MODELING DEPENDENCE BETWEEN REPORT LAG AND CLAIM AMOUNTS USING COPULA MODELS

SHARON ORINA AMAYI

I56/75993/2014

A RESEARCH PROJECT SUBMITTED TO THE DEPARTMENT OF MATHEMATICS IN PARTIAL FULFILMENT OF THE REQUIREMENTS OF THE AWARD OF MASTER OF SCIENCE IN ACTUARIAL SCIENCE OF THE UNIVERSITY OF NAIROBI

March 2016

DECLARATION

This research project is my original work and has never been presented and will never be presented for a degree award in any other university.

Sign ……………………… Date ………………………

Sharon Orina Amayi

I56/75993/2014

This research project has been submitted for examination with my approval as University Supervisor.

Sign …………………….. Date ………………………...

Prof. Patrick Weke

DEDICATION

I dedicate this study to all the people who have played a role in aiding me to reach this far in my studies including my family, my lecturers and my friends.

ACKNOWLEGEMENT

I wish to acknowledge the great help I have received in preparing this study. I owe a huge debt of gratitude to my family and in particular my parents; Dr. Jairus Boston Amayi and Emily Pheminah Amayi who have always offered their support in terms of provision on moral support throughout my studies. My brother and sister; Maureen and Aaron and my cousins for all the moral support they have accorded me.

I also wish to acknowledge all my lecturers and in particular my supervisor; Professor Patrick Weke for the advice and invaluable input throughout my studies

Last but not least I wish to give my regards to the Almighty God without which all this would be impossible to begin with.

ABSTRACT

Relationships between two or more variables are considered a phenomenon of interest in a world where modelling risk is becoming more and more popular. This is especially important for the insurance sector whose core business is protecting individuals from occurrences of risk. Some of the risks insurance companies face includes holding inefficient reserve amounts for claims policyholders take time to report.

Having a variable that can explain the behavior of another can prove an important aid in understanding the variable of interest. In the case of insurance companies; establishing the relationship that claim amounts have to the time policyholders take to report the claim could help establish how much should be kept aside for claims not yet reported.

This relationship is described as dependence between variables. The most common measure used to quantify dependence between variables is the Pearson's correlation coefficient. This is a measure that requires the use of the covariance between the variables and their individual variances. However, the Pearson's correlation coefficient is a measure that assumes linear dependence between variables. This limits the effectiveness of its use as a measure; since it cannot explain dependence in the case of a non-linear relationship. Furthermore the Pearson's correlation coefficient is only a single figure and therefore limits the amount of information we can derive from it concerning the dependence between the variables. This leads us to the use of copulas as a measure of dependence between variables. Copulas; being distributions themselves; have the advantage of being able to portray more information concerning the dependence structures between the two variables.

The following is a study that seeks to establish the relationship between claim amounts and the report delay period for claims in an insurance company using copulas.

TABLE OF CONTENTS

List of Tables

List of Figures

ACRONYMS

IBNR	Incurred But Not Reported
ALAE	Allocated Loss Adjustment Expenses
FML	Full Maximum Likelihood
IFM	Inference Function for Margins
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
SIC	Schwarz Information Criterion

CHAPTER ONE: INTRODUCTION

1.1 Background of the Study

A simple definition of an insurance company is a company that makes a contract with an individual whereby in exchange for a certain payment or series of payments will indemnify the individual on occurrence of a given event that causes a certain loss. Brown (1988) defines a loss reserve as the amount an insurance company sets aside to settle claims. Qaiser (2006) argues that working out adequate and appropriate reserves is a very important aspect in the functioning of an insurance company. Qaiser (2006) goes on to emphasize that an insurance company at any given point in time must be in a position to honor its promise of indemnifying losses as can be seen foreseen reasonably.

A possible result of inefficient reserves held by an insurance company could lead to its insolvency. This has been demonstrated by a study done by A M Best insolvency report in 2004 concerning the sources of insurance companies' involuntary exit in America. The study identified that 37% of failures of insurance companies is attributed to deficient loss reserves and inadequate pricing. The financial services authority in UK also analyzed experiences of failed insurance companies across 15 countries in the European Union based on the Sharma report (2002) and concluded 60% of the companies showed poor underwriting and reserving as a contributing factor.

There are different types of reserves that insurance companies hold in order to remain solvent. Gile (1994) classifies claims reserves in insurance companies into three types; case estimates for reported claims, reserves for additional development of reported claims and reserves for claims incurred but not reported. Gile (1994) goes on to demonstrate that for calculation of the incurred but not reported reserve for an insurance company, three variables play a huge role. These variables include; the count of claims the claim amounts and the report lags.

Correlational research is used to see if two variables are related and to make predictions based on this relationship. Wolley (1997) argues that one way of explaining how certain events predict an outcome is by measuring how predictive variables are, when measured together. With this argument in mind, establishing a relationship between two of the variables suggested by Gile

(1994) could go a long way into developing predictive methods of estimating the incurred but not reported claims reserve in an insurance company.

The idea of Copulas can be dated back to 1940s where Hoeffding (1940) studied standardized distribution functions with support in the square $[-1/2, 1/2]^2$ and margins in the interval [-1/2, 1/2]. He obtained the best-possible bounds for these functions and also studied dependence measures that are invariant under strictly increasing transformations. Fretchet (1950) also obtained the same results; and from these came the Fretchet-Hoeffding bounds.

Sklar (1959) first used the term ‘copula’ when he proposed the Sklar theorem; which shows the relationship between multi-variate distributions and copulas by joining the marginal distributions of the individual variables. Sklar and Schweizer (1986); later went on to develop the Archimedean copulas; which are the most common type of copulas. Since that time a lot of studies have been done and many copulas proposed; to study the different types of dependence structures.

While a lot of work has been done on the copulas that explain positive dependence between variables, not as much work has been done on the negative dependence between variables. Some of the work that has been done on negative dependence includes; Hua (2015) who proposed a copula that explained negative tail dependence. However; it is important to note that the paper focused on variables where the extreme values; referred to as the tail values, rather than all the values, showed negative dependence. Stander (2011) also contributed to the work on negative dependence by seeking to find multi-variate copulas that explain negative dependence between more than two variables from bivariate copulas.

This study sought to add to the work already done on copulas by focusing on copulas that explain negative dependence between variables. The paper focused on Archimedean copulas already established to be able to explain the dependence between the time it takes an individual to report a claim and the amount of loss in that claim. In particular, the research focused on copulas used to model; weak negative dependence, those that only allow negative dependence and those that model all types of dependence

The study proposed that in the insurance industry setting; when a claim amount is large; the delay period for reporting is expected to be short and when the claim amount is small; the delay period is expected to be longer. The study sought to establish to what extent this is true and how strong or weak the dependence will be if any.

1.2 Statement of the Problem

The main problem that this study sought to address was; finding a way to understand claim amounts based on their relationship with their report lag. Such a relationship would aid in improvement of the predictive techniques in existence. One of the major issues facing the insurance industry today is the inability to effectively predict claims. Due to this, there exists a huge challenge in determining efficient reserves for claims that are yet to be reported or settled. Various studies have been done with regard to this problem to determine to what extent under reserving would cost an insurance company. One such study was conducted by A.M. Best and Financial Services Authority which showed; under reserving exposes insurance companies to the risk of being insolvent. This implies that there is a need to improve on the existing predictive techniques to help take care of this issue in the insurance industry.

1.3 Objectives of the Study

1.3.1 General Objective

The general objective of the study was to ascertain an effective copula that explains the dependence structure between the delay it takes for an insured party to report a claim and the claim amount.

1.3.2 Specific Objectives

The study was guided by the following specific objectives;

i. To fit marginal distributions to the report lag and claim amount variables for an insurance company.

ii. To fit different copulas to the report lag and claim amount variables using the marginal distributions obtained.

iii. To test the goodness of fit of the copulas in order to determine the copula that best fits the data.

1.4 Hypothesis of the study

The study proposed the null hypotheses that; for a typical insurance company,

i. The report lags and claim amount variables can be fit to individual marginal distributions.
ii. There exists dependence between the report lag and claim amount variables that can be modeled using copulas
iii. Different copulas represent the report lag and claim amount dependence to different extents and therefore some copulas can be considered to explain the dependence better than others.

1.5 Scope and Limitations of the Study

This study was aimed at the insurance sector and in particular the general insurance industry. The study sought to establish a relationship between the report lag variable and claim amounts in a given insurance company. As such the study was not necessarily limited geographically; and as such can be applied anywhere in an insurance setting; however it is expected that different insurance companies may result in different results in accordance to their respective claims experience. For purposes of illustration this study considered one insurance company in Kenya in which the methodology was applied.

One of the limitations experienced in the study was the definition of the claim amount variable. An ideal definition of the claim amount variable for the study would be the exact amount expected to be paid once a claim is reported. However; in practice it is quite difficult to ascertain this exact amount and as such most insurance companies apply a certain random estimate based on their judgement and the type of claim and revise the amount later once assessments and valuations have been completed. Data can only be retrived for the initial estimate; since it may prove difficult to obtain information on the revised amounts given that revisions for a specific claim could occur more than once. However; given that the study sought to establish a relationship between the report lag and the claim amount variable based the initial perception of the loss estimate; the initial estimates provided by the insurance companies during reporting of the claim proved to be useful estimates. Hence this study made use of the initial estimate as the claim amount variable.

It is worth noting that study was also not able to cater for any human errors at the data entry level and the models proposed are also subject to model and parametric errors.

1.6 Justification of the study

This study sought to find a suitable explicit copula that can be used to explain the relationship between report delay period and claim amount and therefore would be useful for further research in trying to find expected claim amounts from the past delay periods information.

Negative dependence is an aspect of dependence that occurs in the business environment. Understanding negative dependence is therefore important to be able to understand and cope with these instances. This study provides an analysis of negative dependence from the 'Archimedean-Copula' point of view and therefore contributes to adding as a valuable reference guide in this field.

As one of the measures of preparedness of the insurance sector, the insurance regulators have put guidelines that compel insurance companies to hold certain amount of reserves. The calculation of such reserves is based on past information and methods such as the Basic Chain Ladder method, the Bornheutter Ferguson method and just recently the Average Delay Method. The Basic Chain Ladder method (and to an extent the Bornheutter Ferguson method) is a method that studies the development patterns of claims to be able to evaluate what claims should be expected in the future. For one to obtain a dependable pattern there is usually need to remove the extreme claim amounts that distort the pattern. However to be able to realistically prepare the company from claims one must also take into account these extreme cases. One way to cater for these claims is to study the relationship between the delay periods and the claim amounts.

Methods such as the average delay method could directly use the results obtained from copulas explaining the relationship between the delay period and claim amount; since copulas tend to explain the dependence structure in a better way than other common measures of dependence.

CHAPTER TWO: LITERATURE REVIEW

Dependency in the insurance sector is well known concept especially from the actuarial perspective. Dependence in the sector has been considered between different variables that may relate to each other in one way or another such as; different claim types that are related in their occurrence, individual risks in a given portfolio where some dependence may occur due to exposure of similar risks and the relationship between loss and allocated loss expenses to mention a few.

