KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY, KUMASI

Survival Analysis Of The Average Time To Handling A Claim In The Insurance Industry: A Case Study Of An Automobile Insurance Company In Ghana.

BY

PATRICIA ADJELEY LARYEA

BSc. Actuarial Science

A THESIS SUBMITTED TO THE DEPARTMENT OF MATHEMATICS, KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY IN PARTIAL FUFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF MSc. INDUSTRIAL MATHEMATICS

JUNE 2015

DECLARATION

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

PATRICIA A. LARYEA(PG8830913)	.....	.....
Student	Signature	Date

Certified by:

Dr. F. T. ODURO	.....	.....
Supervisor	Signature	Date

Certified by:

Prof. S. K. AMPONSAH	.....	.....
Head of Department	Signature	Date

DEDICATION

I dedicate this work to my father James O. Laryea, my mother, Mercy Tetteh and beloved Emmanuel J. Hanson who were my rock and constant inspiration who supported my study.

Also, to the Late Emmanuel S. Laryea and Abigail Adjeley Laryea. Rest in Perfect Peace.

ABSTRACT

Motor Insurance provides protection for vehicles that operate on the roads in Ghana and it is mandatory. A large sample from an insurance data with a significant proportion of censored observations was used to determine the average time it takes for losses to occur and be paid by an insurance company in Ghana using the Kaplan-Meier approach. An analysis of this portfolio presented using the Cox proportional hazard model to determine if the type of insurance affects the time it takes for a claim to be paid and to establish which variables contribute significantly to the time for a claim to be settled. The study revealed that age, gender, marital status, are significant risk factors that affect the occurring of a loss but not significant in the payment of claims using the log rank test and Cox proportional hazards regression model. Type of policy and type of vehicle were significant factors that influence the survival duration of settling claims. This clearly indicates that in the quoting (calculation) of premiums these risk factors are considered in Ghana.

ACKNOWLEDGMENT

I render my sincere and fatherly thanks to the Gracious Almighty God for his divine protection and guidance throughout the course. His name be exalted, honoured and glorified.

I express my profound gratitude to my supervisor Dr. F. T. Oduro for his time, continuous support, patience, motivation and immense knowledge towards the writing of this thesis and also thanks to the Statistical Consulting and Quantitative Skill Development Unit of the Maths Department.

I also express my heartfelt gratitude to my uncles, Nii Otinkorang Ankrah and John N. O. Ankrah, sister, Zilla Laryea, brother, Emmanuel K. Botchway, the Laryea family, and the Tetteh family.

Special thanks also go to Rev. Fr. Dr. Augustine H. K. Abasi and Rev. Fr. Luke Atanga Abugre. Mr. Yinn Luu, Mr. Abdulai Abdul-Malik and Mr. Anang, Clottey Richard also deserve special thanks.

I sincerely thank Solomon K. Essuman Antwi, Enoch D. Mwini, Shibu Osman, George Adjavon, Samuel Ampofo Dateh, Emmanuel Mfum-Mensah, Albert F. Attakora, John K. Poku and Emmanuel Omenako-Danquah and to all my friends and course mates who have assisted me I am equally grateful, and all those who helped me in diverse ways yet not mentioned I say God richly bless you and prosper you in all your endeavors.

CONTENTS

LIST OF ABBREVIATION

ACE ..... Average Causal Effect

AFT ..... Accelerated Failure Time Model

AIC ..... Akaike Information Criterion

BIC ..... Bayesian Information Criterion

CI ..... Confidence Interval

CACE ..... Complier Average Causal Effect

CTAs ..... Commodity Trading Advisors

DEB ..... Dynamic Energy Budget

DME ..... Diabetic Macular Edema

EC ..... Small-Effect Concentrations

EM ..... Expectation and Maximisation algorithm

F-R ..... Frangakis-Rubin model

GIA ..... Ghana Insurers Association

GIBA ..... Ghana Insurance Brokers Association

HIV ..... Human Immune Virus

LTC ..... Long Term Care

NIC ..... National Insurance Company

MCMC ..... Markov Chain Monte Carlo

ML ..... Maximum Likelihood

NOEC ..... No-Observed-Effect Concentration

PPO-survival model .....Parametric Potential-Outcome survival model

QIS5 .....Quantitative Impact Study 5

RCT .....Randomised controlled trial

SCR .....Solvency Capital Requirement

SPSS software .....Statistical Package for Social Sciences Software

TP .....Third Party Policy

UI .....Unemployment Insurance

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background of the Study

The provision of protection against a possible eventuality such as damage, illness, death or a specified loss in return for payment of a specified premium is known as **Insurance**. A motor insurance policy is a mandatory policy issued by an insurance company as part of prevention of public liability. The 'Act only' policy (i.e., third party liability) and the Comprehensive policy are broadly the two types of insurance in Ghana, (NIC, 2009).

In the auto insurance market the factors considered before rating includes; age (Crocker and Snow, 1986) driving history (being an accident-free driver discounts are applicable), marital status (married drivers have fewer accidents than single drivers), vehicle type (with vehicles purchased under a hire-purchase agreement, the financiers insist upon a comprehensive policy to take care of their interest as collateral security), where one lives, and gender (men under the age of 25 are involved in more accidents than women under the age of 25 and have more than three times as many fatal accidents), (Kiebach, 2014).

The demand for payment of a loss by a policyholder or by an injured third party is considered a **claim**. The claims are organized by **accident date** - i.e. the date on which the accident occurred, leading to the claim and by **report date** - i.e. the date on which the insurer is notified of the claim, (McClenahan, 2001).

Section 44 (4) of Ghana's insurance Act, 2006 [Act724] states that 'The

Commission in consultation with the insurance industry shall by Regulations prescribe a formula to compute the compensation in respect of injury and deceased claims arising out of a motor accident'. The claims are recorded and computed, (NIC, 2011).

Claims settlements are indeed one of the fastest priorities that make an insurance company unique. The insured and insurer must share agreements designed by the policy provided to avoid the need for argument or blame. The insured must be able to provide the following to the insurer after which effective action will be taken:

1. Police Report Form
2. Pictures of the accident
3. Estimates of items damaged in the vehicle
4. Driving License
5. Proof of doctors' report if any.

The National Insurance Commission (NIC), the Ghana Insurers Association (GIA) and the Ghana Insurance Brokers Association (GIBA) in response to legal requirements on 17th August, 2011, set up the Motor Insurance Compensation Guidelines, which states that, "All Motor Insurance claims are to be settled and paid within sixty (60) working days upon receipt of all relevant documents. Reasons for delays should be clearly stated in the claim file for the inspection of the Regulator. Where claims are unreasonably delayed, appropriate sanctions shall be applied", (NIC, 2011)

In Ghana, the payment of claims by some insurance companies has been of great challenge especially in the automobile industry. Therefore, this study applies Cox regression analysis to determine the type of insurance and variables that affect the time it takes for a claim to be settled.

1.2 Problem Statement

It is by law required that every car owner will have some sort of insurance cover for the vehicle that is at least the third party insurance which seeks to provide insurance for the innocent pedestrian and other vehicles. Insurers compete among themselves in order to attract customers by tending to quote premiums below or under coverage. Even under such circumstances, people are not patronizing well enough, since apparently, customer satisfaction refers to the provision of quality services in which time is key. Customers go through a lot of stress before receiving their claims. Ofori-Attah (2012) conducted a study on the effects of slow claims settlement on the sales and marketing of insurance products by administering questionnaires to both customers and staff. The results obtained from the data collection were cross tabulated and subjected to descriptive analysis. In his cross tabulation of the effect of time claim is settled and satisfaction derived revealed that a total of 133 respondents had their claims settled. 80 claims were settled within 3 months, 18 claims were settled in 4-6 months, whilst 5 claims were settled within 7-12 months, 6 and 24, were settled within 1 year and 2 years respectively. It was worthy to note that, from the valid respondents, none of the claims took more than two years before settlement was effected. Out of the 133 valid respondents who made a claim, 112 indicated that they were satisfied with the duration of settlement while 21 indicated they were not satisfied. He saw that 64% of claimants whose settlement lagged from 4 months to 2 years were not willing to take another policy. This indicates that the rate of dissatisfaction increased with increasing period of claim settlement.

1.3 Objectives of the Study

The objectives of the present study are as follows:

• To determine the average time for a claim to occur and be settled using the Kaplan-Meier approach.

• To determine if the type of insurance (policy) affects the time it takes for a claim to be settled and to establish which variables contribute significantly to the time for a claim to be settled using the Cox-Regression.

1.4 Methodology

The study utilized administrative data, gathered from the National Insurance Commission and some insurance companies in Ghana. The data contained the underwriting of 1,000 insureds that purchased insurance policy for their automobiles commencing from January, 2010 and expiring on December, 2010. Some insureds were involved in motor accidents collisions which resulted to damage to either their vehicles or a third-party individual during their cover of insurance. The variables under study in this project were age, gender, marital status, the type of vehicle involved in the accident, type of insurance policy bought, and the nature of the claim. Another variable used in this study was time: this variable was defined as the length of having an insurance policy until a loss occurs and when it was paid.

To facilitate the analysis of the data the variable age was grouped into four categories, i.e., 21-29years, 30-45years, 46-59years and 60years and above. The nature of the claim as to whether a third party policyholder or a comprehensive policyholder was involved in an accident were considered. The data entry and preliminary analysis were done using the statistical software package for social scientist (SPSS) version 17. Further analysis was then done using R.

1.5 Justification

One of the key values to customer loyalty to a company or product is customer satisfaction. Day in and out companies and institutions which render services lose their customers due to the fact that customers are not satisfied for services

rendered to them. This does not exclude insurance companies who lose most of their customers due to some few challenges customers face with the insurer. Some of these challenges include inability to render claim settlement in due time. However, the regulator NIC is making every effort to improve the delivery of insurance as well as the settlement of claims by setting up guidelines and conditions to monitor the payment of these claims in Ghana. The aim of this project is to provide an estimate for the average time to claims settlement and also determine the survival and hazard rates of cases reported at an insurance company from January, 2010 to December, 2010. The findings of the study will inform insurers, policy makers, law enforcement agencies, academicians, and the country at large in setting of priorities, and formulation of policies to address issues related to delay's in claim settlement in Ghana.

1.6 Organization of the Thesis

The first chapter of the thesis is made up of the introduction, which comprises the background of the study, problem statement, and objectives of the study, as well as the methodology and justification of the study. The second chapter comprises the literature review, that is, scholarly work done by other people on the topic. These are empirical evidences of the topic been studied. The third chapter talks about the methodologies employed in the study. This includes secondary data collection, models to be used, the tests to be used and the software to be used in the analysis of the data collected. The fourth chapter deals with data analysis and discussion of results and the fifth chapter concludes the study and offers recommendations.

CHAPTER 2

LITERATURE REVIEW

This chapter deals with the review of related literature on the topic under study. The review includes concepts, theories found in literature, and empirical studies documented in journals and on the Internet. Motor insurance is a major concern in our part of the world due to the increasing nature of vehicles on our roads and its hazards. Czado and Rudolph (2002) saw that a large claims portfolio with significant proportion of censored observations availed was due to the introduction of compulsory long term care (LTC) insurance in Germany in 1995. They used the (Cox, 1972) and estimated transition intensities that computed premiums for LTC insurance plans using the multiple state Markov model, (Haberman and Pitacco, 1999). The estimation of the survival rate of cases in some insurance companies in Ghana will likely explore all useful information that may help policy makers and stakeholders in their quest to improve delays in claim settlement and also compute premiums in automobile insurance.

Harrison and Ansell (2002) used survival analysis to predict cross-selling opportunities that would retain a customer in the insurance industry. They determined who is likely to buy additional product from the same company, what the next product is likely to be and when the purchase is likely to be made using a sample of 9,000 customers selected at random. Harrison et al. (2007) further demonstrated how lifestyle segmentation and survival analysis can identify cross-selling opportunities in life insurance and pension products. They applied the lifestyle analysis and Cox's regression model to behavioral and demographic data describing 10,979 UK customers of a large international insurance company. They observed that "mature" segments appear to offer

greater opportunities for retention and cross-selling than the "younger" segments from the company and the time frames within which that is likely to take place.

Nasvadi and Wister (2009) used survival analysis of insurance data from British Columbia to determine if restricted driver's licenses lower crash risk among older drivers. They used a cohort study design and examined licensing and insurance claims crash records of all drivers aged 66 years and older for the years 1999-2006. Nonparametric and Cox proportional hazards survival analyses were used to compare restricted vs. unrestricted drivers and to estimate crash risks. They saw that the risk of causing a crash was 87% lower for restricted drivers compared with unrestricted drivers after controlling for age and gender. The most common restriction was a combination of daylight driving only plus a speed maximum of 80 km/hr. Restricted drivers retained a driver's license for a longer period of time than unrestricted drivers and continued to drive crash free longer than unrestricted drivers. There was no difference in severity of collisions, and results suggest a high level of compliance with daylight-only restrictions. Therefore they concluded that driving restrictions may be effective for prolonging the crash-free driving of some aging drivers, thus supporting their continued independence and delaying institutionalization.

Raftery et al. (1995) researched on how accounting for model uncertainty improves predictive performance and can be clinically useful in survival analysis. They said model uncertainty can be substantial but it is ignored in the model-building process where predictor variables are selected. Meyer and Laud (2002) using a stepwise procedure to test for their significance to select a single model, and then make inference conditionally on the selected model. They reviewed the standard Bayesian model averaging solution to the problem (Kalbfleisch, 1978) and extended it to survival analysis, introducing partial Bayes factors (Kass and Raftery, 1995) to do so for the Cox proportional hazards model.

Aslanidou et al. (1995) researched on Bayesian analysis of multivariate survival data using Markov Chain Monte Carlo methods. Metropolis along with Gibbs algorithm (Metropolis et al., 1953; Muller, 1991) was used to calculate some of the marginal posteriors hence proposed multivariate survival model because survival times within the same 'group' are correlated as a consequence of a frailty random block effect; (Vaupel and Stallard, 1979). The conditional proportional hazards model of Clayton and Cuzick (1985) was used with a martingale structured prior process, (Arjas and Gasbarra, 1994) for the discretized baseline hazard. Besides the calculation of the marginal posteriors of the parameters, they used Bayesian EDA diagnostic techniques to detect model adequacy of a kidney infection data where the times to infection within the same patients are expected to be correlated. Further Sinha and Dey (1996) analyzed some common types of survival data from different medical studies. They used semi-parametric Bayesian analysis of survival data.

Gepp (2005) used an evaluation of decision tree and survival analysis techniques for business failure prediction. He saw that the potential value of an accurate business failure prediction model has been emphasized by the extremely costly failure of high profile businesses in both Australia and overseas, such as HIH (Australia) and Enron (USA). He said there has been a significant increase in interest in business failure prediction, from both industry and academia. He used survival analysis and decision trees to review various statistical models that attempt to predict the failure or success of a business based on publicly available information about that business (or its industry and the overall economy), such as accounting ratios from financial statements compared with the use of discriminant and logit analysis approaches. Overall, the decision tress provided the most accurate predictions of business failure while survival analysis techniques are slightly less accurate, they provided more information that can

be used to further the understanding of the business failure process.

Brockett et al. (2008) said "Customer-side influences on insurance have been relatively ignored in the literature". They researched on how much time a customer has to stop total defection using survival analysis of a household portfolio of insurance policies. They focused on the behavior of households having multiple policies of different types with the same insurance company, and who cancel their first policy. They considered the time after the household's cancellation of the first policy that the insurer have to retain the customer and avoid customer defection on all policies to the competition: - and, what customer characteristics are associated with customer loyalty. Using logistic regression and survival analysis techniques they assessed the probability of total customer withdrawal, and the length of time between first cancellation and subsequent customer withdrawal. A European database spanning 54 months of household multiple policyholder behavior, resulted in the fact that cancellation of one policy is a very strong indicator that other household policies would be canceled. Further, the insurer can have time to react to retain the customer after the first cancellation. However, this time was significantly dependent on the method used to contact the company, household demographics, and the nature of the household's insurance policy portfolio. Surprisingly, core customers having three or more policies in addition to the canceled policy were more vulnerable to total defection on all policies than noncore customers. Further, the potential customer repelling effects of premium increases seemed to wear out after 12 months.

Volinsky and Raftery (2000) also investigated the Bayesian Information Criterion (BIC) for variable selection in models for censored survival data. Kass and Wasserman (1996) showed that BIC provides a close approximation to the Bayes factor when a unit-information prior on the parameter space is used. They revised the penalty term in BIC so that it is the number of uncensored events

instead of the number of observations which corresponded to a more realistic prior on the parameter space and shown to improve predictive performance for assessing stroke risk in a Cardiovascular Health Study. For the censored data model the exponential distributions of survival times (i.e. a constant hazard rate) resulted in a better approximation to the exact Bayes factor based on a conjugate unit-information prior. In the Cox proportional hazards regression model they used the maximized partial likelihood.

Moncrief et al. (1989) examined data from a large national insurance firm and introduced methodologies of survival analysis to trace the retention of insurance sales agents over a two year period. They saw that existing empirical research findings rely on "cross sectional window designs" and therefore designed the flaws in the traditional approaches and demonstrated how the flaws may have affected the validity of previous studies. Survival analysis was a valuable tool for sales force turnover research and it also traced the survival function of new hires over time. The effect of independent variables on retention examined effect of sales productivity on retention rates.

