SURVIVAL ANALYSIS OF ACCIDENT RISKS OF CAR DRIVERS IN NORTHERN REGION, GHANA

A THESIS SUBMITTED TO THE DEPARTMENT OF STATISTICS, UNIVERSITY FOR DEVELOPMENT STUDIES, GHANA IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTER OF SCIENCE

By:
Alhassan Faisal
UDS/MAS/0003/08

July 2010

Dedication

This work is dedicated to my parents, Afa Alhassan Kpalbe and Alhassan Fatima, for their support and encouragement throughout my undergraduate and graduate studies.

Abstract

The accident risks of drivers and the factors affecting the risks were analyzed using survival models. The data consisted detailed records of fatal accidents in the years 2007 - 2009, generated by the Motor Traffic and Transport Unit (MTTU) of the Ghana Police Service, Northern region. The data on fatal accidents contained 398 drivers. The objectives of the study included both analysis and explanation of driver accident risk factors and investigation of how survival modeling method can be used in traffic accident analysis. The research questions that were addressed include: can driver involvement in accidents and exposure be examined with survival models? Are age and sex major risk factors? Do vehicle characteristics contribute to accident risk? Does the use of unworn out tyres reduce accident risks? The analysis of the data using the SAS package confirmed that the most significant variables to the risks of accident were driver characteristics (age, driver behaviour, kilometerage driven, speed and use of safety belt), drivers' state (driving under influence of alcohol) and vehicle characteristics (vehicle age and weight, and condition of tyres). Young and inexperienced on one hand, or old and experience drivers on the other hand, had the highest fatal accident risks. Drivers using worn out tyres had a somewhat greater accident risk than drivers using unworn out tyres, but the difference was not statistically significant. Survival modeling was used to analyze the data, and it was concluded that survival modeling promises to be a useful tool for road safety analysis.

Summary

The study examined the effects of human factors, kilometerage driven, vehicle characteristics and roadway factors on amount of accidents and risks of accident associated with driving. The study was based on accident data from the MTTU regional office of Ghana Police Service. This data included information on 398 drivers who had been involved in fatal accidents during the period under consideration.

Drivers' accident risks and their dependence on various factors were examined using survival accident models. Survival modeling is commonly applied in medicine to the study of serious diseases and treatment methods. The present study tries to assess the development needs of survival models in the area of traffic accident analysis.

Several factors have bearing on driver's accident risks. This can be explained by driver's age and sex, annual kilometerage driven, age of vehicle, driving under influence of alcohol, the use of safety belt, speed at the site of the accident, the weight of the vehicle and the tyre treated depth.

More specifically, the analysis confirmed the following answers to the main questions of the study;

- The kilometerage driven was negatively related to hazard of accident. The more kilometers are driven, the lower the probability of being involved in an accident. This particular outcome needs further probing.
- Driver age proved to be a statistically significant accident risk factor, however there was no clear sex-related differences between accident risks of male and female drivers.
- Some vehicles characteristics can be separated into own individual risk factors but they are strongly interrelated to many other risk factors.
- Drivers of worn out tyres had a somewhat larger accident risk than the users of unworn out tyres, but the obtained differences were not statistically significant.

The effects of driver's motives and attitudes were not directly modeled in this study, but some of their influence can be assumed to underlay the effects of driver age, sex or experience, as well as determine their choosing speed, drive under influence of alcohol, or not use of safety belt.

The survival modeling analysis included model formulation, parameter estimation and validation. The main survival model type used was Cox Proportional model. It however revealed that some explanatory power is lost when survival models are based on random time period. The results proved that better exposure data e.g. time spent in traffic used in the analysis could improve the models.

Generally, the application of survival models to the accident data appears to be a promising approach but still needs further development.

List of Tables

List of Figures

Contents

Acknowledgments

I would first and foremost thank the Almighty God for seeing me through so much that I have finally arrived at this stage. May this be a testimony of your grace and greatness.

Next, I would like to thank my supervisor Mr. Mamadou Lamine Diedhiou for his guidance, insights, patience, gentle encouragement and advice which made this thesis possible. From him I have learned a great deal and most importantly he has taught me how to research. His enthusiasm and strives for excellence has made working with him a joy. I am deeply grateful to Prof. Kaku Sagary Nokoe, the former Ag. Vice -Chancellor of the University for Development Studies for not only approving my study leave but also granted me the scholarship to pursue the course, words cannot be used to express my heartfelt gratitude to him, what I can only say is that may the Almighty God richly bless him.

Next, I would like to express my special thanks to Professor Okafor, who taught me the Survival analysis course during the taught part of the MSc. programme. The way and manner he handled the course and the relevant text books he made available to us made me develop interest in this area, and consequently I have been able to apply this knowledge to real life situation. I will forever remain grateful to him.

I would also like to thank Dr. Oyetunji, the Dean of the Faculty of Mathematical Sciences for taking the class through on how to conduct a meaningful research work. The knowledge acquired has helped me tremendously in carrying out this project. The same token of appreciation goes to Dr. Seidu Alhassan, Head of Department of Applied and Business Mathematics of the University for imbibing in us on how to develop a very good proposal. I am deeply grateful to my able and very active Post graduate coordinator Mr. Kasim Gbolagade for working administratively hard to ensure that we complete the course on time. On behalf of my colleagues I will say we are very much grateful.

I owe a debt of gratitude to Ghana Education Trust Fund (GET Fund) for providing funding to undertake this study, the assistance is very much appreciated, without this support it would have been very difficult on my part to successfully carry out this research.

I would like to thank all my friends and colleagues who had helped in one way or the other in making this thesis possible. To my parents, Afa Alhassan Kpalbe and Mma Fatima, I am humbled and blessed to have parents like you. The support you gave me throughout the years I will forever remain grateful to you.

And lastly, I am grateful to my wife, Maliatu for giving me the moral support.

Thesis Layout

The research is organized into five chapters. Chapter one contains the introduction. Chapter two is devoted to literature review. Chapter three presents the research methodology. Chapter four contains data presentation and analysis. Chapter five is devoted to conclusions and suggestions for future work.

CHAPTER 1

INTRODUCTION

1.1 Background of the study

Drivers are faced with risky situations and potential accidents every time they are on the road. Counter measures are actions taken by society to prevent accidents or moderate their consequences. Such measures are based on ideas regarding why and how dangerous situations and accidents evolve.

The occurrence of traffic accidents can be explained only to a limited extent by a deterministic causal relationship, in which the occurrence of certain conditions will always lead to accident consequences. Due to this, the occurrence of accidents is often explained with a probabilistic causal relationship, in which the occurrence of the cause will increase the probability of the occurrence of the effect (Elvik and Vaa 1990).

This research project will focus on the person-vehicle-traffic environment system. According to (Hakkinen 1978), the successful functioning of the system requires that its parts- the person, the vehicle and the traffic environment, as well as the "functions" which connect them- should remain within certain limits of variation. According to the theory, accidents happen when road users cannot adapt their actions to the varying demands of the traffic environment. Consequently, the risk of accident can be lowered by improving road users' performance in traffic or by reducing system demands on road users (Elvik 1996). Put another way, humans inevitably make errors but by altering the circumstances in which they operate one can minimize the frequency of errors or moderate their consequences (Elvik and Vaa 1990).