In order to effectively predict possible future occurrences, it is important to understand how different variables interact with each other. Different approaches have been proposed to model different relationships between the variables of interest in the sector. The copula method is one of the various ways of modeling dependence and as such is not new in the insurance sector applications. Through time, studies have been done into how efficient copulas are as a measure of dependence. The following is a review on the modeling of dependence in the insurance sector and attention is given to the report lag variable that is being considered in this research. The report lag variable also referred to as the report delay variable; is the variable that represents the time difference between when a claim was incurred and when it was reported to the insurance company. The review also considers the role that copulas have played in dependence in insurance.

2.1 Theoretical Literature Review

The concept of dependence between variables has been extensively studied throughout the years with a number of theories being developed for its study. For purposes of this study; we review two theories that are a favorite for studying dependence; the Bayesian Inference theory and the Sklars theorem-copula theorem.

2.1.1 Bayesian Inference Theorem

The Bayes theorem was first introduced in the 1740s by Thomas Bayes who wanted to know how to infer causes from effects. The theorem was later rediscovered independently by Pierre Simon Laplace who gave it its modern mathematical form and scientific application. The theorem demonstrates how the probability of a given occurrence given some information can be expressed in terms of the initial estimate of the probability of the information given the occurrence and the probability of the occurrence. The theorem is stated in terms of joint, conditional and marginal distributions of the variables of interest. Many studies in different disciplines have proposed use of this theorem to study dependence. Examples of studies using the method in the insurance sector are demonstrated in the empirical review.

2.1.2 Sklars Theorem

Sklars theorem was introduced in 1959, by Abe Sklar. This theorem suggests use of copula models to demonstrate dependence between different variables. The theorem proposes that the joint distribution of two or more variables can be obtained by combining the individual marginal distributions with function s known as copulas that satisfy certain criteria. A lot of studies have been proposed to study dependence in different sectors using this theorem since its introduction. The following study is based on this theorem as the aim is to seek a suitable copula to explain dependence between the report lag and claim amount variables in insurance.

2.2 Empirical Literature Review

An application of dependence in insurance is demonstrated by Dhaene and Goovaerts (1996), who propose a dependence model that is able to calculate stop loss premiums while catering for any dependence between risks in a portfolio. Stop loss insurance is insurance that protects insurers against large claims. The insurers do not pay beyond a given retention limit of a loss that occurs. The authors estimate the expected maximal stop loss claim amounts based on assumptions that if a given individual survives, then all individuals with a probability of survival greater than that given individual will surely survive and vice-versa. This idea enables us to modify the possible events of the sum of individual risks posed by the portfolio and creates a pattern from which we can deduce the distribution function for this modified sum of risks. The modified sum enables us to determine the stop loss premium since it gives rise to expected maximal stop loss claim amounts. The dependency between individual risks is modeled in this manner and stop loss premium is determined from that dependence. The model discussed is applicable for a fixed number of two point distributed risks.

One of the contributions to the literature of modeling the report lag of claims independently in an insurance company is made by Weissner (1999); who proposes a truncated exponential distribution to model report lag. He proposes that the report lag of claims observed could be modeled with an exponential distribution conditioned on the unknown parameter of the distribution. To cater for claims incurred but not yet reported; and therefore assumed to have a longer report lag than observed; a distribution function of the report lag is used; whose upper limit is the maximum report lag of the claims observed. Using Bayes result the distribution of the report lag conditioned on both the unknown parameter and the maximum report lag of claims observed is estimated. This resulting distribution is a conditional truncated exponential distribution and using the method of maximum likelihood applied to claims data an estimate of the unknown parameter of the exponential distribution is obtained. It is mentioned that this method can also be applied to other distributions such as the log-normal distribution.

The concept developed by Weissner (1999) on use of the exponential distribution for the report lag variable is further considered by De Souza and Veiga (2014); who propose a stochastic model to estimate IBNR. The model is expressed in terms of the distribution of the report lag of claims with the same occurrence day given the total number of claims in that day. The joint distribution of the reported delay, claims reported and claims occurred is obtained from the Bayes result; with the delay distribution conditioned on claims reported in the period and total claims that occurred in the period, the distribution of frequency of claims observed in the period and the distribution of total claims occurred in the period. While the Poisson distribution is used to model the frequency of claims and a binomial distribution to model claims reported given the number of claims that occurred, two different distributions are considered to model the time to report a claim (report lag); Simple exponential distribution and a mixture of two exponential distributions.

Estimation of parameters for the simple exponential distribution was based on maximizing the likelihood function of the delay distribution. From the three distributions used to obtain the report delay distribution; it can be observed that the number of occurred claims and delays of claims not reported are not known; this necessitated the application of the Expectation Maximization method, which is an iterative method to find estimates of maximum likelihood parameters in a statistical model where the model depends on unknown variables. The mixture of exponential distribution entailed; fitting a mixture of exponential distribution to the delay distribution rather than a simple exponential and maximizing the log likelihood to obtain the parameters. Having obtained the parameters, the IBNR is estimated as the expectation of the difference between the claims occurred and claims reported in an occurrence period and the total of these is the IBNR. A difference between the paper by De Souza and Veiga (2014) and Weissner (1999) is that while De Souza and Veiga (2014) proposed modeling the report lag jointly with variables that may have a correlation with it, Weissner (1999) considered modeling the report lag independently.

A paper that combines the concept of De Souza Veiga (2014) and Weissner (1999) by modeling the report lag independently but using this independent distribution to model the various joint distributions the report lag may have with other variables is by Walther Neahaus (2004); who proposes an approach to the problem of estimating outstanding claims based on three dimensions taking into account the time it takes to report an incurred claim and the time since reported to valuation separately. Most two dimension models combine these two delays into one development triangle. The author proposes modeling occurrence of accidents, with a certain delay period using Poisson process whose parameter depends on the reporting delay; which is assumed to have a fixed pattern of delay probabilities. He proposes using credibility estimates of the chain ladder method and a prior mean of the parameter of the Poisson distribution to obtain the credibility estimate of the parameter for the Poisson process. Having this parameter the Poisson distribution is fully determined and the IBNR claim number becomes the portion of claims in the distribution with a report delay larger than time already observed. The claim severities are assumed independent with a distribution that may also depend on the reporting delay distribution. The aggregate claim severities for claims not yet reported is taken to be a compound Poisson with frequency parameter dependent on the Poisson process mentioned for claims occurred with given reporting delays and a mixed severity distribution mentioned above. The estimate of the IBNR therefore becomes the mean of this compound distribution. The idea in this method shows dependence may not only exist between the report lag and the claim severities, but also between the report lag and the claim frequencies.

Another study of how the report lag variable may be modeled jointly with another variable is considered by Verrall et al (2010), who model incurred but not reported claims and reported but not settled claims by focusing the possible delays a claim suffers when incurred; the report delay and the settlement delay. Run off triangles are developed for claims paid data and number of claims reported and a stochastic model is proposed using assumptions of individual claims. The number of claims incurred but not reported is obtained from the claims reported runoff using the usual development factors. However the data for RBNS needs to consider the number of claims paid and their delay period. A multinomial distribution is proposed to model the number of claims reported and paid with a given delay period conditioned on the number of claims reported. An over-dispersed Poisson model is then fitted to the paid claims and parameters estimated using the maximum likelihood method. The estimates of outstanding claims are obtained by summing the predicted values of incremental claims and this is done using bootstrapping and Bayesian methods.

A limitation posed by Verrall et al (2010) in modeling RBNS is that claims were assumed to be paid in one lump sum which is not usually the case. Schiegl (2010) took this into account when modeling the relationship between report delay and number of claims. Schiegl (2010) proposed a three dimensional stochastic model to estimate incurred but not reported and the reported but not settled claims. The 3-dimensional approach involves looking at the computation of the reserves required from three aspects; claims occurrence, claims reporting and claims payment. The number of claims in an occurrence year is modeled using a Poisson distribution. She then opted to model the number of claims occurred conditioned on the report delays using a multinomial distribution with a parameter for each delay period. To take into account active claims which include; claims incurred and reported but not yet fully settled, a survival process is proposed resulting in a binomial distribution for the number of active claims with a particular delay period. The claims paid amounts was then modeled using a gamma distribution and the aggregate claims paid modeled using the individual risk model. The aggregate claims paid data and the number of active claims paid with a given report delay distributions allowed for computation of expected claims paid for claims incurred but not reported with report delay beyond the date of valuation. She went on to demonstrate the application of the model using data from a Monte Carlo simulation.

As mentioned copula is a favorite for modeling dependence. One application of copula in modeling dependence in insurance is by considering claim types and the influence they might have with each other. This is illustrated by Frees and Valdez (2007); who propose a hierarchical approach to modeling loss insurance. The hierarchical arrangement consists of three components; frequency of claims, types of claims and severity of claims. The authors proposes that rather than the normal univariate modeling of claim loss, consider cases where the event can trigger more than one type of claim and treat it as a multivariate model. The main idea expressed in the paper is that claim amounts can be modeled as a joint distribution of the claim frequency, claim types and the claim severity. The authors propose that this joint distribution can be obtained as a product of the claims frequency distribution, the claim type distribution conditioned on the claim frequency and the claim severity distribution conditioned on the claim type and the claim severity which is the Bayes result.

Frequency of claims is modeled with the negative binomial regression model with use of covariates that may affect the frequency of claims. The type of claims was modeled using the multinomial logit model taking into account all possible combination of the claim types. The severities of claims; after being categorized into their respective claim types, was modeled using the generalized beta of the second kind long tailed distribution for each claim type. Then taking into effect the influence which can be recognized as dependence that certain claim types have on others; copula models are considered. Specifically the Normal copula and the t distribution copulas were considered with the generalized beta distributions as the margins. An illustration of the model described is then applied to detailed micro-level automobile insurance records with the help of maximum likelihood method to fit the models. The different types of claims we have in the automobile insurance sector include; injury to third party, third party property damage, insured’s own vehicle damage and within this context one claim can give rise to more than one type of claim.

Another illustration of the application of copula in the insurance sector is by Christian Genest et al (2002); who show how the Compound Poisson distribution can be used to approximate total claim amount in the context of individual risk models; where dependency between individual contracts arises. They show how dependence can be incorporated in the individual risk model using three approaches; General multi-class shock model where the number of classes; also referred to as risks; in which the portfolio can be categorized and shock variables depending on whether the entire portfolio or a given risk would be affected by the shock are applied. Our interest is the second approach which is the single class risk model based on copulas. In this case, they considered a portfolio whose individual policies are under the same risk. The risk variable is represented as a product of the possible loss amount and a Bernoulli variable dependent on whether the risk occurs or not. The concept behind the second approach is modeling dependence between the Bernoulli variables which is the dependence between chances of an event in two different policies using copula. Focus is given to the Frechet-mixture model and Archimedean copulas. The final approach considered is the multi class risk models based on copulas; this is an extension of the previous single case approach.