Chuang and Yu's (2010) study incorporated the survival analysis of unemployment duration into the insurance pricing framework to measure the fairly-priced premium rate for Taiwan's unemployment insurance (UI) program. They saw that the fair premiums range from 0.2041% to 0.2436% under the 1999-2002 scheme and from 0.1388% to 0.1521% under the 2003-2009 scheme for various possible levels of average unemployment duration in Taiwan, and they are all lower than the current UI premium rate 1%, explained why there is a persistent surplus in the UI program. The sensitivity analysis results indicated that the fair premium rate decreases with the hazard rate of exiting from unemployment and increases with the probability of entering into unemployment. The effect of the entering probability is found to be larger than

that of the exiting probability. Hence they provided a wide range of systematic risk coefficient (β) values generated from three alternative methods to measure its impact on fair premium rates and found that the effect of β on premium rates is stronger under the 1999-2002 scheme than that of the 2003-2009 scheme.

Zhang (2008) researched on parametric mixture models in survival analysis with applications. Kouassi and Singh (1997) methodology, estimated the hazard function of a weighted linear mixture of parametric and nonparametric models and their semiparametric mixture model provided flexibility in estimation by assigning more weight to the component in the mixture that fits the data better. Zhang (2008) extended this to the estimation of survival function that minimizes the mean-squared-error. In Zhang dissertation an Expectation-Maximization algorithm was implemented to achieve the maximum likelihood estimation of mixture model and a model selection statistic based on Bayesian Information Criterion was applied to find the mixture form that best fits the data. He exploited the asymptotic properties of the maximum likelihood method for statistical inference about the parameters. The parametric mixture model was extended to a regression framework for analyzing the survival data with covariates and to assess their effects on the joint distribution of survival time and type of failure. Compared the results to a real datasets saw that the applications indicated that the parametric mixture model with its flexibility was a good alternative tool in the analysis of survival data.

Zhang (2010) worked on, "Regression survival analysis with dependent censoring and a change point for the hazard rate: With application to the impact of the Gramm-Leach-Bliley Act to insurance companies' survival." The events of interest were bankruptcy and acquisition, which were correlated and censored. They first assessed the effect of assuming independent censoring on the regression parameter estimates in Cox proportional hazard model then applied the copula

function to model the dependent censoring. Next an extended partial likelihood function maximized with an iteration algorithm was used to estimate the regression parameters and to derive the marginal survival functions under a dependent censoring setting. Lastly, they tested the existence and identified the location of a change-point in a hazard function. The application of their methodology to real insurance companies' survival data disclosed important influence of the GLB Act on insurance companies' survival.

Louzada et al. (2010) researched on the bivariate long-term distribution survival model based on the Farlie-Gumbel-Morgenstern copula model applied to a Brazillian HIV data. This model allows for the presence of censored data and covariates in the cure parameter. For inferential purpose a Bayesian approach via Markov Chain Monte Carlo (MCMC) was considered. They developed a Bayesian case deletion influence diagnostics based on the Kullback-Leibler divergence. Their newly developed procedures were illustrated on artificial and real HIV data.

Grohn et al. (1998) worked on the effect of seven diseases on the Culling of 7523 Holstein Dairy cows in New York State. The cows were from 14 herds and had calved between January 1, 1994 and December 31, 1994; all cows were followed until September 30, 1995. Survival analysis was performed using the Cox proportional hazards model to incorporate time-dependent covariates for diseases. Different intervals representing stages of lactation were considered for effects of the diseases. Five models were fitted to test how milk yield and conception status modified the effect of diseases on culling. Covariates in the models included parity, calving season, and time-dependent covariates measuring diseases, milk yield of the current lactation, and conception status. Data were stratified by herd. The seven diseases and lactational risks under consideration were milk fever (0.9%), retained placenta (9.5%), displaced abomasum (5.3%, ketosis (5.0%), metritis (4.2%), ovarian cysts (10.6%), and mastitis (14.5%).

Older cows were at a much higher risk of being culled. Calving season had no effect on culling. Higher milk yield was protective against culling. Once a cow had conceived again, her risk of culling dropped sharply. In all models, mastitis was an important risk factor throughout lactation. Milk fever, retained placenta, displaced abomasum, ketosis, and ovarian cysts also significantly affected culling at different stages of lactation. Metritis had no effect on culling. The magnitude of the effects of the diseases decreased, but remained important, when milk yield and conception status were included as covariates. Results indicated that diseases have an important impact on the actual decision to cull and the timing of culling. Parity, milk yield, and conception status are also important factors in culling decisions.

Reproduction can be affected directly by toxic chemicals, or indirectly via effects on feeding, growth or maintenance because these processes are intimately linked to each other. The Dynamic Energy Budget (DEB) theory provides a mechanistic basis that has been tested against many experimental data, Kooijman (1993) and Kooijman and Bedaux (1996) presented a statistical analysis of routine toxicity tests on Daphnia survival and reproduction based on insights from the DEB theory. They compare a formulation in terms of effects on survival during oogenesis to various direct and indirect effects on the energetics of reproduction. All formulations characterize the effects by a no-effect concentration, a tolerance concentration and the elimination rate. They conclude that all options lead to similar no-effect levels and compare the analysis to the standard NOEC/EC50 analysis which concluded that their analysis is both simpler and more effective.

Svard and Price (2001) report the long-term survival rate of the Oxford Knee in a series of patients with anteromedial osteoarthritis in which the operations were performed by three surgeons in a non-teaching hospital in Sweden. All the knees had an intact anterior cruciate ligament, a correctable varus deformity

and full-thickness cartilage in the lateral compartment. Thirty-seven patients had died; the mean time since operation for the remainder was 12.5 years (10.1 to 15.6). Using the endpoint of revision for any cause, the outcome for every knee was established. Six had been revised (4.8%). At ten years there were 94 knees still at risk and the cumulative survival rate was 95.0% (95% confidence interval 90.8 to 99.3). This figure is similar to that reported by the designers of the prosthesis below; Goodfellow and OConnor (1978) introduced the Oxford prosthesis, with congruent mobile bearings, for arthroplasty of the knee. In 1982, the first unicompartmental replacement with the Oxford prosthesis was performed. The implant was designed in the belief that the large areas of contact provided by the congruous articulation would diminish polyethylene wear and improve the long-term survival of unicompartmental arthroplasty. White and Ludkowski (1991) defined the clinicopathology of 'anteromedial osteoarthritis' and suggested that its anatomical features made it suitable for unicompartmental replacement. Murray et al. (1998) reported a cumulative survival rate of 98% (confidence interval (CI) 95 to 100) at ten years for the designer's own series of 144 arthroplasties performed for anteromedial osteoarthritis. Lewold et al. (1995) described the results of 699 Oxford replacements (medial and lateral) enrolled in the Swedish Knee Arthroplasty Register between 1983 and 1992. They found a cumulative survival rate at six years of only 89%. Most of the failures (70%) had occurred in the first two years after surgery and dislocation of the bearing was found to be the commonest cause of failure. They cast doubt on whether the good results reported by the designer could be achieved elsewhere and suggested that to validate this a well-documented series from an independent center was required.

Wintrebert et al. (2005) worked on Joint Modelling of Breeding and Survival in the Kittiwake Using Frailty Models. They saw that assessment of population dynamics is central to population dynamics and conservation. In structured populations, matrix population models based on demographic data have been

widely used to assess such dynamics. Although highlighted in several studies, the influence of heterogeneity among individuals in demographic parameters and of the possible correlation among these parameters has usually been ignored, mostly because of difficulties in estimating such individual-specific parameters. Several approaches have been used in the animal ecology literature to establish the association between survival and breeding rates. However, most are based on observed heterogeneity between groups of individuals, an approach that seldom accounts for individual heterogeneity. Few attempts have been made to build models permitting estimation of the correlation between vital rates. For example, survival and breeding probability of individual birds were jointly modelled using logistic random effects models by Cam et al. (2002). Wintrebert et al. (2005) therefore adopted the survival analysis approaches from epidemiology. They model the survival and the breeding probability jointly using a normally distributed random effect (frailty). Conditionally on this random effect, the survival time is modelled assuming a lognormal distribution, and breeding is modelled with a logistic model. Since the deaths are observed in yearly intervals; - they also took into account that the data are interval censored. The joint model is estimated using classic frequentist methods and also MCMC techniques in Winbugs. The association between survival and breeding attempt is quantified using the standard deviation of the random frailty parameters. They applied joint model on a large data set of 862 birds, that was followed from 1984 to 1995 in Brittany (France). Survival was positively correlated with breeding indicating that birds with greater inclination to breed also had higher survival.

Pocock et al. (1982) illustrated methods of survival analysis for long-term follow-up studies, by a study of mortality in 3878 breast cancer patients in Edinburgh followed for up to 20 years. The problems of life tables, advantages of hazard plots and difficulties in statistical modelling are demonstrated by studying the relationship between survival and both clinical stage and initial

menopausal status at diagnosis. To assess the 'curability' of breast cancer, mortality by year of follow-up is compared with expected mortality using Scottish age-specific death rates. Techniques for analysing such relative survival data include age-corrected life tables, ratio of observed to expected deaths and excess death rates. Finally, an additive hazard model was developed to incorporate covariates in the analysis of relative survival and curability.

Singer and Willett (1991) saw that psychologists studying whether and when events occur face unique design and analytic difficulties. The fundamental problem was how to handle censored observations, the people for whom the target event does not occur before data collection ends. The methods of survival analysis overcame these difficulties and allow researchers to describe patterns of occurrence, compare these patterns among groups, and build statistical models of the risk of occurrence over time. A unified description of survival analysis that focuses on the study design and data analysis was presented, showing how psychologists have used the methods during the past decade and identifying new directions for future application. The presentation was based on the experiences with the methods in modeling employee turnover and examples drawn from research on mental health, addiction, social interaction, and the life course.

Abeysundara (2010) said estimating bivariate and marginal densities of paired survival data becomes more challenging when only one component is censored. If both components are censored or both are not censored, a bivariate version of Kaplan-Meier remains as a consistent estimator. But if only one variable is censored, Kaplan-Meier fails to take advantage of the information of the remaining variable. The method proposed by Akritas and Keilegom considered the case of single censoring as well as double censoring, a situation that is typical in medical studies. He therefore estimated the correlation between two variables in paired survival data at the presence of double and single censoring

via nonparametric approaches. He used the estimates of nonparametric bivariate distribution and marginal distribution of each variable proposed by Akritas and Keilegom. These estimates were based on conditional distribution functions considering only those pairs where the value of the conditioning variable is uncensored. He then applied the method on Diabetic Macular Edema (DME) data to estimate densities and correlation between time to cure for right and left eye.

Arnold (2013) studied the performance, performance persistence, survival and flow of Commodity Trading Advisors, also known as CTAs or Managed Futures Funds. She identified two main trading styles: Systematic and Discretionary CTAs which were the main focus of her thesis. She separated Systematic CTAs into trend-followers with differing trading horizon.

Firstly, she investigated the differences in mortality between Systematic and Discretionary CTAs. She saw that Systematic CTAs have a higher median survival than Discretionary CTAs, 12 vs. 8 years. Therefore proposed new filters that would better identify real failures among funds in the graveyard database. Separating graveyard funds into real failure she re-examine the attrition rate of CTAs. The real failure rate was 11.1%, lower than the average yearly attrition rate of 17.3% of CTAs. The effect of various covariates including several downside risk measures was investigated in predicting CTA failure. Controlling for performance, HWM, minimum investment, fund age and lockup, funds with higher downside risk measures had a higher hazard rate. Compared to other downside risk measures, the volatility of returns was less able to predict failure. Funds that received larger inflows were able to survive longer than funds that do not. Large Systematic CTAs have the highest probability of survival.

The second part studied the performance and performance persistence of Systematic and Discretionary CTAs. Controlling for biases, after fees the average CTA was able to add value. These results were strongest for large Systematic

CTAs. She then extended the seven factor model of Fung-Hsieh (2004a) and found that this model was better able to explain the returns of Systematic rather than Discretionary CTAs. Found three structural breaks in the risk loadings of CTAs different to hedge fund breaks: September 1998, March 2003 and July 2007. Using these breaks showed that systematic CTAs were able to deliver significant alpha in every sub-period. Also found evidence of significant performance persistence. However, these findings were heavily contingent on the strategy followed: the persistence of Discretionary CTAs was driven by small funds whereas large funds drive the performance persistence of Systematic funds. These results had important implications for institutional investors who faced capital allocation constraints. They also suggest that contrary to the previous findings, the CTA industry does not appear to be heading towards zero alpha.
The final section looked at the relationship between fund-flows and performance. Investors chase past performance, the fund-flow-performance was significant and concave for some strategies. Although there was some long-term performance persistence of Systematic funds with the highest inflows, there was no smart money effect in the CTA literature. She found no evidence of capacity constraints among Systematic CTAs. Investors were not able to smartly allocate funds to future best performers and take full advantage of the liquidity that CTAs offered.

Dalby (2011) thesis described, analyzed and applied the Solvency II on life and pension insurance by using the standard formulas in the Quantitative Impact Study 5 (QIS5) to calculate the Solvency Capital Requirement (SCR). He specifically examined the consequences for the Norwegian occupational defined benefit schemes, primarily for the private sector. The standard formulas in QIS5 to some extent specify stress scenarios without giving explicit formulas as they should be exact for the application. He therefore outlined exact formulas for the Norwegian occupational defined benefit schemes, both for the net expected cash flows and for the stressed cash flow. Latter he gave a method for calculating

the stressed survival and hazard rate functions. He also priced the embedded interest rate guarantee using market consistent prices from the Norwegian swaption market. Bonds specifically and redistribution of cash flows generally were used to improve the precision. Using the contract boundary principle in Solvency II he based his calculations on that all policies were converted to paid up policies. This may primarily be relevant for pension schemes in the private sector. However, formulas for active policies were also given. At the end he performed a full QIS5 consequence study for a Norwegian pension fund, by developing algorithms in Mathematica to perform the necessary calculations.

Gustafsson (2009) thesis was the application of survival analysis to predict policy churns in a non-life insurance industry. Models and methods were applied to estimate survival probabilities on customer-level in a competing risk setting, where churns occur of different types of causes. By following motor policy holders over a 3-year period, probabilites are estimated which enables scoring of customers, especially those likely to churn within this time period. Cause-specific semiparametric hazard functions were modelled with Cox regression given customer data at the beginning of the study period. The models were estimated from data on private customers in the Danish insurance company Codan. The main conclusion was that time-fixed covariate and time-invariant effect models that were used for prediction might be an over-simplification of churns on customer-level, as they disregard the impact of customers-specific events during followup. This suggested more flexible models when analysing churns.

Tukan (2012) ran two logit regressions analyzing quality of emergency room care as measured by survival rate and a wellness indicator of patient return within 72 hours of the initial visit in the United States healthcare industry. The data for these regressions represents January through October 2011 individual level

data, about 57,000 observations of patient visits gathered from one emergency room in a low socioeconomic, urban demographic region. These logit results illustrated that emergency room quality of care as measured by survival rate was most affected by acuity, age, being Hispanic, and having Medicaid insurance. For the wellness regression, the presence of a primary physician, being African American or Asian, having no insurance, primary insurance, or Medicaid, and age were considered.

Gong (2008) said, estimating causal effects in clinical trials is often complicated by treatment noncompliance and missing outcomes. In time-to-event studies, estimation is further complicated by censoring. Censoring is a type of missing outcome, the mechanism of which may be non-ignorable. While new estimates have recently been proposed to account for noncompliance and missing outcomes, few studies have specifically considered time-to-event outcomes, where even the intention-to-treat (ITT) estimator is potentially biased for estimating causal effects of assigned treatment. He developed a series of parametric potential-outcome (PPO) survival models, for the analysis of randomised controlled trials (RCT) with time-to-event outcomes and noncompliance. Both ignorable and non-ignorable censoring mechanisms were considered. He approached model-fitting from a likelihood-based perspective, using the EM algorithm to locate maximum likelihood estimators. He also gave new formulations for the average causal effect (ACE) and the complier average causal effect (CACE) to suit survival analysis. He re-analysed the HIP breast cancer trial data (Baker, 1998); (Shapiro et al., 1988) using the specific PPO-survival models, the Weibull and log-normal based PPO-survival models, and assumed that the failure time and censored time distributions both follow Weibull or log-normal distributions. Furthermore, an extended PPO-survival model was derived, which permitted investigation into the impact of causal effect after accommodating certain pre-treatment covariates. Finally he compared the Frangakis-Rubin (F-R) model

(Frangakis and Rubin, 1999) to the HIP breast cancer trial data.

Alberts (2006) saw that mobile telecommunication market in the Netherlands has changed from a rapidly growing market, into a state of saturation and fierce competition. The focus of telecommunication companies has therefore shifted from building a large customer base into keeping customers 'in house'. Customers who switch to a competitor are so called churned customers. Churn prevention, through churn prediction, is one way to keep customers 'in house'. In his study he focused solely on prepaid customers. In contrast to postpaid customers, prepaid customers are not bound by a contract. The central problem concerning prepaid customers is that the actual churn date in most cases is difficult to assess. This is a direct consequence of the difficulty in providing a unequivocal definition of churning and a lack of understanding in churn behavior. To overcome the problem he presented the predictive churn model based on the theory of survival analysis. Also, to compare the performance of the extended Cox model with the established predictive models he used a decision tree. Both models performed approximately similar with a sensitivity ranging from 93% to 99% and a specificity ranging from 92% to 97%, depending on the model and the churn definition.

