ISEG
Lisbon School
of Economics
& Management
Universidade de Lisboa

MASTER
ACTUARIAL SCIENCE

MASTER´S FINAL WORK
DISSERTATION

COMBINED LOSS RESERVING AND PREMIUM RATING BY GLM

JOEL AGBO MENSAH

FEBRUARY - 2023

MASTER
ACTUARIAL SCIENCE

MASTER´S FINAL WORK
DISSERTATION

COMBINED LOSS RESERVING AND PREMIUM RATING BY GLM

JOEL AGBO MENSAH

SUPERVISION:
WALTER A.H NEUHAUS

FEBRUARY - 2023

DEDICATION

This dissertation is dedicated to Patience A. Ami

& Wisdom Agbo

GLOSSARY

AIC	Akaike Information Criterion.
AY	Accident Year.
BF	Bornhuetter–Ferguson.
BIC	Bayesian Information Criterion.
C.V	Coefficient of Variation.
CDF	Cumulative Distribution Function.
CL	Chain Ladder.
CUM	Cumulative.
DEV	Development.
EDM	Exponential Dispersion Models.
EF	Exponential Family.
EU	European Union
EXP	Exponential.
GLM	Generalised Linear Models.
IBNR	Incurred but not Reported.
MSEP	Mean Square Error of Prediction.
PDF	Probability Distribution Function.
PP	Probability – Probability.
QQ	Quantile – Quantile.
S.E	Standard Error.

ABSTRACT

Generalised linear models(GLM) are routinely used in two different areas of actuarial work: Loss Reserving and Premium rating. There is little overlap between the two areas: Loss Reserving models attempt to model the development of claims but pays little attention to effect of risk variables. Premium Rating model attempt to model the effect of risk variables on claim patterns (frequency and/or severity), but usually assumes that the claims analysed are fully developed.

In this dissertation, we aim to bridge the gap between these two areas of actuarial work by developing a Premium Rating model that incorporates risk variables. Specifically, we will consider demographic characteristics such as gender on claim patterns. By doing so, we hope to provide a more comprehensive understanding of the factors that contribute to insurance claims and improve insurers' ability to accurately price their policies, something which can be done in GLM but not in the original Chain Ladder or Bornheutter-Ferguson methods.

The GLM approach is applied to real-life statistics of a professional health insurance that is sold to two risk groups, females and males. The results show that with the inclusion of the risk_group variable in the GLM model framework, females have higher claim cost per insured than males, plus that the number of females is increasing while the number of males is falling. The increase of the proportion of females is partly explained by the fact that more females are entering the profession. In a competitive market, the insurance company could risk adverse selection, if at the same time as more women enter, the lower risk group (males) starts falling because premiums are becoming too high. EU regulation does not allow insurers to differentiate premiums by sex. Therefore, the insurance will have to find other ways than premium differentiation, to prevent or reduce adverse selection. It is not my purpose to suggest what the company could do. The purpose of this dissertation is to demonstrate that the use of a GLM in loss reserving may show up facts that would remain concealed if one only used a simple chain ladder method on the aggregate statistics.

The theoretical base of the work is standard; its challenges lies in applying GLM to realistic datasets and studying the results.

KEYWORDS: Chain ladder, Bornhuetter-Ferguson, GLM, risk group.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENT

First, I want to thank the Almighty God for his wisdom and guidance He has given me during this master's program. I want to express a heartfelt gratitude to Professor Walter Neuhaus for his encouragement and guidance. I am also grateful to Dr. Johnson Adamaley for his financial support throughout my master's program. Finally, I am also thankful to my friends.

1.INTRODUCTION

The insurance industry makes promises rather than sells actual products. An insurance contract is a guarantee given by the insurer to the policyholder (insured) to pay for future losses or damages in exchange for a premium paid in advance.

Therefore, insurers must determine a reasonable price for their contract based on their best judgment after analysing and evaluating past data. Most types of home, liability, and auto insurance, as well as other forms of general insurance, have an annual policy term. The time it takes to settle these claims, however, can be years. This means that the actuary frequently lacks information crucial to the completion of the insurance contract, such as the date of settlement. For instance, the payments of claims arising from liability insurance might drag on for a longer period. This may be due to investigations, dispute, litigation, or other processes leading to determination of the claim amount.

Figure 1.1 gives a single claim's history. It demonstrates that t_1 was the date of damage occurred. Then sometime later say, at t_2 the insurer is notified of the loss/claim arising from the damage. Usually, the claim will not be paid instantly.

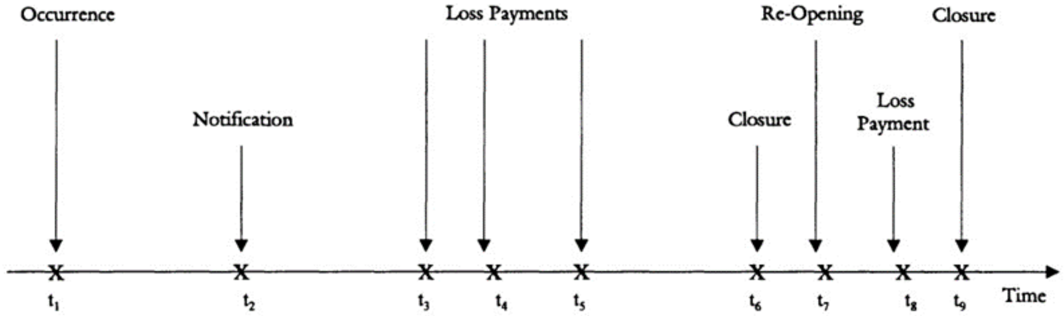


Figure 1.1 Timeline for a single claim payment

After a number of loss payments, the insurer determined that the claim's activity was finished at time t_6, and closes the file. If this decision was later proven to be incorrect at time t_7, the claim file was reopened, another payment was paid (at t_8), and it was once more closed at t_9, since there has been no further action. The duration required to complete a claim payment is subject to fluctuation based on several factors, including the intricacy of the claim, the insurer's internal procedures, and the extent of negotiation needed. Nevertheless, it is crucial for insurers to ensure timely and precise payment of claims to sustain policyholder contentment and guarantee financial soundness.

The single claims payment process plays a crucial role in loss reserving, as each payment impacts the insurer's reserves and can affect their ability to accurately estimate future claim liabilities. However, it is challenging and often impossible to model reserves assuming a normal distribution of data due to the presence of various risk factors. However, the advent of more sophisticated statistical tools, such as the generalized linear model (GLM), has helped to overcome this challenge. Unlike conventional methods like the Chain Ladder (CL) method, which is used to estimate final losses and does not rely on known distributions, GLM can assume various known distributions. Additionally, GLM can reveal insights that would remain hidden if one only used a simple chain ladder method on the aggregate statistics and other basic conventional methods.

The insurance world is changing at a very fast pace. In the Insurance world pricing, underwriting, and claims triage have been metamorphosed by predictive analytics. GLM are routinely extended in two different areas of actuarial work: loss reserving and premium rating. There is little overlap between the two areas: loss reserving models attempt to model the development of claims but normally pay little attention to the effect of risk variables. Premium Rating models attempt to model the effect of risk variables on claiming patterns (frequency and/or severity), but usually assume that the claims analysed are fully developed.

The purpose of this work aims to integrate premium rating factors and loss reserving methodologies by developing parsimonious models that incorporate both risk and development variables. The models will be applied to actual data to test their effectiveness in predicting claims patterns and associated losses, providing insurers with better insights for risk management and pricing strategies.

The current practice of traditional loss reserving estimates claims reserves based only on two variables: an accident year effect and a development year effect. However, this approach ignores the impact of other relevant factors such as premium rating factors. In order to address this gap, this study aims to explore the use of Generalized Linear Models (GLM) for claim reserving that incorporates premium rating factors such as risk group variables. By doing so, the study seeks to extend the applicability of the results to pricing or designing insurance products for each risk group. This approach has been suggested as more accurate and efficient in claim reserve estimation as compared to traditional methods, which only consider accident and development year effects. The results of the study are expected to be relevant to policymakers and actuaries in making better decisions regarding investment for claim settlement and evaluating loss reserves respectively.

The objective of this study is to develop more accurate and reliable loss reserving models that can assist insurers in making better-informed decisions regarding risk management and pricing. To achieve this goal, the study will focus on four key objectives:

First, the study will evaluate the suitability of various probability distributions for modeling claims paid data. This will involve analyzing historical claims data to identify the most appropriate distribution for the data set.

Second, the study will estimate claims reserves using traditional loss reserving methods. These methods involve analyzing historical claims data to predict future claims patterns, based on two key variables: an accident year effect and a development year effect. By estimating claims reserves using these methods, the study will provide a benchmark against which to compare other modeling approaches.

Third, the study will quantify the uncertainty associated with loss reserve estimates by applying the (Mack) stochastic model to calculate the standard error of the estimate. This will provide insurers with a measure of the reliability of their loss reserve estimates, allowing them to adjust their risk management strategies accordingly.

Finally, the study will develop a generalized linear model within a specified model framework to estimate reserves for each risk group, considering relevant risk variables such as policyholder gender.

The research aims to address two important questions in the area of actuarial science and insurance. Firstly, the study will investigate the added value of the stochastic loss reserving technique, specifically the Generalized Linear Model (GLM), compared to other traditional loss reserving techniques, in terms of its impact on decision-making. Secondly, the research will demonstrate how the reserve estimates for each risk group can help insurance companies design and price their products.

The study aims to introduce additional techniques to the traditional loss reserving approach, giving actuaries more options to evaluate loss reserves. This can be beneficial in situations where traditional methods may not be adequate or where more accurate reserve estimates are required. By incorporating generalized linear models and stochastic loss reserving techniques, this research offers more comprehensive and robust tools for estimating future claims liabilities.

Overall, this study's findings have the potential to improve the accuracy of loss reserve estimates, enhance financial planning and stability in insurance companies, and provide additional techniques for actuaries to better evaluate reserves.

The various chapters consist of the following: chapter 2 consists of a literature review of loss reserving methods, including its applications to various insurance data, chapter 3 is dedicated to models and methods in loss reserving, chapter 4 presents the data analysis, first fitting the claim data to various probability distribution and then using packages in R to determines reserve estimate for both the conventional method and GLM and lastly chapter 5 presents the main conclusions.

2. LITERATURE

This chapter looks at the relevant literature and shows how loss reserving methods are used in different areas of actuarial science.

Loss reserving using the chain ladder technique is a traditional method that estimates loss reserves by calculating a consistent ratio of losses between subsequent development years, as described by Mack (1993). The chain ladder method does not rely on known distributions and avoids assumptions seen in other traditional methods. However, recent studies have shown that by assuming claim amount distributions are known, other models can produce estimates that are comparable to the chain ladder method. For example, Wüthrich and Merz (2008) demonstrate that the distribution-free chain ladder estimates and the Poisson model for claim counts can both produce the same projected claim amount. For further details, please refer to Wüthrich and Merz (2008).

In the context of loss reserving, another approach is the Bornhuetter-Ferguson method. Schmidt and Zocher (2008) demonstrate how this method can be extended to encompass loss reserving techniques based on run-off triangles, such as the chain-ladder method. They found that the Bornhuetter-Ferguson principle provides indicators that can be considered the best predictors of ultimate losses. Schmidt and Zocher (2008) provide a comprehensive explanation of this approach and related topics. For further study, refer to Schmidt & Zocher (2008) and Schmidt (2006).

In their paper, Kočović et al. (2018) conducted a comparative analysis of the Chain Ladder (CL) and Bornhuetter-Fergusons (BF) methodologies to gain a better understanding of their benefits and drawbacks. The authors noted that it is crucial to evaluate the results of each method to identify the causes of any discrepancies in the reserve estimates. Their study revealed that the chain ladder method is suitable when there is a constant pattern of loss development and a substantial number of reported claims. However, for cases where an inconsistent pattern of reported claims exists, such as in their example, the Bornhuetter-Fergusons method was more appropriate.

The authors also cautioned that these methods should be used with caution, considering their respective benefits and drawbacks, and combined with the subjective assessments of actuaries based on their experience and knowledge. They highlighted that accurately estimating claim reserves using these traditional methods is challenging and prone to inconsistencies. As such, they recommend future studies focus on utilizing stochastic models to estimate claim reserves, which would address the drawbacks of these traditional methods. Kočović et al.'s study adds to the existing literature on the effectiveness of these methods in accurately estimating claim reserves and the need for more advanced modelling techniques to improve accuracy and efficiency.

Building on the previous discussion about the limitations of traditional loss reserving methods, it is important to note that these methods are deterministic and may not accurately capture the variability inherent in the underlying claims data. This is where stochastic modelling comes into play. Unlike deterministic methods, stochastic models consider the uncertainty and variability associated with claims data and can provide more accurate and reliable reserve estimates. One such model is the Mack stochastic model, which extends the chain ladder method by incorporating the calculation of the standard error to a particular reserve estimate. For further information on the Mack stochastic model, please see Mack (1993) and Mack et al (2000).

The use of stochastic models for loss reserving has become increasingly popular in recent years. One area of emphasis for such models is the Generalized Linear Model (GLM). Schmidt (2004) provides a remarkably simple method to estimate the number of claims required for a tariff computation based on the number of risk factors and the number of levels for each element. Frees and Valdez (2008) proposed a conceptual framework for three components that relate to the rate, nature, and intensity of damages. These components allow actuaries to consider a wide range of factors when estimating losses. Another study by Klein et al. (2014) examines a scenario where the assumption that the response variable has an exponential family of distribution is relieved. This allows the actuary to consider risk factors not only in the mean but also in parameters that affect the behavior of the individual who files loss claims. By considering a wider range of factors, such as the nature and intensity of damages, and incorporating them into stochastic models like the GLM, actuaries can improve the accuracy and efficiency of their loss reserve estimates.

Taylor and McGuire (2002), their paper presented a case study in the application of Generalised linear models to loss reserving. The study was initially approached from the perspective of an actuary with a predisposition to the application of the chain ladder (CL). They saw that the data set used in their study violate the conditions for application of the Chain ladder in many ways. These difficulties of attuning the Chain ladder to allow for these features of their data set to be captured was overcome by the introduction of GLM regression as a well systematized and rigorous form of data analysis. This helped them in modelling and investigating a number of complex features of data responsible for the violation of the CL assumptions. Their paper concluded that the complexity of the data set is seen in the model of claim sizes fitted to it, which entails the following, in addition to the expected variation with operational time: a seasonal effect, which will be extremely difficult to accommodate such trend within the CL framework and estimate them efficiently.

De Jong and Heller (2008) used the vehicle insurance data set to calculate insurance premiums using features of the insured drivers and vehicle. Their goal was to estimate the mean claim frequency while considering various explanatory variables and employing models supplied by the generalized linear model methodology. Their paper offers a theoretical exposition of GLMs, with a particular emphasis on Poisson regression models, as well as an emphasis on the influence of different explanatory factors on the number of claims, using various descriptive approaches in R.