Studies have also been conducted in the area of negative dependence modelling using Copulas. Stander (2011) was a research paper aimed at construction of multi-variate Archimedean copulas from bivariate copulas. The paper focused on multivariate copulas capable of modelling negative dependence. This idea was first proposed by Joe (2007) though was not applied for any specific Archimedean copulas. Stander (2011) aimed at deriving the necessary constraints to ensure that when building multivariate copulas from bivariate copulas that capture negative dependence that the multivariate copula will also be able to capture negative dependence. She considers a method proposed by Joe (1997) using nested generator functions to model the multivariate copula. The parameters of the generator developed are then determined through numerical techniques. The method is applied to two Frank generator functions, two Clayton copulas, and two of the copulas proposed by Nelson (2006). The model is also extended to consider two different copulas to nest. An application of the model is illustrated in the finance sector; in computation of asset allocation and in wrong way risk in counter-party credit exposure. The paper describes in detail regarding copulas and in particular; copulas that model negative dependence. Five of the copulas described are considered to model the report lag and claim amount relationship in this research.

Perhaps the research closest home to the focus of this paper was a method proposed by Weke and Ratemo (2013) on use of copula in calculating IBNR. The dependence between the report delay and claim amount was modeled with Archimedean copulas; Clayton, Frank and Gumbel. Having fit log-normal distributions to the two variables of interest; report delay and claim sizes, the three copulas were fit to the data. The Clayton copula proved the best fit for the data and an estimation of the IBNR was then proposed to be the product of the average claim size obtained from the copula and the average number of claims obtained by fitting a distribution to the number of claims.

Modeling the relationship between report lag and claim severities one should also take into account the different characteristics of the variables. Insurance companies do not necessarily suffer the entire losses of all claims. It is common insurance practice to have policy limits. This is the maximum amount a policyholder can claim, previously referred to as stop loss insurance. Losses in an insurance company are also usually cushioned by re-insurers and therefore insurance companies do not necessarily settle large claims themselves. This brings in the idea that claims data should be censored to the limit. Hence the claim severities in insurance for instance should take into account censoring for such cases. Two papers that have taken this concept into modeling this dependence are by Frees and Valdez (1997) and Denuit et al (2006)

Frees and Valdez (1997) introduce the aspect of copula in the actuarial sector as a useful tool for understanding the relationships among multivariate outcomes. They apply the concept of copulas to modeling the loss-allocated loss adjustment expenses (ALAE) relationship. ALAE are the expenses associated with settling of claims in insurance Frees and Valdez fitted the marginal distributions using the Pareto distribution for the ALAE and to cater for censoring in the loss variable; the Kaplan Meier empirical distribution for the loss variable. They fitted the Frank Gumbel Hougaard copula and they used the inference maximum likelihood method (IML) to find parameters of the copula.

Taking censoring into account when modelling dependence using copula was also illustrated by Michel Denuit et al (2006) who also proposed use of Archimedean copulas to model the relationship between loss and the allocated loss insurance in general insurance. They improve on a concept proposed by Genest and Rivest (1993) that involved obtaining the best fit of copula for data by comparing a parametric estimate of a function of the generator function of the Archimedean copula with a non-parametric estimate of that same function from the data. They argue that it is possible to obtain a non-parametric estimate of the function of the generator that takes a given limit into account. He applied the methodology described to fit the Clayton, Gumbel, Frank and Joe copulas to insurance data.

The copula method has also been extended to combining conditional distributions rather than just marginal distributions. Applications of such copulas have been demonstrated by Purwono (2005) in the insurance sector; specifically life insurance. Copula has been proposed as a better tool in modelling joint life assurances. The assumption of independence between lives has been disputed especially in cases where the two lives being considered are spouses; since they are exposed to many similar risks as a result of living together and having the same lifestyle. Purwono (2005) was a paper that showed how the theory of copulas could be applied to multiple lives analysis. He proposed use of a conditional Bayesian copula model whose construction is based on application of copula mixing to conditional rather than to marginal survival functions.

Copulas with conditional marginals have played a huge role in their application even in the finance sector. One such application of copula in finance was by; Patton (2004) which was a research paper that sought to explore the application of copula especially in economics and finance. The paper included a study on conditional copulas; conditioning a copula on one variable. He later applied this concept in modeling the Deustch mark – US dollar and the Yen – US dollar mark; where the copula was conditioned on the US dollar. The method proposed in the research by Patton (2004) could also be applied to the calculation of IBNR in insurance. By conditioning the claim amount and the number of claims on the report delay, one could find a copula that describes the relationship between claim amounts and number of claims given a certain report lag.

In summary dependence in the insurance sector is an area where much has been done. However modifications of the various ideas proposed could further the field of study. For instance Weissner (1999) proposed an exponential distribution to study the distribution of the report lag based on the method of maximum likelihood. A similar concept is proposed in this study. However this study sought to apply a discrete distribution to model the report lag variable in this study taking the lag as days-hence discrete. The truncation method demonstrated by Weissner (1999) could also be studied for the negative binomial distribution. A lot has been done in modeling distributions of related variables even without the copula method. Neahaus (2010) for instance proposed use of the compound distribution to model dependence between number of claims unreported and claim amounts. The copula method could also be applied to the two variables instead. However a number of papers reviewed have demonstrate the use of copula in dependence; Frees and Valdez (2007) for instance relate two variables with a copula based on their individual marginal distributions. The following research is an application of this concept to the variables report delay and claim amount in the insurance sector. The idea of copulas could also further be extended in calculation of the IBNR rather than just the dependence. The model developed by Stander (2011) for instance could also be considered to model IBNR claims in the insurance sector especially if three variables; report lag, claim amount and number of claims are used to develop the multivariate copula. The conditional copula described by Purwono (2004) and Patton (2002) could also be applied by conditioning the number of claims and the claim amounts on the report lag. This would propose a model that demonstrates the claim amount and number of claims with a given report delay.

Various limitations have also been observed in modeling the dependence between report delay and claim amounts. An example; Verrall et al (2010) describes the limitation of using reported claims paid rather than paid claims data; reported claims are merely estimates of the loss suffered while claims paid is the actual loss suffered. However there is also another disadvantage with the claims paid data; payments of claims are rarely made as a lump sum, rather they are made as a series of payments. It is due to this disadvantage that the model proposed in this research opts for reported claims data rather than claims paid. The disadvantage of reported claims amount is catered for by taking into consideration that the model that we seek to develop is purely meant to estimate how long an individual would take to report a claim based on the initial perception of the claim loss which is the initial estimate. It is important to note a model that used current estimates; estimates adjusted with payments made, would be an even better model than both the cases described above.

The Schiegl (2010) model proposed poses a solution for claims that are not paid as per a lump-sum with the survival probability in the binomial distribution. It also offers use of claims paid data rather than report paid data which are merely estimates and therefore a better realistic model. However it may prove difficult though not impossible to obtain data aggregated to the extent described in the model from insurance records. Furthermore, the dependence implied in this model was between the number of claims and the report delay; our model on the other hand seeks to establish the dependence between the report delay and the claim amount.

2.3 Conceptual Framework

As has been mentioned, the main aim of this study was to determine the relationship between variables report lag and claim amount using copula. This involved obtaining the marginal distribution of the report lag and that of the claim amount. Then using a copula; the two distributions were joined to obtain a distribution that shows how the two variables relate with each other.

CHAPTER THREE: COPULA AND DEPENDENCE MEASURES

3.1 Dependence

Dependence is a statistical term used to explain the relationship that exists between two or more given variables. Variables are said to be dependent if the movement of one affects another. This can be in the case that an increase (decrease) in one variable results in an increase (decrease) in another; this is referred to as positive dependence. The case where an increase (decrease) in one variable leads to a decrease (increase) in another is referred to as negative dependence. Independent variables are variables whose movement has no effect on the other.

In order to understand just how dependent two or more variables are, one has to be able to measure that dependence. Common measures of dependence include;

- Concordance measures
- Quadrant dependence
- Tail dependence

3.1.1 Concordance Measures

Concordance is described as the agreement between two or more variables. A pair of random variables is concordant if large (or small values) tend to occur together. Consider the case of a pair of variables; say (X_1, Y_1) and (X_2, Y_2)

If $(X_1 - X_2)(Y_1 - Y_2) > 0$, then the pair (X_1, Y_1) and (X_2, Y_2) are said to be concordant.

If $(X_1 - X_2)(Y_1 - Y_2) < 0$, then the pair (X_1, Y_1) and (X_2, Y_2) are said to be discordant.

If $(X_1 - X_2)(Y_1 - Y_2) = 0$, then the pair (X_1, Y_1) and (X_2, Y_2) are said to be neither concordant nor discordant.

The measures used to determine concordance between variables include;

- Pearson's correlation coefficient
- Kendall's tau
- Spearman's rho

3.1.1.1 Pearson's correlation coefficient

This is the simplest and most common measure of concordance in use. It simply requires the covariance between two variables and their individual variances. It measures the linear correlation between the variables and this has proved to be one of its major weaknesses. It is expressed as;

$$\rho = \frac{cov(x, y)}{\delta x \delta y}$$

(3.1)

It will also prove impossible to obtain the Pearson's correlation coefficient if either of the variables has an infinite variance. This is another weakness that the Pearson's correlation coefficient has.

3.1.1.2 Kendall's tau

Another common measure of concordance is the Kendall's tau. The population version of Kendall's tau is the probability of concordance minus the probability of discordance for a pair (X_1, Y_1), (X_2, Y_2). The population version of Kendall's tau is;

$$\tau = prob[(X_1 - X_2)(Y_1 - Y_2) > 0] - prob[(X_1 - X_2)(Y_1 - Y_2) < 0]$$

(3.2)

Computation of Kendall's tau for a given sample requires one to find count of the concordant and discordant pairs. It is expressed as;

$$\tau = \frac{N_c - N_d}{\frac{1}{2}n(n-1)}$$

(3.3)

3.1.1.3 Spearman's rho

Spearman's correlation coefficient is a statistical measure of the strength of a monotonic relationship between paired variables. A monotonic relationship is a causal relationship between two variables whereby if one increases the other will either increase or decrease.

Let (X_1, Y_1), (X_2, Y_2) and (X_3, Y_3) be independent random vectors with a common joint distribution function H. The population version of the Spearman's rho is the difference between the probabilities of concordance and discordance of the vectors (X_1, Y_1) and (X_2, Y_3) (Kruskal 1958).

The calculation of a samples' Spearman's rho requires one to first find the ranks of the variables; and then compute the Pearson's correlation coefficient on the ranked values (ranks).

Spearman's rho of X and Y is in-fact the linear correlation of $F_1(x)$ and $F_2(y)$.

3.1.2 Positive Quadrant Dependence

Positive quadrant dependence is a specific type of dependence such that;

$$Prob(X > x, Y > y) \geq Prob(X > x)\, Prob(Y > y)$$

In simple terms, this implies that the probability that two random variables are jointly large is greater than or equal to when they are looked at independently.

3.1.3 Tail dependence measure

Tail dependence measure is a dependence measure that looks at the concordance between extreme values of the two variables. It is a measure of dependence in the upper right and lower left quadrant of the distribution. The definition is divided into two parts; one for the upper and one for the lower.