Road traffic accidents, for the most part, have been examined from the standpoint of either traffic environment, weather conditions or the behavior of individual drivers. Such studies have not sufficiently taken into consideration the impact of possible differences between driver groups or interactions between factors. This present study will examine which characteristics are connected with drivers and vehicles, together with the prevalent road way conditions, influence the occurrence of road traffic accidents.

1.2 Traffic accident risk factors

Factors may be considered accident causes if they either increase or decrease the probability of accident occurrence. Therefore, in order to prevent accident, one must know which of the numerous traffic risk factors have a real strong influence on the number and probability of accidents. The risk factors that will be considered in this study will be human factors and mobility, vehicle factors and traffic environment factors. In the following, an overview of the risk characteristics of each sub-group of factors is provided. However, literature review will describe in more detail previous findings about the risks of several of the factors mentioned here.

1.3 Human factors

According to studies in the 1970's and 1980's, factors associated with the road users were the direct cause of about 95% of accidents investigated. Factors associated with the traffic environment were the direct cause of 28 - 34%, and factors associated with the vehicle directly caused only 8-12% of accidents (Elvik 1996). The factors are overlapping which explains that the total exceeds 100%. All of the studies also acknowledged the contribution of background or indirect factors to the causation of accidents.

Regardless of which evaluation model was used to assess the causes of accidents, human factors have been given a prominent position in it. Human factors relevant to driving behavior can be classified in different ways. One classification, suggested by (Elvik and Vaa 1990), divides the factors into four groups (Table 1.1).

Table 1.1: Human psychological and physiological factors that have been studied as possible accident risk factors (Elvik and Van 1990)

Permanent or slowly changing physiological factors	Permanent or slowly changing psychological factors	Temporary physiological factors	Temporary psychological factors
Age	Intelligence	Fatigue	
Sex	Personality factors	Stress	Emotional state
Vision	Attitudes	Alcohol	Breakdown in concentration
Hearing	View of own skills	Drugs	
Reaction time	Recognition of dangerous situations	Acute illness	
Physical disorders and deficiencies	Mental illness	The menstrual cycle	
A heart condition		Pregnancy	
Diabetes			
Epilepsy			

Elvik and Vaa (1990) further reports on the relative impact of the above human factors on accident that the impact of driver's age, vision, reaction time, knowledge of traffic regulations, actual skills, fatigue and use of alcohol on traffic accidents have been proven by research. However, the impacts of driver's sex, personality, and use of drugs are less conclusive.

1.3.1 Traffic and road way factors

Traffic environment can support and promote safe behavior, but it can also encourage or lead to risky behavior, the possible accident risk and exposure factors associated with traffic and its environment are:

- Amount of traffic
- Characteristics of the traffic

- Road networks and land use
- Roadway conditions and
- Management and control of traffic.

The amount of traffic relates to the amount of mobility and therefore, represents exposure to accidents. Accident risks are dependent on the amount of traffic. Various studies (e.g. Kulmala 1995) have ascertained that accident type distribution depends on the amount of traffic. The amount of traffic affects other traffic characteristics such as speed, headway distribution and the amount of overtaking, all of which can influence accident risks.

Roadway networks are made up of various types of junctions and sections, which have different risks associated with them. Traffic control and management are needed to make sure that traffic is as fluent and safe as possible on the network. Control devices and management procedures regulate drivers' mobility and behaviour in traffic. Therefore, they directly influence exposure as well as accident risks.

1.3.2 Vehicle factors

Case studies of traffic accidents have found that in a small proportion of vehicle factors (such as mechanical failures) were direct and primary causes of the accidents. Sudden failures associated with vehicles were just 1 % of the "key events" in accidents investigated by Finnish teams. This is compared to 70% to 80% of the events attributed to drivers' errors in vehicle control and operation, anticipation and judgment (Koomstra, 1992).

The role of vehicle factors appears to be larger in fatal accidents. In fatal collisions investigated by the teams, 22- 26% of the "key events" evaluated were associated with the vehicle, 15-28% with the traffic environment, 43-56% with human factors and about 1% with traffic regulations. The most cited vehicle factors were unsuitable or worn out tyres (20 - 42%), misuse or malfunction of safety belts (11 - 27%), poor crashworthiness (27-43%), and defective steering (Koomstra, 1992).

1.3.3 Environmental conditions factors

Up to 20% of all the risks events analysed by accident teams in Finland (Koomstra, 1992) were related to characteristics of the traffic environment - mainly road surface, weather and lighting conditions.

Accident risk is usually 1 - 2 times as great in the dark as it is in the light (Kulmala and Peltola

1985). Darkness is often associated with weather and road surface conditions and with the amount of traffic as these factors may vary with time of day. The accident rate on wet roads is usually higher than on dry roads.

1.4 Road Traffic Accident Situation in Ghana

Road safety has become a major national issue receiving front-page coverage in the press and National TV news on a regular basis. Road accidents are common in this country to the extent that in 1995, Ghana ranked 2nd to Mexico in terms of road fatalities worldwide. In 1997, it ranked 2nd to Nigeria in West Africa as far as road accidents are concerned.

Ghana has lost 602 lives through road accidents in the first quarter of 2009, almost double the figure recorded within the same period in 2008. The National Road Safety Commission (NRSC) said the first quarter of 2008 saw 339 deaths and over 60 percent of the fatalities were caused by speeding.

According to the National Road Safety Commission annual report for 2005, road crashes kill an average of four persons daily in Ghana. The regions Ashanti, Eastern, Gt. Accra, Central and Brong Ahafo Regions accounted for more than 70% of the total number of crash fatalities. 70% of crashes occur on flat and straight roads. Speeding is a major cause of crashes, accounting for over 50% of reported crashes. Road users between 16-45 years are the most vulnerable group and account for 58% of total road crash fatalities from 2002-2005. 70% of persons killed in road crashes are males. The age groups from 0-5, 46-65 and over 65 years also accounted for a 20.8%, 16.7% and 4.6% respectively of the total fatalities during the same period. 18% of the accidents occur between 6 and 8pm; Saturday is the most accident prone day; April is the most accident-prone month.

The statistics are alarming despite the interventions made by governments. What we experience is the equivalent of the population of an average-sized town in Ghana being killed every year in road accidents. Despite occasional outrage, little seems to be done to ensure safety on Ghanaian roads. The statistics should even be viewed with caution as the quality of the national data is affected by under-reporting.

Every day we lose our friends, colleagues and relatives by what we call road accidents even though

most of these so-called accidents are preventable and cannot be called accidents in the strictest sense of the word. “Accident” is defined as a happening that is not expected, foreseen or intended. Our roads have become slaughterhouses where we are butchered in great numbers. We all know that most of these accidents can be prevented with very little effort, discipline and respect for other road users.