N. Naufal, S. Devila and D. Lestari (2019) presented a paper on using GLM to determine life Insurance premiums. According to them the risk of mortality for each individual is determined by various risk factors which includes gender, marital status, alcohol consumption, age, smoking status, geographical location, profession and education. This risk factors affects the premium paid by each individual to ensure fairness. For this reason, insurers need a model that will measure the effect on mortality of these risk factors. According to them GLM provides important insights in insurance data analysis. In their study they used GLM to model the risk of mortality caused by various risk factors and then calculate the premium for each individual. The data they used in their study are life claim data which comprises of risk factors that affect mortality rates in Indonesia. Based on the discussion and case studies, the risk factor with a significant effect on the probability of mortality is gender. The issue of age and smoking status do not affect the probability of mortality. The probability of individual female mortality which is greater than that of male individual.

Having reviewed various literature, little work has been found that applies a stochastic model (GLM) to loss reserving, and includes other risk variables than the accident year and development year effect as covariates.

3. METHODS AND MODELS

According to Neuhaus (2014), several lines of research are currently en vogue in the actuarial profession. For want of better terms, Neuhaus (2014) also referred to these lines of research as “fitting to method”, and “fitting to data”. Before explaining what the two terms meant, he proceeded to define certain key words which are mostly used interchangeable, especially in the actuarial profession.

- Model: a simplified mathematical description of the claim development mechanism.
- Method: an algorithm for turning observed data into projections of future data.

3.1 Fitting to method

Several academics and actuaries are attempting to analyze the statistical properties of the heuristic methods, most often the Chain-ladder method. The seminal paper is Mack (1993), see also England and Verrall (2002) and Wüthrich & Merz (2008) for comprehensive descriptions. This line of research involves finding a model within which a given method is optimal or at least justifiable, for example, because its predictions coincide with maximum likelihood estimates. Thereafter the statistical properties of the method are computed within constraints of that model. As a result, the actuary will be able to produce an estimate of predictive uncertainty.

3.2 Fitting to data

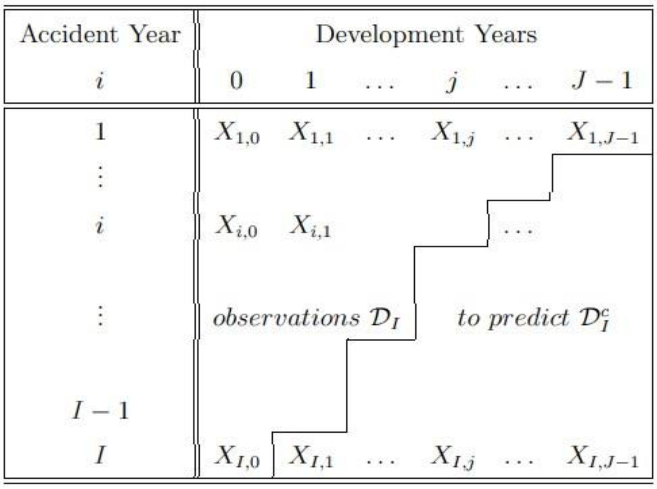
Other authors fit models not to methods, but to data. An extensive treatment can be found in Taylor (2000). According to Neuhaus (2014), the main difference between fitting to method and fitting to data is that in the former approach the method being studied puts à priori constraints on the admissible models, while in the latter approach the model is built with the objective of capturing important aspects of the mechanism that underlies claim development.

3.3 Classical methods

The chain ladder method

The chain ladder approach to loss reserving is used by most insurance firms for loss reserving. It is a particular fundamental reserve strategy that is employed to foretell final losses. According to Christofides (1997), the key idea behind the chain ladder technique is that previous payments are good predictors of the ones to come. According to Taylor and McGuire (2004), the triangle was first studied from the standpoint of the inclination to apply the CL. Wüthrich (2019), gives the annotation to the various elements of the total claim reserves for the future payments where $X_{i,j}$ stands for the payments made for claims with accident year *i* in development year *j*. Thus, the vertical axis *i* = year of accident and *j* = development year on the horizontal axis. $C_{i,j} = \sum_{k=0}^{j} X_{i,k}$

as the total payments made for claims from accident year i until development year j. According to Wüthrich (2019), all the observations are located in the upper left of the triangle $\mathcal{D}_I$, and the lower part of the triangle D_I^C is future observation that actuaries would like to predict.



Accident Year	Development Years					
i	0	1	...	j	...	$J-1$
1	$X_{1,0}$	$X_{1,1}$	...	$X_{1,j}$	...	$X_{1,J-1}$
$\vdots$						
i	$X_{i,0}$	$X_{i,1}$			...	
$\vdots$	*observations* $\mathcal{D}_I$			*to predict* $\mathcal{D}_I^c$		
$I-1$						
I	$X_{I,0}$	$X_{I,1}$	...	$X_{I,j}$	...	$X_{I,J-1}$

Table 3.1: The Run – off Triangle

According to Merz and Wüthrich (2008), the CL technique is one of the most popular claims loss reserving techniques. Classical actuarial literature often classifies the CL method as a purely computational algorithm for estimating reserves.

Assumptions of the chain ladder algorithm

1. Various accident years' $i = 1, \dots I$ cumulative claims $C_{i,j}$ are independent from one another.

2. It further assumes the presence of developmental factors $g_o, \dots, g_{j-1} > 0$ such that $\forall\ 0 \leq i \leq I$ and $\forall\ 1 \leq j \leq J$. We obtain;

$$\mathbb{E}\left[C_{i,j} | C_{i,0} \dots C_{i,j-1}\right] = \mathbb{E}\left[C_{i,j} | C_{ij-1}\right] = g_{J-1} . C_{ij-1} \qquad \textbf{(3.1)}$$

Merz and Wüthrich (2008) also remarks that the first moment is assumed in Equation (3.1) which is already adequate (and hence gives the CL algorithm) for calculating the conditional expectation of future claims. Merz and Wüthrich (2008), then proposed that by using the cumulative claim $C_{i,0}, C_{i,1}, \dots, C_{i,j}$ in accident year i then it forms a Markov chain, which is a stronger assumption in addition to the first two assumption. Thus,

$$C_{ij} . \prod_{k=0}^{j-1} g_k^{-1} \qquad \textbf{(3.2)}$$

The factors g_j are known as the CL age-to-age ratio, CL factors, or CL link ratios, and they serve as the CL method's main point of interest.

Merz and Wüthrich (2008) further assumed $\mathcal{D}_I = \{C_{ij}; i + j \leq I, 0 \leq j \leq J\}$ } to be the collection of observations in the upper part of the run-off triangle. Using the model assumption in (3.1) $\forall\, 1 \leq i \leq I$. Thus,

$$\mathbb{E}[C_{i,j}|D_I] = \mathbb{E}[C_{ij}|C_{i,I-i}] = C_{i,I-i} \, . \, g_{I-i} \, \dots \, g_{j-1} \qquad \textbf{(3.3)}$$

Equation (3.3) represents the best estimate of reserves for accident year i, predicated on the Upper part of the run-off triangle ($\mathcal{D}_I$) and observed CL factors. In this vein, predicting the outcome of the random variable $C_{i,j} - C_{i,I-i}$ give the observation $\mathcal{D}_I$, using the conditionally expected value (3.3) becomes possible. Sadly, in most real-world situations, the CL factors are unknown, hence the need to estimate them. Equation (3.4) gives the formular to estimate the CL factors.

$$\hat{g}_j^{CL} = \frac{\sum_{i=1}^{I-J-1} C_{i,j+1}}{\sum_{i=1}^{I-J-1} C_{i,j}} = \frac{\sum_{i=1}^{I-j-1} C_{i,j}}{\sum_{k=1}^{I-J-1} C_{k,j}} \cdot \frac{C_{i,J+1}}{C_{i,j}} \qquad \textbf{(3.4)}$$

Now the CL estimator for $\mathbb{E}[C_{i,j}|\mathcal{D}_I]$ is $\hat{C}_{i,j}^{CL} = \hat{\mathbb{E}}[C_{i,j}|\mathcal{D}_I] = C_{i,I-i} \, . \, \hat{g}_{I-i} \dots \hat{g}_{j-1} \;\; for\; i + j > I$. This is the CL computational algorithm that results in the CL reserves.

Bornhuetter – Ferguson Method

The Bornhuetter Ferguson loss reserving approach is another straightforward technique. Bornhuetter and Ferguson (1972) proposed a method that combines the chain-ladder forecast with previous knowledge of anticipated loss costs. Thus, The BF method constructs a loss reserve considering the insurance company's exposure to loss, as opposed to the basic chain ladder, which relies on the concept of experience. The advantage of the BF reserving method is that it does not change when the number of claims goes up or down. It also does not consider the "run-up" of each claim cohort and assumes that its future will follow a model pattern.

In their publication from 1972, Bornhuetter and Ferguson state that the BF technique as a mechanical procedure for estimating reserves. As it ignores outliers in the observations, it is regarded as a robust strategy according to Merz and Wüthrich (2008). The CL assumes that the observation $\mathcal{D}_I$ are extended into the lower part of the run-off triangle, whereas the Bornhuetter-Ferguson (BF) assumes a different position by suggesting that the lower $\mathcal{D}_I^C$ is extended independently from $\mathcal{D}_I$ utilizing professional expertise.

Assumptions of the BF method

1. Various accident years' $i = 1, \dots I$ cumulative claims $C_{i,j}$ are independent from one another.

2. It also assumes that the presence of parameters $\mu_0, \ldots, \mu_1 > 0$ and a pattern $\beta_0, \ldots, \beta_j > 0$ with $\beta_J = 1$,$\forall\ i\ \in\ \{0,\ \ldots I\}$, $j \in\ \{0, \ldots J-1\}$ and $k\ \in\ \{1, \ldots J-1\}$

$$\mathbb{E}\left[C_{i,0}\right] = \ \beta_{\boldsymbol{i}} \cdot \mu_0$$

$$\mathbb{E}\left[C_{i,j+k} | C_{i,0}, \ldots, C_{i,j}\right] = \ C_{i,j} + \ \mu_{\boldsymbol{i}}\ .\ \left(\beta_{j+k} - \beta_{\boldsymbol{j}}\right) \qquad \textbf{(3.5)}$$

From the above equation we have

$$\mathbb{E}\left[C_{i,j}\right] = \ \mu_i \,.\, \beta_j \quad \text{and} \quad \mathbb{E}\left[C_{i,j}\right] = \ \mu_i$$

According to Merz and Wüthrich (2008), the sequence $(\beta_j)_{j=0,\ldots,J}$ depicts the overall trend of claim development. If $C_{i,j}$ are the cumulative payment, then $\left(\beta_j\right)_j$ depicts the cumulative pay-out pattern. Based on the right hand of Equation (3.5); the BF estimator for $\mathbb{E}\left[C_{i,j}|\mathcal{D}_I\right]$ is given by $\hat{C}_{i,j}^{BF}= \widehat{\mathbb{E}}\left[C_{i,j}|\mathcal{D}_I\right] = C_{i,I-i} + \left(\ 1 - \hat{\beta}_{I-i}\right)\hat{\mu}_i\ ,\ for\ 1 \le i \le I$, this represent the BF method which give the computational algorithm that results to the BF reserves.

3.4 Mack Stochastic Model

The chain-ladder procedure only provides a single point estimate of the outstanding claims and gives no hint of the expected variations of the actual outcome around the reserves. In any loss reserving exercise, it is important to understand the data and the best loss reserving method for which it is suitable. To go beyond the chain-ladder technique's straightforward reserve estimations, a stochastic model must be specified. Mack (2000) introduced a stochastic model of claim development that allows loss reserves to be calculated with a given confidence interval level. According to Mack (2000), this approach will also compute the process variance, parameter variance, and the standard error of the reserve estimate.

Estimation of the mean square error

The mean square error $MSE\left(\hat{C}_{ij}\right)$ estimator is:

$$MSE\left(\hat{C}_{ij}\right) = \ \mathbb{E}\left(\left(\hat{C}_{ij} - C_{ij}\right)^2 | D_i\right)$$

Next, with $\mathbb{E}(X-a)^2 = Var(X) + (\mathbb{E}(X) - a)^2$ we get:

$$MSE\left(\hat{C}_{ij}\right) = Var\left(C_{ij}|D_i\right) + \ \left(\mathbb{E}\left(C_{ij}|D_i\right) - \hat{C}_{ij}\right)^2$$

The Mack stochastic model is a widely used technique for estimating loss reserves. One of the key advantages of this model is that it takes into account the uncertainty inherent in loss reserving. To quantify the uncertainty of a particular reserve estimate, the mean square error (MSE) is used. This measures the difference between the estimated and actual values.

Specifically, the $MSE\ (\hat{C}_{ij})$estimator is defined as the expected value of the squared difference between the estimated and actual claim amounts $\hat{C}_{ij}$ and C_{ij}, respectively, given the development data D_i. The estimator can then be decomposed into the variance of the claim amount conditional on the development data $Var\ (C_{ij} \mid D_i)$ and the squared difference between the expected value of the claim amount conditional on the development data $\mathbb{E}\ (C_{ij} \mid D_i)$ and the estimated value $\hat{C}_{ij}$. This allows for a more accurate and reliable estimation of loss reserves, taking into account the inherent uncertainty in the claims data.

3.5 Generalized Linear Models

A Generalised Linear Model consists of two parts: the random component and the systematic component.

The random component assumes that the response's distribution belongs to the family of exponential dispersion distributions. The response Y has a distribution in the EDF, with density function taking the form:

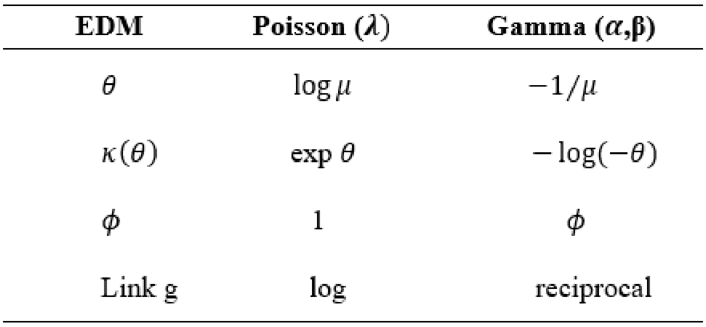
$$f(y|\theta,\phi) = c(y,\phi)\exp\left\{\frac{y\theta-\kappa(\theta)}{\phi}\right\},\text{ where:}$$

- θ is the canonical parameter;
- κ is a known function, and is called the cumulant function;
- $\phi > 0$ is the dispersion parameter;
- $c(y,\phi)$ is the normalising function: it ensures that $\int f(y|\theta,\phi)dy = 1$, if y is continuous.

The systematic component assumes that the function g links the linear predictor $\eta = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$ and the mean response: g(μ) = η; in the linear regression, $g(\mu)$ = μ.

The systematic component connects predictor variables to the response variable. It assumes a function g links the predictor variables to the mean response value μ. The linear predictor η is the sum of the intercept β_0 and the product of the predictor variables (represented by x) and their coefficients β. The function g links η to μ, telling us how the predictor variables contribute to the response variable.