Definition: Upper tail dependence

Consider two random variables X and Y with cumulative distribution functions F_1 and F_2 respectively. The coefficient for upper tail dependence of X and Y is;

$$\lambda_u = limit_{u \to 1^-} (Prob[Y > F_2^{-1}(u) / X > F_1^{-1}(u)])$$

(3.4)

Provided a limit λ_U exists and is between 0 and 1, then X and Y are said to be asymptotically dependent in the upper tail. If the limit is equal to 0, then X and Y are said to be asymptotically independent.

Definition: Lower tail dependence

Consider again two random variables X and Y with distribution functions F_1 and F_2 respectively. The coefficient for lower tail dependence of X and Y is;

$$\lambda_l = limit_{u \to 0^+} \left(Prob[Y \leq F_2{}^{-1}(u) / X \leq F_1{}^{-1}(u)] \right)$$

(3.5)

Provided a limit λ_L exists and is between 0 and 1, then X and Y are said to be asymptotically dependent in the lower tail. If the limit is equal to 0, then X and Y are said to be asymptotically independent.

3.2 Copula

A copula is a function which joins or "couples" a multivariate distribution function to its one dimensional marginal distribution functions (Roger B. Nelson (2006)).

Sklar (1953) defined a copula by showing the relationship between the copula and the joint distribution function the copula by joining the marginal distribution functions.

I.e. Let *H* be an n dimensional distribution function with margins; $F_1, F_2, \ldots\ldots\ldots F_n$. Then a copula *C* exists such that for all 'x' in $\mathbb{R}^n$ it follows that:

$$H(x_1, \ldots, x_n) = C(F_1(x_1), \ldots, F_n(x_n))$$

(3.6)

Similarly, we can express the copula in terms of the distribution function as;

$$C(u_1, \ldots, u_n) = H(F_1^{-1}(u_1), \ldots, F_n^{-1}(u_n))$$

(3.7)

Definition: Quasi-inverse

The quasi-inverse of a function is defined as $F^{[-1]}$ such that;

$$F(F^{[-1]}(t)) = t$$

(3.8)

$$F^{[-1]}(t) = inf(x|F(x) \geq t) = sup(x|F(x) \leq t)$$

(3.9)

In this form, the copula is expressed as a joint distribution of random variables uniformly distributed between 0 and 1.

It is also worth noting that the density function $c\left(u_1, u_2, \ldots\ldots\ldots, u_n\right)$ associated with a copula $C\left(u_1, u_2, \ldots\ldots\ldots, u_n\right)$ is defined as;

$$c(u_1, \ldots, u_n) = \frac{\partial^n C(u_1, \ldots, u_n)}{\partial u_1 \ldots . \partial u_n}$$

(3.10)

3.2.1 Properties of copulas

To explain the properties of copulas; we consider a bivariate case where the copula is derived from two variables. For the bivariate case, a copula C is a function $C: I^2 \rightarrow I$ such that;

1. Copula functions are grounded

   This property implies that $C(0, v) = C(u, 0) = 0$

   Consider a graphical representation;

2. Copula functions are 2-increasing

   This is basically a two dimension version of non-decreasing functions in one dimension. Mathematically, this property can be represented as; C is 2-increasing: for a, b, c, d in I with $a \leq b$ and $c \leq d$ if;

   $C(b, d) - C(a, d) - C(b, c) + C(a, c) > 0$

3. When one variable is equal to 1, the copula is equal to the other variable

   I.e. $C(1, x) = C(x, 1) = x$ for all x in I

4. Copula functions must lie between the Frechet-Hoeffding bounds

$$W(u, v) = max\{u + v - 1, 0\} \leq C(u, v) \leq min\{u, v\} = M(u, v)$$

Frechet-Hoeffding bounds

The last property is an inequality known as the Frechet-Hoeffding bound inequality. Hoeffding (1940) and Fretchet (1950) proposed that for a copula to be considered valid, it must be within these bounds;-

$$W(u, v) = max(u + v - 1, 0) \leq C(u, v)) \leq min(u, v) = M(u, v)$$

(3.11)

The functions *W* and *M* are known as the lower and upper bounds respectively.

W and *M* are also copulas by themselves and along with another copula $\pi = UV$ known as the product copula plays an important role in explaining copulas. These three copulas have the following interpretations;

- The copula of *U* and *V* is *M (u, v)* if and only if *U* and *V* is almost surely an increasing function of the other.
- The copula of *U* and *V* is *W (u, v)* if and only if *U* and *V* is almost surely a decreasing function of the other.
- The copula of *U* and *V* is π if and only if *U* and *V* are independent.

These three copulas form a class of copulas known as Fundamental copulas. With the fact that copulas have to be within the Frechet-Hoeffding bounds, we note that while some cover both bounds, others fail to cover all. Copulas that are capable of modelling all three copulas mentioned above are described as comprehensive copulas.

Consider the following graphs for the three fundamental copulas:

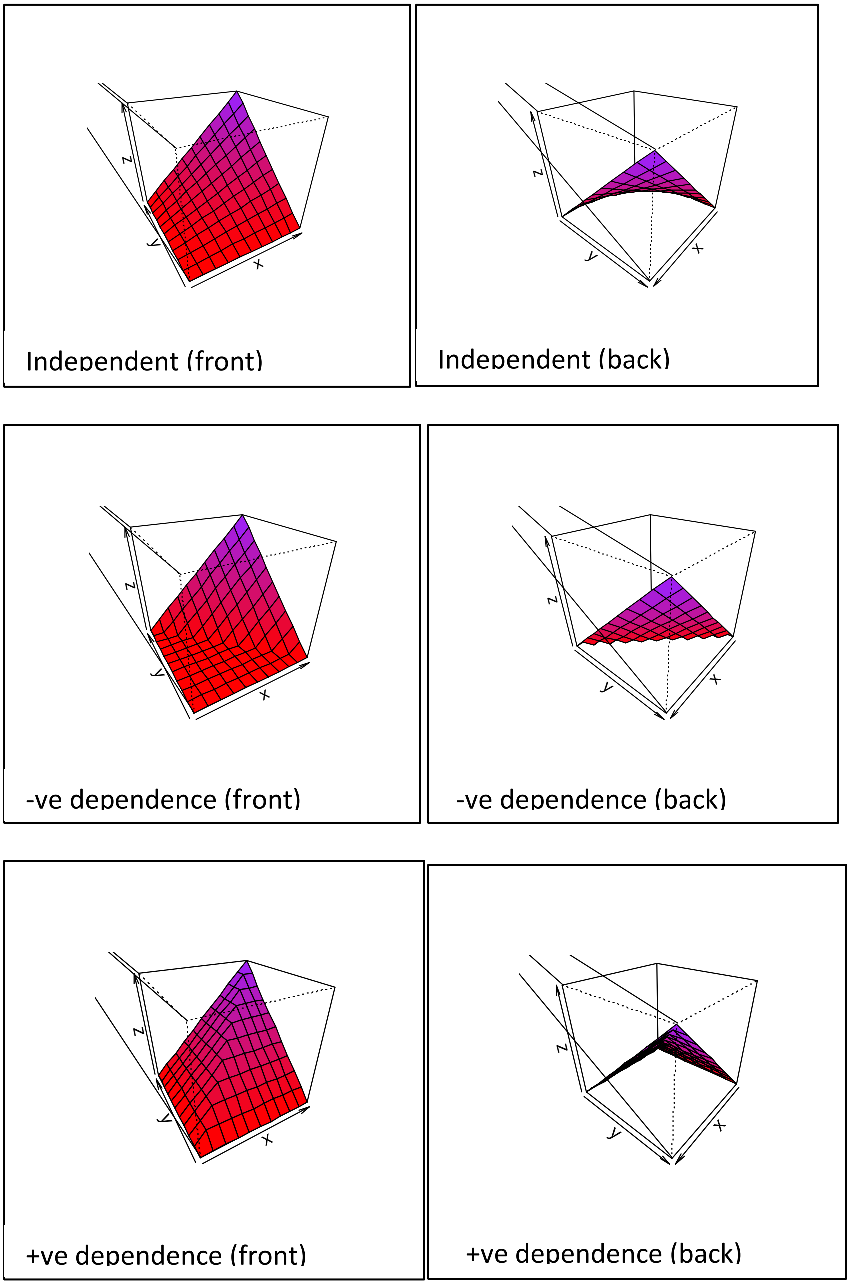

Figure 3.1: Graphical representation for the Frechet Hoeffding bounds and the independence copula

Copulas are considered better in modelling dependence because of their flexibility and the various types (variety) of dependence they allow for.

3.2.2 Copulas and dependence measures

It is important to note that one of the reasons that copulas are a favorite for measuring dependence is that copulas can easily be connected to the dependence measures already discussed. Later it is mentioned how these connections are improved for the specific case of Archimedean copulas.

Just to mention a few of these connections, and considering the dependence measures earlier discussed, we have;

Kendall's Tau: - Note that Kendall's tau is related to copulas in general in the following way;

$$\tau = 4\int\int C(u,v)dC(u,v) - 1$$

or

$$\tau = 4E[C(u,v)] - 1$$

(3.12)

Spearman's *rho:* - Spearman's rho is related to copulas in the following ways;

$$\rho s = 12\int\int uvdC(u,v) - 3$$

$$\rho s = 12\int\int C(u,v)\,dudv - 3 = 12E[UV] - 3$$

$$\rho s = \rho(u,v) = \rho(F(x), F(y))$$

(3.13)

Positive Quadrant dependence: - In terms of copula; two variables are considered positive quadrant dependent if;

$$C(u,v) \geq uv$$

(3.14)

Upper tail dependence: - Consider the definition of upper tail dependence with regards to copula; if a bivariate copula *C* exists such that;

$$\lambda_U = \lim_{u->1^-} \frac{1 - 2u + C(u,u)}{1-u}$$

(3.15)

If λ_U exists and is greater than 0 but less than or equal to 1 then *C* has upper tail dependence and if λ_U is 0 then *C* has upper tail independence.

Lower tail dependence: - If a bivariate copula *C* exists such that;

$$\lambda_L = \lim_{u->0^+} \frac{C(u,u)}{u}$$

(3.16)

If λ_L exists then C has lower tail dependence as long as λ_L is greater than 0 but less than or equal to 1 and has lower tail independence if λ_L is equal to 0.

3.2.3 Types of Copulas

Copulas can be categorized into three types:

- Fundamental - These copulas represent perfect positive dependence, independence and perfect negative dependence. These are the Frechet-Hoeffding bounds along with the independence copula.
- Implicit - These copulas are extracted from well-known multivariate distributions and do not have closed form expressions. An example is the Gaussian copula.
- Explicit (Archimedean) - These are simple closed form expressions and follow general **mathematical construction** to yield copulas.

For purposes of this study, we focus on the Archimedean family of copulas.

3.3 Archimedean Copulas

These are copulas that are derived from mathematical functions known as generator functions. Generator functions are represented by $\psi(t)$ and have specific properties as discussed below.