Ghana is one of the countries where road signs and speed limits are not respected at all. An independent urban speed study was conducted by GRSP Ghana in November 2006. The results indicate that drivers are exceeding posted limits (50kph) by as much as 50kph and that vulnerable road users are at extreme risk of severe injuries from high urban speeds.

The issue of badly worn tyres and other dangerous tyre products needs to be looked at. We usually import tyres that have been discarded by users in other countries, in the name of home-used products. We need a strong system of regulation at our ports to ensure that importation of used tyres and other categories of used car parts attract taxes that would discourage cheap and dangerous imports by unscrupulous businessmen.

Sometimes the accidents are caused by a vehicle’s mechanical problems. However, a high proportion of accidents are directly blamable on human factors. The clearest example includes drunken drivers, intoxication with drugs and alcohol, speeding and fatigue. Drink driving cases usually increase during festive times, Easter and Christmas.

Many roads in Ghana have become death traps with potholes dotted along the length and breadth of the roads. Broken-down vehicles could be seen blocking roads. Road building equipment are left on roads without warning signs. Road signs, on the very few occasions they are present, are often unhelpful if not deceptive. Some of the new roads are very poorly built and soon lapse into disrepair, posing danger to road users.

There are too many unlicensed/unqualified drivers with little or no knowledge of road rules in Ghana. Some people even get their licenses without having even taken that flawed driving test. This means the wrong people are acquiring licenses for which they are not qualified to hold. There are a number of drivers on the roads with very poor eyesight or an insufficient field of vision e.g. poor vision for night driving, and there is no mechanism to check this.

ineffective vehicle examination and enforcement of road rules. The attitude of traffic policemen has not changed. Road offenders go free by a simple payment of bribes to the policeman. Many drivers have eye problems but are holding driving licenses. They can neither see what is happening around them nor drive in darkness. In the fear of losing their jobs they continue to drive even though they have very poor eye-sights. This problem has also contributed to many accidents on our roads.

The DVLA or GHA or Ministry of Transport should sponsor more research into the causes of road accidents. The aim of this study therefore is to further describe the determinants of accident risks of vehicle drivers in the country using survival analysis. The findings can serve as a basis for health care professionals and policy makers to create preventive measures for traffic accidents.

To solve this problem of extra-Poisson variation, several authors such as Miaou (1994) developed Bernardo and Ivan (1997) studied the prediction of the number of crashes versus the crash rate using Poisson regression. They suggested that the Poisson distribution allows for the relationship between exposure and crashes to be more accurately modeled as opposed to the linear relationship assumed in crash rate prediction.

Predictive models for accidents have been researched intensively in the world. For example, Oppe (1989) used multiple linear regression models. In these models the dependent variable (either number of accidents or accident rate) is a function of a series of independent variables such as speed or traffic volume. Accident occurrence in these models is assumed to be normally distributed. These models generally lack the distributional property that is necessary to describe adequately the random and discrete vehicle accident events on the road and they are inappropriate for making probabilistic statements about accident occurrence.

Saccomanno and Buyco (1988) and Blower *et al* (1993) used a Poisson loglinear model to explain variations in accident rates. This Poisson regression model is especially suitable for handling data with large numbers of zero counts. However, this model could be inappropriate for road accident counts, since it fails to account for extra-Poisson variation (the value of the variation could exceed the value of the mean) in the observed accidents counts.

two types of negative binomial models, one using a maximum likelihood method and one using a method of moments. The maximum likelihood model was found to be more reliable than the Poisson regression model in predicting accidents where over dispersion is present. In 1949, R. J. Smeed also developed a regression model (log-linear model) and he found an inverse (or negative) relationship between the traffic risk (fatality per motor vehicle) and the level of motorisation (number of vehicles per inhabitant). This regression represented the best estimates of the mean values of traffic risk for each given value of motorisation (what is called least square). This shows that with annually increasing traffic volume, fatalities per vehicle decrease. Smeed concluded that fatalities (F) in any country in a given year are related to the number of registered vehicles (V) and population (P) of that country by the following equation.

$$\frac{F}{V} = \alpha \left(\frac{V}{P}\right)^{-\beta} \qquad (1.1)$$

Where
F = number of fatalities in road accidents in the country
V = number of vehicles in the country
P = population
α = 0.003, β = 2/3

This formula became popular and has been used in many studies. It is often called as Smeed's formula. This nonlinear relationship can be translated to a linear one by taking the logarithms of the two sides: logY = logα – βlogX, where Y is F/V and X is V/P.

The number of fatalities can be derived Smeed's formula as: $F = c.V^{\alpha}.P^{\beta}$, where c, α, β are parameters and they are estimated from data by using the least square method.

Some authors followed the equation of estimating the regression parameter (α, β) of the data by calculating the country road safety performance in comparison to other countries; Jacobs and Hutchinson (1973). They found that Smeed's formula can give a close estimation of the actual data and it can be applied to different sample sizes of countries and years with the use of different values of *alpha* and *β*. Some authors have tried to develop Smeed's formula and its accuracy further by including several socio-economic variables in the model. Fieldwick (1987) has included speed

limits in the same model. The number of registered vehicles has been replaced by the total vehicle kilometre driven in many studies (e.g. Silvak, 1983). This measure (vehicle kilometre driven) was not available at the time of Smeed's study.

At the same time, many studies have criticised Smeed's model because it only concentrates on the motorisation level of country and ignores the impact of other variables, (e.g. Broughton, 1988), where according to Smeed's model, population and vehicles are the only country values, road fatalities can simply be predicted from population and vehicle numbers in any country and any year. Adams, J. (1987) criticised the model's accuracy because there would always be a decline in traffic risk for any increase in the number of vehicles, but generally in non-linear way. These models are not able to incorporate the effect of risk factors on accident involvement.

Oppe (1989) assumes that fatality rates follow a negative exponential learning function in relation to the number of vehicle kilometers and time. This method has been found to be most effective when the components describing the time series behave slowly over time as follows:

$$\ln\left(\frac{F_t}{V_t}\right) = \ln(R_t) = \alpha t + \beta \qquad (1.2)$$

or equivalently $R_t = e^{\alpha\ t\beta}$ Where the *ln* function is the natural logarithm, F_t is the number of fatalities for some country in a year t,V_t the number of vehicle kilometers traveled in that year. R_t is F_t and α, β are constants. This means that the logarithm of the fatality rate decreases (sign of improvement) if a: is negative proportional with time. This model is called the negative exponential learning model, where? is supposed to be less than zero. Both α and β are the parameters to fit. This formula shows that countries with a large? should have a fast growth in traffic. The traffic volume will increase quickly first and at the end it will reach its saturation level, which differs from country to country.

Adams (1987) has stated a similar relation between fatalities (F) and vehicle kilometers (V), which was present: $Log(F/V) = \alpha + \beta * y$, where y = year - 1985

1.6 Problem Statement/Motivation

When I started reviewing what people have done on accident prediction models, I realized that the on data on individual behavior. I found some articles on accident prediction models but all of them assume the occurrence of accidents to follow a distribution and therefore, use the classical approach in their analysis and modeling. This study intends to apply survival theory models, since this approach has not been widely adopted by other researchers in this field. These models will be used to explain the impact of different factors on driver's conditional accident risk (hazard function). The conditional accident risk refers to the probability of a vehicle being involved in an accident at time t given that the vehicle had survived until that time. The essence is to see how survival modeling can be developed and used in explaining also the normal accident process based on accident data. Therefore, the study seeks to examine driver's accident risks and their dependence on various factors using statistical accident (survival) models. Survival modeling is commonly applied in medicine to the study of serious diseases and treatment methods. This study will assess the development needs of survival models in the area of traffic accident analysis.