The table below gives the different model component of the GLMs most used in insurance data for observed claim count or count severities.



EDM	Poisson (λ)	Gamma (α,β)
θ	$\log \mu$	$-1/\mu$
$\kappa(\theta)$	$\exp \theta$	$-\log(-\theta)$
ϕ	1	ϕ
Link g	log	reciprocal

Table 3.2 : EDMs

3.6 Fitting probability distribution to the data.

Claims paid amounts are mostly positively skewed. Therefore, the probability distributions that fit claim paid data, according to Klugman et al. (2008), are the Gamma, Lognormal, Weibull, Beta, Pareto, Burr, Normal, and Inverse Gaussian distributions. See Appendix 1 for the selected probability distribution used to fit the claims paid amount data, their parametric estimates, and log-likelihood.

4. DATA ANALYSIS

In this chapter, we present various analyses to the data. To enable a thorough understanding of the nature of the claims paid amount, probability distribution that fitted the claims paid amount the best was identified. To model the insurance data to obtain reserves, the conventional CL approach, and the Mack stochastic model, respectively, were utilized. Lastly, we used the GLM with specified frameworks to determine both the CL and BF estimates with the inclusion of the risk variable. All the data analysis were done in R and Microsoft Excel.

4.1 Plotting the claims paid amount.

The analysis of the claims paid data begins with the scatter plot of the claims paid amount. figure 4.1 the original scatter plot did not show a clear pattern, but it is clustered below the claim amount less than 1,000,000. For figure 4.2 the logged scatter plot also shows no clear pattern in the claim paid data with a reduction in the variability. The scatter plot does not give enough information pertaining to the claims paid data.

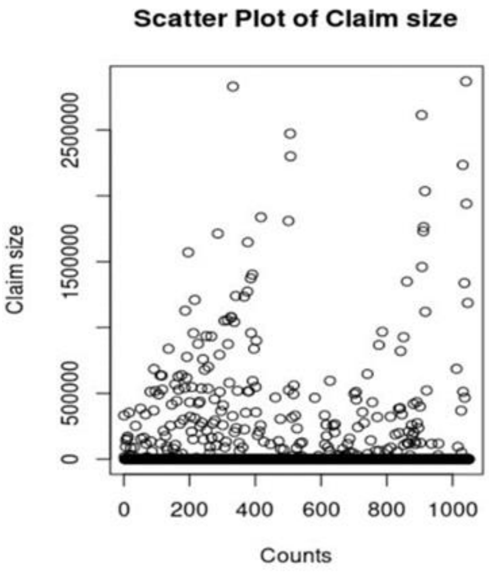


Figure 4.1: Scatter plot of claims paid

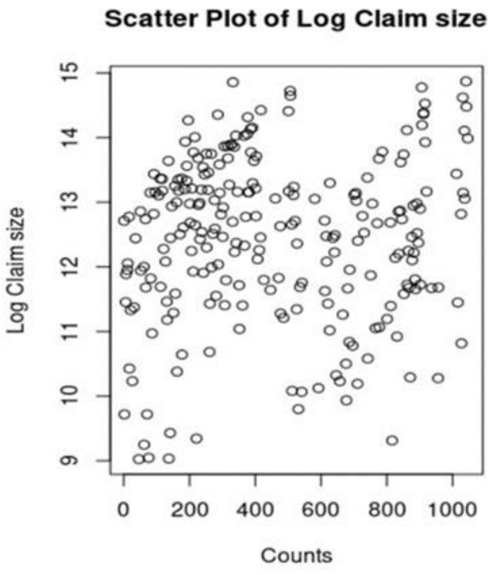
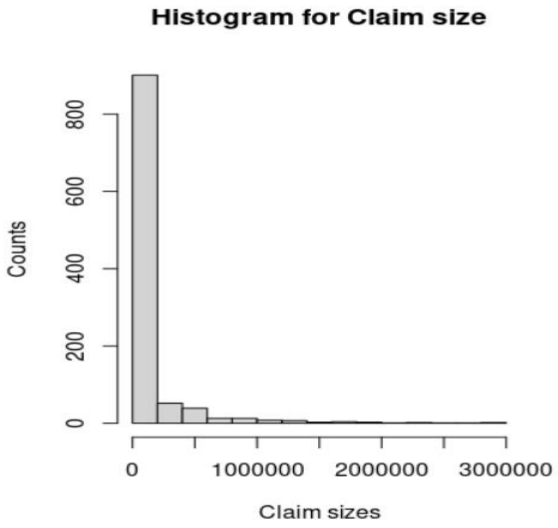


Figure 4.2: Logged scatter plot of claims paid.

The next plot is the Histogram, from figure 4.3 we could see that the original claim size is right tailed. However, with figure 4.4 taking natural logarithm of the claim amount, gives a clear picture of how the claims are distributed. The original claim size is positively skewed which is suggestive of a gamma distribution. We could see that the histogram gives us more insight into the data.



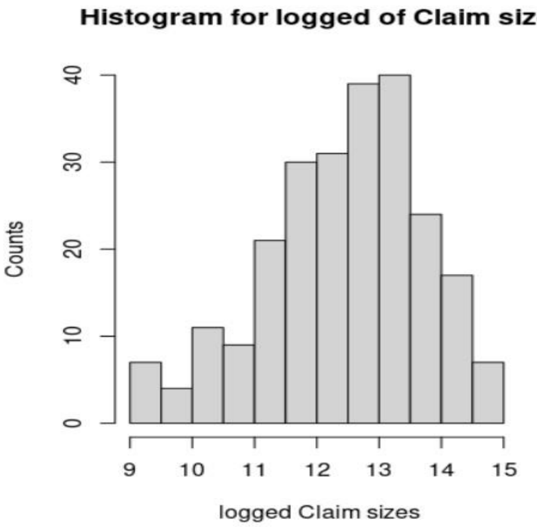
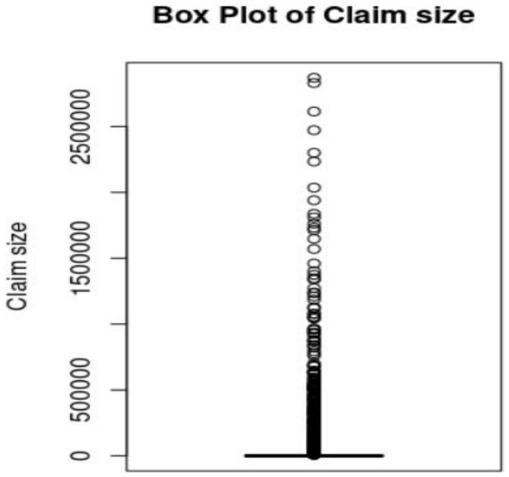


Figure 4.3 Histogram for Claims paid

Figure 4.4 Histogram for logged claims paid.

Lastly, we plotted the claims paid amount using the boxplot. figures 4.5 & 4.6 represents the boxplot. The boxplot summarizes large amount of data by displaying the data along a number line. A small distance between the extremes and the quartiles shows that the data is clustered together, the opposite is however true. We can see that about 75% of the claim amount is below 1,310,000.



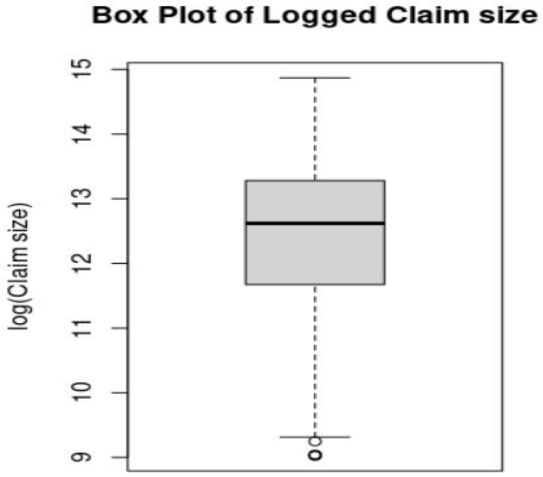


Figure 4.5 Box plot of claims paid

Figure 4.6 Box plot of logged claims paid.

4.2 Fitting probability distribution to the claims paid data.

According to Neuhaus (2014), the main objective of fitting probability distribution is to capture important aspects of the mechanism that underlies claim development.

The gamma distribution is a continuous probability distribution that is widely used to model continuous variables that are positively skewed distributions.

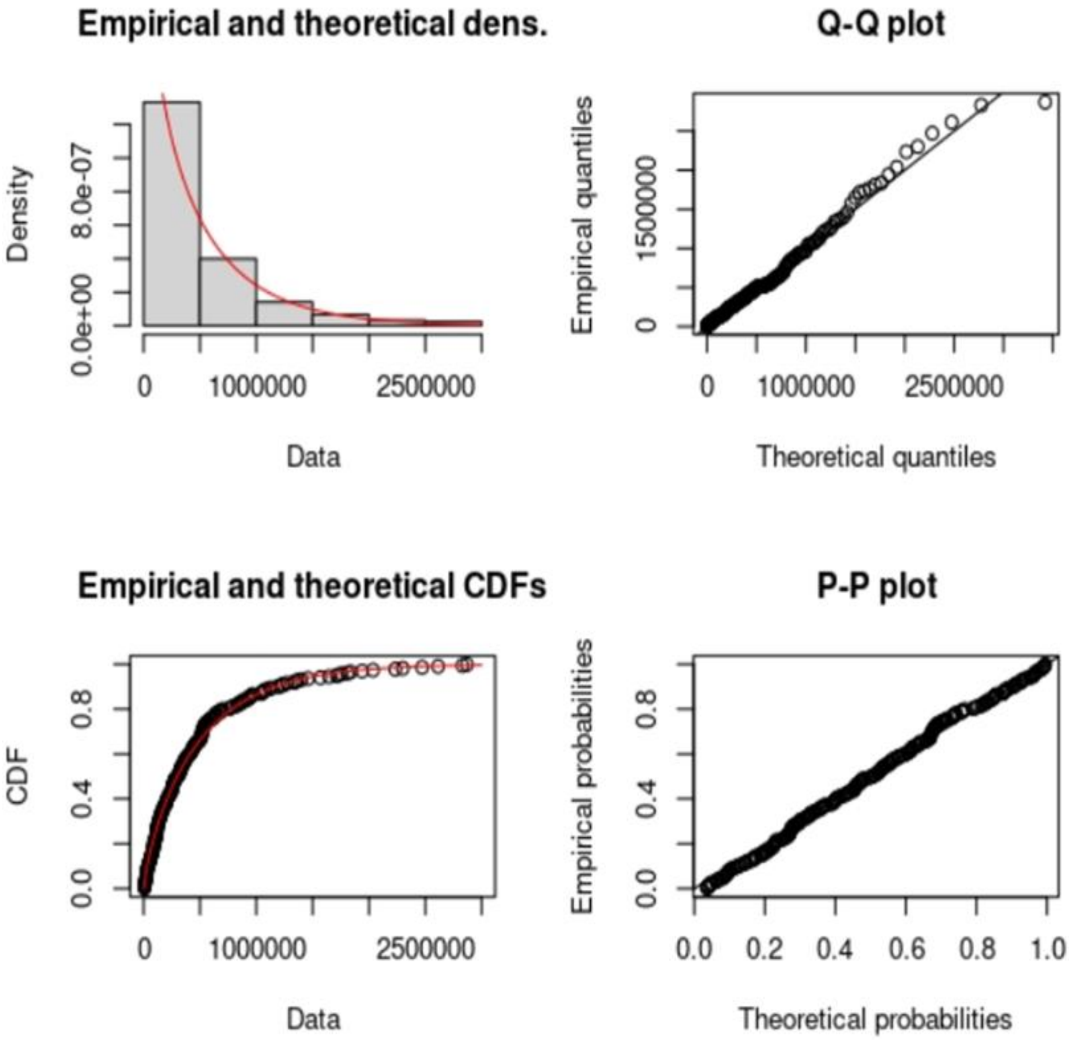


Figure 4.7: The diagram for the fitted gamma distribution to the claims paid.

Consideration would be given to the (QQ) and the PP plots when determining whether a distribution in question better fitted the claim amount data. The data points on the QQ plot should all be on the 45° line for a symmetric distribution. It can be seen from the QQ plot in the diagram that the data points are on the 45° line. Consequently, the gamma distribution is a better fit to the claims paid data than the exponential and lognormal distribution.

Exponential distribution

The exponential distribution describes the arrival time of a randomly recurring independent event sequence. It is a special case of the gamma distribution.

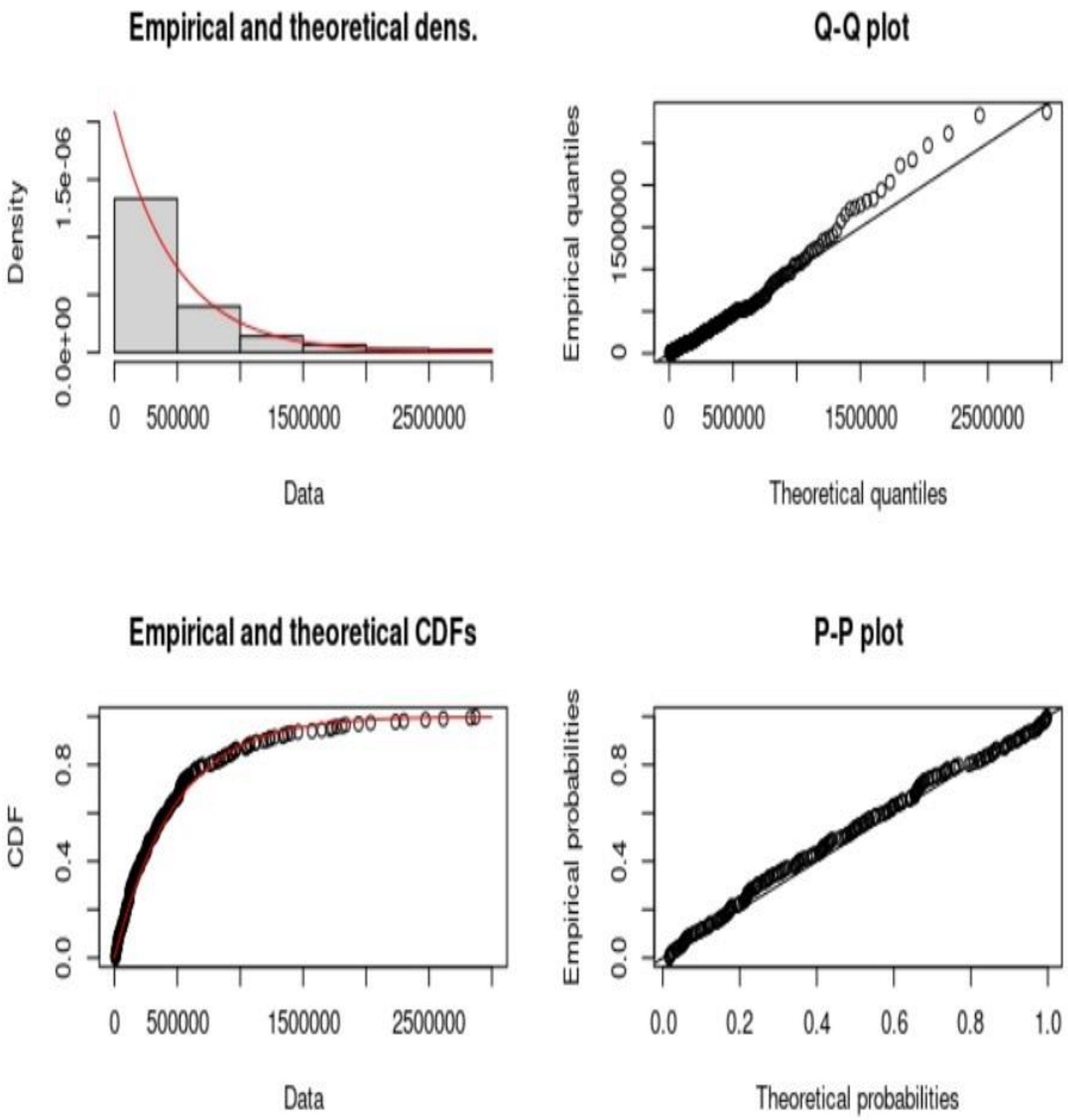


Figure 4.8: The diagnostics diagrams for the fitted exponential distribution to the claims paid.