3.3.1 Properties of the generator function include;

- Continuous
- Decreasing
- Convex function with $\psi(1) = 0$

Consider $\psi(t)$; a continuous and strictly decreasing generator function from *I* to *[0, ∞)* that $\psi(1) = 0$. The Archimedean copula function *C* is then given by;

$$C(u, v) = \psi^{[-1]}(\psi(u) + \psi(v))$$

(3.17)

Definition: pseudo-inverse

The pseudo inverse function is defined as;

$$\psi^{[-1]}(t) = \left\{\psi^{(-1)}(t), 0 \leq t \leq \psi(0)\right\}$$

(3.18)

Note that if $\psi(0) = \infty$, then ψ is said to be a strict generator and it follows that the resulting copula will be a strict Archimedean copula.

3.3.2 Construction of the Generator Functions

There are different ways to come up with the generator functions; a common one is the Laplace transform approach.

Consider the inverse of the generator function; $\psi^{[-1]}(t)$, and let *G* be a distribution function for a non-negative random variable *(Z)*. Then the relationship between the two is the Laplace transform;

$$\psi^{[-1]}(t) = \int_0^\infty e^{tz} dG(z)$$

(3.19)

The inverse $\psi^{[-1]}(t)$, of the generator has the following properties;

- $\psi^{[-1]}(-t)$, is the moment generating function of the random variable Z.
- $\psi^{[-1]}(t)$, is continuous and strictly decreasing with $\psi^{[-1]}(0) = 1$ and $\psi^{[-1]}(\infty) = 0$
- $\psi^{[-1]}(t)$, is a completely monotonic function.

The Clayton copula generator is considered to have been derived from the Laplace transform approach where the random variable Z was from a gamma distribution with parameters (α, 1/α).

Genest and Rivest (1993) pointed out a common feature of some copulas that appeared to be special cases of the independence and minimum copulas.

3.3.3 Properties of Archimedean copulas

- Archimedean copulas are symmetric in nature; i.e. $C(u, v) = C(v, u)$
- They are associative in nature; i.e. $C(C(u, v), w) = C(u, C(v, w))$
- For any constant $k>0$, $k\,\psi(t)$ is also a generator

3.3.4 Archimedean copulas and dependence measures

One of the attractive features of Archimedean copulas is that they are easily related to dependence measures.

For the special case of Archimedean copula, the ***Kendall's tau*** is related to the generator function in the

Following way;

$$\tau = 1 + 4\int_0^1 \frac{\psi(t)}{\psi'(t)}\,dt$$

(3.20)

It is also worth noting that for Archimedean copulas, tail dependence can be expresses in terms of the generators. This was demonstrated by Joe (1997). If ψ is a strict generator such that ψ^{-1} belongs to the class of Laplace transforms for strictly positive random variables. If $\psi'(0)$ is finite and different from zero then the Copula does not have tail dependence. If $\psi^{-1'}(0)$ is finite, then the copula from that generator function does not have upper tail dependence.

If $\psi^{-1'}(0) = -\infty$ then the copula has ***upper tail dependence*** and it's coefficient of upper tail dependence is given by;

$$\lambda_U = 2 - 2 \lim_{s->0^+} \frac{\psi^{-1'}(2s)}{\psi^{-1'}}$$

(3.21)

Similarly for ***lower tail dependence***, consider the same generator function ψ as above, then, the coefficient of lower tail dependence is;

$$\lambda_L = 2 \lim_{s->\infty} \frac{\psi^{-1'}(2s)}{\psi^{-1'}(s)}$$

(3.22)

However, as mentioned previously, we focus on Archimedean copulas that model negative dependence. Hence we consider the following;

- The Clayton Copula
- Frank's Copula
- The Gumbel-Barnett Copula
- Nelson No. 7
- Nelson No. 10

3.3.5 The Clayton Copula

This is an Archimedean copula also known as the Pareto family of copulas. It was first introduced by Clayton (1978). Its generator function is taken to be;

$$\psi(t) = \frac{1}{\alpha}\left(t^{-\alpha} - 1\right)$$

(3.23)

The Clayton Copula distribution is;

$$C(u, v) = max\left\{\left(u^{-\alpha} + v^{-\alpha} - 1\right)^{-\frac{1}{\alpha}}, 0\right\}$$

(3.24)

This copula;

- Is strict for $\alpha > 0$
- Is comprehensive

The Clayton copula is related to Kendall's tau in the following way; $\tau = \frac{\alpha}{\alpha+2}$.

The Clayton copula does not have upper tail dependence but has lower tail dependence with coefficient; $\lambda_L = 2^{-1/\alpha}$.

3.3.6 Frank's Copula

The Frank's copula (1979) is an Archimedean copula whose generator function is;

$$\psi(t) = -ln\frac{\left(e^{-\alpha t} - 1\right)}{\left(e^{-\alpha} - 1\right)}$$

(3.25)

The distribution function for the Frank's copula is;

$$C(u, v) = -\frac{1}{\alpha}ln\left(1 + \frac{\left(e^{-\alpha u} - 1\right)\left(e^{-\alpha v} - 1\right)}{\left(e^{-\alpha} - 1\right)}\right)$$

(3.26)

The Frank's Copula;

- Is a strict copula
- Is comprehensive
- Is radially symmetric

Earlier, we had mentioned that Archimedean copulas are symmetric. However, the Frank Copula has the property of is radially symmetric in addition to the fore-mentioned plain symmetry property. Radial symmetry implies that; $C\ (u,\ v) = u + v - 1 + C\ (1 - u,\ 1 - v)$. A radially symmetric copula has equivalent upper and lower tail dependence.

The Frank's copula allows for negative dependence when α is negative and positive dependence when α is positive. The value of α also affects the level of dependence for the Frank's copula. A higher level of α implies more dependence.

The connection between Spearman's rho and the Franks copula is;

$\rho s = 1 - 12[D_2(-\alpha) - D_1(-\alpha)]/\alpha$ where $D_k(\alpha)$ is the Debye function represented as;

$D_k(\alpha)= \frac{k}{\alpha^k}\int_0^{\alpha}\frac{t^k}{e^{k}-1}dt$

Note the connection between Kendall's tau and the Franks copula is; $\tau = 1 - 4[1 - D_1(\alpha)]/\alpha$ with the same Debye function.

3.3.7 The Gumbel-Barnett Copula

This is an Archimedean copula whose generator function is;

$$\psi\,(t) = ln\,(1 - \alpha ln\,(t))$$

(3.27)

The copula distribution is;

$$C\,(u,v) = uv \cdot exp\,\{-\alpha ln\,(u)\,ln\,(v)\}$$

(3.28)

The Gumbel-Barnett copula;

- Is strict
- Is not comprehensive with only one limit; $C_0 = \pi$

This copula does not exhibit tail dependence. The Gumbel-Barnett copula also only has the capability of modelling weak negative dependence.

3.3.8 The Nelson No. 7

This is an Archimedean copula whose generator function is;

$$\psi(t) = -ln(\alpha t + 1 - \alpha)$$

(3.29)

The copula distribution function is

$$C(u,v) = max\{(\alpha uv + (1-\alpha)(u+v-1)), 0\}$$

(3.30)

This copula is;

- Not strict
- Not comprehensive since it only has two limits; $C_1 = \pi$ and $C_0 = W$

The limits of this copula imply that this copula only allows for negative dependence.

3.3.9 The Nelson No. 10

This is an Archimedean copula whose generator function is;

$$\psi(t) = ln\left(2t^{-\alpha} + 1\right)$$

(3.31)

The copula distribution function is;

$$C(u,v) = \frac{uv}{1 + (1-u^{\alpha})(1-v^{\alpha})}$$

(3.32)

This copula is;

- Strict
- Not comprehensive with only one limit $C_0 = \pi$

This copula does not exhibit any tail dependence and only allows for weak negative dependence.

3.4 Estimation of parameters

Having mentioned the copulas considered for this study, the study then sought to estimate the parameters for the marginal distributions and the copulas. There are different approaches that can be considered in estimating the parameters. One can either use non-parametric methods or parametric methods.

3.4.1 Non-Parametric (Empirical) method

This refers to estimation of parameters from the observations themselves rather than making initial assumptions about the distributions of the data and later checking for the validity of the assumption.

The main advantage of this method is that aspect of it does not make prior assumptions but rather based its findings purely on the data.

3.4.2 Parametric methods

There are various parametric methods that can be applied to data and to mention a few, we will describe two; full maximum likelihood estimation (FML) and Inference for margins (IFM).

Full maximum likelihood estimation (FML)/Exact maximum likelihood method

In this approach we maximize the log likelihood with an aim of finding the parameters of the marginal distributions and the copula simultaneously.

This method is considered computationally exhausting and we therefore opt for the inference for margins method.

Inference functions for margins (IFM)

In this method parameters for the marginal distributions are computed individually and the results are used to compute the parameters for the copula.

3.5 Goodness of Fit Tests

Once the parameters for the marginal distributions and for the copulas are determined, we the study then sought to find the goodness of fit of the different copulas. This helped in determining the best copula for the data and hence the properties that the data has.

For the goodness of fit, we consider two tests;

- The Akaike Information Criterion (AIC)
- The Schwarz Information Criterion (SIC)

3.5.1 The Akaike Information Criterion

The Akaike information criterion was proposed in Akaike (1973) as a simple and versatile method for determining the suitability of statistical models. It is a method based on earlier works such as Neyman and Pearson (1928, 1933), Wald (1943) and Kullback (1959) among others.

Kullback (1951) proposed determining the fitness of a model using the Kullback-Leibler (1951) information quantity. This is a measure of the distance between the model of interest and the true model (observations). He proposed that minimizing this distance would help determine just how helpful the model would be in explaining the observations.

The Kullback-Leibler information quantity is a measure based on the expectation of the log likelihood of the distribution proposed and that of the log likelihood of the true distribution.

Akaike (1977) proposed a new measure (mean log likelihood) as an estimate of the Kullback-Leibler information quantity.

From these the Akaike Information Criterion statistic was introduced;

$$AIC = -2Likelihood + 2k$$

(3.33)

Where k represents the number of parameters.

3.5.2 The Schwarz Information Criterion

This criterion was derived by Schwarz (1978). It is also commonly known as the Bayesian Information Criterion. The SIC computation is linked to Bayesian statistics but is considered more favorable at times since it does not require the input of the prior.

The Schwarz Information Criteria statistic is given by;

$$SIC = -2Likelihood + k \cdot ln\left(n\right) \quad \textbf{(3.34)}$$

Where k represents the number of parameters.

CHAPTER FOUR: NUMERICAL RESULTS

The method proposed was applied to data from an insurance company. The data considered was the claims reported in a period of one month. The data was arranged on a per claim basis and therefore the date of loss, date reported and amount per claim was easily attainable.

The report delay was obtained as the number of days between when the claim was incurred and when the claim was reported.

The amount considered in this case was the original estimate of the claim. The original estimate of the claim was preferred to claim amount paid since the premise of our hypothesis is that individuals who are under the impression of a large loss report earlier. Therefore the report delay is dependent on the first estimate of the loss.