1.7 Research Objectives

The objectives of the research will be to investigate the following issues;

- Which factors influence accident risks of drivers.
- How do principles of survival modeling apply to investigating accidents and accident risks.
- What are the data requirements in using survival models.
- How can accident risks be evaluated by survival models when only basic information on the parties involved in an accident is available.

1.8 Research Questions

The main problems that are formulated according to the objectives of this research emphasized the accident risk factors. They are based on previous research results that traffic accident risks vary with driver characteristics, vehicle attributes, and the time and place of driving.

The main questions of this study are;

1. Can driver involvement in road traffic accident be examined with survival models on the basis of exposure over time?
2. Do age and sex major factors?
3. Do vehicle characteristics contribute to accident risk beyond their interaction with exposure and speed?
4. Can driver involvement in accident be examined with survival models?
5. Does the use of unworn out tyres reduce accident risks?

The research objectives are not formulated as main problems. They will be discussed in the context of data analysis and interpretation.

CHAPTER 2

DETAILED REVIEW OF PREVIOUS RESEARCH RESULTS

The purpose of this literature review is to give a literature survey of the most important risk factors and concepts in road accidents. The literature for this study will be focused on the principal risk factors that influence traffic accidents globally, although there are other possible risk factors that are not included in this literature review, variables that were measured in this study have been specifically focused upon.

2.1 Concepts of accident risk

Many factors affect driver accident involvement. Nevertheless, at any given time, driver accident risk is affected by personal risk factors (e.g., hours of sleep the previous night), vehicle risk factors (e.g., brake adjustment), environmental factors (e.g., weather and roadway features), and, perhaps most important, risks created by other drivers and traffic. Driver errors can be violations of rules, mistakes of judgment, inattention errors, or inexperience errors. Common driver errors resulting in accidents include recognition errors (failure to perceive a crash threat) and decision errors (risky driving behavior such as tailgating), or poor decision-making in dynamic traffic situations (such as trying to cross a stream of traffic). (Michael S. et al, 2004).

Another common classification for driver errors resulting in accident is as follows (Dewer and Olson 2002): Rule-based (failure to obey rules or regulations), Knowledge-based (failure to understand required safe behavior), Skill-based (lack of proper skills to perform the task). Drivers can also make mistakes without obvious misbehaviors, such as failure to see another vehicle or misjudgment of a gap in the traffic stream. Red-light running may be regarded as a rule based misbehavior if it is intentional, a skill-based mistake if it is not. (Michael S. *et al*, 2004).

Reason (1990) proposed three error categories: violations (deliberate deviation), mistakes (intended action with unintended consequences), and lapses/slips (execution of unintended action). Rimmo (2002) has expanded this by splitting the lapses/slips category into inattention errors (unintended action resulting from recognition failure) and inexperience errors (unintended action resulting from lack of knowledge or skill). Rimmo's classification, with examples, follows:

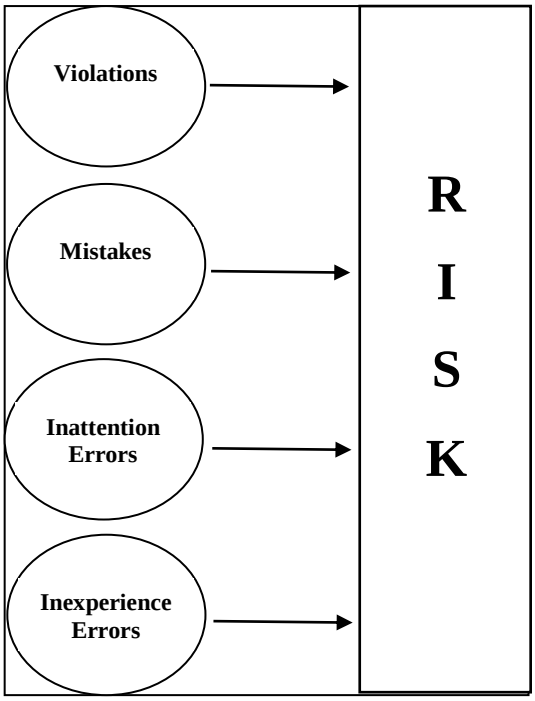
Violations (Deciding to drive when known to be very fatigued, Deliberately exceeding speed limits, Accelerating at green-to-yellow signal change.

Mistakes (Misjudging gap when crossing traffic, misjudging speed of oncoming vehicle, and misjudging stopping distance).

Inattention Errors (Failing to notice red light at intersection, Failing to see that vehicle has stopped in lane ahead, and Failing to notice sign).

Inexperience Errors (Having to check gear with hand, Driving in too low a gear and Switching on wrong appliance in truck).

Figure 2.1: Schematic drawing of the four error types contributing to driver risk



2.1.1 Risk and Road User Behaviour

Different studies indicate that the human factor (road users) is the major contributory factor to accidents. At the same, any error in the system and on roads will lead to unsafe road user behavior:

2.1.1.1 Speed and risk of accident involvement

Speed has been identified as a highly important influencing factor concerning road safety risk and consequences. An increase in average speed results in a higher risk of involvement in an accident and greater severity. In many countries, speed contributes to a significant percentage of all deaths on the roads. Leaf and Preusser (1999), for example, concluded that reducing vehicle speeds could have a highly significant influence on pedestrian accidents and injuries. Garber and Gadiraju (1988) determined that accident rates increased with increasing variance of speed.

2.1.1.2 Alcohol and risk of accident involvement

Drivers with high BAC (Blood Alcohol Content) in their blood have more chance of being killed than those with zero BAC (sober drivers). Hakkert and Braimaister (2002) provided a review of many studies and reported that the risk in traffic will increase rapidly with BAC. Such results have given the basis for setting BAC limits in many countries (e.g. 0.8 g/dl).

Evans (1991) has estimated that in 1982 about 53% of traffic accident fatalities in the USA had alcohol in their blood. This means that the total elimination of alcohol use would have decreased the amount of fatalities by over 50%.

Elvik and Vaa (1990) state as a summary of many studies that no other single human factor affects the occurrence of accidents as dramatically as alcohol. The risk of a fatal accident is over 100 times as high for drivers under the influence of alcohol, as it is for sober drivers. Drivers' accident risk grows exponentially as blood alcohol level increases.

Studies have shown that alcohol can increase the seriousness of injuries in traffic crashes as well as the chance of being involved in a crash (Segui-Gomez et al., 2007). In 2005, about 39% of all traffic fatalities and 254,000 injuries were attributed to crashes in the United States in which alcohol was involved (NHTSA, 2007).