CHAPTER 3

METHODOLOGY

3.1 Survival Analysis

Survival analysis is a statistical technique used to describe and quantify time to an event data, (Stevenson, 2007). Survival data is a term used for describing data that measure the time to a given event of interest. The name survival data arose because originally events were most often deaths. The term survival data is now used for all kind of events. In all cases, the event can be seen as a transition from one state to another, (Wintrebert, 2007). The response is often referred to as a failure time, survival time, or event time. The term 'survival time' specifies the length of time taken for failure to occur. Situations where survival analyses have been used in epidemiology include:

- Survival of patients after surgery.
- The length of time taken for cows to conceive after calving.
- The time taken for a farm to experience its first case of an exotic disease.

3.2 Describing Time to an Event

In this section, the probability tools usually encountered in survival analysis and their properties are described.

Let T be the time variable, considered as a positive real valued variable, having a continuous distribution with finite expectation. For applications, this variable represents the time being in a given state or the time between two events. Several functions characterize the distribution of T:

3.2.1 Probability Density Function

In here, the variable under consideration is the length of time taken for an event to occur (e.g. death). It is also known as the death density function and denoted by f(t). The proportion of individuals who have died as a function of t is known as the cumulative death distribution function and is called F(t).

F(t) = Pr(an individual fails before t)

$$F(t) = P[T \leqslant t], t \geqslant 0 \tag{3.1}$$

When T is a survival time, F(t) was the probability that a randomly selected subject from the population will die before time t. If T is a continuous random variable, then it has a density function f(t), which is related to F(t) through the following equations

S(t) =Pr(an individual survives longer than t)

$$F(t) = P[T \leqslant t] = 1 - S(t), \tag{3.2}$$

If F is differentiable, then the derivative, which is the density function is denoted as;

$$f(t) = \frac{dF(t)}{dt} = F' \tag{3.3}$$

The function f is sometimes called the event density; it is the rate of death or failure events per unit time.

3.2.2 Survival Function

Survival Function gives the probability of surviving or been event-free beyond time t. It is denoted by S(t) and given by;

$$S(t) = Pr(T > t) = \int_t^{\infty} f(x)dx = 1 - F(t) \tag{3.4}$$

Survival Function, S(t) is a non-increasing function over time taking on the value 1 at t = 0, i.e., S(0) = 1. For a proper random variable T, S(∞) = 0, which means that everyone will eventually experience the event. However, there is the possibility that S(∞)>0. This corresponds to a situation where there is a positive probability of not "dying" or not experiencing the event. For example, if the event of interest is the time from response until disease relapse and the disease has a cure for some proportion of individuals in the population, then S(∞)>0, where S(∞) corresponds to the proportion of cured individuals, (Tsiatis and Zhang, 2005)

Similarly, a survival event density function can be defined as

$$s(t) = S' = \frac{dS(t)}{dt} = \int_t^{\infty} f(x)dx = \frac{d}{dt}[1 - F(t)] = -f(t) \tag{3.5}$$

3.2.3 Hazard Function

Hazard Function represents the probability that an individual alive at t experiences the event in the next period. The instantaneous rate at which a randomly-selected individual known to be alive at time (t-1) will die at time t is called the conditional failure rate or instantaneous hazard, h(t). Mathematically, instantaneous hazard equals the number that fail between time t and time t+$\triangle$(t) divided by the size of the population at risk at time t, divided by $\triangle$(t). This gives the proportion of the population present at time t that fail per unit time.

Instantaneous hazard is also known as the force of mortality, the instantaneous

death rate, or the failure rate. The **mortality rate** at time t, where t is generally taken to be an integer in terms of some unit of time (e.g., years, months, days, etc), is the proportion of the population who fail (die) between times t and t+1 among individuals alive at time t, , i.e.,

$$m(t) = P[t \leqslant T < t+1 | T \geqslant t] \tag{3.6}$$

The hazard rate λ(t) is the limit of a mortality rate if the interval of time is taken to be small (rather than one unit). The hazard rate is the instantaneous rate of failure (experiencing the event) at time t given that an individual is alive at time t. Specifically, hazard rate λ(t) is defined by the following equation

$$\lambda(t) = \lim_{\triangle(t) \longrightarrow 0} \frac{P[t \leqslant T < t + \triangle t \mid T \geqslant t]}{\triangle t} \tag{3.7}$$

Therefore, if λ(t) is very small, we have

$$P[t \leqslant T < t + \triangle\,(t) | T \geqslant t] \approx \lambda(t) \triangle\,(t) \tag{3.8}$$

The definition of the hazard function implies that

$$\lambda(t) = \frac{\lim\limits_{\triangle(t) \longrightarrow 0} \frac{P[t \leqslant T < t+\triangle t]}{\triangle t}}{P[T \geqslant t]} = \frac{f(t)}{S(t)} \tag{3.9}$$

$$= -\frac{S^{'}(t)}{S(t)} = -\frac{dlog\{S(t)\}}{dt} \tag{3.10}$$

From this, we can integrate both sides to get

$$\wedge(t) = \int_0^t \lambda(t)dt = -logS(t) \tag{3.11}$$

where $\wedge$(t) is referred to as the cumulative hazard function. Here we used the fact that S(0) = 1.

Hence

$$S(t) = e^{-\wedge(t)} = e^{-\int_0^t \lambda(t)dt} \tag{3.12}$$

3.3 Censoring

In longitudinal studies exact survival time is only known for those individuals who show the event of interest during the follow-up period. For others (those who are disease free at the end of the observation period or those that were lost) all we can say is that they did not show the event of interest during the follow-up period. These individuals are called censored observations. Therefore, Censoring is present when we have some information about a subject's event time, but we don't know the exact event time. An attractive feature of survival analysis is that we are able to include the data contributed by censored observations right up until they are removed from the risk set, (Stevenson, 2007). There are generally three reasons why censoring might occur:

• A subject does not experience the event before the study ends

• A person is lost to follow-up during the study period

• A person withdraws from the study because of death (if death is not the event of interest) or some other reason.

The following terms are used in relation to censoring: right censoring, left censoring and interval censoring. Right censoring is the case where an individual may experience the event of interest after the given time t; we know only that the individual is alive (not failed) up to the given time. Left censoring is where an individual has experienced the event of interest prior to the start of the study. Interval censoring is where the only information is that the event occurs within some interval of time.

Let T and C represent the failure and censoring time respectively. Then the three types of censoring can be expressed mathematically as the following:

Right censoring : T $\in (C_r, \infty)$ and it is known only that the failure time T is greater than the observed censoring time Cr, but exact value of failure time is unobservable.

Left censoring : T $\in (0, C_l)$ and it is known only that the failure time T is less than the observed censoring time C_l, but its exact value is unobservable.

Interval censoring : T $\in (C_l, C_r)$ and it is known only that the failure time T is less than the observed right censoring time C_r and greater than the observed left censoring C_l, but its exact value is unobservable.

For example, if individuals are right censored at time C_i we know that their failure time would be at least greater than t, that is T > t, Klein (1997).

Besides censoring, there is another feature, called truncation, which may also be present in some time-to-event studies. In this thesis, the causal effect and potential-outcome approach does not consider truncated data. In fact, the approach proposed in this thesis is based only on right censoring.

3.3.1 Censoring Mechanisms

There are several types of censoring schemes which lead to different likelihood functions for inference, Cox and Oakes (1984). These are delineated below.

• **Type I Censoring**: For Type I censoring, the event is observed only if it occurs prior to some pre-specified time. Censoring time may vary from individual to individual. Owing to cost or time considerations, the investigators may terminate the study or report the results before all subjects realize their events. If no accidental loses or subject withdrawals, censored observations

have times equal to the length of study time period; the censored time for each individual is the same and can be treated as a fixed time for a certain trial.

• **Type II Censoring**: For Type II censoring, the study continues until r individuals experience the pre-specified event of interest. This number r may be some predetermined integer. Experiments for testing equipment failure time often involve this type of censoring. In this case, the censored time for each individual may be different and can be treated as a random variable. However, Type II censoring rarely occurs in clinical trials involving human subjects.

• **Random censoring**: Random censoring involves what is called competing risks scenario. In this, individuals experience other competing events which may cause them to be removed from the study, and the primary event of interest is then not observed.

3.4 Estimation of survival functions

Survival analysis can be based either on an assumption about survival following a certain distribution or on direct observation based on the actual data. Both procedures require dealing with censored and uncensored observations. The most commonly used survival distributions are the negative exponential distribution, the Weibull distribution, the Gumbel distribution, the Logarithmic normal distribution or their combinations. Which type of function is best at describing the survival distribution is mainly dependent on the data and can be carried out with the Kaplan-Meier method.

3.4.1 Kaplan-Meier

Kaplan-Meier (KM) estimator, also known as the product-limit estimator, is an estimator for estimating the survival function from lifetime data, (Kaplan and

Meier, 1958). It is a non-parametric maximum likelihood estimate of S(t), and incorporates information from all of the observations available, both censored and uncensored, by considering any point in time as a series of steps defined by the observed survival and censored times. In the medical research, it might be used to measure the fraction of patients living for a certain amount of time after treatment. An Insurer might measure the time of purchase of an insurance policy until loss occurs or time of compensation after notification of a loss. The economist might measure the length of time people remain unemployed after job losses. An engineer might measure the time until failure of machine parts. Survival function is a series of horizontal steps of declining magnitude which when a large enough sample is taken approaches the true survival function for the population. The value of the survival function between successive distinct sampled observations is assumed to be constant. An important advantage of Kaplan- Meier curve is that, the method can take into account some types of censored data, particularly right-censoring, which occurs if an insured withdraws from a study, i.e., is lost from the sample before the final outcome is observed. On the plot, small vertical tick-marks indicate losses, where an insured survival time has been right-censored. When no truncation or censoring occurs the Kaplan Meier curve is equivalent to the empirical distribution function.

Suppose $t_1 \leqslant t_2 \leqslant, ..., \leqslant t_k$ are the ordered failure times.

For $t_k \leqslant t \leqslant t_{(k+1)}$, the probability of surviving beyond time t is;

$$S_{KM}(t) = P(T > t) = P(T \geqslant t_{(k+1)}) \tag{3.13}$$

Implies $S_{KM}(t) = P(T \geqslant t_1, T \geqslant t_2, \ldots, T \geqslant t_k)$

$$S_{KM}(t) = P(T > t_1) \prod_{j=1}^{k} P\{T \geqslant t_{j+1} \mid T > t_j\} \tag{3.14}$$

But P(T>t) = S(0) = 1,

Implies; $S_{KM}(t) = \prod_{j=1}^{k} P\{T \geqslant t_{j+1} \mid T > t_j\}$

$S_{KM}(t) = \prod_{j=1}^{k}[1 - P\{T = t_j \mid T > t_j\}]$

$\hat{S}_{KM}(t) = \prod_{i=1}^{j}(1 - \hat{\lambda}_j)$

But $\hat{\lambda}_j = d_j/r_j$

Hence, the Kaplan-Meier estimator of the survival function S(t) is given as;

$$\hat{S}_{KM}(t) = \prod_{i=1}^{j}(1 - \frac{d_j}{r_j}), for 0 \leqslant t \leqslant t^{+} \tag{3.15}$$

where;

$\hat{S}_{KM}(t)$ = Kaplan-Meier estimator of survival at time t

d_j= Number of failures (claims) at time t_j

r_j= Number of individuals alive (at risk) just before the time t_j, including those who will die(claim) at t_j

t_j, j= 1, 2, ..., n is the total set of failure(claims) times recorded (with t^{+} the maximum failure time).

3.4.2 Variance of the Kaplan Meier estimator (Greenwood formula)

The Kaplan-Meier estimator is a statistic, and several estimators are used to approximate its variance. One of the most common of such estimators is the Greenwood's formula. For $t_k \leqslant t \leqslant t_{(k+1)}$,

From

$$log\hat{S}_{KM}(t) = \sum_{r_j<k}^{j} log(1 - \frac{d_j}{r_j}) \tag{3.16}$$

$$Var[log\hat{S}_{KM}(t)] = \sum_{r_j<k}^{j} Var[log \prod_{i=1}^{j}(1 - \frac{d_j}{r_j})] \tag{3.17}$$

$$Var[log\hat{S}_{KM}(t)] = \sum_{r_j<k}^{j} Var[log(1 - \lambda_j)] \tag{3.18}$$

Applying the Delta Method,

$$= \sum_{r_j<k}^{j} [(-1)/(1 - \hat{\lambda}_j)]^2 Var(\hat{\lambda}_j) \tag{3.19}$$

By large sample theory,

$$\hat{S}(t) \sim N[S(t), (S(t)(1 - S(t))/n)] \tag{3.20}$$

But if,

$\hat{\lambda}_j \sim N[\lambda_j, (\lambda_j(1 - \lambda_j))/r_j]$

Then

$$Var(\hat{\lambda}_j) = (\lambda_j(1 - \lambda_j))/r_j \tag{3.21}$$

Var[log $\hat{S}_{KM}(t)$] = $\sum_{r_j<k}^{j} \hat{\lambda}_j/((1 - \lambda_j))/r_j)$, but $\hat{\lambda}_j - d_j/r_j$

$$Var[log\hat{S}_{KM}(t)] = \sum_{r_j<k}^{j} \frac{d_j/r_j}{(1 - d_j/r_j)r_j} \tag{3.22}$$

If Var[log $\hat{S}_{KM}(t)$] = $\sum_{r_j<k}^{j} \frac{d_j}{(1 - d_j/r_j)r_j}$

Then Var[log $\hat{S}_{KM}(t)$] = $\sum_{r_j<k}^{j} \frac{d_j}{(r_j - d_j)r_j}$

Therefore,

$$Var[log\hat{S}_{KM}(t)] = \sum_{r_j<k}^{j} \frac{d_j}{(r_j - d_j)r_j} \tag{3.23}$$

Also,

$\hat{S}_{KM}(t) = exp[log\hat{S}_{KM}(t)]$

And Var$[\hat{S}_{KM}(t)] = [\hat{S}_{KM}(t)]^2 Var[log\hat{S}_{KM}(t)]$

Hence, the Greenwood Formula is;

$$Var[\hat{S}_{KM}(t)] = [\hat{S}_{KM}(t)]^2 \sum_{r_j<k}^{j} \frac{d_j}{(r_j - d_j)r_j} \tag{3.24}$$

In some cases, one may wish to compare different Kaplan-Meier curves. This may be done by several methods including: the log rank test and the Cox proportional hazards test.

3.4.3 Confidence interval

The standard error of a large sample for $\hat{S}_{KM}(t)$ is given by;

$$Se[\hat{S}_{KM}(t)] = \sqrt{Var[\hat{S}_{KM}(t)]} \tag{3.25}$$

$$= \sqrt{[\hat{S}_{KM}(t)]^2 \sum_{r_j<k}^{j} \frac{d_j}{(r_j - d_j)r_j}} \tag{3.26}$$

$$= \hat{S}_{KM}(t) \sqrt{\sum_{r_j<k}^{j} \frac{d_j}{(r_j - d_j)r_j}} \tag{3.27}$$

The point-wise confidence interval for the Kaplan-Meier estimate is;

$$\hat{S}_{KM}(t) \pm Z_{1-\propto/2} \hat{S}_{KM}(t) \sqrt{\sum_{r_j<k}^{j} \frac{d_j}{(r_j - d_j)r_j}} \tag{3.28}$$

3.5 Survival Curves

In survival curves symbols represent each event time, either a death (or claim) or a censored time. Survival curves estimate the probability that a participant survives past a certain period by locating the period on the X axis and reading up and over to the Y axis. The median survival is estimated by locating 0.5 on the Y axis and reading over and down to the X axis.

3.5.1 Comparing Survival Curves

We are often interested in assessing whether there are differences in survival (or cumulative incidence of event) among different groups of participants. For example, in a clinical trial with a survival outcome, we might be interested in comparing survival between participants receiving a new drug as compared to a placebo (or standard therapy). In an observational study, we might be interested in comparing survival between men and women, or between participants with and without a particular risk factor (e.g., hypertension or diabetes). There are several tests available to compare survival among independent groups. This thesis compares the survival among different covariates using the log rank test.

3.6 The Log Rank Test

The log rank test is a popular test to test the null hypothesis of no difference in survival between two or more independent groups. The test compares the entire survival experience between groups and can be thought of as a test of whether the survival curves are identical (overlapping) or not. Survival curves are estimated for each group, considered separately, using the Kaplain-Meier method and compared statistically using the log rank test. The log rank test presented in this thesis is linked to the chi-square test statistic and compares observed to expected numbers of events at each time point over the follow-up period using

R statistical package. Mathematically the test statistic for the log rank test is represented as;

$$\chi^2 = \sum \frac{(\sum O_{jt} - \sum E_{jt})^2}{\sum E_{jt}} \tag{3.29}$$

where $\sum O_{jt}$ represents the sum of the observed number of events in the jth group over time (e.g., = 1,2) and $\sum E_{jt}$ represents the sum of the expected number of events in the jth group over time.

The sums of the observed and expected numbers of events are computed for each event time and summed for each comparison group. The log rank statistic has degrees of freedom equal to k-1, where k represents the number of comparison groups. In this thesis, k = 2 so the test statistic has 1 degree of freedom.