The data statistical results for the two variables under consideration; report delay and claim amount were as follows;

	Report delay (days)	Claim amount (KShs)
Mean	56.24	93786.44
Standard deviation	138.30	94866.09
Skewness	5.03	1.28
Kurtosis	28.98	2.62
Sample size	79.00	79.00

Table 4.1: General Statistics for the Variables

The correlational results for the two variables were as follows;

Pearson's correlation coefficient: **-0.07965**

Kendall's tau: **-0.2971**

Spearman's rho: **-0.4068**

4.1 Fitting Marginal Distributions

4.1.1 Report Lag Distribution

The report delay was considered as a discrete distribution in terms of days. With this in mind the two favorable discrete distributions based on the comparison of their pdf curves and the histogram of the data were; the negative binomial distribution and the geometric distribution.

The negative binomial distribution

Using the maximum likelihood method to fit the negative binomial distribution to the report delay data, the following results were obtained;

	Parameter estimate	Standard error
Size parameter	0.4725	0.0634
Mu parameter	56.2458	9.2457

Table 4.2: Results for the negative binomial fit to the report lag variable

The fit that the negative binomial distribution had to the data could also be illustrated by super imposing the negative binomial density curve to the histogram of the observed data as shown in the following graph;

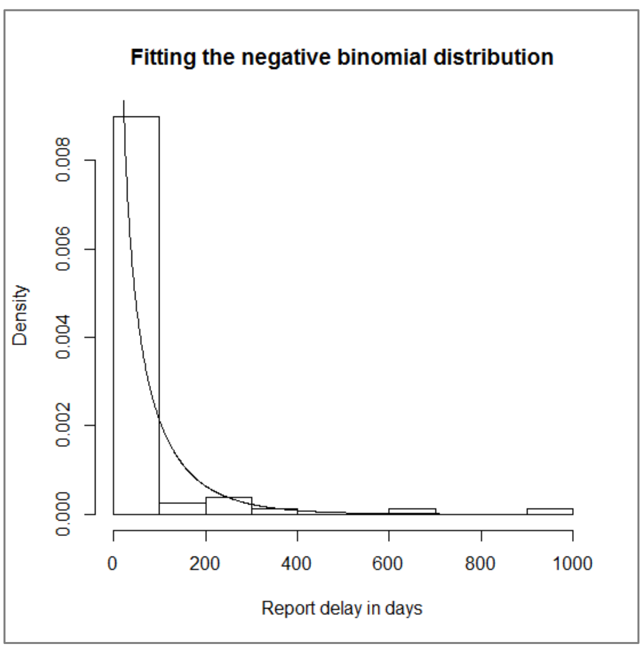

Figure 4.1: Graphical illustration for the negative binomial fit to the report lag variable

The geometric distribution

Using the maximum likelihood method to fit the geometric distribution, the following results were obtained;

	Parameter estimate	Standard error
Probability	0.0175	0.0019

Table 4.3: Results for the geometric distribution fit to the report lag variable

The graphical fit of the geometric distribution to the report delay data could also be illustrated as shown;

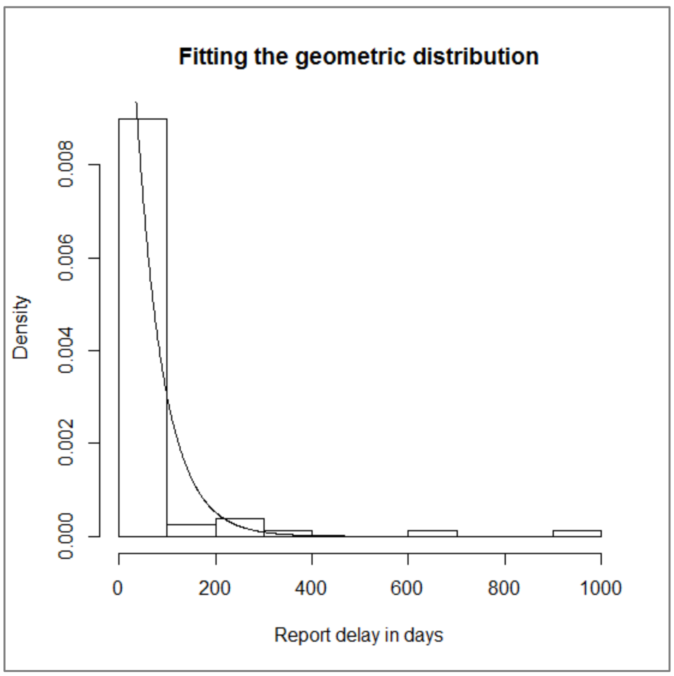

Figure 4.2: Graphical illustration for the geometric distribution fit to the report lag variable

Summary of report delay distribution marginal fit

From graphical illustrations it clearly shows that both distributions could be considered a reasonable fit for the report delay data. Comparison between the log likelihoods, AIC and BIC for the two distributions was as follows;

Distribution	Log likelihood	AIC	BIC
Negative Binomial	-379.1867	762.3734	767.1123
Geometric	-398.0396	798.0791	800.4486

Table 4.4: Results for the AIC and BIC statistics for the fitted marginal distributions

To be able to determine which fit was better, the Kolmogorov Smirnov test was also carried out on both distributions and the results were as follows;

Distribution	Kolmogorov Statistic	Kolmogorov p-value
Negative Binomial	0.1529	0.049731
Geometric	0.2877	0.000006

Table 4.5: Results for the Kolmogorov test for the fitted marginal distributions

It was noted that both the AIC and the BIC values for the negative binomial were less than that of the geometric distribution implying that the negative binomial distribution was the better fit for the data. The Kolmogorov statistic further confirmed this by showing the p-value of the Negative binomial to be greater than that of the Geometric distribution. Hence the better distribution for the marginal of the report delay data was taken to be the Negative binomial distribution.

4.1.2 Claim Amount Distribution

Two continuous distribution curves were considered for fitting the claim amount; the log-normal distribution and the Weibull distribution.

The lognormal distribution

Using the maximum likelihood method to fit the lognormal distribution, the following results were obtained;

	Parameter estimates	Standard error
Mean-log	10.71	0.0019
Sd-log	1.42	

Table 4.6: Results for the lognormal distribution fit to the claim amount variable

The graphical fit of the lognormal distribution to the claim amounts data is illustrated as shown;

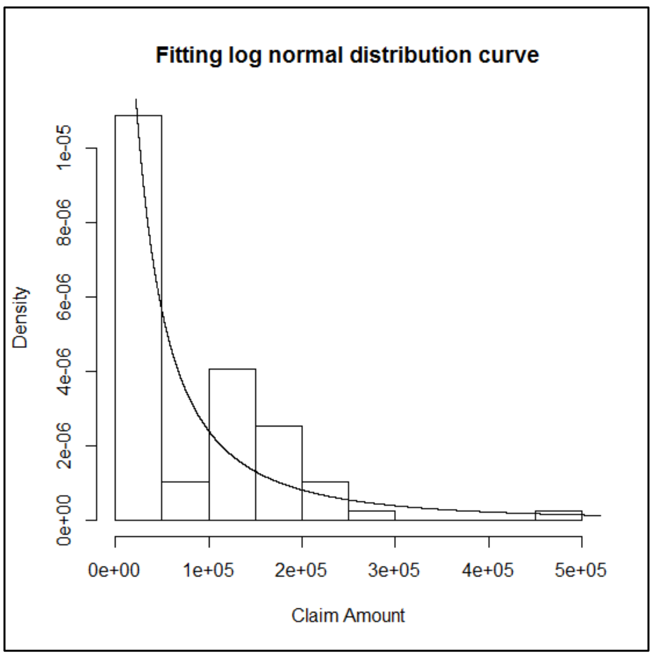

Figure 4.3: Graphical illustration for the lognormal distribution fit to the claim amount variable

The Weibull Distribution

Using the maximum likelihood method to fit the Weibull distribution, the following results were obtained;

	Parameter estimate	Standard error
Shape	8.71	0.0019
Scale	8.81	

Table 4.7: Results for the Weibull distribution fit to the claim amount variable

Summary of claim amount distribution marginal fit

The summary results for fitting the distributions were as follows;

Distribution	Log likelihood	AIC	BIC
Lognormal	-985.6014	1975.20	1979.94
Weibull	-982.1955	1968.39	1973.13

Table 4.8: Results for the AIC and BIC statistics for the fitted marginal distributions

The Kolmogorov test statistic results on the claim amount data was as follows;

Distribution	Kolmogorov Statistic	Kolmogorov p-value
Lognormal distribution	0.2084	0.002093
Weibull distribution	1	0.000000

Table 4.9: Results for the Kolmogorov test for the fitted marginal distributions

From Kolmogorov tests results the lognormal distribution is a better fit of the data.

4.2 Fitting the Copula Distributions

The maximum likelihood method was employed to the copulas given the results obtained for the marginal distributions.

The copula distributions first had to be converted to their respective probability density functions in order to obtain the likelihood function. The log likelihood function was then obtained for easier optimization.

It is important to note that the likelihood and the log likelihood functions were not easily maximized and therefore for some of the copulas; a numeric iterative method – The Newton Raphson method was applied to obtain the parameter estimates.

The variables used in the copula functions were the distribution values of the observed instances using the distribution functions that were maximized in the previous section.

The results for the copulas were as follows;

4.2.1 Gumbel Barnett Copula

The main results for the Gumbel Barnett Copula were as follows:

Method of likelihood maximization	Newton Raphson (7 iterations)
Parameter estimate	1.4481
Standard error	0.1968
Log likelihood	65.29457
AIC	-128.5891

Table 4.10: Results for the Gumbel Barnett Copula

We note that the parameter value obtained is outside the bounds required for the Gumbel Barnet copula; (0, 1).

In such a case; we find the maximum likelihood estimator within these bounds by plotting the graph of parameter values against the log likelihood to obtain the parameter value that maximizes the copula. The graph obtained is as below;

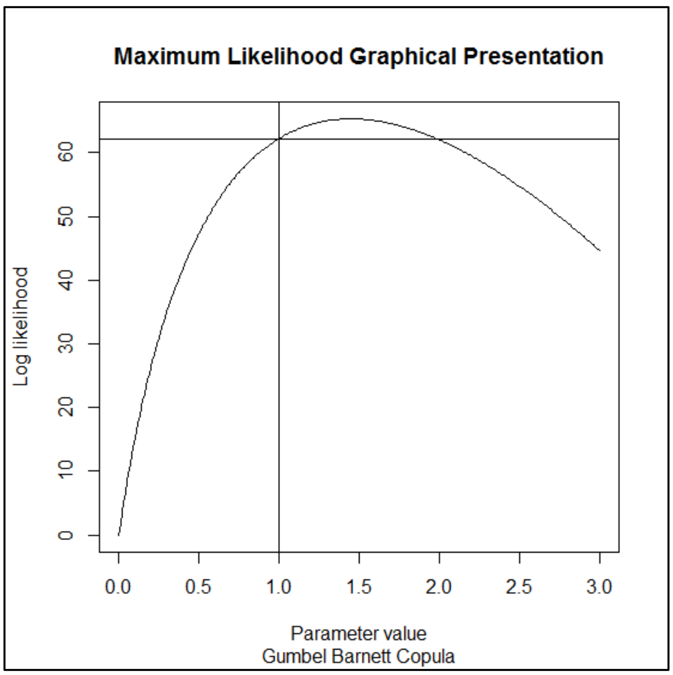

Figure 4.4: The graphical representation of the maximum likelihood estimate for the Gumbel Barnett

This shows that the parameter value that maximizes the log likelihood within the bounds is 1. Hence rather than 1.44 as obtained by the numerical analysis we settle for 1.