2.1.1.3 Age of drivers and risk of accident involvement

Road accidents are the leading cause of death for young drivers and motorbike riders. The risk by age group per kilometer traveled and per hour exposed to traffic is higher among young people (15-24) and old (65+). However the exposure for young is higher than old people. Evans (1991) reported that young male drivers are overrepresented in accidents in the US. Page (2001) concluded from a survey in OECD countries that the higher the proportion of young people in the population,

the higher the number of road accident fatalities.

For the overall driver population, age is one of the strongest personal factors affecting crash involvement (NHTSA, 2000). Teenage drivers, especially males, have crash involvement rates per mile traveled that are several times higher than those of the adult population. Driver errors seen in teenage drivers include both risk-taking behaviors and misjudgments (Mayhew and Simpson 2003).

Hanowski *et al.* (2000) analyzed factors (including both personal driver factors and situational factors) predicting truck driver involvement in critical incidents (caused by the truck driver). They evaluated driver age, ambient illumination, prior night's sleep, current drowsiness rating, physical work requirements for the day, and several other factors. They found that driver age was the strongest predictor of critical incident involvement.

In Finland one study showed that there were differences amongst drivers, depending on their age. Driving for fun or leisure as well as driving during evenings, at night, and with passengers was more typical for young drivers than for middle-aged drivers. The most typical driving for middle aged drivers was going to and from work.

Also, a European study (Lynam and Twisk 1995) listed the special factors that may underlie the association between young age and accidents. They include: psycho-biological immaturity, a limited recognition of danger, the acceptance of risk, an excessive belief in one's own abilities, lack of experience, driving culture, and lifestyle induced risky type of exposure such as night driving.

2.1.1.4 Gender of drivers and risk of accident involvement

A study from the United Kingdom reported that male drivers between the ages of 17-20 are involved in over 4 times as many injury crashes as males overall and almost twice as many accidents as girls their own age (Clarke *et al*., 2006). In the United States, fatal crashes for young male drivers between the ages of 15 to 20 years of age increased by 5% between 1995 and 2005, while 15 to 20 year old females involved in a fatal crash decreased by 1% during the same time period (NHTSA, 2007).

As the number of female drivers and their amount of driving has increased in many western countries the number of accidents by female drivers has increased (Laapotti and Keskinen, 2003) even

among the high risk 18-24 year old group. One study showed that males and females have approximately the same amount of involvement in injury crashes per million miles traveled. The number of males involved in fatal crashes per million miles traveled, however, far outnumbered females (Williams, 2003). Therefore, when taking into account exposure (miles traveled), young males and females have almost the same amount of involvement in overall driver crashes with males being more likely to be involved in serious crashes.

Many studies have concentrated on some of the possible reasons for the differences that exist among male and female drivers. Meadows and Stradling (1999) concluded that females are more safety oriented drivers and Wells-Parker et al. (1996) concluded that women tend to have lower risk profiles. A study from Sweden compared young female and males learning how to drive and found that females study more theory than males and that females practice more driving skills in different environments during supervised training than males (Nyberg and Gregersen, 2007).

In the Singh (2003) study of driver attributes and rear-end crash involvement mentioned above, both young males and young females 18 to 24 years old were more likely than older drivers to be involved in rear-end crashes. Comparing genders, young males and young females had about the same risk of involvement in the struck vehicle role. For the striking vehicle role, however, young males had about a 50% higher risk of involvement than did young females.

2.1.1.5 Use of seat belts

Road accident research has found that seat belts reduce the fatal injuries significantly and it can reduce the risk of fatal injury to front-seat passengers. The use of seatbelts varies from country to country. In high-income countries, the usage rate tends to be high. In Sweden for instance, seatbelt usage exceeds 90% (K.oomstra et al., 2002). The use of seat belts reduces the probability of being killed by 40-50% for drivers and front-seat passengers and by 25% for passengers in the back seats as shown in (Elvik and Vaa, 2004). Regarding the use of safety seats for children and infants studies (WHO, 2004) have shown that reduce infant deaths in cars by 70% and deaths of small children by 50%. Mandatory seat belt use proves to provide strong protection against fatalities in accidents in different countries according to various studies.

The effectiveness of seat belts has been established in various studies (e.g. Rivara *et al*., 2000). Seat belts are estimated to reduce serious injuries from motor vehicle crashes by 55% and fatalities

by 50% (Forjuoh, 2003). In the United States the use of safety belts has increased by 65% since the early 1980s (National Highway Traffic Safety Administration, 2004). Various studies have attributed this large increase in seat belt usage to seat belt enforcement laws, principally primary enforcement laws that allow for enforcement officers to issue a citation whenever they observe an unbelted passenger or driver (Eby *et al*., 2003). Such a large increase in lower income countries would be very difficult to obtain because of the lack of resources that lower income countries have. It is believed that over half of vehicles in many lower income countries do not have functional seat belts and that some vehicles imported from high income countries lack functional seat belts as well (Forjuoh, 2003). In this environment, legislation to require the use of seat belts would be ineffective.

However, there is also evidence among the general population of drivers that non-safety belt use is associated with various risky driver attitudes and behaviors. Lancaster and Ward (2002) reviewed studies indicating that driver non-safety belt use is associated with speeding, short headways (tailgating), alcohol use, red light running, more previous traffic violations, and sensation-seeking personalities. Eby *et al*. (2003) observed that safety belt use among drivers using hand-held cell phones was lower in every age group studied than among comparable non-cell phone users. Thus, non-belt use by a driver should probably be regarded as a safety "red flag".

2.1.1.6 Vulnerable Road Users

A vulnerable road user is considered to be anyone using a bicycle or a motorcycle, as well as all pedestrians. Several studies have shown that these groups tend to be over-represented among crash victims (Peden *et al*., 2004) and are also at high risk of disability resulting from traffic crashes (Mayou and Bryant, 2003). Much of the burden of vulnerable road user traffic crashes is within the low and middle income countries with pedestrians being the most frequently injured road users in Latin America and the Caribbean (Hijar *et al*., 2004).

An example of the types of vulnerable road users that are most affected by traffic crashes comes from a study in Mexico. This study found that none of the pedestrians injured in the study had insurance, which typically left payment responsibilities up to the family because the injured person was most often the bread winner of the family (Hijar *et al*., 2004). Under these circumstances, family members of the victim only stop paying for treatments when all of their assets are gone,

consequentially leading the entire family into long-term debt. The individual injured could also lose their job and continually suffer from health problems because of the inability to continue their treatments (Over *et al*., 1992).

2.1.1.7 Driver fatigue

Drivers who are sleep deprived have significant deficits in vigilance and other cognitive abilities related to driving. McCartt *et al*. (2000) identified factors associated with why long-distance truck drivers reported falling asleep at the wheel. They found six underlying, independent factors, including (1) greater daytime sleepiness, (2) more arduous schedules, with more hours of work and fewer hours' off-duty, (3) older, more experienced drivers, (4) shorter, poorer sleep on the road, (5) symptoms of sleep disorder, and (6) greater tendency to nighttime drowsy driving.