There are several variations of the log rank statistic as well as other tests to compare survival functions between independent groups. For example, a popular test is the modified Wilcoxon test which is sensitive to larger differences in hazards.

3.7 Cox-Regression Model

One of the most popular regression techniques for survival outcomes is Cox proportional hazards analysis. There are several important assumptions for appropriate use of the Cox proportional hazards regression model, including

1. Independence of survival times between distinct individuals in the sample,
2. A multiplicative relationship between the predictors and the hazard and,
3. A constant hazard ratio over time.

Cox regression (or proportional hazards regression) allows analyzing the effect of several risk factors on survival. The probability of the endpoint (death, or any other event of interest, e.g. making of a claim) is called the hazard. A probability

must lie in the range 0 to 1. However, the hazard represents the expected number of events per one unit of time. As a result, the hazard in a group can exceed 1. The hazard is modeled as:

$$\lambda(t; z) = exp(z\beta)\lambda_0(t) \tag{3.30}$$

For the j_{th} individual let the values of z be $z_j = (z_{1j}, \ldots, z_{pj})$
Taking the natural logarithm (ln) of both sides, to produce the following which relates the log of the relative hazard to a linear function of the predictors or risk factors;

$$ln\{\lambda(t)/\lambda_0(t)\} = \beta_1 z_{j1} + \beta_2 z_{j2} + \cdots + \beta_k z_{jk} \tag{3.31}$$

where $\lambda(t)$ is the expected hazard at time t, $\lambda_0(t)$ is the unknown baseline hazard function at time t, β is a p x 1 vector of unknown parameters. The z's are assume to be independent of time(constant covariates), and j is the number of variables considered in the study.

3.7.1 Estimation of the Cox Proportional Hazard Model

Here, we describe how estimates are obtained for the parameters of the Cox model. The parameters are the $\beta's$ in the general Cox model formula. The corresponding estimates of these parameters are called Maximum Likelihood (ML) Estimates.

The Cox proportional hazards model is called a semi-parametric model, because there are no assumptions about the shape of the baseline hazard function. A Cox model was explicitly designed to be able to estimate the hazard ratios without having to estimate the baseline hazard function in this study. There are however, other assumptions as noted above (i.e., independence, changes in predictors produce proportional changes in the hazard regardless of time,

and a linear association between the natural logarithm of the relative hazard and the predictors). Therefore, the likelihood function is actually called a "partial" likelihood function Cox (1975) because the likelihood formula considers probabilities only for those subjects who fail.

3.8 Hypothesis

To compare the survival between variables the hypothesis to be used is as follows:

H_o : The survival curves are identical (or $S_{1t} = S_{2t}$)

H_1 : The survival curves are not identical (or $S_{1t} \neq S_{2t}$, at any time t)(p<0.05)

The dependent variable is the time it takes for claim to occur and to be paid (survival time). Gender, age, marital status, type of policy, type of vehicle are the independent variables over time t and nature of the claim is a time-dependent covariate incorporated into the survival analysis model since the risk changes over time. The hypothesis then undertakes the form:

$$h(t, z) = \beta_o + \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3 + \beta_4 z_4 + \beta_5 z_5 + \beta_6 z_6 \tag{3.32}$$

For motor insurance policy holders who claimed and are paid.

The $\beta_i' s$ are estimated coefficients of the regression model where,

z_1 : gender,

z_2 : age,

z_3 : marital status,

z_4 : type of policy,

z_5 : type of vehicle,

z_6 : nature of claim.

These variables are represented by dummy variables(0 or 1) using SPSS software package. The required model for the research is however, dependent on the significance of each of these factors at a level of significance of 0.05.

3.9 Proportionality Assumption

A very important assumption for the appropriate use of the log rank test and the Cox proportional hazards regression model is the proportionality assumption.

Specifically, we assume that the hazards are proportional over time which implies that the effect of a risk factor is constant over time. There are several approaches to assess the proportionality assumption, some are based on statistical tests and others involve graphical assessments.

In the statistical testing approach, predictor by time interaction effects are included in the model and tested for statistical significance. If one (or more) of the predictor by time interactions reaches statistical significance (e.g., $p<0.05$), then the assumption of proportionality is violated. An alternative approach to assessing proportionality is through graphical analysis. There are several graphical displays that can be used to assess whether the proportional hazards assumption is reasonable. These are often based on residuals and examine trends(or lack thereof) over time, Hosmer and Lemeshow (1999).

If either a statistical test or a graphical analysis suggest that the hazards are not proportional over time, then the Cox proportional hazards model is not appropriate, and adjustments must be made to account for non-proportionality. One approach is to stratify the data into groups such that within groups the hazards are proportional, and different baseline hazards are estimated in each stratum.

3.10 Competing Risks

The competing risks issue is one in which there are several possible outcome events of interest. For example, this thesis is to determine if the type of insurance affects the time it takes for a claim to be settled and to establish which variables contribute significantly to the time for a claim to be settled in the insurance industry in Ghana. This variables include; gender, age, marital status, type of policy, type of vehicle, and nature of the claim. The investigator measures whether each of the component outcomes occurs during the study observation period as well as the time to each distinct event. The goal of the analysis is to determine the risk factors for each specific outcome and the outcomes are correlated; (Kalbfleisch and Prentice, 2002).

CHAPTER 4

DATA ANALYSIS AND RESULTS

4.1 Introduction

In this chapter, the analysis and results were obtained by using the various statistical tools and procedures described in the previous chapter. It includes a brief descriptive analysis of the raw data summarized in tables 4.1, 4.2, 4.3, 4.4, 4.5 and 4.6. The main results were achieved by the Kaplan-Meier (product limit) approach and Cox regression model based on a 5 % level of statistical significance.

Table 4.1 shows the number of insureds that bought motor insurance. The columns shows the covariates, insureds, frequencies and percentages respectively.

Table 4.1: Frequency Distributions of Insureds that bought Motor Insurance

Covariates	Insureds	Frequency	Percent(%)
Gender	Male	658	65.8
	Female	342	34.2
	Total	1000	100
Age	21 - 29	370	37.0
	30 - 45	328	32.8
	46 - 59	237	23.7
	≥ 60	65	6.5
	Total	1000	100
Marital Status	Single	583	58.3
	Married	417	41.7
	Total	1000	100
Policy Type	Comprehensive	559	55.9
	Third Party	441	44.1
	Total	1000	100
Vehicle Type	Saloon	407	40.7
	Station Wagon	219	21.9
	Truck	73	7.3
	Pickup	127	12.7
	Motor Cycle	26	2.6
	Minibus	61	6.1
	Bus	86	8.6
	Total	1000	100.0

Insureds that bought Motor Insurance and the number that Claimed:

Figure 4.1 shows the frequency distribution of insureds that claimed and those that did not claim.

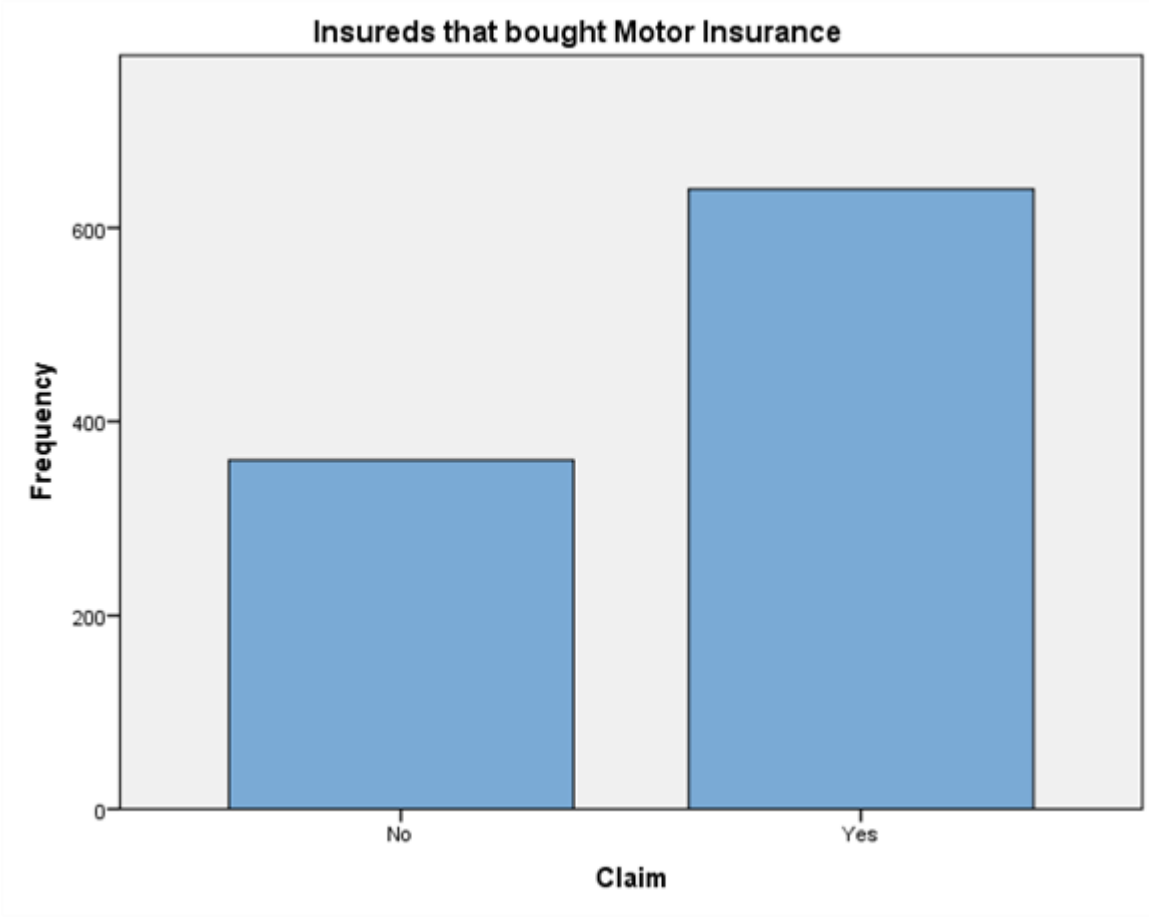

Figure 4.1: A Bar Chart showing Insureds that Claim

Figure 4.1 shows that out of 1,000 insureds that bought motor insurance 640 claimed and 360 are not claimants.

Insureds that Claimed Motor Insurance:

Table 4.2 shows the frequency distribution of insureds that Claimed. The first column shows the covariates while the second shows the claimants and the third and fourth columns shows the frequency and percentages respectively.

Table 4.2: Frequency Distributions of Insureds who Claimed Motor Insurance

Covariates	Claimants	Frequency	Percent(%)
Gender	Male	427	66.7
	Female	213	33.7
	Total	640	100.0
Age	21 - 29	233	36.4
	30 - 45	205	32.0
	46 - 59	159	24.8
	≥ 60	43	6.7
	Total	640	100.0
Marital Status	Single	378	59.1
	Married	262	40.9
	Total	640	100.0
Policy Type	Comprehensive	388	60.6
	Third Party	252	39.4
	Total	640	100.0
Vehicle Type	Saloon	283	44.2
	Station Wagon	151	23.6
	Truck	43	6.7
	Pickup	79	12.3
	Motor Cycle	19	3.0
	Minibus	34	5.3
	Bus	31	4.8
	Total	640	100.0
Nature of Claim	Own Damage	139	21.7
	Own Damage-Total loss	52	8.1
	Theft	36	5.6
	Collision	179	28.0
	Breakage of Windshield	91	14.2
	Third Party Damage & Injury	59	9.2
	Third Party Damage & Injury(Fatal)	84	13.1
	Total	640	100.0

The number of Claimants whose losses have been Paid:

Figure 4.2 shows the frequency distribution of claimants whose losses have been paid and those not paid.

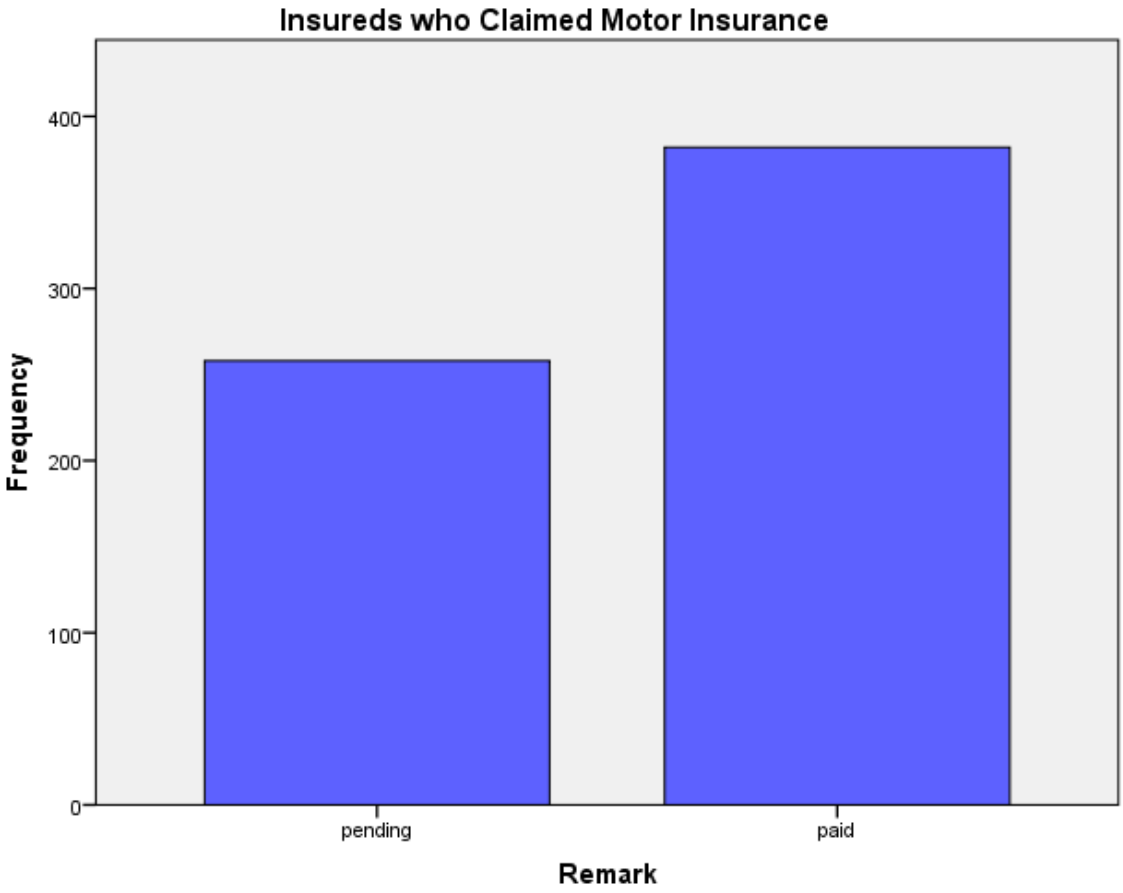

Figure 4.2: A Bar Chart showing Remark of Insureds who Claimed

Figure 4.2 shows that out of the 640 claimants 382 have been paid and 258 have not been paid. This shows that 59.7 % have been paid and 40.3 % have not been paid.

Motor Insurance Claimants by Age:

Table 4.3 shows the frequency distributions of claimants by age. The first column shows the covariates and the second shows the claimants followed by the various categories of age and their percentages respectively.

Table 4.3: Frequency Distributions of Claimants by Age

Covariates	Claimants	21 - 29	(%)	30 - 45	(%)	46 - 59	(%)	≥ 60	(%)	Total	(%)
Gender	Male	154	66.1	126	61.5	118	74.2	29	67.4	427	66.7
	Female	79	33.9	79	38.5	41	25.8	14	32.6	213	33.3
	Total	233	100	205	100	159	100	43	100	640	100
Marital Status	Single	139	59.7	124	60.5	88	55.3	27	62.8	378	59.1
	Married	94	40.3	81	39.5	71	44.7	16	37.2	262	40.9
	Total	233	100	205	100	159	100	43	100	640	100
Policy Type	Comprehensive	144	61.8	123	60.0	99	62.3	22	51.2	388	60.6
	Third Party	89	38.2	82	40.0	60	37.7	21	48.8	252	39.4
	Total	233	100	205	100	159	100	43	100	640	100
Vehicle Type	Saloon	160	68.7	75	36.6	40	25.2	8	18.6	283	44.2
	Station Wagon	62	26.6	53	25.9	29	18.2	7	16.3	151	23.6
	Truck	1	0.4	11	5.4	22	13.8	9	20.9	43	6.7
	Pickup	5	2.1	39	19.0	25	15.7	10	23.3	79	12.3
	Motor Cycle	4	1.7	11	5.4	4	2.5	0	0.0	19	3.0
	Minibus	0	0.0	9	4.4	20	12.6	5	11.6	34	5.3
	Bus	1	0.4	7	3.4	19	11.9	4	9.3	31	4.8
	Total	233	100	205	100	159	100	43	100	640	100
Nature of Claim	Own Damage	49	21.0	48	23.4	35	22.0	7	16.3	139	21.7
	Own Damage-Total loss	13	5.6	25	12.2	14	8.8	0	0.0	52	8.1
	Theft	16	6.9	8	3.9	10	6.3	2	4.7	36	5.6
	Collision	62	26.6	62	30.2	38	23.9	17	39.5	179	28.0
	Breakage of Windshield	27	11.6	26	12.7	27	17.0	11	25.6	91	14.2
	Third Party Damage & Injury	28	12.0	12	5.9	16	10.1	3	7.0	59	9.2
	Third Party Damage & Injury(Fatal)	38	16.3	24	11.7	19	11.9	3	7.0	84	13.1
	Total	233	100	205	100	159	100	43	100	640	100

Table 4.3 shows that the age group with the highest claims was 21 - 29 years with 233 claim cases. The age group 30 - 45 years reported 205 claims and 46 - 59 years also had 159 claims. Lastly the age group 60 years and above had the least number of claims as 43 claims.