4.2.2 Nelson Number 10

The results under the Nelson number 10 copula were as follows;

Method of likelihood maximization	Newton Raphson (7 iterations)
Parameter estimate	0.6900
Standard error	0.2983
Log likelihood	5.078408
AIC	-8.156817

Table 4.11: Results for the Nelson No. 10 Copula

The parameter estimate obtained from numerical approach this time is within the required bounds (0, 1].

This result can be complemented by the graph of the log likelihood function where the estimate is observed as the maximum point of the curve.

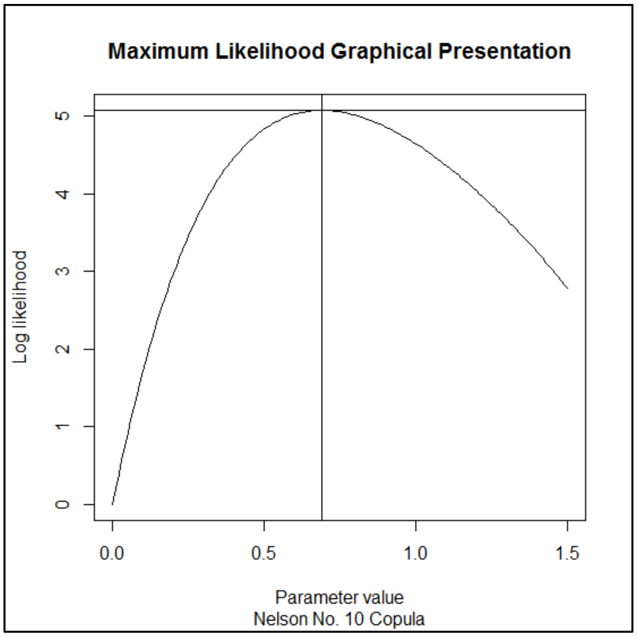

Figure 4.5: The graphical representation of the maximum likelihood estimate for the Nelson No. 10

4.2.3 Franks Copula

The results for Franks Copula were as follows:

Method of likelihood maximization	Newton Raphson (3 iterations)
Parameter estimate	-2.7171
Standard error	0.7505
Likelihood	6.775623
AIC	-11.55125

Table 4.12: Results for the Franks Copula

The result for the parameter estimate from the numerical approach was further complemented using the graphical approach as shown below;

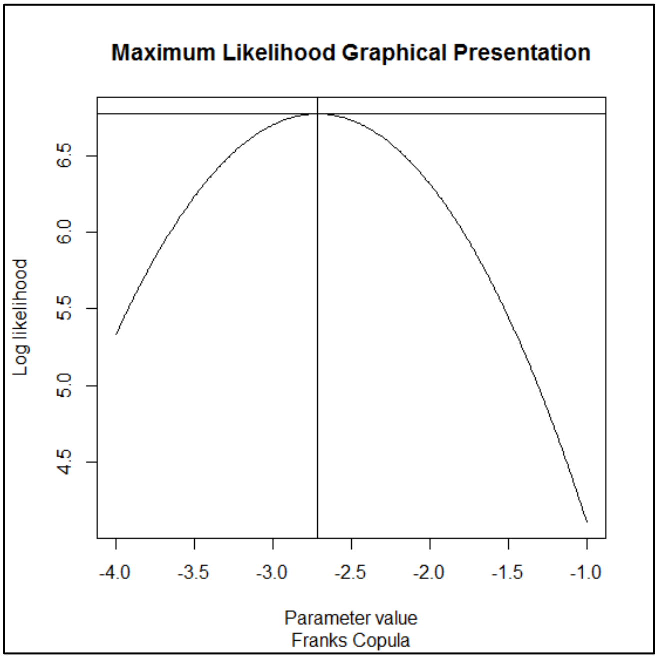

Figure 4.6: The graphical representation of the maximum likelihood estimate for Franks Copula

4.2.4 Clayton Copula

The results for Clayton Copula were as follows:

Method of likelihood maximization	**Newton Raphson (6 iterations)**
Parameter estimate	-0.27312
Standard error	0.03786
Log likelihood	4.477831
AIC	-6.955661

Table 4.13: Results for the Clayton Copula

When the maximum likelihood graphical approach was considered, the results were as below;

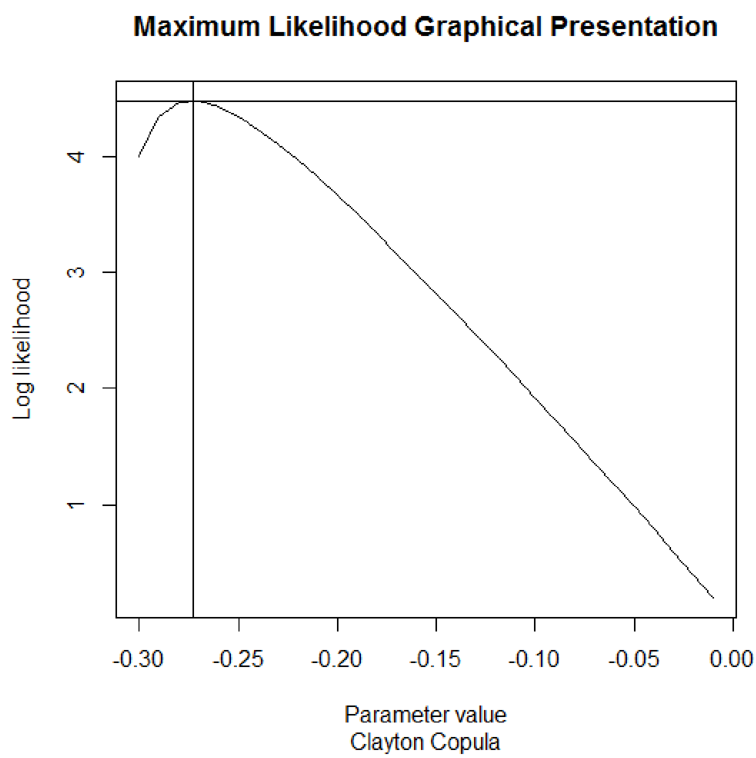

Figure 4.7: The graphical representation of the maximum likelihood estimate for Clayton Copula

Furthermore; Kendall's tau as implied by the Clayton copula is: **- 0.1585**

4.2.5 Nelson Number 7 Copula

The results for Nelson Number 7 Copula were as follows:

Method of likelihood maximization	**Estimation**
Parameter estimate	0.9968323
Likelihood	-0.2506438
AIC	-2.5012876

Table 4.14: Results for Nelson Number 7 Copula

4.3 Summary of the copula fits to the data

The results for the AIC test statistics for the copulas compared as follows;

Copula	AIC
Gumbel Barnett	-128.5891
Clayton	-6.955661
Nelson number 7	-2.5012876
Nelson number 10	-8.156817
Franks Copula	-11.55125

Table 4.15: Results for the AIC statistics for the five copulas

CHAPTER FIVE: DISCUSSIONS AND CONCLUSIONS

5.1 Discussions

From the general results obtained for the two variables claim amount and report lag; it was observed that there exists a negative dependence. This was illustrated by the correlation measures; Pearson's correlation coefficient, Kendall's tau and spearman's rho. In addition to the correlation measures being negative; it was also noted that the magnitude of the measures was significantly small. This implied existence of weak dependence.

When fitting the Gumbel Barnett copula, the resulting parameter estimate was 1. The Gumbel Barnett copula is a favorite among the copulas in measuring weak negative dependence between variables. A result by Charpentier (2004); shows that when the parameter is equal to 0 the copula becomes the independence copula and as long as the parameter is less than 1; it remains within the Frechet Bounds. Hence from our result along with the nature of the Gumbel Barnett Copula; we can deduce that the obtained parameter implies existence of weak negative dependence between the variables.

The Nelson number 10 copula is another copula suited to measure weak negative dependence. The nature of the copula itself implies existence of weak negative dependence between the two variables.

Franks copula is comprehensive in nature and hence can measure all types of dependence. As earlier mentioned a negative value of the parameter implies a negative dependence and the higher the absolute value the higher the dependence. In this case the value obtained was a negative value indicating negative dependence and its low absolute nature implied weak dependence between the variables. This also showed the existence of weak negative dependence between the two variables

The Clayton copula with parameter -1 implies a perfect negative dependence and with parameter 0 implies perfect independence between the variables. The Clayton copula parameter estimate lies between -1 and 0 implying that there is indeed negative dependence between the two variables. One could also infer from the low absolute value of the parameter that the negative dependence is weak in nature. This was further emphasized with the Kendall's tau estimate obtained as per the Claytons' parameter. The Kendall's tau estimate was found to be negative and had a small absolute value implying existence of weak negative dependence between the two variables.

The Nelson number 7 copula is suited to measure negative dependence between variables. The limits of the copulas' parameter estimate are; 1 which implies independence and 0 which implies negative dependence. The parameter estimate obtained in our results was extremely close to 1 implying that if indeed there exists negative dependence between the two variables; it is quite weak in nature tending to independence.

5.2 Conclusion

The correlational measures for the sample data; the Pearson's correlation coefficient, Kendall's tau and the Spearman's rho all implied existence of weak negative dependence between the two variables; claim amounts and report lag. Two of the copulas considered in the study were comprehensive in nature. This implied that they could measure all types of dependence and therefore based on the parameter estimate obtained, they could be able to show what kind of dependence existed between the variables. The two comprehensive copulas were the Clayton and Frank's copula. Results from both copulas confirmed that indeed weak negative dependence existed between the variables.

The remaining three copulas are copulas suited to measure negative dependence and therefore could only aid in showing how strong or weak the negative dependence is if any. While the Nelson number 7 is suited for modelling negative dependence in general, the Gumbel Barnett and the Nelson number 10 are specifically suited for modelling weak negative dependence between variables.

As already mentioned all the five copulas that were considered in the study showed that there exists weak negative dependence between the report lag variable and claim amount variable. However each copula measured this negative dependence to different extents. A comparison of the AIC values of each of the copulas was conducted. The Gumbel Barnett copula had the smallest AIC value while the Nelson number 7 copula had the largest AIC value. This result implied that according to the comparisons of the AIC values for the different copulas, the Gumbel Barnett is the best option in modeling dependence between the variables report delay and claim amount.

As per the results obtained from the sample correlational measures that implied weak negative dependence; the copula best suited to measure dependence between claim amounts and report lag based on the AIC results, is the Gumbel Barnett copula; which in itself is a copula suited to measure weak negative dependence.

REFERENCES

A.M. Best Company (2004), "Best's Insolvency Study/Property Casualty U.S. Insurers 1969-2002", *Oldwick, NJ: A.M. Best Company*.