The authors also suggest that if these six factors were to be ranked, a tendency toward daytime sleepiness was most highly predictive of falling asleep at the wheel, followed by an arduous work schedule and older, long-time drivers. Michael S. *et al*, (2004) found similar findings in an analysis they conducted on 567 professional drivers that included five different commercial driver types (long-haul drivers, short-haul drivers, bus drivers, drivers transporting wood, and drivers transporting dangerous goods). They found that regardless of the commercial driver type, sleepiness-related problems was strongly related to prolonged driving, sleep deficit, and driver's health status. Sagberg (1999) conducted a study of crashes caused by drivers falling asleep. The study showed that fatigue was a strong contributing factor in nighttime accidents, run-off-road accidents, and accidents after driving more than 150 km on one trip. Although his study was conducted on non-commercial drivers, many findings were consistent with McCartt *et al*.'s study. In addition, Sagberg found that more males than females were involved in sleep-related accidents. Sagberg also suggests that drivers' lack of awareness of important precursors of falling asleep in addition to the reluctance to discontinue driving despite feeling tired contributed to sleep-related accidents.

2.1.2 Risk and Road Conditions

Motorways have the lowest risk on injury accidents compared to other types of roads because of the separation between vehicle movements according to their speed (no high speed variance). (Elvik and Vaa, 2004) show that the rate of injury accidents per million vehicle kilometers of travel on motorways is about 25% of the average for all the public roads. Road surface conditions, poor road

surface, defects in road design and maintenance contribute to an increase in the risk of accidents. Bester (2001) reported that countries with more paved roads will lead to lower fatality rates.

2.1.3 Risk and vehicle related factors

New cars tend to have more safety and protection features, such as air bags, anti-brake system (ABS), etc. There is relation between vehicle age and risk of a car crash. One study (in WHO, 2004) showed that occupants in cars manufactured before 1984 have almost three times the risk of new cars. Many developed countries improved vehicle crash worthiness and safety, which means the protection that a vehicle gives its passengers (and to the Vulnerable Road Users) from a crash. Many countries in the European Union (EU) as well as USA have set out legislation for safety standards in motor vehicles, for instance the New Car Assessment Program (NCAP), where vehicle crash performance is evaluated by rating the vehicles models according to their safety level for occupant protection, child protection and pedestrian protection. Vehicle defects increase the risk of accident. The size of vehicle is crucial; the greater the mass of the vehicle (e.g. heavy trucks), the more protection people have inside the vehicle (their occupants) and the more involved in fatal accidents to others. It is known that poor vehicle maintenance and technical conditions can also contribute to accidents. In terms of periodic vehicle inspection, different research shows different results. (Elvik and Vaa, 2004) concluded in the review of macro-studies that there is no clear evidence that periodic vehicle inspection has an effect on the number of accidents, while (Hakim *et al*. 1991) presented in another review of macro-studies that the periodic inspection of motor vehicles reduces the number of road fatalities.

2.14 Risk and post-crash injury outcome

Different studies have shown that fatality rates are correlated with the level of medical facilities available in the country expressed in terms of population per physician and population per hospital bed, (Jacobs and Fouracre, 1977). A review of a European study, in (WHO, 2004), showed that about half of deaths from road accidents occurred at the spot of the accident or on the way to the hospital. Noland (2003) concludes that medical care has led to reductions in traffic-related fatalities in developed countries over time (1970-1996).

2.1.5 Socio-Economic Factors and Risks

There are many socioeconomic factors that contribute to the causes of accidents. Some of the major factors are the following;

2.1.5.1 Gross National Product (GNP).

It is widely known that the motorisation rate (vehicle per population) increases with income (GNP per capita). This may affect both exposure and the risk of fatal accidents. Many studies (e.g. World Bank, 2003) have shown that the fatalities per vehicle appear to decline rapidly with income. Maybe this reflects the shift from vehicles with high risk (motorbikes, foot) to safer and protected vehicles (e.g. four-wheelers) or it may show more funds and expenditure being spent by the country on its road safety measures. There is a negative relationship between income growth and the number of road accidents in the long term (Hakim *et al.* 1991). The increase in income leads to safer vehicles and more investment in road infrastructure, which leads to fewer road accidents and casualties. However, it should be clear that the improvement of income could also increase the travel distance (higher exposure) and more alcohol consumption (higher risk).

2.1.5.2 Unemployment

Few studies have used the unemployment factor as a risk factor for accidents. It appears to be negatively related to accidents and casualties. Hakim et al. (1991) has shown in the literature review he made that an increased unemployment rate in country might reflect on the ability to pay for a single journey and a reduced exposure to the whole journey. Page (2001) included employment (percentage of population in employment) into his model in the study conducted for the OECD countries. The higher employment figures showed an increase in the number of fatalities.

2.1.5.3 Urban population

Urban roads will have more accidents and fewer fatalities or severity per kilometer traveled than rural roads, because of the density of vehicles and the lower speeds of travel. Hakkert and Braimaister (2002) have shown in one macro-study that countries with a high level of urbanisation will have higher population densities and they may experience lower levels of fatalities and serious injuries.

Page (2001) has found that the population who live in urban areas have fewer road accident fatalities than other places. Bester (2001) also reported similar results that countries with higher road densities will have fewer fatality rates. Shorter distances to medical services can explain this.

In 2006, the fatality rate per 100 million vehicle miles traveled was reported as being 2.4 times higher in rural areas of the United States as compared to urban areas (NHTSA, 2008). Investigations that were carried out in the 1980s in several different states found that rural traffic injury fatality rates were higher than urban rates. Studies from other countries have documented higher occurrences of traffic crash fatalities in urban areas as well. A study in Ghana reported that the majority of traffic fatalities and injuries occurred in rural areas and that the crashes that occurred on rural roads were generally more severe (Afukaar et al., 2003) and a study in Quebec, Canada found that severe crashes are more common in rural areas (Thouez et al., 1991).

Many different explanations are available in the literature as to why such differences between urban and rural areas occur. One such explanation is that rural crash victims may not receive medical attention as quickly as their urban counterparts because the crashes often occur in remote areas (Clark, 2003). Another explanation is that rural roads may not be as safe as urban roads, with some lacking guardrails, traffic control devices, and traffic law enforcement. Combining the lack of sufficient traffic law enforcement with higher speed limits often times means that traffic crashes in rural areas are more severe than those in urban areas (Zwerling, 2005).

2.1.5.4 Illiteracy

Bester (2001) has analysed socio-economic factors in different countries and he found that the illiteracy percentage has a statistically significant effect on the national fatalities rate. He explained that a country that can read and write is expected to influence the ability of road users to understand rules of the road and road signs.

2.1.5.5 Technology level

Few studies have described the decline in the number of fatalities in all industralised countries as a result of the increase of technology use in level and road infrastructure (i.e. Evans, 1991).