Motor Insurance Claimants by Gender:

Table 4.4 shows the frequency distributions of claimants by gender. The first column shows the covariates and the second shows the claimants. The third column shows the number of male claimants and their percentages and the fourth column shows the number of female claimants and their percentages. The last column shows the totals and their percentages.

Table 4.4: Frequency Distributions of Claimants by Gender

Covariates	Claimants	Male	(%)	Female	(%)	Total	(%)
Age	21 - 29	154	36.1	79	37.1	233	36.4
	30 - 45	126	29.5	79	37.1	205	32.0
	46 - 59	118	27.6	41	19.2	159	24.8
	≥ 60	29	6.8	14	6.6	43	6.7
	Total	427	100	213	100	640	100
Marital Status	Single	266	62.3	112	52.6	378	59.1
	Married	161	37.7	101	47.4	262	40.9
	Total	427	100	213	100	640	100
Policy Type	Comprehensive	264	61.8	124	58.2	388	60.6
	Third Party	163	38.2	89	41.8	252	39.4
	Total	427	100	213	100	640	100
Vehicle Type	Saloon	191	44.7	92	43.2	283	44.2
	Station Wagon	94	22.0	57	26.8	151	23.6
	Truck	19	4.4	24	11.3	43	6.7
	Pickup	56	13.1	23	10.8	79	12.3
	Motor Cycle	18	4.2	1	0.5	19	3.0
	Minibus	25	5.9	9	4.2	34	5.3
	Bus	24	5.6	7	3.3	31	4.8
	Total	427	100	213	100	640	100
Nature of Claim	Own Damage	90	21.1	49	23.0	139	21.7
	Own Damage-Total loss	31	7.3	21	9.9	52	8.1
	Theft	25	5.9	11	5.2	36	5.6
	Collision	115	26.9	64	30.0	179	28.0
	Breakage of Windshield	70	16.4	21	9.9	91	14.2
	Third Party Damage & Injury	37	8.7	22	10.3	59	9.2
	Third Party Damage & Injury(Fatal)	59	13.8	25	11.7	84	13.1
	Total	427	100	213	100	640	100

Table 4.4 shows that the number of males involved in motor accident was 427 claimants and the number of females involved was 233 claimants.

Motor Insurance Claimants by Marital Status:

Table 4.5 shows the frequency distributions of claimants by marital status. The first column shows the covariates and the second shows the claimants. The third column shows the number of married claimants and their percentages and the fourth column shows the number of claimants who are not married and their percentages. The last column shows the totals and their percentages.

Table 4.5: Frequency Distributions of Claimants by Marital Status

Covariates	Claimants	Married	(%)	Single	(%)	Total	(%)
Age	21 - 29	94	35.9	139	36.8	233	36.4
	30 - 45	81	30.9	124	32.8	205	32.0
	46 - 59	71	27.1	88	23.3	159	24.8
	≥ 60	16	6.1	27	7.1	43	6.7
	Total	262	100	378	100	640	100
Gender	Male	161	61.5	266	70.4	427	66.7
	Female	101	38.5	112	29.6	213	33.3
	Total	262	100	378	100	640	100
Policy Type	Comprehensive	154	58.8	234	61.9	388	60.6
	Third Party	108	41.2	144	38.1	252	39.4
	Total	262	100	378	100	640	100
Vehicle Type	Saloon	108	41.2	175	46.3	283	44.2
	Station Wagon	67	25.6	84	22.2	151	23.6
	Truck	22	8.4	21	5.6	43	6.7
	Pickup	27	10.3	52	13.8	79	12.3
	Motor Cycle	6	2.3	13	3.4	19	3.0
	Minibus	18	6.9	16	4.2	34	5.3
	Bus	14	5.3	17	4.5	31	4.8
	Total	262	100	378	100	640	100
Nature of Claim	Own Damage	50	19.1	89	23.5	139	21.7
	Own Damage-Total loss	18	6.9	34	9.0	52	8.1
	Theft	18	6.9	18	4.8	36	5.6
	Collision	74	28.2	105	27.8	179	28.0
	Breakage of Windshield	35	13.4	56	14.8	91	14.2
	Third Party Damage & Injury	24	9.2	35	9.3	59	9.2
	Third Party Damage & Injury(Fatal)	43	16.4	41	10.8	84	13.1
	Total	262	100	378	100	640	100

Table 4.5 shows that the number of married claimants involved in motor accident was 262 and those not married was 378.

Motor Insurance Claimants by Policy Type:

Table 4.6 shows the frequency distributions of claimants by policy type. The first column shows the covariates and the second shows the claimants. The third column shows the number of comprehensive policy claimants and their percentages and the fourth column shows the number of third party policy claimants and their percentages. The last column shows the totals and their percentages.

Table 4.6: Frequency Distributions of Claimants by Policy Type

Covarites	Claimants	Comprehensive	(%)	Third Party	(%)	Total	(%)
Age	21 - 29	144	37.1	89	35.3	233	36.4
	30 - 45	123	31.7	82	32.5	205	32.0
	46 - 59	99	25.5	60	23.8	159	24.8
	≥ 60	22	5.7	21	8.3	43	6.4
	Total	388	100	252	100	640	100
Gender	Male	264	68.0	163	64.7	427	66.7
	Female	124	32.0	89	35.3	213	33.3
	Total	388	100	252	100	640	100
Marital Status	Single	234	60.3	144	57.1	378	59.1
	Married	154	39.7	108	42.9	262	40.9
	Total	388	100	252	100	640	100
Vehicle Type	Saloon	174	44.8	109	43.3	283	44.2
	Station Wagon	95	24.5	56	22.2	151	23.6
	Truck	28	7.2	15	6.0	43	6.7
	Pickup	48	12.4	31	12.3	79	12.3
	Motor Cycle	14	3.6	5	2.0	19	3.0
	Minibus	13	3.4	21	8.3	34	5.3
	Bus	16	4.1	15	6.0	31	4.8
	Total	388	100	252	100	640	100
Nature of Claim	Own Damage	137	35.3	2	0.8	139	21.7
	Own Damage-Total loss	52	13.4	0	0.0	52	8.1
	Theft	36	9.3	0	0.0	36	5.6
	Collision	1	0.3	178	70.6	179	28.0
	Breakage of Windshield	91	23.5	0	0.0	91	14.2
	Third Party Damage & Injury	22	5.7	37	14.7	59	9.2
	Third Party Damage & Injury(Fatal)	49	12.6	35	13.9	84	13.1
	Total	388	100	252	100	640	100

Table 4.6 shows that 388 comprehensive policies claims were reported and 252 of third party claim were also reported. This shows that comprehensive policies are sold mostly by our insurers to insureds compared to that of the third party liability.

4.2 Estimation of Survival Time using the Kaplan-Meier(product limit) approach on Motor Insurance Policies

The Kaplan-Meier approach use observed event times and censoring times by computing their survival probabilities. The Kaplain-Meier estimate for insurance policies been sold out of which some insureds claimed whilst others did not is

shown in Table 4.7 below. The first column shows insureds that claimed and those that did not. The 25th, 50th and 75th percentile are presented in the second, third and fourth columns. The mean and 95% confidence interval are presented in the fifth and sixth columns respectively.

Table 4.7: Summary of Time from the start of an insurance policy to claim occurring for the entire data

	Percentiles						Mean			
Claim	25%		50%		75%				95% Confidence Interval	
	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error	Lower Bound	Upper Bound
No	365.000	.000	365.000	.000	365.000	.000	365.000	.000	365.000	365.000
Yes	270.000	5.677	207.000	7.378	43.000	6.530	179.819	4.531	170.937	188.700
Overall	365.000	.000	270.000	7.115	126.000	15.636	246.484	4.039	238.567	254.401

The quartile estimates shows that the time interval for purchasing an insurance **without a claim is 365 days**. Also the time interval for purchasing an insurance with a claim is 270 days for the 25th percentile whilst at the 75th percentile is 43 days. **The 50th percentile (average time) for a loss to occur to a customer is within 207 days** with a 95% confidence interval of 170.937 and 188.700 days. The mean for both groups was reported as 246.484 days.

Log Rank Test for Claim:

Figure 4.3 shows the log rank test for the different levels of claim reported. The columns shows the chi-square test of the reported claims recorded, the degree of freedom and their significance respectively using SPSS.

Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	950.190	1	.000
Breslow (Generalized Wilcoxon)	714.651	1	.000
Tarone-Ware	829.785	1	.000

Test of equality of survival distributions for the different levels of Claim.

Figure 4.3: Test of equality of survival distributions for the different levels of Claim.

Figure 4.3 compares the survival of insurance between insureds with and without claim. For this test the decision rule is to Reject H_o if χ^2>3.84. We observe X^2 = 950.190 on 1 degree of freedom for the log-rank test, which exceeds the critical value of 3.84. Therefore, we reject H_o.

4.3 Estimation of Survival Time using the Kaplan-Meier(product limit) approach on Motor Insurance Claim Policies

Table 4.8 shows the average time for claimants to be paid their losses. The columns shows the insureds that had a loss, the 25th, 50th and 75th percentile, the means and 95% confidence interval respectively.

Table 4.8: Summary of Time from the start of a motor claim report date to period of payment

	Percentiles						**Mean**			
Claim	25%		50%		75%				95% Confidence Interval	
	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error	Lower Bound	Upper Bound
Yes	346.000	11.282	270.000	9.287	126.000	21.006	255.725	6.437	243.108	268.342
Overall	346.000	11.282	270.000	9.287	126.000	21.006	255.725	6.437	243.108	268.342

The time interval for purchasing an insurance with a claim and paid is 346 days and 21 days for the 25th and 75th percentile respectively. And the **average time for a loss to be paid is 270 days** at the 50th percentile. The mean for settlement of a claim is 255.725 days with a 95% confidence interval of 243.108 and 268.342 days.

Therefore the average time to claim settlement for this data holding all variables constant is 270 days with a mean of 255.725 days.

Kaplan-Meier survival curve for a claim to be paid:

Figure 4.4 shows the cumulative incidence(probabilities) of claim paid to claimants enrolled in the study. The horizontal axis represents time in days, and the vertical axis shows the probability of survival.

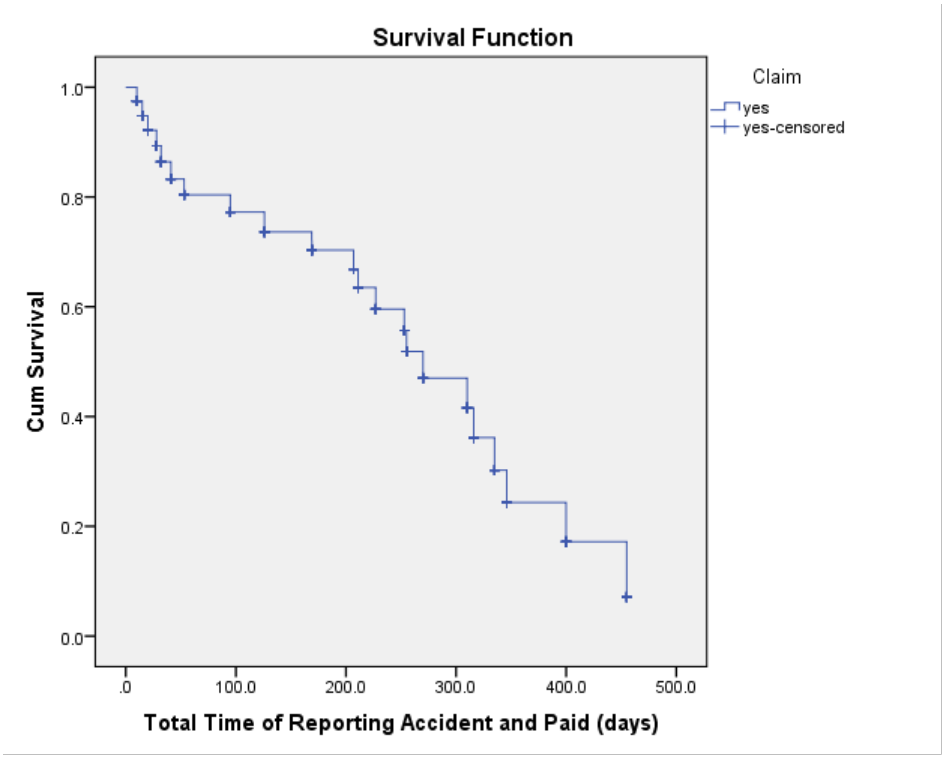

Figure 4.4: Plot of Survival Function for a claim to be paid

Figure 4.4 shows the Kaplan-Meier survival curve. In the survival curve the symbols represent each event time, either a claim paid or a censored time. The median survival is estimated by locating 0.5 on the Y axis and reading over and down to the X axis. The median survival is approximately 270 days.

4.4 Analysis on Whether Survival Time to Payment of Motor Insurance Claims Differs for Age and Marital Status

Figure 4.5 shows the log rank test for the different levels of Marital Status against Age. The columns shows the chi-square test, the degree of freedom and their significance respectively using SPSS.

Overall Comparisons[a]

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	1.830	1	.176
Breslow (Generalized Wilcoxon)	3.141	1	.076
Tarone-Ware	2.765	1	.096

Test of equality of survival distributions for the different levels of Marital Status.[a]

a. Adjusted for Age.

Figure 4.5: Test of equality of survival distributions for the different levels of Marital Status against Age.

Comparing marital status against age using the log rank test suggest that do not reject H_o because $1.830 < 3.84$. We do not have statistically significant evidence at $p < 0.05$ to show that the time to pay claimants is different between groups. This is shown in Figure 4.5

Kaplan-Meier survival curve for a claim to be paid for Marital Status against Age:

Figure 4.6 shows the cumulative survival of claim paid to claimants. The horizontal axis represents time in days, and the vertical axis shows the probability of survival.

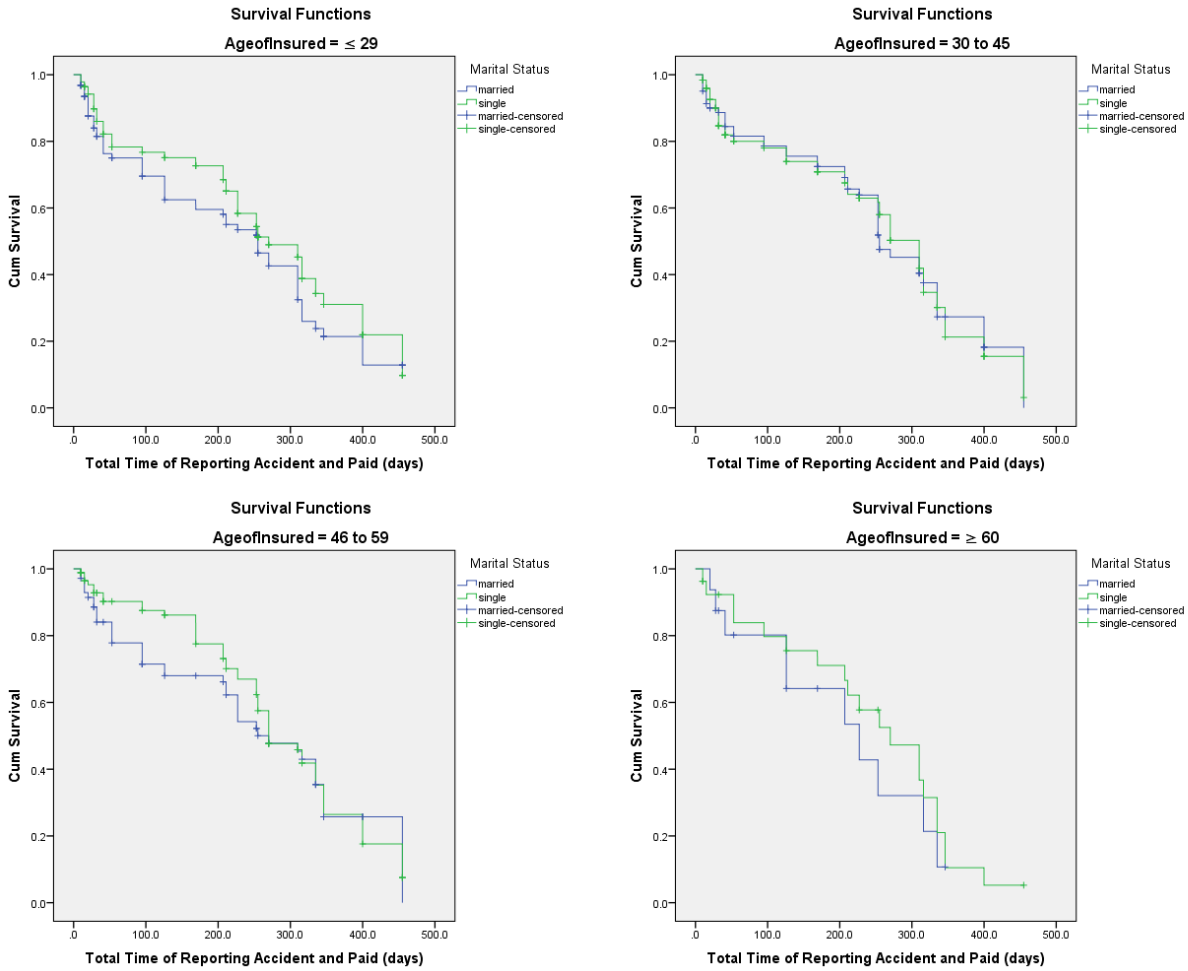

Figure 4.6: Plots of Survival Functions for the average time for a claim to be paid for Marital Status against Age.