Brown B. Z. (1988). Considerations in setting loss reserves. *Casualty Loss reserve seminar, retrieved from; https://www.casact.org/pubs/CLRSTrans/1988/1.pdf*

Charpentier A. (2004); Dynamic dependence ordering for Archimedean copulas and distorted copulas; *Kybernetika Vol. 44, pg 777-794*

De Souza L. G. and Veiga A. (2014), A Stochastic Model to Estimate the Amount of IBNR Claims Using Micro-data; *Paper presented at 8th Conference in Actuarial Science & Finance on Samos, Greece*

Denuit M., Purcaru O. and Keilegom I. V. (2006), Bivariate Archimedean Copula Models for Censored Data in Non-Life Insurance, *Journal of Actuarial Practice, Vol 13 pg 5-32*

Dhaene, J., and Goovaerts, M.J. (1996). Dependency of risks and stop-loss order. *ASTIN Bulletin Vol 26, pg 201-212.*

Fréchet, M. R. (1951). Sur les tableaux de corrélation dont les marges sont données. *Ann. Univ. Lyon Sci. 4, 53–84.*

Frees E. W. and Valdez E. A. (2007), Hierarchical Insurance Claims Modeling, *Journal of the American Statistical Association 103, pg 1457-1469*

Genest C., Marceau E. and Mesfioui M. (2002), Compound Poisson approximations for individual models with dependent risks, *Insurance: Mathematics and Economics Vol. 32 pg 73–91*

Gile B. S. (1994), A statistical approach to IBNR Reserves, *Journal of Actuarial Practice vol 2, No. 2, pg 257-272*

Hoeffding, W. (1940). Scale-invariant correlation theory. In N. I. Fisher and P. K. Sen (Eds.), *The Collected Works of Wassily Hoeffding, pp. 57–107. New York: Springer-Verlag*

Hua L. (2015), Higher order tail densities of copulas and hidden regular variation, *Journal of Multi variate analysis volume 138, pgs 143 - 155*

Nelsen, R. B. (1998). An Introduction to Copulas. *New York: Springer-Verlag*

Neuhaus W. (2004), On the Estimation of Outstanding Claims, *Australian Actuarial Journal Vol 11, pg 549-587*

Patton A. J. (2002), Applications of Copula Theory in Financial Econometrics, *PhD dissertation, Department of Economics, University of California Sandiego*

Purwono Y. (2004), Copula Inference For Multiple Lives Analysis–Preliminaries, *PhD dissertation, Department of Management, Faculty of Economics, University of Indonesia*

Qaiser R. (2006) Reserving in general insurance business in india and its implications. *National insurance academy, Pune*

Schiegl M. (2010), A three dimensional stochastic Model for Claim Reserving, *Astin 2009 colloquium, Helsinki*

Schweizer, B., Sklar, A. (1983): Probabilistic metric spaces. *North-Holland Series in Probability and Applied Mathematic*s. New York

Sharma P. (2002), Prudential supervision of insurance undertakings, *retrieved from: http://ec.europa.eu/internal_market/insurance/docs/solvency/impactassess/annex-c02_en.pdf*

Sklar, A. (1959). Fonctions de r´epartition ´a n dimensions et leurs marges. *Inst. Statist. Univ. Paris, 8, pg 229-231.*

Stander Y. S. (2011), Multivariate copulas in financial market risk with particular focus on trading strategies and asset allocation, *PhD dissertation; Department of Economics and Econometrics, University of Johannesburg*

Verrall R., Nielsen J. P. and Jessen A. H. (2010), Prediction of RBNS and IBNR claims using claim amounts and claim counts, *Astin bulletin, Vol 40, pg 871-887*

Weissner E.W. (1999), Estimation Of The Distribution Of Report Lags By The Method Of Maximum Likelihood, *Casualty Actuarial Society Publication, Vol LXV Part1, No. 123*

Weke P. and Ratemo C. (2013), Estimating IBNR Claims Reserves for General Insurance Using Archimedean Copulas, *Applied Mathematical Sciences, Vol. 7, pg 1223 - 1237*

Wolley K. K. (1997), How Variables Uncorrelated with the Dependent Variable Can Actually Make Excellent Predictors: The Important Suppressor Variable Case*; Texas A&M University publication*

APPENDIX

R Code:

Construction of fundamental copulas:

```
x<-seq(from=0, to=1, by=0.1)
y<-x
f<-function(x,y){r=x*y;r}
z<-outer(x, y, f)
#####################################################
r<-matrix(nrow=11, ncol=11)
for (j in 1:11){
for (i in 1:11){r[i,j]=min(x[i],y[j])}                    }
z<-r
######################################################
r<-matrix(nrow=11, ncol=11)
for (j in 1:11){
for (i in 1:11){r[i,j]=max(x[i]+y[j]-1,0)}                    }
z<-r
######################################################
nrz<-nrow(z)
ncz<-ncol(z)
jet.colors<-colorRampPalette(c("red", "purple"))
nbcol<-100
color<-jet.colors(nbcol)
zfacet <- z[-1, -1] + z[-1, -ncz] + z[-nrz, -1] + z[-nrz, -ncz]
facetcol <- cut(zfacet, nbcol)
persp(x, y, z, col = color[facetcol], phi = 30, theta = 130)->res
```

Fitting the marginal distributions

```
library(fitdistrplus)
library(maxLik)

####### REPORT DELAY ############
table<-read.csv("C:/Users/samayi/Documents/Project.csv", header = TRUE)
hist(table$Report, prob=TRUE, xlab="Report delay in days", main="Fitting the distribution")

summary(fitdist(table$Report, "nbinom"))
lines(x, dnbinom(x,size=0.4724853, mu=56.2457902), col="black")
summary(fitdist(table$Report, "geom"))
lines(x, dgeom(x,prob=0.01747015), col="orange")
ks.test(table$Report, "pnbinom",size=0.4724853, mu=56.2457902)
ks.test(table$Report, "pgeom" ,prob=0.01747015)

######## CLAIM ESTIMATES ###############
table<-read.csv("C:/Users/samayi/Documents/Project.csv", header = TRUE)
hist(table$Estimate, prob=TRUE, xlab="Claim Amount", main="Fitting log normal distribution curve")
x<-seq(from=10000,to=1000000, by=1)
summary(fitdist(table$Estimate, "lnorm"))
lines(x, dlnorm(x,meanlog=10.706871, sdlog=1.419291), col="black")
summary(fitdist(table$Estimate, "weibull"))
lines(x, dweibull(x,shape=0.8709962, scale=0.0008184679), col="blue")
```

Fitting the Gumbel Barnett Copula

```
############# GUMBEL BARNETT ###################

x<-c(pnbinom(table$Report, size=0.4724853, mu=56.2457902))
y<-c(plnorm(table$Estimate, meanlog=10.706871, sdlog=1.419291))
n<-length(table$Estimate)
GB<-function(alpha){
z<-alpha[1]
zu<-(-1*z*log(x, base=exp(1))*log(y, base=exp(1)))
zuu<-log( (1-z*(log(x, base=exp(1))+log(y, base=exp(1))-(z*log(x, base=exp(1))*log(y, base=exp(1)))-1))
,base=exp(1))
return(sum(zu)+sum(zuu))
}
ml <- maxLik( GB, start = 0)
summary(ml)
AIC(ml)
z<-seq(from=0, to=3, by=0.01)
h<-NULL
for (i in 1:301){ h[i]<-GB(z[i])}
plot(z,h, type="l", xlab="Parameter value", ylab="Log likelihood", main="Maximum Likelihood Graphical
Presentation", sub="Gumbel Barnett Copula")
abline(h=GB(1), v=1)
```

Fitting the Nelson No. 10 Copula

```
############# NELSON NUMBER 10 #########

x<-c(pnbinom(table$Report, size=0.4724853, mu=56.2457902))
y<-c(plnorm(table$Estimate, meanlog=10.706871, sdlog=1.419291))
NT<-function(alpha){
z<-alpha[1]
m<-(1/(1+((1-x^z)*(1-y^z))))
n<-m^(1/z)
o<-(x^z+y^z-((x^z)*(y^z)*(2+z)))*(m^((1+z)/z))
p<-(x^z*y^z*(1-x^z)*(1-y^z)*(z+1)*m^((1+(2*z))/z))
r <-n+o+p
zu<-(log(r ,base=exp(1)))
return(sum(zu))
}
ml <- maxLik( NT, start = 0.5)
summary(ml)
z<-seq(from=0, to=1.5, by=0.01)
h<-NULL
for (i in 1:151){ h[i]<-NT(z[i])}
plot(z,h, type="l", xlab="Parameter value", ylab="Log likelihood", main="Maximum Likelihood Graphical
Presentation", sub="Nelson No. 10 Copula")
abline(h=NT(0.69), v=0.69)
```

Fitting Franks Copula

```
########### FRANKS COPULA   ################

x<-c(pnbinom(table$Report, size=0.4724853, mu=56.2457902))
y<-c(plnorm(table$Estimate, meanlog=10.706871, sdlog=1.419291))
FC<-function(alpha){
z<-alpha[1]
a<-(z*exp(-z*(x+y)))
b<-((exp(-z*x)-1)*(exp(-z*y)-1))
c<-((exp(-z)-1)+b)
d<-(a*((b/(c*c))-(1/c)))
zu<-(log(d, base=exp(1)))
return(sum(zu))
}
ml <- maxLik( FC,start=-3)
summary(ml)
z<-seq(from=-4, to=-1, by=0.01)
h<-NULL
for (i in 1:301){ h[i]<-FC(z[i])}
plot(z,h, type="l", xlab="Parameter value", ylab="Log likelihood", main="Maximum Likelihood Graphical
Presentation", sub="Franks Copula")
abline(h=FC(-2.7171), v=-2.7171)
```

Fitting the Clayton Copula

```
########### CLAYTON COPULA   ################

x<-c(pnbinom(table$Report, size=0.4724853, mu=56.2457902))
y<-c(plnorm(table$Estimate, meanlog=10.706871, sdlog=1.419291))
CC<-function(alpha){
z<-alpha[1]
a<-(1+z)
b<-(x^(-z-1))*(y^(-z-1))
c<-((x^(-z)+y^(-z)-1)^((-1-(2*z))/z))
d<-a*b*c
zu<-log(d, base=exp(1))
return(sum(zu))
}
ml <- maxLik( CC ,start=-0.1)
summary(ml)
AIC(ml)
z<-seq(from=-0.3, to=-0.01, by=0.01)
h<-NULL
for (i in 1:30){ h[i]<-CC(z[i])}
plot(z,h, type="l", xlab="Parameter value", ylab="Log likelihood", main="Maximum Likelihood Graphical
Presentation", sub="Clayton Copula")
abline(h=CC(-0.27312), v=-0.27312)
```

Fitting the Nelson No. 7 Copula

```
############# NELSON NUMBER 7 #########

x<-c(pnbinom(table$Report, size=0.4724853, mu=56.2457902))
y<-c(plnorm(table$Estimate, meanlog=10.706871, sdlog=1.419291))
z<-max((1-x-y)/((x*y)-x-y+1))
z
length(table$Estimate)*log(z, base=exp(1))
```