2.1.6 Risk and other factors

Different macro studies have shown that the risk of crash will increase by other factors such as: poor visibility, using hand-held mobile telephones, dark conditions, wet roads and roads that are covered with snow or ice (Elvik and Vaa, 2004), (Evans, 1991). There is an inverse relationship between accidents and the average gasoline prices (Hakim *et al*., 1991). It seems that an increase in the price of gasoline reduces the number of trips and the exposure. Similarly, there is an inverse relationship between number of accidents and the number of driving licenses delivered (Van and Wets, 2003). Moreover, the road safety audit process is shown to have a clear impact on the number of accidents (Proctor *et al*., 2001). However, there is lack of data concerning all these factors and they are not available in many countries.

2.1.6.1 Stress

Stress is generally seen as a human response to an aversive or threatening situation. Heightened stress has been implicated in increasing the risk of vehicle crashes. Brown and Bohnert (1968) reported that 80% of drivers involved in fatal crashes, but only 18% of controls, were under serious stress involving interpersonal, marital, vocational, or financial areas prior to the crash. Finch and Smith (1970) reported similar findings.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Source of data

The data for this project were solely secondary data which was taken from Motor Traffic and Transport Unit of the Ghana Police Service, Northern regional office, Tamale.

3.2 Population

Three years of data, containing detailed information on accident – involved drivers for the period of 2007 to 2009, was used in this study. The data contained those fatal accidents that had occurred in the region within the period under consideration.

3.3 Study Variables

The Traffic Accident Form of the unit collects data surrounding the occurrence of the traffic fatality. The following data is collected data on the form: Name, Sex, Age, Date, Time, and Address where the event occurred, Number of fatal victims in the same event; Type of traffic accident (including but not limited to a collision with a fixed or moving object, pedestrian struck, or a overturned vehicle); Role of the victim during the event (Including but not limited to: Driver of the vehicle, passenger, pedestrian); Use of safety equipment; Type of vehicles, bicycle, motorcycle, pedestrian); and Alcohol level of the victim. Other variables include conditions, consequences, traffic and specific information about each “party” to the accident, age and type of driving and vehicle characteristics. All these variables were considered in the study.

3.4 Statistical Analysis

Drivers' accident risks and their dependence on various factors were examined using statistical (survival) accident models. SAS software package was used in the analysis. The analysis included model formulation, parameter estimation and model evaluation. The main survival model type used was the Cox type distribution - free proportional model. Statistical hypothesis testing was used as a tool for testing the various hypotheses. The statistical decisions were based on the developed survival models and the estimated parameters.

3.5 MODELING APPROACH

3.5.1 Overview of the approach

Survival analysis is a class of statistical methods for studying the occurrence and timing of events. These methods are most often applied to the study of deaths. In fact, they were originally designed for that purpose, which explains the name survival analysis. That name is somewhat unfortunate, because it encourages a highly restricted view of the potential applications of these methods. Survival analysis is extremely useful for studying many different kinds of events in both social and natural sciences (Paul D. Allison, 2008). Survival can simply be defined as time-to-event. Examples: time to die from disease say cancer. time to first marriage, time to promotion to next position, time to earthquake, time to first arrest after release from prison, time to birth after first marriage, time to divorce after first marriage, etc. Survival analysis was originally designed for prospective/longitudinal data on the occurrence of events. Longitudinal data are generated when you begin observing a set of individuals at some well - defined point in time, and you follow them for some substantial period of time, recording the times at which the events of interest occur. It is not necessary that every individual experience the event. Apart from recording the time of occurrence of events, you might also record other events related to time.

One can perform survival analysis when the data consist only of the times of events, but a common aim of survival analysis is to estimate causal or predictive models in which the risk of an event depends on auxiliary variables or covariates. If this is the goal, the data set must obviously contain measurements of the covariates. Some of these covariates like tribe, sex, religion may be constant

over time. Others, like income, marital status, or blood pressure, may vary with time. For time - varying covariates, the data set should include as much detail as possible on their variation across time.

Survival analysis is frequently also used with retrospective data. These are data in which a group of individuals are, instead of being followed in time as described above, asked to recall the dates of events like marriages, births, promotions, etc. In using methods of survival analysis for retrospective data, one should recognise the potential limitations. For one thing, people may make substantial errors in recalling times of events and they may forget some events entirely. They may also have difficulty providing accurate information on time - dependent covariates. A more difficult problem is that the sample of people who are actually interviewed may be different from those who actually were at the risk of the event. For example, people who have died or moved away will not be included. Nevertheless, although prospective (longitudinal) data are certainly preferable, much can be learnt from retrospective. Survival data have two common features that are difficult to handle with conventional statistical methods: censoring and time - varying covariates. Therefore, an important issue in survival research is how to deal with cases whose survival cannot be followed during the entire research period. Such individuals are called censored individuals (observations). There are generally three reasons why censoring may occur:

(1) A person does not experience the event before the study ends;

(2) A person is lost to follow-up during the study period;

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

Censoring can happen in the following three ways;

- Type I: the duration of the study is fixed to a chosen period. The study includes cases that are monitored from a set starting point for as long as the phenomenon under examination occurs, or until the individuals are lost to the monitoring, or the entire monitoring period has passed. Individuals whose monitoring does not provide information about the occurrence of the phenomenon under examination until they drop out, or at the end of the study period, are censored observations.

- Type II: The length of the monitoring period depends on the desired number or proportion of uncensored observations. The length of the period is the same as the survival of the individual with the longest life span. Individuals, who are removed from the study for various reasons or survive less than the monitoring period, are censored observations.
- Type Ill: The duration of monitoring is fixed. However, individuals may enter the study at different starting points. Censored observations are the ones whose survival period continues after the overall monitoring period has ended.

Survival studies can be divided into the following two groups:

- Monitoring censored to the right: the investigation has begun at a certain selected moment when the individuals entering the examination are exposed to the phenomenon under investigation, e.g. a medicine or treatment, the investigation is continued from that moment on for a certain length of time.
- Monitoring censored to the left: the investigation has begun at a certain selected moment, but includes individuals whose exposure to the phenomenon under investigation has begun before the examination period started (as in the present study).

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 (Bunday, 1991).

3.5.2 Principles of Survival Modeling

In survival analysis, we usually refer to the time variable as survival time, because it gives the time that an individual has "survived" over some follow up period. We also typically refer to the event as a failure, because the event of interest usually is death, disease incidence, or some other negative individual experience. Time = survival time and Event = failure

In survival analysis, the basic mathematical terminology and notation is the random variable (T) which represents the person's survival time. Since T denotes time, its possible values include all

nonnegative numbers; that is, T can be any number equal to or greater than zero. Now a small letter t denotes any specific value of interest for the random variable T. We also have the probability density function which gives the probability of an individual experiencing the event of interest at time t denoted by ***f(t).*** Other important functions in survival analysis, besides the density function, are the survival function denoted by S(t), and hazard function also denoted by h (t). The survival function S(t) gives the probability that a person survives longer than some specified time t, that is, S(t) gives the probability that the random variable T exceeds the specified time t. The survivor function is fundamental to a survival analysis, because obtaining survival probabilities for different values oft provides crucial summary information from survival data. For instance, at time $t = 0$, $S(t) = S(0) = 1$; that is, at the start of the study, since no one has gotten the event yet, the probability of surviving past time 0 is one; at time $t = \infty$, $S(t) = S(\infty) = 0$; that is, theoretically, if the study period increased without limit, eventually nobody would survive, so the survivor curve must eventually fall to zero. The hazard function ***h(t)*** , on the other hand , gives the instantaneous potential per unit time for the event to occur, given that the individual has survived up to time t. In contrast to the survivor function, which focuses on not failing, the hazard function focuses on not failing, the hazard function focuses on failing, that is, on the event occurring. Thus, in some sense, the hazard function can be considered as giving the opposite side of the information given by the survivor function. In other words, the hazard function can also represent the probability of an occurrence to end survival at point t on the condition that the individual or other object of examination has been "alive" until point t.