The Figure 4.6 above shows the survival of claim payment among the various age groups either married or not. It is observed that survival curves do not show much separation, consistent with the non-significance findings in the test of hypothesis.

4.5 Analysis on Whether Survival Time to Payment of Motor Insurance Claims Differs for Gender and Type of Vehicle

Figure 4.7 shows the log rank test for the different levels of Gender against Type of Vehicle. The columns shows the chi-square test, the degree of freedom and

their significance respectively using SPSS.

Overall Comparisons[a]

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	.270	1	.604
Breslow (Generalized Wilcoxon)	2.597	1	.107
Tarone-Ware	1.656	1	.198

Test of equality of survival distributions for the different levels of Gender.[a]

a. Adjusted for TypeofVehicle.

Figure 4.7: Test of equality of survival distributions for the different levels of Gender against Type of Vehicle.

In here p=0.604 > 0.05 we say that there is really no difference among survival curves between all the groups hence not statistically significant. Thus Fail to reject H_o

Kaplan-Meier survival curve for a claim to be paid for Gender against Type of Vehicle:

Figure 4.8 shows the cumulative survival of claim paid to claimants. The horizontal axis represents time in days, and the vertical axis shows the probability of survival.

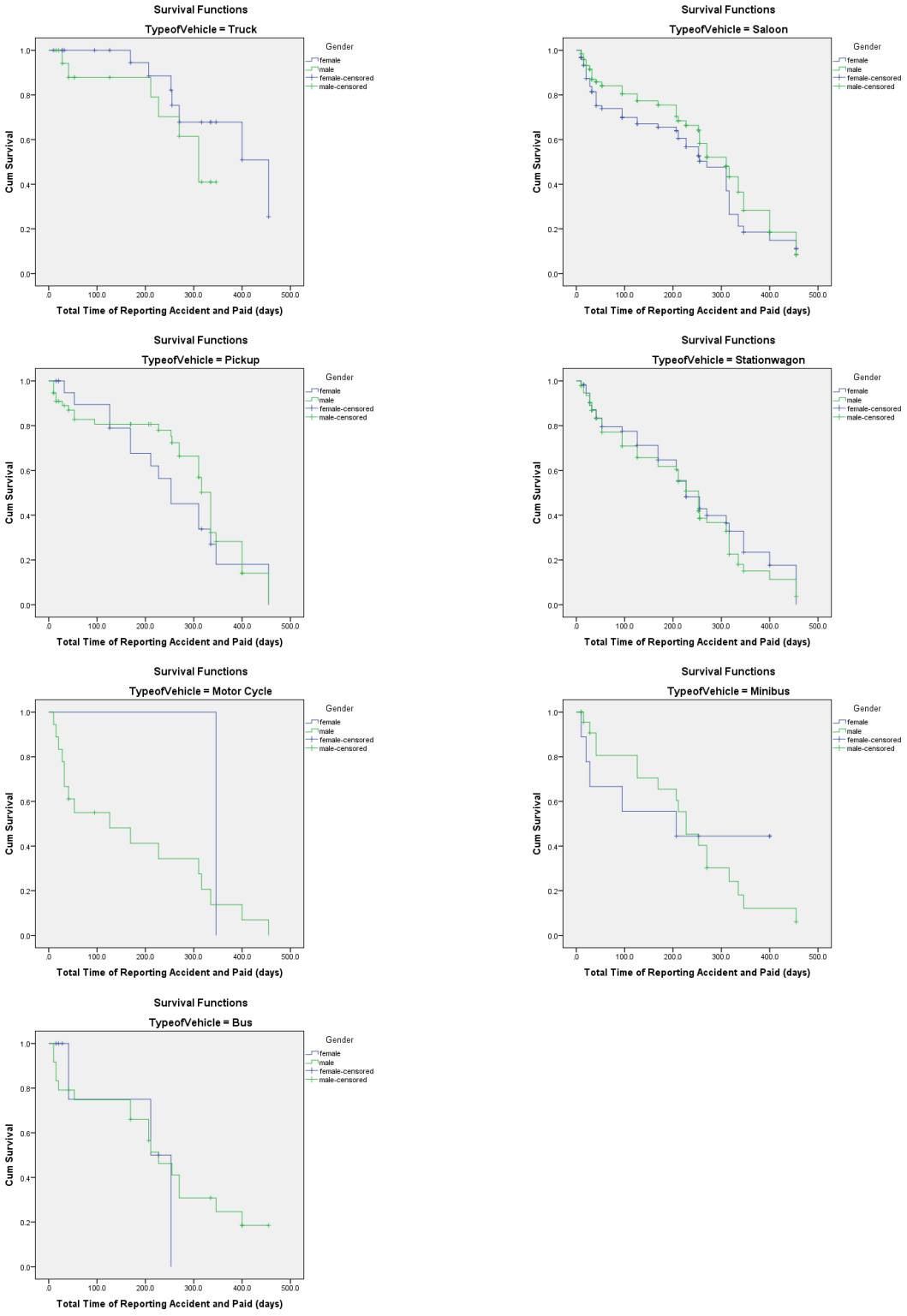

Figure 4.8: Plots of Survival Functions for the average time for a claim to be paid for Gender against Type of Vehicle

Figure 4.8 shows the survival of claim payment for gender against type of vehicle.

4.6 Analysis on Whether Survival Time to Payment of Motor Insurance Claims Differs for Type of Policy and Age

Figure 4.9 shows the log rank test for the different levels of Type of Policy against Age. The columns shows the chi-square test, the degree of freedom and their significance respectively using SPSS.

Overall Comparisons[a]

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	21.623	1	.000
Breslow (Generalized Wilcoxon)	21.170	1	.000
Tarone-Ware	22.168	1	.000

Test of equality of survival distributions for the different levels of TypeofPolicy.[a]

a. Adjusted for Age.

Figure 4.9: Test of equality of survival distributions for the different levels of Type of Policy against Age

The test statistic is approximately distributed as chi-square with 1 degree of freedom. Thus, the critical value for the test can be found in the table of critical values of the X^2 Distribution. For this test the decision rule is to Reject H_o as seen in Figure 4.9. This is because $X^2 = 21.623 > 3.84$ showing there is difference among survival curves between the groups hence statistically significant, $p < 0.05$.

Kaplan-Meier survival curve for a claim to be paid for Type of Policy against Age:

Figure 4.10 shows the cumulative survival of claim paid to claimants. The horizontal axis represents time in days, and the vertical axis shows the probability of survival.

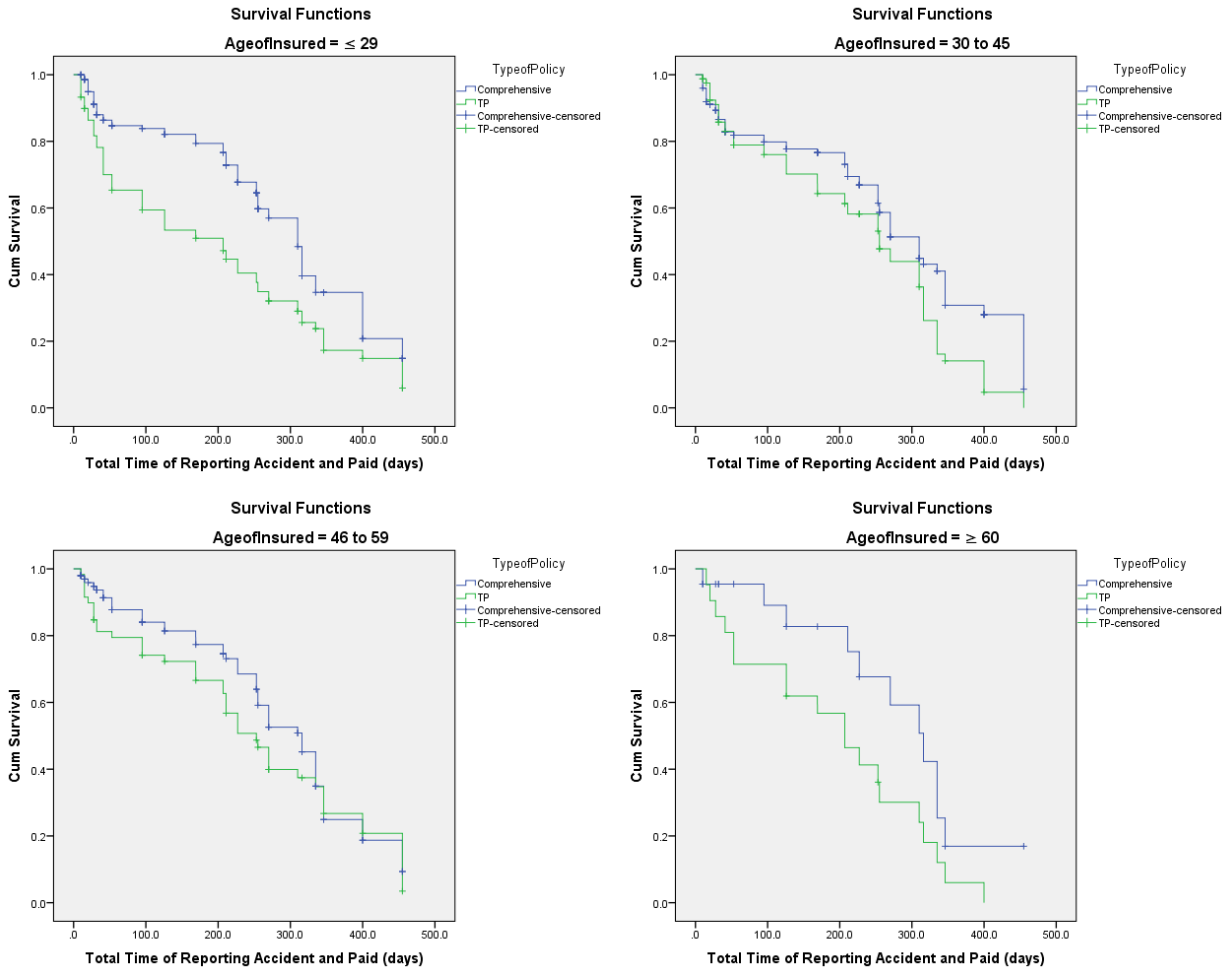

Figure 4.10: Plots of Survival Functions for the average time for a claim to be paid for Type of Policy against Age.

Figure 4.10 shows the survival of claim payment for type of policy against age.

4.7 Analysis on Whether Survival Time to Payment of Motor Insurance Claims Differs for Type of Policy and Gender

Figure 4.11 shows the log rank test for the different levels of Type of Policy against Gender. The columns shows the chi-square test, the degree of freedom and their significance respectively using SPSS.

Overall Comparisons[a]

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	21.933	1	.000
Breslow (Generalized Wilcoxon)	19.052	1	.000
Tarone-Ware	21.605	1	.000

Test of equality of survival distributions for the different levels of TypeofPolicy.[a]

a. Adjusted for Gender.

Figure 4.11: Test of equality of survival distributions for the different levels of Type of Policy against Gender

Comparing the Type of Policy issued to an insured against covariate Gender notice that the survival curves show separation indicating they are statistically significant. This is supported when performing a log rank test, which gives a test statistic $X^2 = 21.933$ on 1 degree of freedom in Figure 4.11

Kaplan-Meier survival curve for a claim to be paid for Type of Policy against Gender:

Figure 4.12 shows the cumulative survival of claim paid to claimants. The horizontal axis represents time in days, and the vertical axis shows the probability of survival.

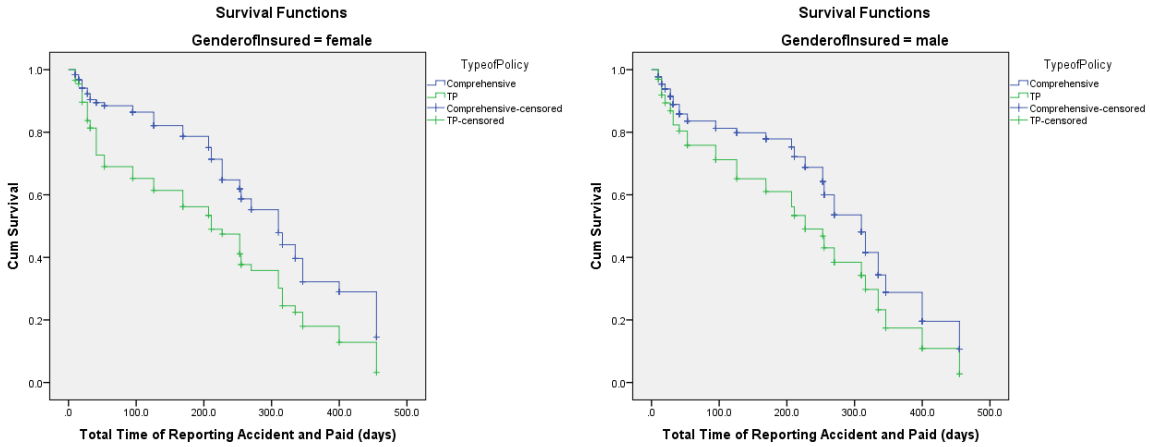

Figure 4.12: Plots of Survival Functions for the average time for a claim to be paid for Type of Policy against Gender.

Figure 4.12 shows the survival of claim payment for type of policy against gender.

4.8 Analysis on Whether Survival Time to Payment of Motor Insurance Claims Differs for Type of Policy and Marital Status

Figure 4.13 shows the log rank test for the different levels of Type of Policy against Marital Status. The columns shows the chi-square test, the degree of freedom and their significance respectively using SPSS.

Overall Comparisons[a]

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	21.547	1	.000
Breslow (Generalized Wilcoxon)	17.720	1	.000
Tarone-Ware	20.331	1	.000

Test of equality of survival distributions for the different levels of TypeofPolicy.[a]

a. Adjusted for Marital Status.

Figure 4.13: Test of equality of survival distributions for the different levels of Type of Policy against Marital Status

The log rank test of covariate Type of Policy against Marital Status, gives a border line significant effect of the number of claims reported and paid prior to the study on survival time, Figure 4.13

Kaplan-Meier survival curve for a claim to be paid for Type of Policy against Marital Status:

Figure 4.14 shows the cumulative survival of claim paid to claimants. The horizontal axis represents time in days, and the vertical axis shows the probability of survival.

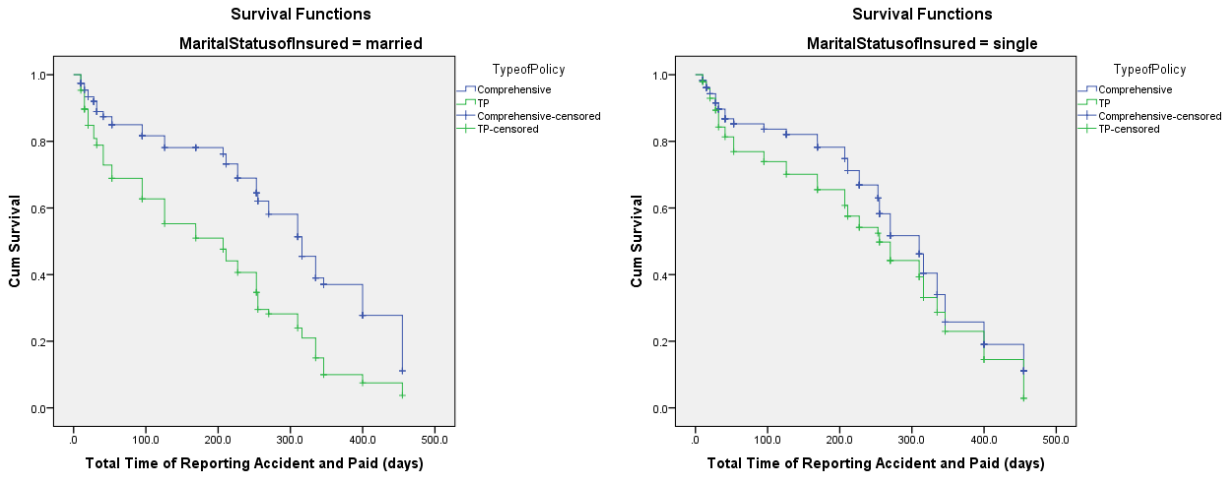

Figure 4.14: Plots of Survival Functions for the average time for a claim to be paid for Type of Policy against Marital Status.

Figure 4.14 shows the survival of claim payment for type of policy against marital status.

4.9 Analysis on Whether Survival Time to Payment of Motor Insurance Claims Differs for Type of Policy and Type of Vehicle

Figure 4.15 shows the log rank test for the different levels of Type of Policy against Type of Vehicle. The columns shows the chi-square test, the degree of freedom and their significance respectively using SPSS.