It is obvious from the QQ plot that some data points stray off the 45° line. Hence the gamma distribution is still chosen to be a better fit to the claims paid data as compared to the exponential distribution.

Log-normal

The log-normal distribution is a continuous probability distribution of a random variable whose logarithm is normally distributed.

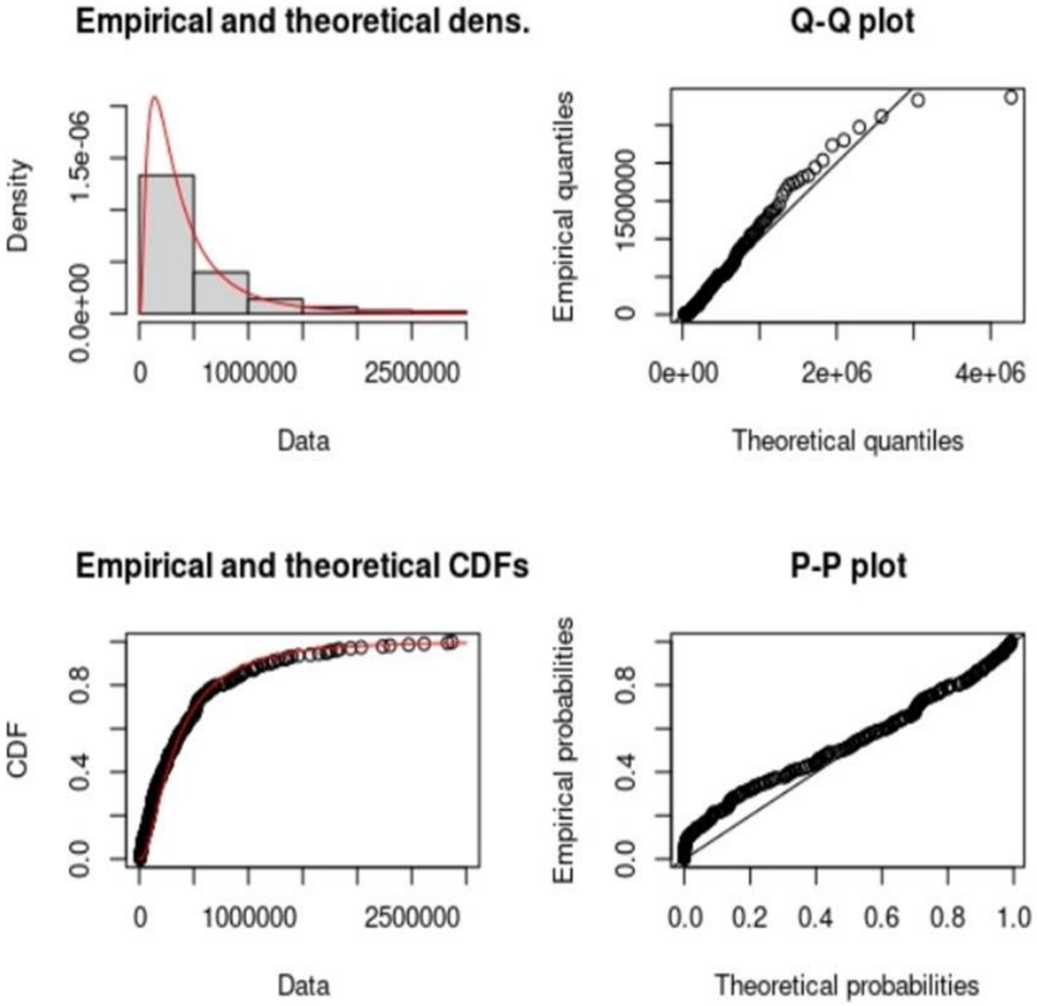


Figure 4.9: The diagnostics diagrams for the fitted log- normal distribution to the claims paid.

It obvious from the QQ plot that some data points stray off the 45° line. Hence the gamma distribution is chosen to be a better fit to the claims paid data as compared to the log-normal distribution.

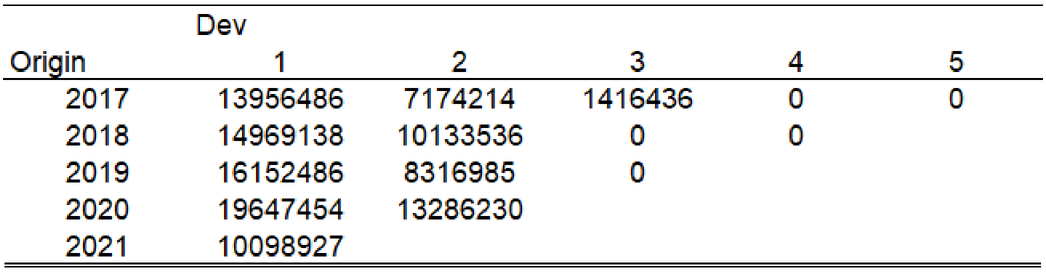
These basic plots (Scatter plot, Histogram & Box plot) and the diagram derived to determine a better probability distribution that fits the observed claims paid amount using R statistical software[1]

[1] See Appendix 2 for the Basic R codes used for estimating the various statistics, the basic plots, and the diagram for fitting of the probability distribution for the claims paid data.

4.3 The Chain ladder method

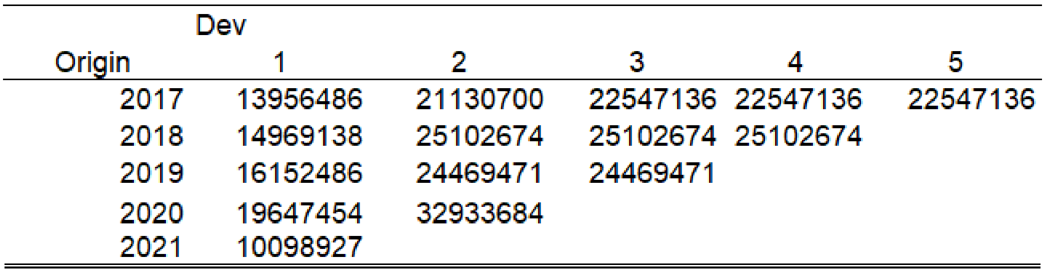
The development period of the claim payments as given by the insurer's data is 1,2,3,4 and 5. The accident years range from 2017 to 2021. The primary objective of the CL loss reserving method is to forecast the amount of reserves that must be set aside to cover projected future claims by projecting past claims experience into the future. Table 4.1 and 4.2 shows the run-off triangle of both the incremental claims payment and the cumulative payments obtained using the ChainLadder packages in R[2].

Incremental and Cumulative loss payment through development years



	Dev				
Origin	1	2	3	4	5
2017	13956486	7174214	1416436	0	0
2018	14969138	10133536	0	0	
2019	16152486	8316985	0		
2020	19647454	13286230			
2021	10098927				

Table 4.1: Incremental claims payments



	Dev				
Origin	1	2	3	4	5
2017	13956486	21130700	22547136	22547136	22547136
2018	14969138	25102674	25102674	25102674	
2019	16152486	24469471	24469471		
2020	19647454	32933684			
2021	10098927				

Table 4.2: Cumulative payments

Now that these cumulative payments have been given, development patterns can be examined. We can estimate the age-to-age loss-development factors from the cumulative payments.

Development pattern in chain ladder

Figures 4.10 and 4.11 represent the incremental and cumulative claims development by origin year. The triangle appears to be well behaved. The years 2020 has a higher incremental payment. For the years 2017 and 2018, the values appears to be relatively stable at the latter part of the development year. In general, the incremental and cumulative payment diagram can be used to

[2] The complete R codes used for this process and for the dissertation is available in this repository: https://joeboy15.github.io/Dissertation--MFW/MFW.

analyze trends in payments over time and to identify any unusual patterns or outliers. It can also be used to predict future losses by extrapolating the trends in payments.

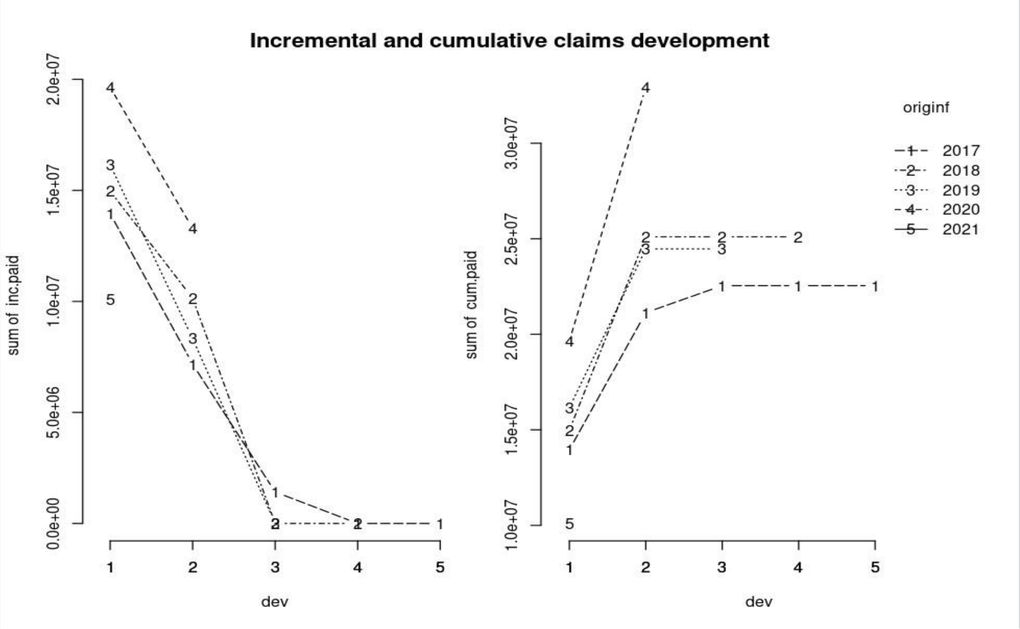


Figure 4.10: Incremental and cumulative of claims development

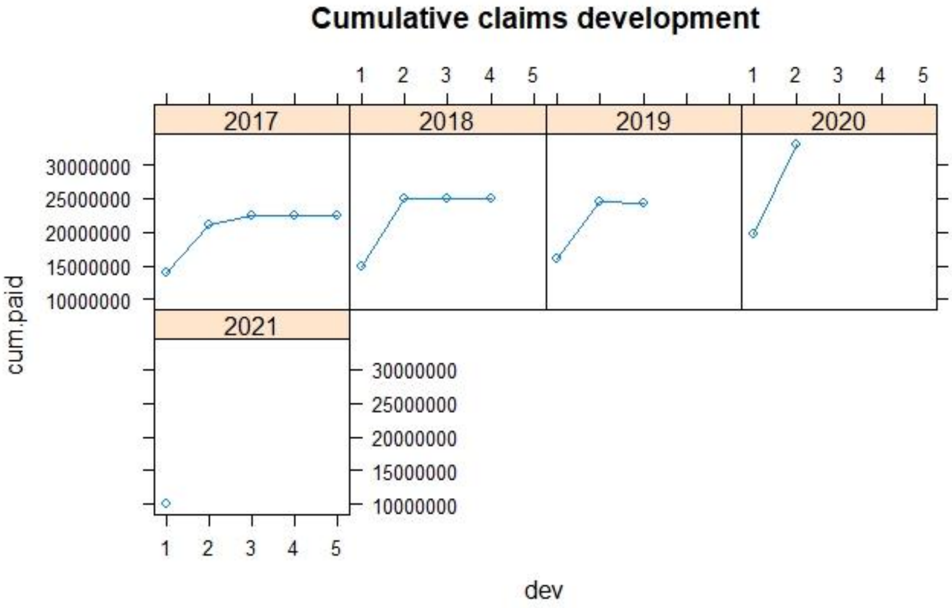
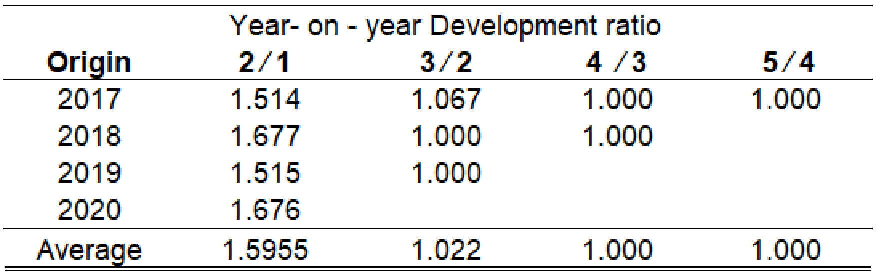


Figure 4.11: Cumulative claims development

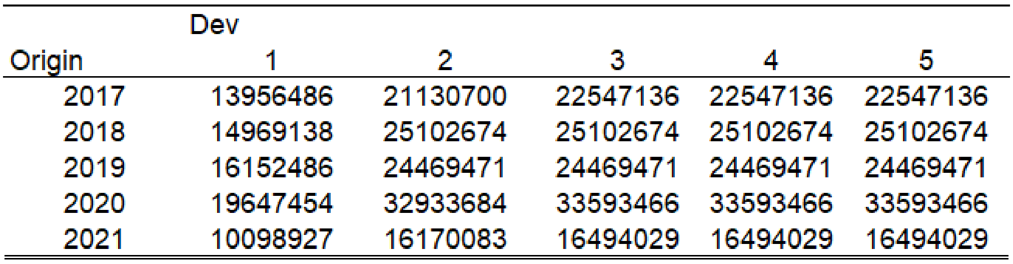
Age-to-Age paid loss-development factors based on cumulative payment.



