison of Competing Statistical Models for Count Responses. AIDS Research and Treatment, 2012, 1-11. https://doi.org/10.1155/2012/593569

Xie, M., He, B., & Goh, T. (2001). Zero-inflated Poisson model in statistical process control.Computational Statistics & Data Analysis, 38(2), 191-201. https://doi.org/10.1016/s0167-9473(01)00033-0

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. https://doi.org/10.1016/j.insmatheco.2004.11.002

Zamani, H., & Ismail, N. (2013). Score test for testing zero-inflated Poisson re-

gression against zero-inflated generalized Poisson alternatives. Journal of Applied Statistics, 40(9), 2056–2068. https://doi.org/10.1080/02664763.2013.804904

Zeileis, A., Kleiber, C., & Jackman, S. (2008). Regression Models for Count Data in R. Journal of Statistical Software, 27(8). https://doi.org/10.18637/jss.v027.i08

Appendix A

Link to the French Motor Third Party Liability Data: `https://www.kaggle.com/datasets/karansarpal/fremtpl2-french-motor-tpl-insurance-claims`

A1: Proof of The scaled Deviance of The Gamma Model

Recall the definition of deviance:

$$D^*(y;\hat{\mu}) = 2(\ell_f(y;y) - \ell_m(\hat{\mu};y))$$

Consider a gamma-theory linear regression model for a single observation with density;

$$f(y,\theta,\phi) = (\frac{\nu}{\mu})^{\nu}\frac{y^{\nu-1}e^{-\frac{\nu y}{\mu}}}{\Gamma(\nu)}$$

The log-likelihood function is defined as;

$$\ell(\hat{\mu};y) = -\sum_{i=1}^{n}\frac{\nu y}{\hat{\mu}} - n\nu\log\hat{\mu} + \sum_{i=1}^{n}(\nu-1)\log y + n\nu\log\nu - n\log\Gamma(\nu) \quad (5.1)$$

Setting (5.1) $\hat{\mu} = y$ gives the maximum achievable log-likelihood, $\ell(y;y)$. Hence:

$$\ell(y;y) = -\sum_{i=1}^{n}\frac{\nu y}{y} - n\nu\log y + \sum_{i=1}^{n}(\nu-1)\log y + n\nu\log\nu - n\log\Gamma(\nu)$$

which becomes

$$\ell(y;y) = -\sum_{i=1}^{n}\nu - n\nu\log y + \sum_{i=1}^{n}(\nu-1)\log y + n\nu\log\nu - n\log\Gamma(\nu) \quad (5.2)$$

Subtracting (5.1) from (5.2), we have;

$$-\sum_{i=1}^{n}\nu+\sum_{i=1}^{n}\frac{\nu y}{\hat{\mu}}-n\nu\log y+n\nu\log\hat{\mu}+\sum_{i=1}^{n}(\nu-1)\log y-\sum_{i=1}^{n}(\nu-1)\log y$$
$$-n\log\Gamma(\nu)+n\log\Gamma(\nu) \quad (5.3)$$

It is apparent that in (5.3), the last four terms are eliminated as each opposite corresponding terms cancels each other out. We are now left with the first four terms

$$\sum_{i=1}^{n}\frac{\nu y}{\hat{\mu}}-\sum_{i=1}^{n}\nu+n\nu\log\hat{\mu}-n\nu\log y \quad (5.4)$$

note that, previously; $n\nu\log\hat{\mu}=\sum_{i=1}^{n}\nu\log\hat{\mu}$ and $n\nu\log y=\sum_{i=1}^{n}\nu\log y$

hence 5.4 becomes

$$\sum_{i=1}^{n}\frac{\nu y}{\hat{\mu}}-\sum_{i=1}^{n}\nu+\sum_{i=1}^{n}\nu\log\hat{\mu}-\sum_{i=1}^{n}\nu\log y \quad (5.5)$$

(5.5) then becomes

$$\sum_{i=1}^{n}\left[\frac{\nu y}{\hat{\mu}}-\nu-(-\nu\log\hat{\mu}+\nu\log y)\right] \quad (5.6)$$

which becomes

$$\sum_{i=1}^{n}\left[\frac{\nu y}{\hat{\mu}}-\nu-(\nu\log y-\nu\log\hat{\mu})\right] \quad (5.7)$$

applying the laws of logarithm on the last two terms of the summands of (5.7), we have;

$$\sum_{i=1}^{n}\left[\left(\frac{\nu y}{\hat{\mu}}-\nu\right)-\nu\log\frac{y}{\hat{\mu}}\right] \quad (5.8)$$

(5.8), we now have

$$\sum_{i=1}^{n}\left[\left(\frac{-\nu\hat{\mu}+\nu y}{\hat{\mu}}\right)-\nu\log\frac{y}{\hat{\mu}}\right] \tag{5.9}$$

factorizing μ out and rearranging (5.9) for each summand, and rearranging we have

$$\sum_{i=1}^{n}\left[\nu\left(\frac{y-\hat{\mu}}{\hat{\mu}}\right)-\nu\log\frac{y}{\hat{\mu}}\right] \tag{5.10}$$

Meaning;

$$\ell_f(y;y)-\ell_m(\hat{\mu};y)=\nu\sum_{i=1}^{n}\left[\frac{y-\hat{\mu}}{\hat{\mu}}-\log\frac{y}{\hat{\mu}}\right] \tag{5.11}$$

And now our scaled deviance, which is $2(\ell_f(y;y)-\ell_m(\hat{\mu};y))$ is

$$D^*(y;\hat{\mu})=2\nu\sum_{i=1}^{n}\left[-\log\left(\frac{y}{\hat{\mu}}\right)+\frac{y-\hat{\mu}}{\hat{\mu}}\right] \tag{5.12}$$

as stated in our methodology in (3.125)

A2: Proof of The unscaled Deviance (Residual Deviance) of The Gamma Model

Since the Gamma regression model has its $\phi \neq 1$, the deviance formula produces the scaled deviance, hence the unscaled deviance, called the residual deviance is established by the following relation;

$$D^*(y;\hat{\mu})=D(y;\hat{\mu})/\phi \tag{5.13}$$

Hence

$$D(y;\hat{\mu}) = \phi D^*(y;\hat{\mu}) \tag{5.14}$$

For the Gamma model, $\phi = 1/\nu$

Therefore, (5.12) becomes

$$\frac{1}{\nu} \times 2\nu \sum_{i=1}^{n} \left[-\log \frac{y}{\hat{\mu}} + \frac{y-\hat{\mu}}{\hat{\mu}} \right] \tag{5.15}$$

cancelling out the ν in 5.15, we now have the residual deviance

$$D(y;\hat{\mu}) = 2 \sum_{i=1}^{n} \left[-\log \left(\frac{y}{\hat{\mu}} \right) + \frac{y-\hat{\mu}}{\hat{\mu}} \right] \tag{5.16}$$

as stated in (3.123)