Overall Comparisons[a]

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	22.189	1	.000
Breslow (Generalized Wilcoxon)	15.330	1	.000
Tarone-Ware	18.966	1	.000

Test of equality of survival distributions for the different levels of TypeofPolicy.[a]

a. Adjusted for TypeofVehicle.

Figure 4.15: Test of equality of survival distributions for the different levels of Type of Policy against Type of Vehicle

Figure 4.15 shows that test statistic $X^2 = 22.189 > 3.84$. Hence conclude that, there is a difference in survival times between them under the null of no difference in survival, a highly significant result.

Kaplan-Meier survival curve for a claim to be paid for Type of Policy against Type of Vehicle:

Figure 4.16 shows the cumulative survival of claim paid to claimants. The horizontal axis represents time in days, and the vertical axis shows the probability of survival.

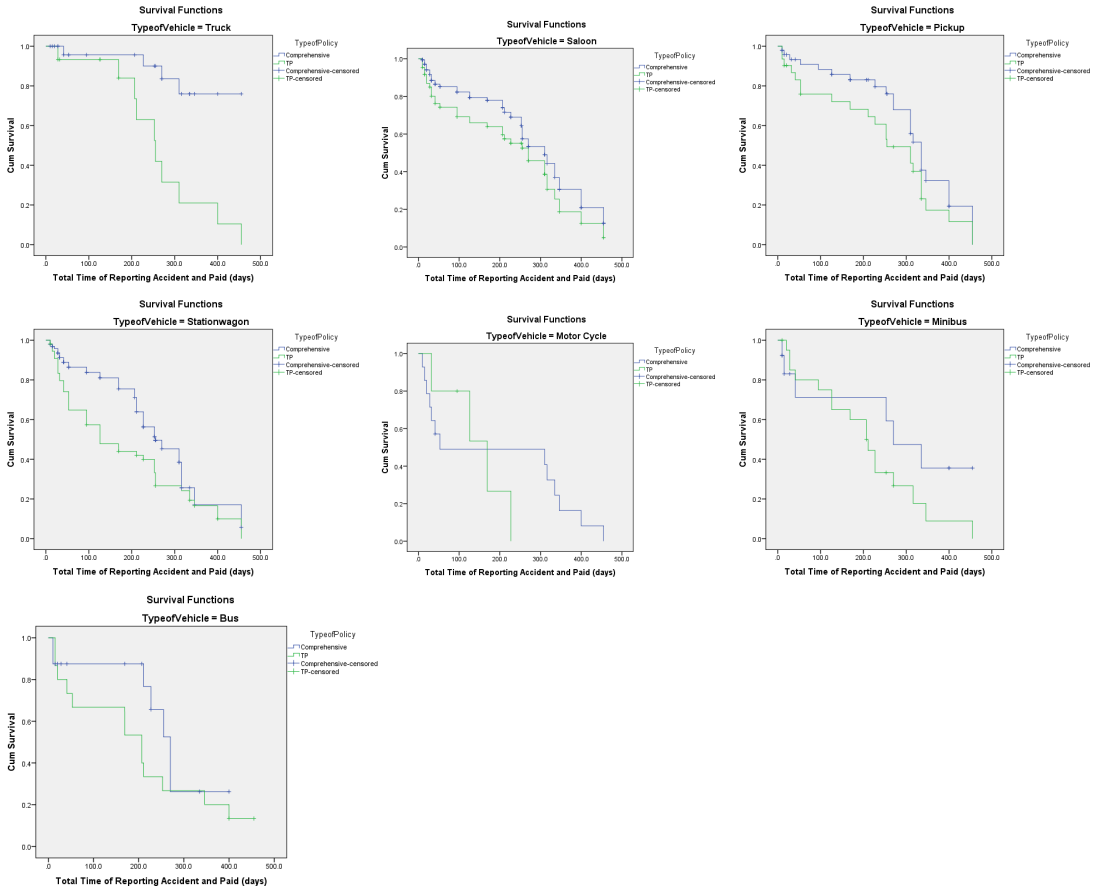

Figure 4.16: Plots of Survival Functions for the average time for a claim to be paid for Type of Policy against Type of Vehicle.

Figure 4.16 shows the survival of claim payment for type of policy against type of vehicle.

4.10 Analysis on Whether Survival Time to Payment of Motor Insurance Claims is Affected by Type of Policy

Table 4.9 shows the average time for claimants to be paid their losses for the type of policy issued. The columns shows the insureds that had a loss, the 25th, 50th and 75th percentile, the means and 95 % confidence interval respectively.

Table 4.9: Summary of Time from the start of a motor claim report date to period of payment (total duration) for the Type of Policy issued.

	Percentiles						**Mean**			
Type of Policy	25%		50%		75%				95% Confidence Interval	
	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error	Lower Bound	Upper Bound
Comprehensive	400.000	16.355	310.000	11.170	211.000	19.875	282.103	8.306	265.824	298.383
TP	335.000	7.276	227.000	14.935	53.000	14.009	219.879	9.849	200.575	239.184
Overall	346.000	11.282	270.000	9.287	126.000	21.006	255.725	6.437	243.108	268.342

The time interval for purchasing a Comprehensive and Third Party Insurance with a claim to be compensated and for a legal liabilities to be paid is 400 days and 335 days for the 25th percentile whilst at the 75th percentile is 211 days and 53 days respectively. The average time for a claimant to be paid is 310 days and 227 days for both policies at the 50th percentile shown in Table 4.9. The mean for both groups was reported as 255.725 days.

Kaplan-Meier survival curve for a claim to be paid for Type of Policy: Figure 4.17 shows the cumulative survival of claim paid to claimants for type of policy issued. The horizontal axis represents time in days, and the vertical axis shows the probability of survival.

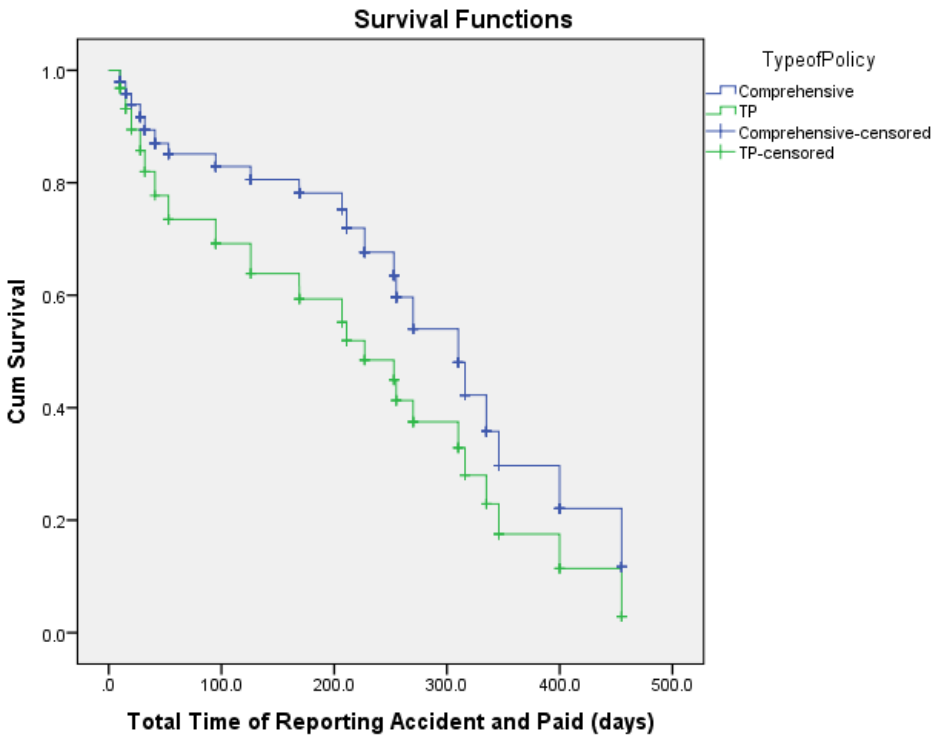

Figure 4.17: Plot of Survival Function for the average time for a claim to be paid for the Type of Policy issued.

Figure 4.17 shows that the survival probabilities for Comprehensive policies are higher than the survival probabilities for Third Party policies, suggesting a survival benefit. The median survival is approximately 270 days.

Log Rank Test for Type of Policy:

Figure 4.18 shows the log rank test for the different levels of Type of Policy. The columns shows the chi-square test, the degree of freedom and their significance respectively using SPSS.

Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	22.212	1	.000
Breslow (Generalized Wilcoxon)	23.174	1	.000
Tarone-Ware	23.746	1	.000

Test of equality of survival distributions for the different levels of TypeofPolicy.

Figure 4.18: Test of equality of survival distributions for the different levels of Type of Policy.

Figure 4.18 compare survival between the groups using the log rank test. Reject H_o because df = k-1=2-1=1 and p<0.05 as in the critical value for the X^2 distribution. Thus $X^2 > 3.84$. Hence the type of policy issued to a clientele is a significant variable in an insurance industry.

4.11 Modeling the Average Time of How Motor Insurance Claims are Handled and the Variables that are Affected.

4.11.1 The Cox Regression Model for Motor Insurance Policy Holders Who Claimed and are Paid.

Using a hazard model (Cox) with the various risk factors mentioned earlier and allowing for time-varying effects, produced the analysis below in Table 4.10 using R. The first column shows the various risk factors with their regression coefficient (β) in the second column. The third column shows the exponential coefficient and their standard error in the fourth column. The fifth column shows the z-values and the sixth column shows the p-values.

Table 4.10: Analysis of Maximum Likelihood Estimate for Cox Regression

Risk Factor	coef (β)	exp(coef)	se(coef)	z	Pr($>|z|$)
Age	-0.04242	0.95847	0.05762	-0.736	0.4616
Gender	0.01549	1.01561	0.11317	0.137	0.8911
Marital Status	-0.11914	0.88769	0.10680	-1.116	0.2646
Type of Policy	0.52178	1.68503	0.11364	4.592	4.4e-06 ***
Type of Vehicle	0.08540	1.08915	0.03330	2.564	0.0103 *
Nature of Claim	-0.05470	0.94677	0.02992	-1.828	0.0675 .

Table 4.10 shows that at df=6, the covariate (Type of Policy, and Type of Vehicle), p<0.05 are highly significant on the effect of survival time of having a motor insurance. Also, the parameter estimates represent the increase in the expected log of the relative hazard for each one unit increase in the predictor, holding other predictors constant. There is 0.52178 unit increase in expected log of the relative hazard for type of policy, holding type of vehicle constant. Lastly a 0.08540 unit increase in expected log of the relative hazard for type of vehicle, holding type of policies constant.

Hence from Table 4.10 the Cox regression model for the study is;

$$log\frac{\lambda(t)}{\lambda_o(t)} = 0.52178z_1 + 0.08540z_2$$

where:

$\lambda(t)$ is the expected hazard

$\lambda_o(t)$ is the baseline hazard function

z_1 is Type of Policy

z_2 is Type of Vehicle

Therefore;

$$\lambda(t) = \lambda_o(t)e^{(0.52178z_1+0.08540z_2)}$$

If;

$$S(t) = e^{-\wedge(t)} = e^{-\int_0^t \lambda(u)du}$$

Then;

$$S(t) = e^{-\int_0^t \lambda_o(u) e^{(0.52178 z_1 + 0.08540 z_2)} du}$$

It is observed that there is a positive association between type of policy, and type of vehicle. This shows there is increased risk of claims for insureds when under insurance coverage. Type of Policies are to pay for legal liabilities (i.e., restitution) as well as material damages, Type of Vehicle is dependent on the make, usage and cubic capacity of the vehicle.

Hazard Ratio Computation (The R Procedure):

Table 4.11 computes the hazard ratios of the variables used in the study. The first column shows the various risk factors. The second shows the regression coefficient (β) with their p-values in the third column respectively. The fourth column shows their hazard ratios with their 95% confidence interval.

Table 4.11: Analysis of Maximum Likelihood Estimate for Cox Regression

Risk Factor	coef (β)	Pr($>|z|$)	Hazard Ratio (HR)(95% CI for HR)
Age	-0.04242	0.4616	0.95847(0.8561 - 1.073)
Gender	0.01549	0.8911	1.01561(0.8136 - 1.268)
Marital Status	-0.11914	0.2646	0.88769(0.7200 - 1.094)
Type of Policy	0.52178	4.4e-06 ***	1.68503(1.3486 - 2.105)
Type of Vehicle	0.08540	0.0103 *	1.08915(1.0203 - 1.163)
Nature of Claim	-0.05470	0.0675	0.94677 (0.8929 - 1.004)

.

Computing hazard ratios by exponentiating the parameter estimates in Table 4.11 for type of policy, there is a 1.68503 times expected hazard, holding vehicle type constant(or there is 68.5% increase in the expected hazard relative to a one year increase in type of policy). Similarly, the expected hazard is 1.08915 times higher for type of vehicle, holding type of policy constant.

All of the parameter estimates are estimated taking the other predictors into account. After accounting for type of policies, and type of vehicle, there are

no statistically significant associations between age, gender, marital status and nature of claim. This is not to say that these risk factors are not associated with claims; their lack of significance is likely due to confounding (interrelationships among the risk factors considered). Notice that for the statistically significant risk factors (i.e., type of policies, and type of vehicle), that the 95% confidence intervals for the hazard ratios do not include 1 (the null value). In contrast, the 95% confidence intervals for the non-significant risk factors (age, gender, marital status and nature of claim) include the null value.

CHAPTER 5

CONCLUSION AND RECOMMENDATIONS

5.1 Introduction

In this chapter, the conclusions were made based on the study findings and the recommendations were also made based on the conclusions drawn.

5.2 Conclusion

In this thesis, the study determined the average time it takes to handle motor claims in the automobile insurance industry in Ghana. Analyzing a given data was able to calculate the changes in the survival function, especially in a competing risk setting.

The Preliminary Analysis indicates that:

• Regarding the age of the insured; young and middle aged drivers were at higher risk. This is what we can expect, since older people drive carefully.

• Regarding the gender of the insured; males are most often involved in accidents compared to females.

• Regarding the marital status of the insured; singles are mostly involved in accidents than the married.

• Comprehensive policies were mostly bought compared to Third Party policies. This was because most individuals want to protect their assets and value it. Also every company would want to grow in revenue and the sales of one comprehensive policy can cover about several fold on Third party policies.

Overall, the main results from the Survival Analysis indicates that;

1. The quartile estimate for purchasing a motor insurance without a claim is 365 days at the 50th percentile.

2. The average time for a loss to occur is within 207 working days.

3. The average time to payment of motor insurance claims is within 270 working days.

4. The average time to payment of motor comprehensive policy claims is within 310 working days.

5. The average time to payment of motor third party policy claims is within 227 working days.

6. The covariates (age, gender, marital status, and nature of claim) were not significant risk factors that affect the processing (payment) of a claim.

7. Type of policy, and type of vehicle were highly significant and influence the payment of claims.

5.3 Recommendations

From the conclusions drawn it is therefore recommended that the regulator, NIC and other stakeholders should ensure the following;

More research on the average time to handling a claim in the motor insurance industry is done particularly, in other insurance companies in Ghana in order to monitor them so that appropriate control measures and strategies could be approved and implemented to control its failure.

Insurers identify the bottleneck causing high number of days to claims payment over and above the benchmark set by NIC.

The average time to payment of motor insurance claims is not efficient. Hence insurers manage the credit risk liquidity by putting in more resources as to how to pay claims promptly and avoid piling up the books.

Actuaries and software programmers come together to design a premium software in which claims processed by all insurance companies can be assessed by the regulator, NIC to prevent poor delivery of claims payment.

5.4 Recommendation for further research

• The actuaries should research into the premium pricing and reserves in the Ghana market as to whether they are viable and adequate.

• Identify those customers that are loyal and less risky and determine how they can be compensated in other to encourage them to stay.

REFERENCES

Abeysundara, H. T. (2010). Non-parametric estimation of bivariate survival function. Statistics, Graduate Faculty of Texas Tech University.

Alberts, L. J. S. M. (2006). Churn prediction in the mobile telecommunications industry; an application of survival analysis in data mining. Engineering & computer science, Maastricht University, Department of General Sciences.

Arjas, E. and Gasbarra, D. (1994). Nonparametric bayesian inference for right-censored survival data, using the gibbs sampler. *Statistica Sinica*, 4(2):505–524.

Arnold, J. (2013). *The Performance Persistence, Flow and Survival of Systematic and Discretionary Commodity Trading Advisors (CTAs)*. Finance, Imperial College Business School, London.

Aslanidou, H., Dey, D. K., and Sinha, D. (1995). Bayesian analysis of multivariate survival data using monte carlo methods. Technical report, University of Connecticut and University of New Hampshire.

Baker, S. G. (1998). Analysis of survival data from a randomised trial with all-or-none compliance: Estimating the cost-effectiveness of a cancer screening program. *Journal of the American Statistical Association*, 93(443):929–934.

Brockett, P. L., Golden, L. L., Guillen, M., Nielsen, J. P., Parner, J., and Perez-Marin, A. M. (2008). Survival analysis of a household portfolio of insurance policies: How much time do you have to stop total customer defection? *Journal of Risk and Insurance*, 75(3):713–737.