	Year- on - year Development ratio			
Origin	**2/1**	**3/2**	**4 /3**	**5/4**
2017	1.514	1.067	1.000	1.000
2018	1.677	1.000	1.000	
2019	1.515	1.000		
2020	1.676			
Average	1.5955	1.022	1.000	1.000

Table 4.3: Development year ratio

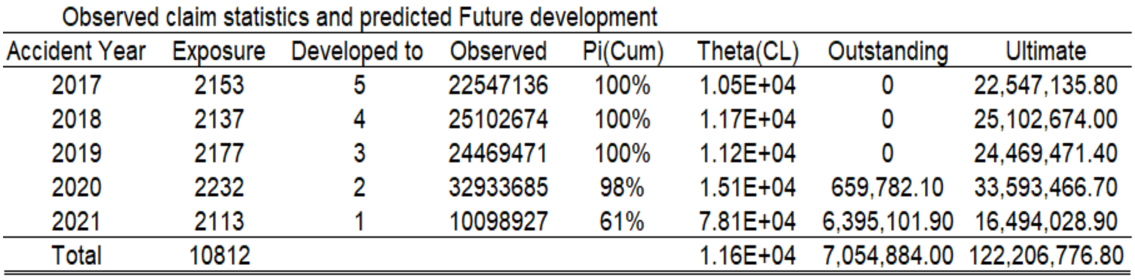
The above Table 4.3 offers insights in development years. The ratio reflects claim payment stability. The average factors will be applied to the last payment points, one for each accident year (thus the diagonals in the cumulative payments table), to calculate the expected ultimate pay-out per accident year.



	Dev				
Origin	1	2	3	4	5
2017	13956486	21130700	22547136	22547136	22547136
2018	14969138	25102674	25102674	25102674	25102674
2019	16152486	24469471	24469471	24469471	24469471
2020	19647454	32933684	33593466	33593466	33593466
2021	10098927	16170083	16494029	16494029	16494029

Table 4.4: Full triangle

The last column contains the forecast ultimate loss cost of 122,206,776.8.

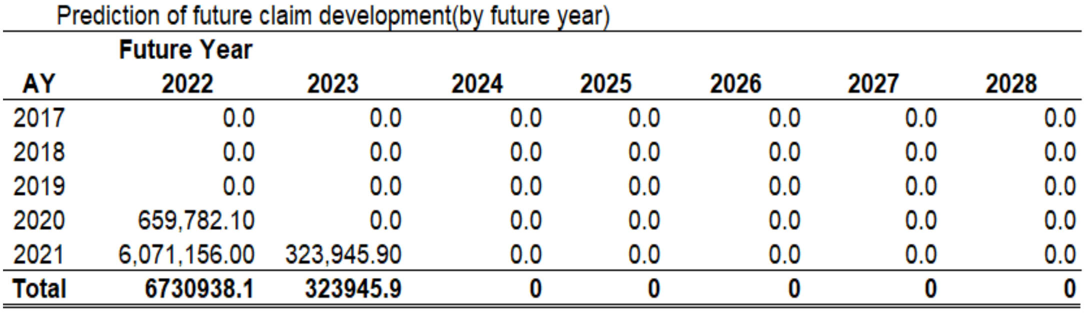


Observed claim statistics and predicted Future development

Accident Year	Exposure	Developed to	Observed	Pi(Cum)	Theta(CL)	Outstanding	Ultimate
2017	2153	5	22547136	100%	1.05E+04	0	22,547,135.80
2018	2137	4	25102674	100%	1.17E+04	0	25,102,674.00
2019	2177	3	24469471	100%	1.12E+04	0	24,469,471.40
2020	2232	2	32933685	98%	1.51E+04	659,782.10	33,593,466.70
2021	2113	1	10098927	61%	7.81E+04	6,395,101.90	16,494,028.90
Total	10812				1.16E+04	7,054,884.00	122,206,776.80

Table 4.5: Observed claims statistics and future claim development.

The chain ladder approach seeks to predict the claims payment amount into the triangle's bottom right corner of the run-off triangle as well as amount after age 4. By the end of 2021, the insurer should set aside 7,054,884 per the Chain Ladder Loss Reserving Method to fulfil all benefit obligations made to the insurance company's various clients.



Prediction of future claim development(by future year)

	Future Year						
AY	2022	2023	2024	2025	2026	2027	2028
2017	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2018	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2019	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2020	659,782.10	0.0	0.0	0.0	0.0	0.0	0.0
2021	6,071,156.00	323,945.90	0.0	0.0	0.0	0.0	0.0
Total	**6730938.1**	**323945.9**	**0**	**0**	**0**	**0**	**0**

Table 4.6: Prediction of future claim development

From Table 4.6 shows the outstanding estimates to be paid in future we see that the total paid in 2022 is **6,730,983.1** and that of 2023 is **323,945.7** which sums to the total outstanding estimate **7,054,884.**

4.4 Mack Stochastic Model Results

The Mack model is regarded as a stochastic structure for the chain-ladder technique, allowing us to compute the mean square error of future payments. Figures 4.12 and 4.13 are plots form the MackChainladder package in R.

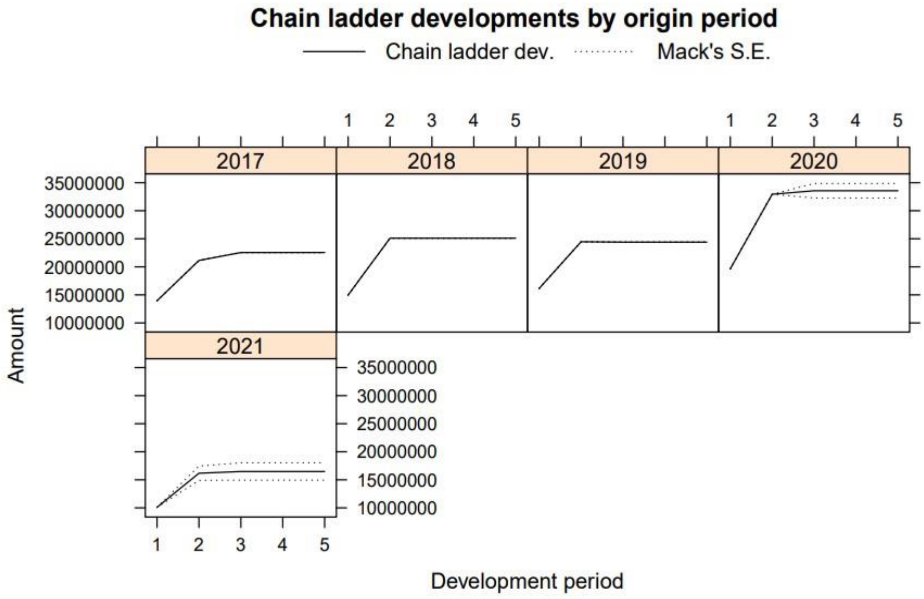


Figure 4.12: Chain ladder development by the origin period

Figure 4.12 depicts the chain ladder plot with Mack's standard error. Each plot depicts the evolution of the cumulative sums paid through time beginning with the year of origin. The solid lines indicate the predicted evolution of cumulative payments for unseen future periods, and the dashed lines depict a plus or minus one standard error as calculated by the Mack technique. As a result, the total sums paid in all years from 2017 to 2021 grew and remained consistent. Furthermore, the standard error for 2017, 2018, and 2019 is not detected when contrasted to 2020 and 2021, which clearly show a large standard error, because the years 2017-2019 are considered to be fully developed.

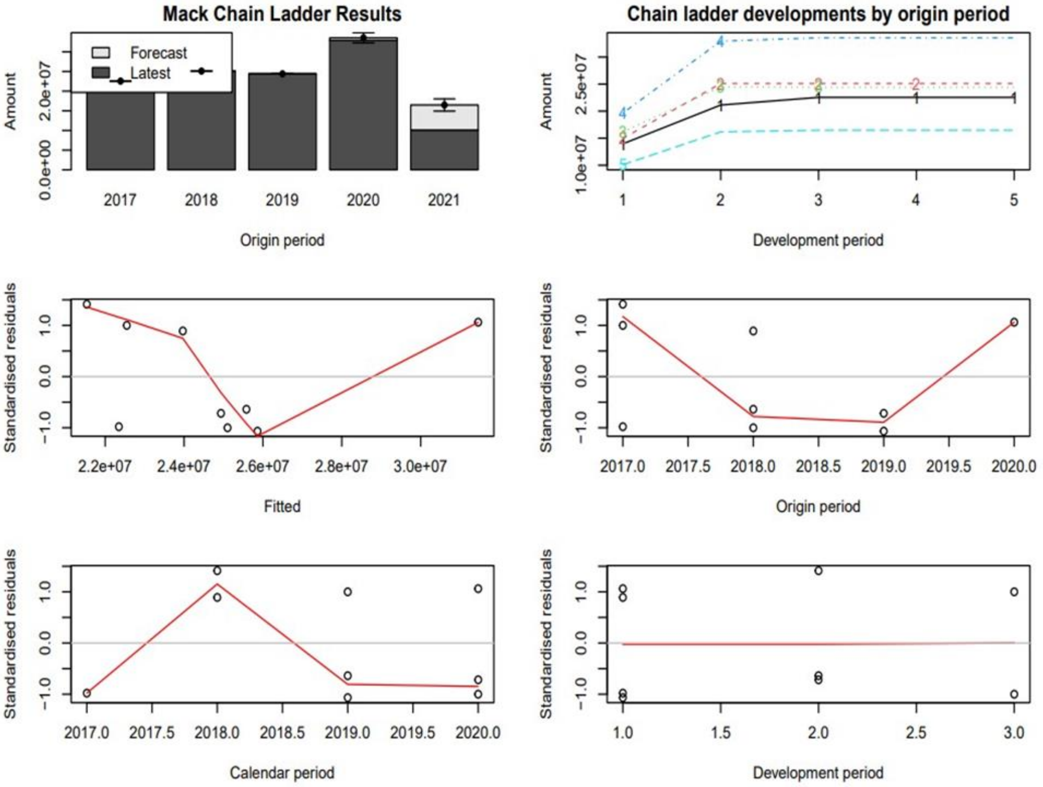
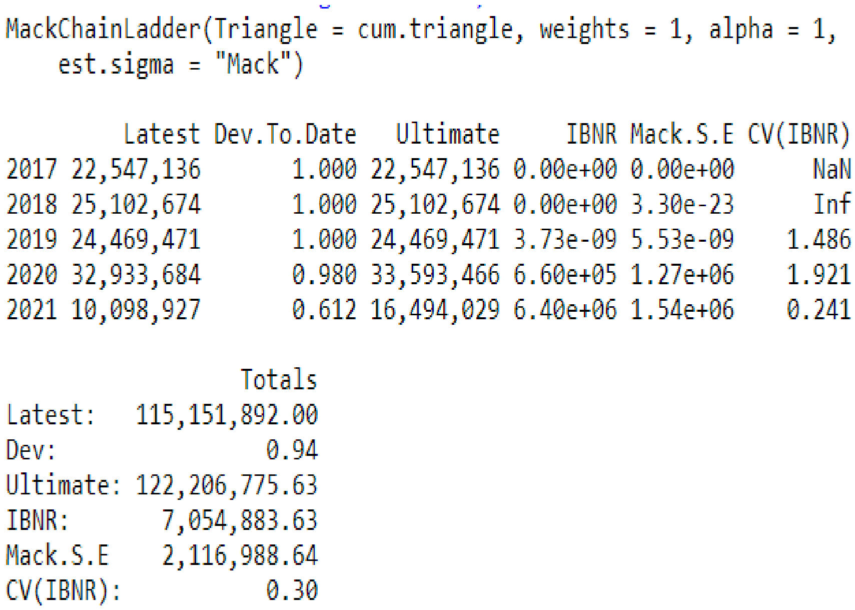


Figure 4.13: Mack Model diagram

The plot of the chain ladder in evaluating the Mack assumptions is shown in Figure 4.13. There are no trends in the four residual plots, indicating that the Mack's assumption is correct. The evolution of the chain ladder by origin period follows a similar pattern for dev 1 to 5. Furthermore, the origin period and forecast amount in the first graph from the left show that there is no forecast region for the origin (2017), indicating that the development years are fully developed.

Mack Reserve Estimate from the R output

The results shown in table 4.7 below provides the statistics of the stochastic Mack model using the R program MackChainLadder.



```
MackChainLadder(Triangle = cum.triangle, weights = 1, alpha = 1,
    est.sigma = "Mack")

         Latest Dev.To.Date   Ultimate      IBNR Mack.S.E CV(IBNR)
2017 22,547,136       1.000 22,547,136 0.00e+00 0.00e+00      NaN
2018 25,102,674       1.000 25,102,674 0.00e+00 3.30e-23      Inf
2019 24,469,471       1.000 24,469,471 3.73e-09 5.53e-09    1.486
2020 32,933,684       0.980 33,593,466 6.60e+05 1.27e+06    1.921
2021 10,098,927       0.612 16,494,029 6.40e+06 1.54e+06    0.241

                  Totals
Latest:   115,151,892.00
Dev:                0.94
Ultimate: 122,206,775.63
IBNR:       7,054,883.63
Mack.S.E    2,116,988.64
CV(IBNR):           0.30
```

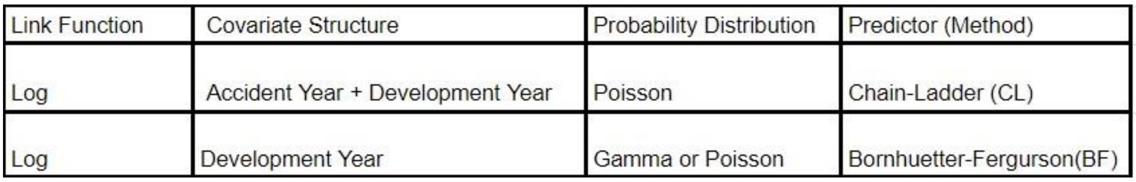
Table 4.7 Mack Reserve Estimate

The incurred but not reported (IBNR) and the mean square error associated with the individual loss payment throughout the years are calculated using the MackChainLadder package. The Mack model's coefficient of variance was plus or minus 30%. This suggests that the prediction or future payment has a standard error of 30% of IBNR.

4.5 Generalized Linear Models in Estimating Loss Reserves

Generalized linear models (GLMs) extend the range of available modelling options, which could include, the following: a calendar year effect like inflation; a trend in the accident year effect; a parametric function for the development year effect; adding risk group effects. It is quite difficult to incorporate and quantify such patterns in the classical CL framework. The GLM is one example of a fully parameterized model that considers all other effects to get a more detailed reserves estimate.

Specialized cases for the Generalized Linear Model



Link Function	Covariate Structure	Probability Distribution	Predictor (Method)
Log	Accident Year + Development Year	Poisson	Chain-Ladder (CL)
Log	Development Year	Gamma or Poisson	Bornhuetter-Fergurson(BF)

Table 4.8: Specialized cases for the Generalized Linear Model

This gives us a stochastic model to justify the CL and BF methods and to analyse their behaviour, using established GLM techniques (diagnostics, confidence intervals etc). However, model assumptions should always be chosen with the primary aim of providing a satisfactory description of the mechanism that generates claim development - even when the resulting estimates differ from those of the CL or BF method. Generalized linear models extend the range of available modelling options.

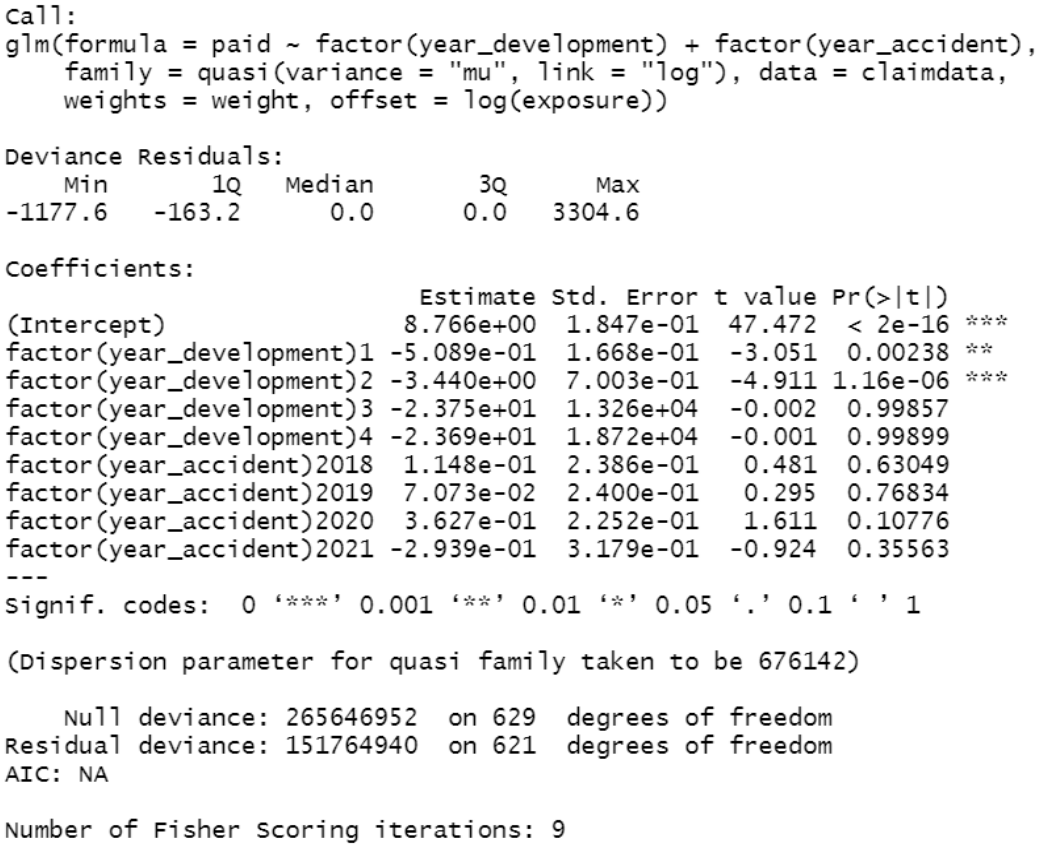
Introduction to the data set

The data set relates short-tailed business, and it includes the following variables:

- **Accident year** is the years in which the accident occurred, ranging from 2017 to 2021. Accident years before 2017 are considered to be fully developed and require no valuation of outstanding claims.
- **Development year** is the delay between the accident year and the payment year, for this dataset the development years are 1 to 5. The development years 4-5 are inactive, as most or all claim payments happen in development years 1-3.
- **Exposure** represents how many persons were exposed to risk in different accident year.
- **Claims Paid** represents incremental payments on the claim between the reporting date and the valuation date.
- **Claims Paid Cumulative** represents the cumulation of the incremental claims paid.
- **Risk group** for this data the risk group is 0 for females and 1 for Males.
- **Calendar year** A 12-month interval beginning in January and ending in December.
- **Valuation year** this is the year in which claims are valued, i.e. 2021.

4.6 Pure CL estimates.

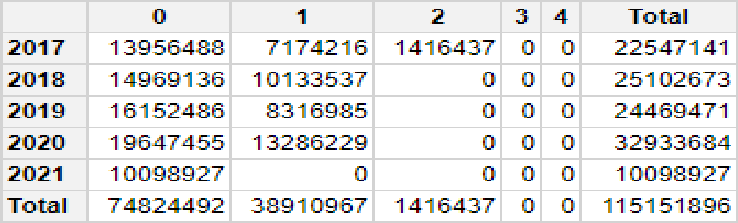
To get the chain ladder estimate via GLM, we model **claims paid** as our response variable and then use the **year_development** and **year_accident** variables in the dataset as independent variables. The family of distribution chosen, is the Poisson distribution, with this distribution the maximum likelihood estimates of the GLM coincides with the estimates from the Chain-ladder method. As a result, a GLM with Poisson distributions is often cited as the model underlying the Chain-ladder method as seen in tables 4.9 and 4.10. The following code shows how to use the glm():



```
Call:
glm(formula = paid ~ factor(year_development) + factor(year_accident),
    family = quasi(variance = "mu", link = "log"), data = claimdata,
    weights = weight, offset = log(exposure))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1177.6   -163.2      0.0      0.0   3304.6

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                8.766e+00  1.847e-01  47.472  < 2e-16 ***
factor(year_development)1 -5.089e-01  1.668e-01  -3.051  0.00238 **
factor(year_development)2 -3.440e+00  7.003e-01  -4.911 1.16e-06 ***
factor(year_development)3 -2.375e+01  1.326e+04  -0.002  0.99857
factor(year_development)4 -2.369e+01  1.872e+04  -0.001  0.99899
factor(year_accident)2018  1.148e-01  2.386e-01   0.481  0.63049
factor(year_accident)2019  7.073e-02  2.400e-01   0.295  0.76834
factor(year_accident)2020  3.627e-01  2.252e-01   1.611  0.10776
factor(year_accident)2021 -2.939e-01  3.179e-01  -0.924  0.35563
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasi family taken to be 676142)

    Null deviance: 265646952  on 629  degrees of freedom
Residual deviance: 151764940  on 621  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 9
```

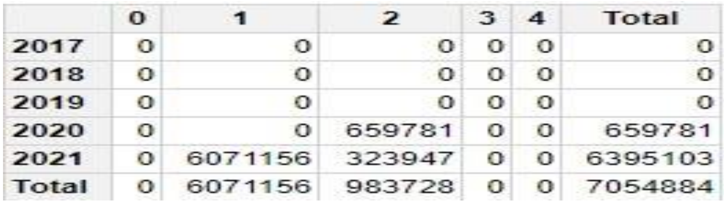
If a predictor variable has a p-value less than the significance level (typically 0.05) in the GLM summary output, it is statistically significant. This indicates that the predictor variable has a significant effect on the response variable, and its coefficient estimate shows the direction and strength of the relationship. In this case, only the intercept, factor(year_development)1, and factor(year_development)2 are statistically significant in the R output.

The estimates derived for the Pure CL Method



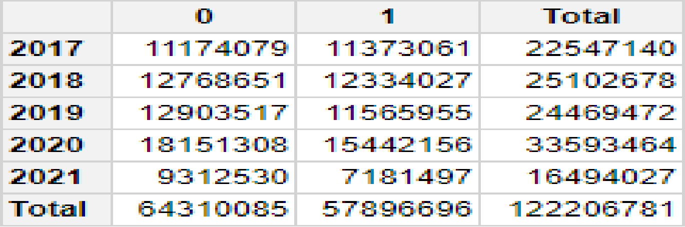
	0	1	2	3	4	Total
2017	13956488	7174216	1416437	0	0	22547141
2018	14969136	10133537	0	0	0	25102673
2019	16152486	8316985	0	0	0	24469471
2020	19647455	13286229	0	0	0	32933684
2021	10098927	0	0	0	0	10098927
Total	74824492	38910967	1416437	0	0	115151896

Table 4.9: Pure Chain ladder using the GLM- cumulative claims paid.



	0	1	2	3	4	Total
2017	0	0	0	0	0	0
2018	0	0	0	0	0	0
2019	0	0	0	0	0	0
2020	0	0	659781	0	0	659781
2021	0	6071156	323947	0	0	6395103
Total	0	6071156	983728	0	0	7054884

Table 4.10: Pure Chain ladder using GLM - outstanding estimates.



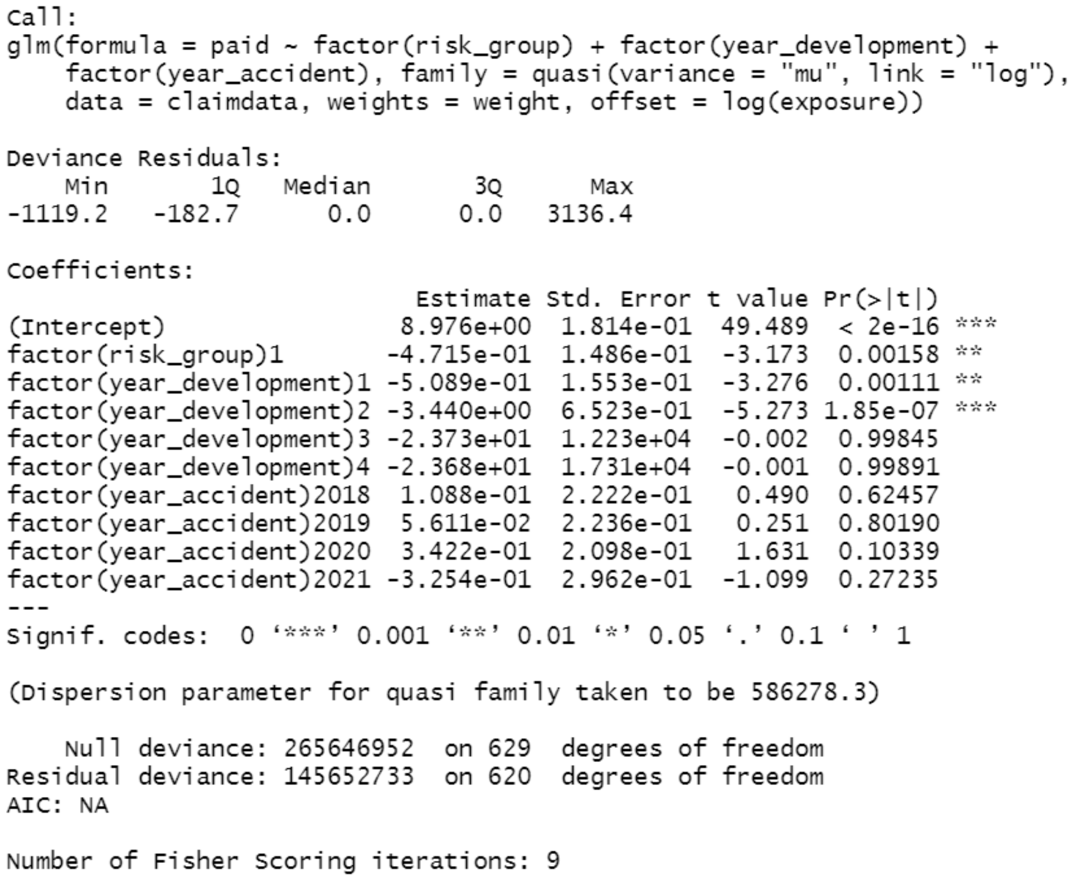
	0	1	Total
2017	11174079	11373061	22547140
2018	12768651	12334027	25102678
2019	12903517	11565955	24469472
2020	18151308	15442156	33593464
2021	9312530	7181497	16494027
Total	64310085	57896696	122206781

Table 4.11 Chain Ladder predictions

Tables 4.9, 4.10, and 4.11 represents the cumulative claims paid run-off triangle, the outstanding estimates, and the ultimate estimates, respectively. In Table 4.11, 0 and 1 represents female and male insureds, respectively. Also, the predictions are “fitted values” and the total column shows a “fitted ultimate cost”. The predictions are the fitted claim cost estimates that are most relevant for premium rating.

4.7 CL estimates with risk_group

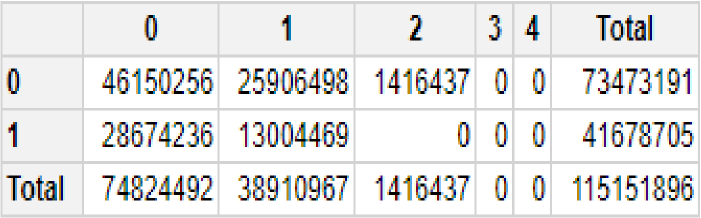
In this model, we model the **claims paid** as our response variable. We then use the **year_ development**, **year_accident** variables and the **risk_group** in the dataset as independent variables. The introduction of the **risk_group** allows the researcher to see which group generated more claims. The result of this model is the CL + risk estimate. The usefulness of the addition of the risk group helps the insurer in designing its products for the various risk group it introduces. This is one of the many useful variations of the GLM method. The following code shows how to use the GLM:



```
Call:
glm(formula = paid ~ factor(risk_group) + factor(year_development) +
    factor(year_accident), family = quasi(variance = "mu", link = "log"),
    data = claimdata, weights = weight, offset = log(exposure))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1119.2   -182.7      0.0      0.0   3136.4

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               8.976e+00  1.814e-01  49.489  < 2e-16 ***
factor(risk_group)1      -4.715e-01  1.486e-01  -3.173  0.00158 **
factor(year_development)1 -5.089e-01  1.553e-01  -3.276  0.00111 **
factor(year_development)2 -3.440e+00  6.523e-01  -5.273 1.85e-07 ***
factor(year_development)3 -2.373e+01  1.223e+04  -0.002  0.99845
factor(year_development)4 -2.368e+01  1.731e+04  -0.001  0.99891
factor(year_accident)2018  1.088e-01  2.222e-01   0.490  0.62457
factor(year_accident)2019  5.611e-02  2.236e-01   0.251  0.80190
factor(year_accident)2020  3.422e-01  2.098e-01   1.631  0.10339
factor(year_accident)2021 -3.254e-01  2.962e-01  -1.099  0.27235
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasi family taken to be 586278.3)

    Null deviance: 265646952  on 629  degrees of freedom
Residual deviance: 145652733  on 620  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 9
```

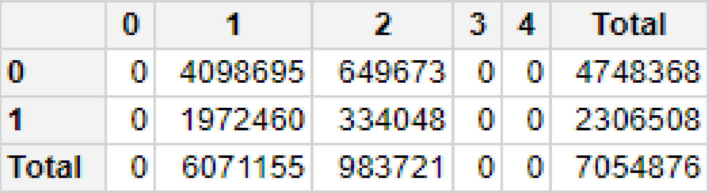
From the GLM summary above, a predictor variable is considered statistically significant if its associated p-value is less than the chosen significance level, typically 0.05. This indicates that the variable has a significant effect on the response variable, and its coefficient estimate can be used to make reliable predictions. In this model, the intercept, factor(risk_group)1, factor(year_development)1, and factor(year_development)2 are statistically significant predictors.

The estimates derived for the Pure chain Ladder with risk_group.