Cam, E., Link, W. A., Cooch, E. G., Monnat, J. Y., and Danchin, E. (2002). Individual covariation in life-history traits: Seeing the trees despite the forest. *American Naturalist, Chicago Journals*, 159(1):96–105.

Chuang, H.-L. and Yu's, M.-T. (2010). Pricing unemployment insurance: An unemployment duration adjusted approach. *Journal of International Actuarial Association*, 40(2):519–545.

Clayton, D. G. and Cuzick, J. (1985). Multivariate generalizations of the proportional hazards model. *Journal of Roy. Statist. Soc., A*, 148:82–117.

Cox, D. and Oakes, D. (1984). *Analysis of Survival Data: Monographs on Statistics and Applied Probability. London,New York.* Chapman and Hall.

Cox, D. R. (1972). Regression models and life tables. *journal of the royal statistical society, series b*, 34(2):187–220.

Cox, D. R. (1975). Partial likehood. *Biometrika*, 62:269–276.

Crocker, K. and Snow, A. (1986). The efficiency effects of categorical discrimination in the insurance industry. *Journal of Political Economy*, 94(2):321–344.

Czado, C. and Rudolph, F. (2002). Application of survival analysis methods to long term care insurance. *Center of Mathematical Sciences, Munich University of Technology, Germany*, 31:395–413.

Dalby, K. (2011). Solvency ii : Qis5 for norwegian life and pension insurance. Mathematics and natural sciences, Faculty of Mathematics and Natural Sciences, University of Oslo.

Frangakis, C. E. and Rubin, D. B. (1999). Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika*, 86(2):365–379.

Gepp, A. C. B. (2005). An evaluation of decision tree and survival analysis techniques for business failure. Economics and financial services, fraud detection and insurance, Bond University Papers.

Gong, Z. (2008). *Parametric Potential-Outcome Survival Models for Causal Inference*. PhD thesis, University of Canterbury.

Goodfellow, J. and OConnor, J. (1978). The mechanics of the knee and prosthesis design. *The Jounal of Bone and Joint Surgery [Br]*, 60-B:358–369.

Grohn, Y. T., Eicker, S. W., Ducrocq, V., and Hertl, J. A. (1998). Effect of diseases on the culling of holstein dairy cows in new york state. *Journal of Dairy Science*, 81(4):966–978.

Gustafsson, E. (2009). Customer duration in non-life insurance industry. Mathematical statistics, dept. of mathematics, Stockholm University, Sweden.

Haberman, S. and Pitacco, E. (1999). Actuarial models for disability insurance. *Insurance: Mathematics and Economics, CRC Press*, 27(3):397–398.

Harrison, T. and Ansell, J. (2002). Cross-selling opportunities that would retain a customer in the insurance industry. *Journal of Financial Services Marketing*, 6(3):229–239.

Harrison, T., Ansell, J., and Archibald, T. (2007). Identifying cross-selling opportunities using lifestyle segmentation and survival analysis. *Marketing Intelligence and amp Planning*, 24(4):394–410.

Hosmer, D. W. and Lemeshow, S. (1999). *Applied Survival Analysis: Regression Modelig of Time to Event Data*. John Wiley and Sons, NewYork.

Kalbfleisch, J. D. (1978). Nonparametric bayesian analysis of survival time data. *Journal of the Royal Statistical Society, Series b*, 40(2):214–221.

Kalbfleisch, J. D. and Prentice, R. (2002). *The Statistical Analysis of Failure Time Data*. Wiley, NewYork.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457–481.

Kass and Wasserman (1996). Formal rules for selecting prior distributions: A review and annotated.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors.

Kiebach, A., editor (2014). *Five Factors That Affect Car Insurance Rate.* Lancaster Red Rose Credit Union.

Klein, J. P. (1997). *Survival Analysis:Statistics for Biology and Health, New York.* Springer-Verlag, 1st edition.

Kooijman, S. A. L. M. (1993). Dynamic energy budgets in biological systems, theory and applications in ecotoxicology. Technical report, Cambridge University Press,pp350.

Kooijman, S. A. L. M. and Bedaux, J. J. M. (1996). Analysis of toxicity tests on daphnia survival and reproduction. *Elsevier Science Ltd, Great Britian*, 30(7):1711–1723.

Kouassi, D. A. and Singh, J. (1997). A semiparametric approach to hazard estimation with randomly censored observations. *Journal of American Statistical Association*, 92(440):1351–1355.

Lewold, S., Goodman, S., Knutson, K., Robertsson, O., and Lidgren, L. (1995). Oxford meniscal bearing knee versus the marmor knee in unicompartmental arthroplasty for arthrosis: a swedish multi-centre survival study. *Journal Arthroplasty*, 10:722–731.

Louzada, F., Suzuki, A. K., Cancho, V. G., Prince, F. L., and Pereira, G. A. (2010). The long-term bivariate survival fgm copula model: An application to a brazilian hiv data. *Journal of Data Science*, 4:511–535.

McClenahan, C. L. (2001). *Ratemaking.* Foundations of Casualty Actuarial Science, fourth edition.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092.

Meyer, M. and Laud, P. (2002). Predictive variable selection in generalized linear models. *Journal of the American Statistical Association*, 97:859–871.

Moncrief, W. C., Hoverstad, R., and Jr., G. H. L. (1989). Survival analysis: A new approach to analyzing sales force retention. *Journal of Personal Selling and Sales Management*, 9(2):19–30.

Muller, P. (1991). A generic approach to posterior integration and gibbs sampling.

Murray, D. W., Goodfellow, J. W., and Connor, J. J. O. (1998). The oxford unicompartmental arthroplasty: a ten-year survival study. *The Jounal of Bone and Joint Surgery [Br]*, 80-B:983–989.

Nasvadi, G. C. and Wister, A. (2009). Do restricted driver's licenses lower crash risk among older drivers? a survival analysis of insurance data from british columbia. *Gerontologistgerontologist.oxfordjournals*, 49:474–484.

NIC, editor (2009). *Motor Vehicles: Third Party Insurance Act 1958*. Powered by Interface Technologies.

NIC (2011). Motor insurance claims. In *Motor Insurance Compensation Guidelines*, pages 1–47. Insurance.

Ofori-Attah, H. E. B. (2012). The effect of slow claims settlement on the sales and marketing of insurance products. Master's thesis, The Institute of Distance Learning, Kwame Nkrumah University of Science and Technology.

Pocock, S. J., Gore, S. M., and Kerr, G. R. (1982). Long term survival analysis: The curability of breast cancer. *Statistics in Medicine*, 1(2):93–104.

Raftery, A., Madigan, D., and Volinsky, C. T. (1995). *Accounting for Model Uncertainty in Survival Analysis Improves Predictive Performance.* In Bayesian Statistics 5, University press.

Shapiro, S., Venet, W., Strax, P., and Venet, L. (1988). *Periodic Screening for Breast Cancer: The Health Insurance Plan Project and Its Sequelae, 1963 - 986.* The Johns Hopkins Series in Contemporary Medicine and Public Health. Baltimore: The Johns Hopkins University Press.

Singer, J. D. and Willett, J. B. (1991). Modeling the days of our lives: Using survival analysis when designing and analyzing longitudinal studies of duration and the timing of events. *Psychological Articles, American*, 110(2):268–290.

Sinha, D. and Dey, D. K. (1996). Semiparametric bayesian analysis of survival data. *Journal of American Statistical Association*, 92:1195–1212.

Stevenson, M. (2007). *An Introduction to Survival Analysis.* PhD thesis, EpiCentre, IVABS, Massey University.

Svard, U. C. G. and Price, A. J. (2001). Oxford medial unicompartmental knee arthroplasty: A survival analysis of an independent series. *The Jounal of Bone and Joint Surgery [Br]*, 83-B(2):191–194.

Tsiatis, A. and Zhang, D. (2005). *Analysis of Survival D ata.* Statistics, North Carolina State University, Department of Statistics.

Tukan, A. (2012). A quantitative analysis of quality of emergency room care. Economics and business, The Colorado College, The Faculty of the Department of Economics and Business.

Vaupel, K. G. M. and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16:439–454.

Volinsky, C. T. and Raftery, A. E. (2000). Bayesian information criterion for censored survival models. *Biometrics*, 56(1):256–262.

White, S. H. and Ludkowski, P. F. (1991). Anteromedial osteoarthritis of the knee. *The Jounal of Bone and Joint Surgery [Br]*, 72-B:582–586.

Wintrebert, C. M. A. (2007). *Introduction: Survival Analysis and Frailty Models*. PhD thesis, Leiden University.

Wintrebert, C. M. A., Zwinderman, A. H., E. Cam, R. P., and Houwelingen, J. C. V. (2005). Joint modelling of breeding and survival in the kittiwake using frailty models. *Ecological Modelling*, 181(2-3):203–213.

Zhang, N. J. (2010). *Regression survival analysis with dependent censoring and a change point for the hazard rate: With application to the impact of the Gramm-Leach-Bliley Act to insurance companies' survival*. Dissertations, Rice University.

Zhang, Y. (2008). *Parametric mixture models in survival analysis with applications*. Dissertation, Temple University.

APPENDIX A

```
> library(foreign, pos=17)
> Dataset <- read.spss("C:/Users/Patricia/Desktop/USE.sav",
+   use.value.labels=FALSE, max.value.labels=Inf, to.data.frame=TRUE)
> <-fixdata set
> Dataset <- within(Dataset, {
+   Claim <- as.factor(Claim)
+ })
> Dataset <- within(Dataset, {
+   TypeofVehicle <- as.factor(TypeofVehicle)
+ })

> CoxModel.1 <- coxph(Surv(TimeofPayment,Remark) ~ TypeofPolicy*TypeofVehicle
+   + strata(Claim), method="breslow", data=Dataset)

> summary(CoxModel.1)
Call:
coxph(formula = Surv(TimeofPayment, Remark) ~ TypeofPolicy *
    TypeofVehicle + strata(Claim), data = Dataset, method = "breslow")

  n= 640, number of events= 382
   (372 observations deleted due to missingness)

                              coef exp(coef) se(coef)      z Pr(>|z|)
TypeofPolicy                1.6882    5.4096   0.5920  2.852 0.004348 **
TypeofVehicle[T.2]          1.3913    4.0201   0.5111  2.722 0.006489 **
TypeofVehicle[T.3]          1.2547    3.5066   0.5436  2.308 0.021000 *
```

```
TypeofVehicle[T.4]                1.6313    5.1103   0.5196  3.140 0.001692 **
TypeofVehicle[T.5]                1.9430    6.9799   0.5720  3.397 0.000681 ***
TypeofVehicle[T.6]                1.2548    3.5070   0.6459  1.943 0.052074 .
TypeofVehicle[T.7]                1.5205    4.5746   0.6271  2.425 0.015318 *
TypeofPolicy:TypeofVehicle[T.2] -1.3374    0.2625   0.6124 -2.184 0.028961 *
TypeofPolicy:TypeofVehicle[T.3] -1.2181    0.2958   0.6616 -1.841 0.065587 .
TypeofPolicy:TypeofVehicle[T.4] -1.2629    0.2828   0.6261 -2.017 0.043676 *
TypeofPolicy:TypeofVehicle[T.5] -1.0696    0.3431   0.8248 -1.297 0.194676
TypeofPolicy:TypeofVehicle[T.6] -0.9010    0.4062   0.7594 -1.187 0.235416
TypeofPolicy:TypeofVehicle[T.7] -1.3403    0.2618   0.7555 -1.774 0.076042 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                                  exp(coef) exp(-coef) lower .95 upper .95
TypeofPolicy                         5.4096     0.1849   1.69539   17.2605
TypeofVehicle[T.2]                   4.0201     0.2487   1.47625   10.9475
TypeofVehicle[T.3]                   3.5066     0.2852   1.20826   10.1769
TypeofVehicle[T.4]                   5.1103     0.1957   1.84577   14.1486
TypeofVehicle[T.5]                   6.9799     0.1433   2.27494   21.4158
TypeofVehicle[T.6]                   3.5070     0.2851   0.98879   12.4384
TypeofVehicle[T.7]                   4.5746     0.2186   1.33839   15.6359
TypeofPolicy:TypeofVehicle[T.2]      0.2625     3.8092   0.07905    0.8718
TypeofPolicy:TypeofVehicle[T.3]      0.2958     3.3807   0.08089    1.0817
TypeofPolicy:TypeofVehicle[T.4]      0.2828     3.5358   0.08291    0.9648
TypeofPolicy:TypeofVehicle[T.5]      0.3431     2.9143   0.06814    1.7279
TypeofPolicy:TypeofVehicle[T.6]      0.4062     2.4621   0.09168    1.7992
TypeofPolicy:TypeofVehicle[T.7]      0.2618     3.8202   0.05954    1.1507

Concordance= 0.609  (se = 0.018 )
```

```
Rsquare= 0.067   (max possible= 0.999 )
Likelihood ratio test= 44.72  on 13 df,   p=2.332e-05
Wald test            = 37.31  on 13 df,   p=0.0003698
Score (logrank) test = 41.52  on 13 df,   p=7.832e-05

> Dataset <- within(Dataset, {
+   NatureofClaim <- as.factor(NatureofClaim)
+ })

> CoxModel.2 <- coxph(Surv(TimeofPayment,Remark) ~ TypeofPolicy
+   *NatureofClaim + strata(Claim), method="breslow", data=Dataset)

> summary(CoxModel.2)
Call:
coxph(formula = Surv(TimeofPayment, Remark) ~ TypeofPolicy *
    NatureofClaim + strata(Claim), data = Dataset, method = "breslow")

  n= 640, number of events= 382
   (372 observations deleted due to missingness)

                                   coef exp(coef) se(coef)      z Pr(>|z|)
TypeofPolicy                     0.3475    1.4156   0.3772  0.921 0.356928
NatureofClaim[T.2]              -1.2971    0.2733   0.3699 -3.507 0.000453 ***
NatureofClaim[T.3]              -0.5692    0.5660   0.2889 -1.970 0.048837 *
NatureofClaim[T.4]              -0.3738    0.6881   1.0063 -0.371 0.710331
NatureofClaim[T.5]              -0.2027    0.8165   0.1775 -1.142 0.253487
NatureofClaim[T.6]               0.2777    1.3201   0.2394  1.160 0.246094
NatureofClaim[T.7]              -0.8282    0.4368   0.3084 -2.685 0.007254 **
TypeofPolicy:NatureofClaim[T.2]      NA        NA   0.0000     NA       NA
```

```
TypeofPolicy:NatureofClaim[T.3]      NA        NA   0.0000     NA       NA
TypeofPolicy:NatureofClaim[T.4]  0.3199    1.3770   1.0723  0.298 0.765465
TypeofPolicy:NatureofClaim[T.5]      NA        NA   0.0000     NA       NA
TypeofPolicy:NatureofClaim[T.6] -0.2009    0.8180   0.4824 -0.416 0.677116
TypeofPolicy:NatureofClaim[T.7]      NA        NA   0.0000     NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                                exp(coef) exp(-coef) lower .95 upper .95
TypeofPolicy                       1.4156     0.7064   0.67581    2.9650
NatureofClaim[T.2]                 0.2733     3.6588   0.13238    0.5643
NatureofClaim[T.3]                 0.5660     1.7669   0.32125    0.9971
NatureofClaim[T.4]                 0.6881     1.4532   0.09574    4.9459
NatureofClaim[T.5]                 0.8165     1.2247   0.57661    1.1563
NatureofClaim[T.6]                 1.3201     0.7575   0.82568    2.1105
NatureofClaim[T.7]                 0.4368     2.2891   0.23866    0.7996
TypeofPolicy:NatureofClaim[T.2]        NA         NA        NA        NA
TypeofPolicy:NatureofClaim[T.3]        NA         NA        NA        NA
TypeofPolicy:NatureofClaim[T.4]    1.3770     0.7262   0.16834   11.2633
TypeofPolicy:NatureofClaim[T.5]        NA         NA        NA        NA
TypeofPolicy:NatureofClaim[T.6]    0.8180     1.2225   0.31777    2.1057
TypeofPolicy:NatureofClaim[T.7]        NA         NA        NA        NA

Concordance= 0.618  (se = 0.018 )
Rsquare= 0.091   (max possible= 0.999 )
Likelihood ratio test= 61.35  on 9 df,   p=7.371e-10
Wald test            = 50.6  on 9 df,   p=8.313e-08
Score (logrank) test = 55.59  on 9 df,   p=9.416e-09
```

APPENDIX B

Model Selection: R is a powerful tool that understands terms involving more than one degree of freedom, so it keeps together dummy variables representing the effects of a factor. Results were obtained by using the cox.zph function in R. A complete overview of used data is seen below.

Covariates	Class	Description
Claim	0	No
	1	Yes
Remark	0	Pending
	1	Paid
Type of Policy	0	Comprehensive
	1	Third Party
Type of Vehicle	1	Truck
	2	Saloon
	3	Pick Up
	4	Station Wagon
	5	Motorcycle
	6	Minibus
	7	Bus
Age	1	21 - 29 years
	2	30 - 45 years
	3	46 - 59 years
	4	$\geq$ 60 years
Gender	0	Female
	1	Male
Marital Status	0	Married
	1	Single
Nature of Claim	1	Own Damage
	2	Own Damage-Total Loss
	3	Theft
	4	Collision
	5	Breakage Of Windshield
	6	Third Party Damage and Injury
	7	Third Party Damage and Injury (Fatal)