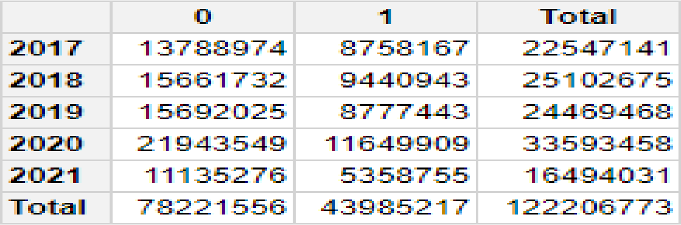
	0	1	2	3	4	Total
0	46150256	25906498	1416437	0	0	73473191
1	28674236	13004469	0	0	0	41678705
Total	74824492	38910967	1416437	0	0	115151896

Table 4.12: Chain ladder using GLM with risk group - cumulative claims paid.



	0	1	2	3	4	Total
0	0	4098695	649673	0	0	4748368
1	0	1972460	334048	0	0	2306508
Total	0	6071155	983721	0	0	7054876

Table 4.13: Chain ladder using GLM with risk group - outstanding estimates.



	0	1	Total
2017	13788974	8758167	22547141
2018	15661732	9440943	25102675
2019	15692025	8777443	24469468
2020	21943549	11649909	33593458
2021	11135276	5358755	16494031
Total	78221556	43985217	122206773

Table 4.14 Chain ladder + risk group prediction

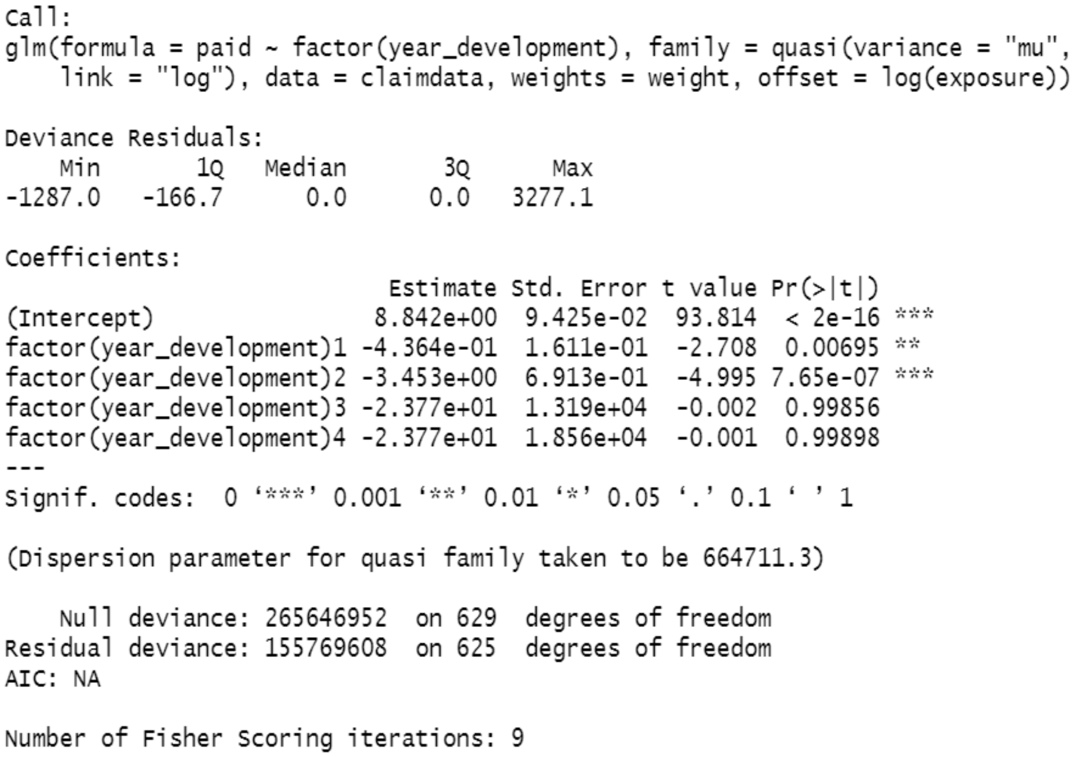
From Tables 4.12, 4.13 and 4.14 represents the cumulative claims paid based on the risk group, the outstanding payments based on the risk group and the prediction based on the risk group, respectively. In Table 4.14, 0 and 1 represents female and male insurers, respectively. Also, the predictions are "fitted values" and the total column shows a "fitted ultimate cost".

4.8 Implication of the Pure (CL) and the CL with risk group

Since there is only one pricing variable (risk_group1), $e^{-0.4715} = 0.62407$ is the factor by which the outstanding payments decreases when we compare male insureds to female insureds. This is true because from table 4.11 and 4.14 we can see that the risk_group 0 which is 1representative of female had higher predictions as compared to risk_group1 which is representative of males. By splitting between risk group the insurer can detect this disparity. Even though the overall predictions are the same in both table 4.11 and table 4.14, the difference in the predicted claim cost by risk group is much clearer in table 4.14 than in table 4.11.

4.9 Pure Bornhuetter – Ferguson estimates

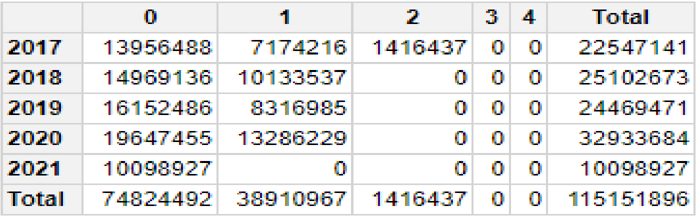
To derive the Pure (BF) estimates, we model the **claims paid** as our response variable. We then use the **year_development** in the dataset as independent variable. The result of this model is the BF estimates. The following code shows how to use the GLM:



```
Call:
glm(formula = paid ~ factor(year_development), family = quasi(variance = "mu",
    link = "log"), data = claimdata, weights = weight, offset = log(exposure))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1287.0   -166.7      0.0      0.0   3277.1

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                8.842e+00  9.425e-02  93.814  < 2e-16 ***
factor(year_development)1 -4.364e-01  1.611e-01  -2.708  0.00695 **
factor(year_development)2 -3.453e+00  6.913e-01  -4.995 7.65e-07 ***
factor(year_development)3 -2.377e+01  1.319e+04  -0.002  0.99856
factor(year_development)4 -2.377e+01  1.856e+04  -0.001  0.99898
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasi family taken to be 664711.3)

    Null deviance: 265646952  on 629  degrees of freedom
Residual deviance: 155769608  on 625  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 9
```

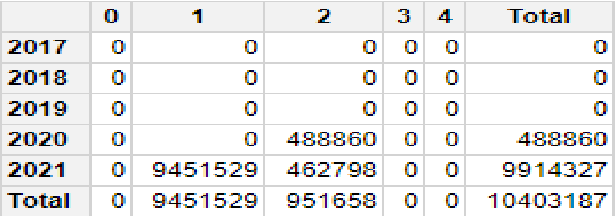
If a predictor variable has a p-value less than the significance level (typically 0.05) in the GLM summary output, it is statistically significant. This indicates that the predictor variable has a significant effect on the response variable, and its coefficient estimate shows the direction and strength of the relationship. In this case, only the intercept, factor(year_development)1, and factor(year_development)2 are statistically significant in the R output.

The estimates derived for the Pure BF Method



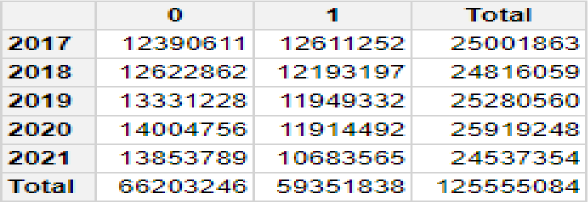
	0	1	2	3	4	Total
2017	13956488	7174216	1416437	0	0	22547141
2018	14969136	10133537	0	0	0	25102673
2019	16152486	8316985	0	0	0	24469471
2020	19647455	13286229	0	0	0	32933684
2021	10098927	0	0	0	0	10098927
Total	74824492	38910967	1416437	0	0	115151896

Table 4.15: Pure Bornhuetter-Ferguson using GLM- cumulative claims paid.



	0	1	2	3	4	Total
2017	0	0	0	0	0	0
2018	0	0	0	0	0	0
2019	0	0	0	0	0	0
2020	0	0	488860	0	0	488860
2021	0	9451529	462798	0	0	9914327
Total	0	9451529	951658	0	0	10403187

Table 4.16: Pure BF using GLM - outstanding estimate.



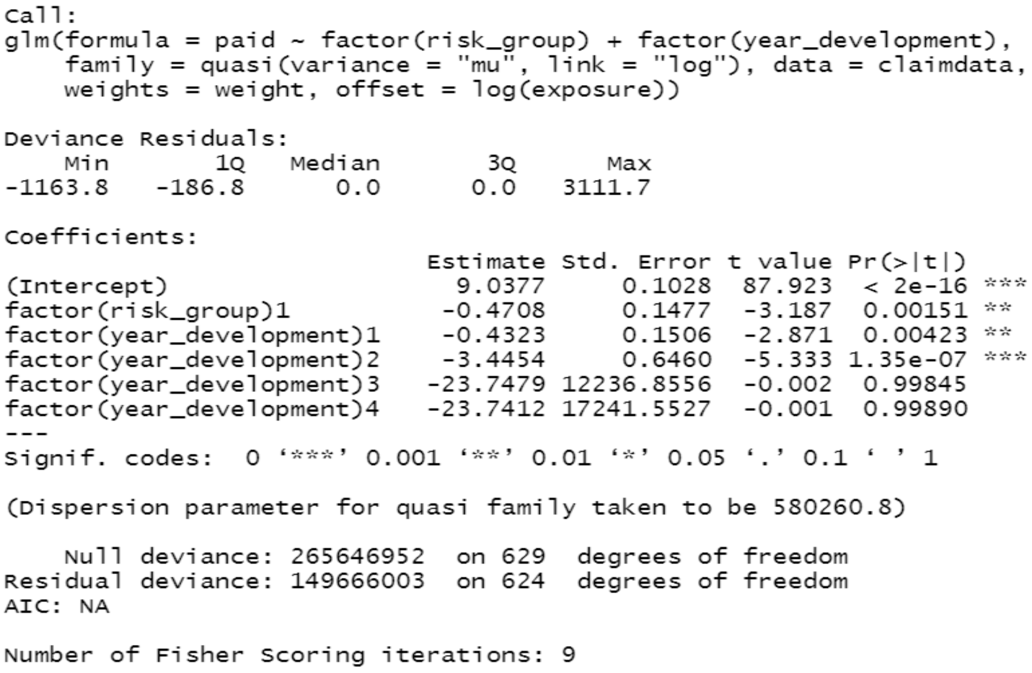
	0	1	Total
2017	12390611	12611252	25001863
2018	12622862	12193197	24816059
2019	13331228	11949332	25280560
2020	14004756	11914492	25919248
2021	13853789	10683565	24537354
Total	66203246	59351838	125555084

Table 4.17: Pure BF Predictions

Tables 4.15, 4.16 and 4.17 the cumulative claims paid run-off triangle, BF outstanding estimates, and the BF predictions, respectively. In Table 4.17, 0 and 1 represents female and male insurers, respectively. Also, the predictions are “fitted values” and the total column shows a “fitted ultimate cost”. The predictions are the fitted claim cost estimates that are most relevant for premium rating.

4.10 BF estimates with risk_group

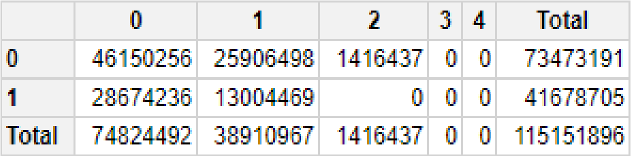
In this last model, we model the **claims paid** as our response variable. We then use the **year_development** and the **risk_group** in the dataset as independent variable. The result of this model is the BF+ risk group estimates. For the insurer to better appreciate the dynamics owing to the claims paid amount, we add the risk group. The following code shows how to use the GLM:



```
Call:
glm(formula = paid ~ factor(risk_group) + factor(year_development), 
    family = quasi(variance = "mu", link = "log"), data = claimdata, 
    weights = weight, offset = log(exposure))

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1163.8   -186.8      0.0      0.0   3111.7  

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                   9.0377     0.1028  87.923  < 2e-16 ***
factor(risk_group)1          -0.4708     0.1477  -3.187  0.00151 ** 
factor(year_development)1    -0.4323     0.1506  -2.871  0.00423 ** 
factor(year_development)2    -3.4454     0.6460  -5.333 1.35e-07 ***
factor(year_development)3   -23.7479 12236.8556  -0.002  0.99845    
factor(year_development)4   -23.7412 17241.5527  -0.001  0.99890    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasi family taken to be 580260.8)

    Null deviance: 265646952  on 629  degrees of freedom
Residual deviance: 149666003  on 624  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 9
```

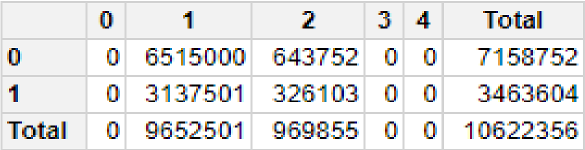
From the GLM summary above, a predictor variable is considered statistically significant if its associated p-value is less than the chosen significance level, typically 0.05. This indicates that the variable has a significant effect on the response variable, and its coefficient estimate can be used to make reliable predictions. In this model, the intercept, factor(risk_group)1, factor(year_development)1, and factor(year_development)2 are statistically significant predictors.

The estimates for BF with risk_group



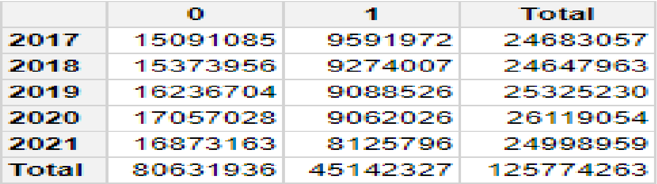
	0	1	2	3	4	Total
0	46150256	25906498	1416437	0	0	73473191
1	28674236	13004469	0	0	0	41678705
Total	74824492	38910967	1416437	0	0	115151896

Table 4.18: BF using GLM with risk group - cumulative claims paid.



	0	1	2	3	4	Total
0	0	6515000	643752	0	0	7158752
1	0	3137501	326103	0	0	3463604
Total	0	9652501	969855	0	0	10622356

Table 4.19: BF using GLM with risk group - outstanding estimates.



	0	1	Total
2017	15091085	9591972	24683057
2018	15373956	9274007	24647963
2019	16236704	9088526	25325230
2020	17057028	9062026	26119054
2021	16873163	8125796	24998959
Total	80631936	45142327	125774263

Table 4.20 BF + risk group prediction

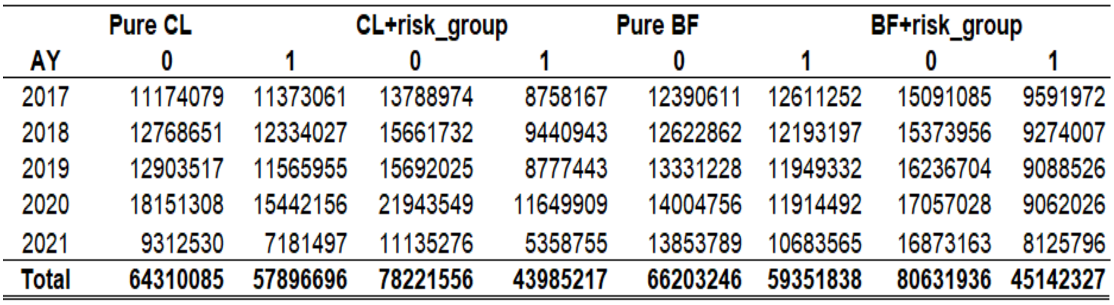
From Tables 4.18, 4.19 and 4.20 represents the BF cumulative claims paid based on the risk group, the BF outstanding payments based on the risk group and the BF prediction based on the risk group, respectively. In Table 4.20, 0 and 1 represents female and male insurers, respectively. Also, the predictions are “fitted values” and the total column shows a “fitted ultimate cost.”

4.11 Implication of the Pure (BF) and the BF with risk group

Since there is only one pricing variable (risk_group), $e^{-0.4708} = 0.62450$ is the factor by which the outstanding payments decreases when we compare male insureds to female insureds. This is true because from Tables 4.17 and 4.20 we can see that the risk_group 0 which is representative of female had higher predictions as compared to risk_group1 which is representative of males. By splitting between risk groups, the insurer can detect this disparity. Even though the overall predictions are similar the same in both table 4.17 and table 4.20 the difference in predicted claim coast by risk group is much clearer in table 4.20 than in table 4.17.

5.CONCLUSION

The table below shows the various estimates based on the model framework specified, that way we demonstrate the value added by including the extra risk variable/s, something which can be done in GLM but not in the original CL or BF methods.

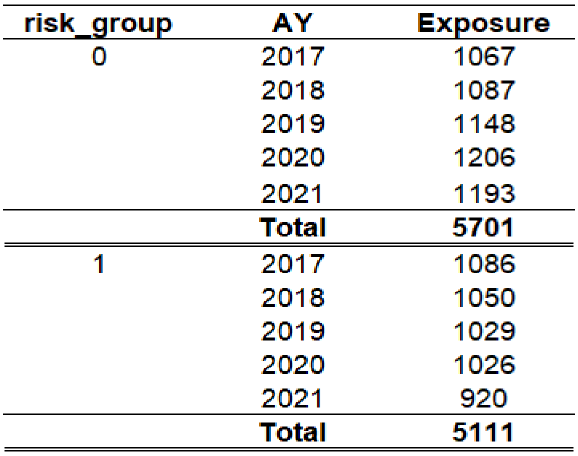


	Pure CL		CL+risk_group		Pure BF		BF+risk_group	
AY	0	1	0	1	0	1	0	1
2017	11174079	11373061	13788974	8758167	12390611	12611252	15091085	9591972
2018	12768651	12334027	15661732	9440943	12622862	12193197	15373956	9274007
2019	12903517	11565955	15692025	8777443	13331228	11949332	16236704	9088526
2020	18151308	15442156	21943549	11649909	14004756	11914492	17057028	9062026
2021	9312530	7181497	11135276	5358755	13853789	10683565	16873163	8125796
Total	**64310085**	**57896696**	**78221556**	**43985217**	**66203246**	**59351838**	**80631936**	**45142327**

Table 5.1 Prediction estimates

The 0 and 1 represents female and male insureds, respectively .We can observe that the BF + risk group gives the strongest distinction between the risk groups predicted ultimate. This is because the BF also takes account of the number of insureds per risk group. Also, the predictions are "fitted values" and the total column shows a "fitted ultimate cost."

In table 5.2 below we see that the number of females is increasing while the number of males is falling, which in part explains the disparity in predicted claim cost.



risk_group	AY	Exposure
0	2017	1067
	2018	1087
	2019	1148
	2020	1206
	2021	1193
	Total	**5701**
1	2017	1086
	2018	1050
	2019	1029
	2020	1026
	2021	920
	Total	**5111**

Table 5.2**:** Number of sums of risk insured.

From the findings of the study, the following conclusion can be made.

We can also see that females have higher claim cost per insured than males, plus that the number of females is increasing while the number of males is falling. It is a typical adverse selection situation that needs to be addressed. By EU regulation, insurers are not allowed to charge different premiums, for males and females buying the same insurance. Insurers can devise different stratagems to circumvent the restriction: for example, by offering products that appeal more to the one sex than the other. Or by charging different premiums for other rating variables that are highly correlated with sex.

In further studies, the model that was generated using GLM approaches can have the age, sum insured per risk and/or premiums added to it to acquire further information about the insurance data.

REFERENCES

Bornhuetter, R. L., & Ferguson, R. E. (1972), 'The actuary and IBNR'. Proc. CAS 59, 181–195.

Christofides, S. (1997), 'Regression models based on log-incremental payments', Claims Reserving Manual, Volume 2, Section D5, Institute of Actuaries, London.

Chambers, M.J., Hothorn, T., Lang, D.T., & Wickham, H. (2015), 'Computational Actuarial Science with R' Taylor & Francis Group, LLC. ISBN-13:978-1-4665-9260-5 (eBook - PDF).

De Jong, P., & Heller, G. Z. (2008), 'Generalized Linear Models for Insurance Data', Cambridge University Press.

England, P. D., and Verrall, R. J. (2002), 'Stochastic claims reserving in general insurance', British Actuary Journal. 8, 443–544.

Frees, E.W. & Valdez, E.A (2008), 'Hierarchical Insurance Claims Modeling', Journal of the American Statistical Association 103(484):1457-1469.

Gesmann M, Murphy D, Zhang Y, Carrato A, Wüthrich M, Concina F, Dal Moro E (2022), 'ChainLadder: Statistical Methods and Models for Claims Reserving in General Insurance', R package version 0.2.16.

URL:<https://mages.github.io/ChainLadder/>.

Gesmann M, Murphy D, Zhang Y, Carrato A, Wüthrich M, Concina F, Dal Moro E (2022), 'ChainLadder: Claims reserving with R'.

URL:https://cran.rstudio.com/web/packages/ChainLadder/vignettes/ChainLadder.html#introduction

Haberman, S., and Renshaw, A. E (1996), 'Generalized linear models and actuarial science'. The Statistician 45, 407–436.

J. Kocovic., Mitrasevic M., Trifunovic, D. (2018), 'Advantages and Disadvantages of Loss Reserving Methods in Non-Life Insurance', Yugoslav Journal of Operations Research.

URL: https://www.researchgate.net/publication/331254258

Klein, N., Denuit, M., Lang, S., Kneib, T. (2014), 'Nonlife ratemaking and risk management with Bayesian generalized additive models for location, scale and shape', Insurance: Mathematics and Economics, Volume 55, pp. 225-249.

Klugman, S., Panjer, H. & Wilmot, G. (2008), 'Loss Models: From Data to Solutions', third. Edition, Wiley, Hoboken, NJ.

Mack, T. (2000), 'A comparison of stochastic models that reproduce chain–ladder reserve estimates', Insurance: Mathematics and Economics. 26, 101-107

Mack, T. (1993), 'Distribution–free calculation of the standard error of chain–ladder reserve estimates', ASTIN Bull. 23, 213–225.

Mack, T., Quarg, G., and Braun, C. (2000), 'The mean square error of prediction in the chain–ladder reserving method', ASTIN Bulletin. 36, 543–552

Mario V. Wüthrich. (2022), 'non-life insurance: Mathematics & Statistics',

URL: https://ssrn.com/abstract=3491790.

Neuhaus, W. (2004), 'On the estimation of outstanding claims', Australia Actuary Journal., 10, 485–518.

Neuhaus, W. (2014), 'Outstanding Claims in General Insurance', Unpublished manuscript. Universidade Técnica de Lisboa (UTL).

Schmidt, K. D. (2006). 'Methods and models of loss reserving based on run–off triangles', A unifying survey. In: CAS Forum Fall 2006, pp. 269–317.

Schmidt, K.D. (2004), 'Credibility Modelle: Grundlagen. In: Handbuch zur Schadenreservierung', pp. 71–80. Karlsruhe: Verlag Versicherungswirtschaft.

Schmidt, K. D., and Zocher, M. (2008). 'The Bornhuetter–Ferguson principle', Variance 2, 85–110.

Sattayatham, P. & T. Talangtam (2012), 'Fitting of finite mixture distributions to motor insurance claims', J. Math. Stat. 2012;8(1):49–56. DOI: 10.3844/jmssp.2012.49.56

S. M. Burney, L.M. Khan, S. Burney & M. Humayoun (2012), 'Data Analysis and Modeling of Claim Amounts of Car Insurance using Big Data', A Study for Pakistan. Asian Journal of Probability and Statistics 19(4): 46-53, 2022; Article no. AJPAS.91702 ISSN: 2582-0230.

Taylor, G. C., & McGuire, G. (2004), 'Loss Reserving with GLMs – A Case Study'. CAS Discussion Paper Program, pp. 327– 391.

Taylor, G. C. (2000), 'Loss Reserving – An Actuarial Perspective', Boston, Dordrecht, London: Kluwer Academic Publishers, cop. 2000

Yakubu M. B. (2019), ‘Stochastic methods loss reserving with individual claim size modelling’. **URL**: http://ugspace.ug.edu.gh

Appendices

A.1 Selected probability distribution used to fit our claims paid amount data and their parametric estimates and loglikelihood

This section of Appendix A.1 is dedicated to outlining the continuous distribution functions (CDFs), and probability density functions (PDFs) of the selected probability distribution for fitting the data.

1. Log-Normal Distribution

The pdf of the lognormal distribution is given as $f(x) = \frac{1}{\sigma\sqrt{2\pi}}\frac{1}{x}e^{\left\{-\frac{1}{2}(\frac{\log x - \mu}{\sigma})^2\right\}}, x > 0, (\mu, \sigma^2)\ \sigma > 0$. Whereas the cdf is given by $\Phi(z),\ z = \left(\frac{log\, x - \mu}{\sigma}\right)$.

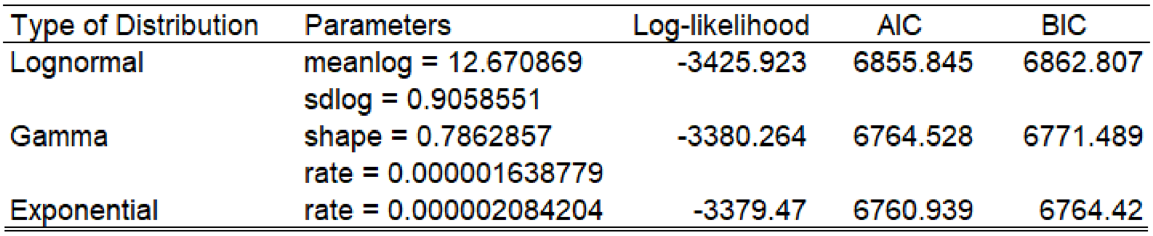
2. Gamma distribution

The gamma distribution pdf is given as $f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x},\ x > 0,\ \alpha > 0,\ \lambda > 0$. The cdf is obtained when 2α is an integer, probabilities for the gamma distribution can be found using the relationship: $2\lambda x \sim \chi^2_{2\alpha}$. α represents the shape and λ is the rate.

3. Exponential distribution

The pdf of the exponential distribution is given as $f(x) = \lambda e^{-\lambda x},\ x > 0,\ \lambda > 0$. The cdf is $F(x) = 1 - e^{-\lambda x}$. λ represents the rate.

Table A.1 summarizes the estimated parameters, log-likelihood, and their respective information criteria values for the selected fitted probability distribution used for claims paid amount data.



Type of Distribution	Parameters	Log-likelihood	AIC	BIC

Lognormal	meanlog = 12.670869
	sdlog = 0.9058551
		-3425.923	6855.845	6862.807

Gamma	shape = 0.7862857
	rate = 0.000001638779
		-3380.264	6764.528	6771.489

Exponential	rate = 0.000002084204	-3379.47	6760.939	6764.42

Table A.1

This table shows the gamma and the exponential distribution as a better fit for the claims paid amount based on their likelihood. Since, the exponential is a particular case of the gamma distribution the gamma distribution is chosen among the selected distribution as the better fit than the exponential distribution. The AIC and BIC also show relatively equivalent results for the gamma and the exponential distribution. AIC and BIC cannot determine how well a model explains data. It is only able to identify if the model balances complexity and predictive ability.

A.2 Basic R codes used for estimating the various statistics, getting the basic plots and the fitting of the probability distribution for the claims paid data.

Table A.2 represents the descriptive statistics obtained using R. This table shows all the codes to derive the mean, standard deviation, and coefficient of variation.

```
# Importing the dataset into R
claims_d <= read.csv("Glm_loss_reserving_csv_20220815.txt", header = T)
# Loading additional Libraries
library(evd)
library(evir)
library(latticeExtra)
library(actuar)
library(TSA)
library(fitdistrplus)
# Getting the claims paid amount only
claim_s <= claims_d$claims_paid_incremental
# Removing zero and negative values*
new_claim_s <= Claim_s
New_Claim_s[New_Claim_s < 0] <= 0
New_Claim_s<-New_Claim_s1
New_Claim_s1[New_Claim_s1 == 0] <= NA
# Calculating the various summary statistic
(mean(New_Claim_s1))
(var(New_Claim_s1))
(sd(New_Claim_s1))
cov <= sd(New_Claim_s1)/mean(New_Claim_s1);cov
```

Table A.2

Table A.2.1 represents the codes used in plotting the scatter plot, histogram and Boxplot of both the original and log of the claims paid amount data.

```
par(mfrow=c(1,2))
plot(x=New_claim_s1, xlab="Counts", ylab="Claim size", main="Scatter Plot of Claim size")
plot(x=log(New_Claim_s1), xlab="Counts", ylab="Log Claim size", main="Scatter Plot of Log Claim size")
hist(x=New_Claim_s1, xlab="Claim sizes", ylab="Counts", main="Histogram for Claim size")
hist(x=log(New_Claim_s1), xlab=" logged Claim sizes", ylab="Counts", main="Histogram for logged of Claim size")
```

```
boxplot(x=New_Claim_s1, ylab="Claim size", main="Box Plot of Claim size")
boxplot(x=log(New_Claim_s1), ylab="log(Claim size)", main="Box Plot of Logged Claim size")
```

Table A.2.1

Table A.2.2 represents the codes used in fitting the probability distribution for the claims paid data.

```
fw1<-fitdist(sort(New_Claim_s1), distr = "lnorm", method = c("mme"), discrete = F); plot(fw1)
fw2<-fitdist(sort(New_Claim_s1), distr = "exp", method = c("mme"), discrete = F); plot(fw2)
fw3<-fitdist(sort(New_Claim_s1), distr = "gamma", method = c("mme"), discrete = F); plot(fw3)
```

Table A.2.2