The following relationship exists between these functions:

$$h(t) = \frac{f(t)}{S(t)}$$

$$H(t) = \int_0^t h(t)dt$$

$$S(t) = exp[-H(t)]$$

Where h (t) is hazard function, H(t) is cumulative hazard function and S(t) is survival function. Finally, the Greek letter delta (δ) denotes a {0, 1} random variable indicating either failure or censorship. That is $\delta = 1$ for failure if the event occurs during the study period, or $\delta = 0$ if the survival time is censored by the end of the study period.

3.6 The Cox Proportional Model and Its Characteristics

This study intends to apply the Cox proportional hazards models. Proportional survival models assume that survival t has its density, hazard and survival functions. There are no special starting assumptions made on the form of the density function of survival t.

The Cox PH model is usually written in terms of the hazard mode1 formula shown below;

$h(t, X) = h_0(t)exp(X\beta)$

where $h(t, X)$ is the hazard function, $h_o(t)$ is the base level of the hazard function and exp(Xβ) is linear function formed by the variables and their parameters. This model gives an expression for the hazard at time t for an individual with a given specification of a set of explanatory variables denoted by X. That is, the X represents a collection of predictor variables that is being modeled to predict an individual's hazard.

The Cox model formula says that the hazard at time t is the product of two quantities. The first of these, $h_o(t)$, is called the baseline hazard function. The second quantity is the exponential expression e to the linear sum of $\beta_i X_i$, where the sum is over the p explanatory X variables. An important feature of this formula, which concerns the proportional hazards (PH) assumption, is that the baseline hazard is a function of t, but does not involve the X's. In contrast, the exponential expression shown here, involves the X's, but does not involve t. The X's here are called time-independent X's.

It is possible, to consider X's which do not involve t. Such X's are called time-dependent variables. If time-dependent variables are considered, the Cox model form may still be used, but such a model no longer satisfies the PH assumption, and is called the extended Cox model. This study will consider time-independent X's only.

The Cox model formula has the property that if the X's entire are equal to zero, the formula reduces to the baseline hazard function. That is, the exponential part of the formula becomes e to the zero, which is 1. This property of the Cox model is the reason why $h_o(t)$, is called the baseline function. Or, from a slightly different perspective, the Cox model reduces to the baseline hazard when no X's are in the model. Thus, $h_o(t)$ may be considered as a starting or "baseline" version of the hazard function, prior to considering any of the X's.

Another important property of the Cox model is the baseline hazard, $h_o(t)$, is an unspecified

function. It is this property that makes the Cox model a semiparametric model.

In contrast, a parametric model is one whose functional form is completely specified, except for the values of the unknown parameters. A key reason for the popularity of the Cox model is that, even though the baseline hazard is not specified, reasonably good estimates of regression coefficients, hazard ratios of interest, and adjusted survival curves can be obtained for a wide variety of data situations. Another way of saying this is that the Cox Proportional Hazard model is a "robust" model, so that the results from using the Cox model will closely approximate the results for the correct parametric model.

The parametric model is preferred if we are sure of the correct model, but if we are not completely certain that a given parametric model is appropriate, thus, when in doubt, as is typically the case, the Cox model will give reliable enough results so that it is a "safe" choice of model, and the user does not need to worry about whether the wrong parametric model is chosen.

Another appealing property of the Cox model is that, even though the baseline hazard part of the model is unspecified, it is still possible to estimate the β's in the exponential part of the model. All we need are estimates of the β's to assess the effect of explanatory variables of interest. The measure of effect, which is called a hazard ratio, is calculated without having to estimate the baseline hazard function.

One point about the popularity of the Cox model is that it is preferred over the logistic model when survival time information is available and there is censoring. That is, the Cox model uses more information – the surviving times – than the logistic model, which considers a (0,1) outcome and ignores survival times and censoring.

The hazard function h(t,X) and its corresponding survival curves S(t,X) can be estimated for the Cox model even though the baseline hazard function is not specified. Thus, with the Cox model, using a minimum of assumptions, we can obtain the primary information desired from a survival analysis, namely, a hazard ratio and a survival curve.

3.7 Estimation of the Cox Proportional Hazard Model

We now describe how estimates are obtained for the parameters of the Cox model. The parameters are the β's in the general Cox model formula shown above. The corresponding estimates of these

parameters are called maximum likelihood (ML) estimates and are denoted as β

The formula for the Cox model likelihood function is actually called a "partial" likelihood function rather than a (complete) likelihood function. The term "partial" likelihood is used because the likelihood formula considers probabilities only for those subjects who fail, and does not explicitly consider probabilities for those subjects who are censored. Thus the likelihood for the Cox model does not consider probabilities for all subjects, and so it is called a "partial" likelihood. The Cox Survival model can be further written for calculations into the form:

$S(t, X) = [S_0(t)]^{\exp(XB)}$

Where in this study S (t, X) is estimate of the share of drivers not involved in accidents until time t, $S_0(t)$ is base level of the survival function formed on the basis of data,

$exp(X\ \beta) = exp(\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)$. The β is the coefficients of the variables x_i (parameters)

When drawing up proportional survival or hazard models, stratification can also be used as an examination method (not considered in this study). The different basic hazards $ho(t)$, are estimated with the model for the different values of the stratifying variable. The survival model explains the effect of the variables on these different base levels of the hazard function.

In this study, the survival, or accident time, will be defined as the number of days counted from the start of the analysis period. The starting point for the accident data will be 1/112007 to 31/12/2009 for the 2007 accidents; 1/11/2008 to 31/1/2009 for the 2008 accidents and 1/1/2009 to 31/12/2009 for the 2009 accidents. Accident time will be the time from the starting time until the moment accident occurs.

The "survival" of drivers who got into accidents will be the number of days from the beginning of the examination period until the day a driver was involved in an accident. If a driver had been involved in more than one accident, the days between the two accidents will be considered the survival days.

One important feature in survival analysis is the idea of censored observations that generate great difficulty when trying to analyze such data using traditional statistical models such as multiple

linear regression.

The accident data taken from the MTTU contains, definition only drivers involved in accidents, this means that they have all experienced the event of interest. Therefore, in order to analyze the data using survival models, drivers that were involved in accidents within the first 244 days of each year were considered as “primary party” (uncensored drivers) and those that were involved after the 244 days were considered as “other involved party” (censored drivers). This approach enabled a distinction among the involved drivers who to be classified as uncensored and hence data was analyzed using survival approach. This situation is depicted in the figure below.

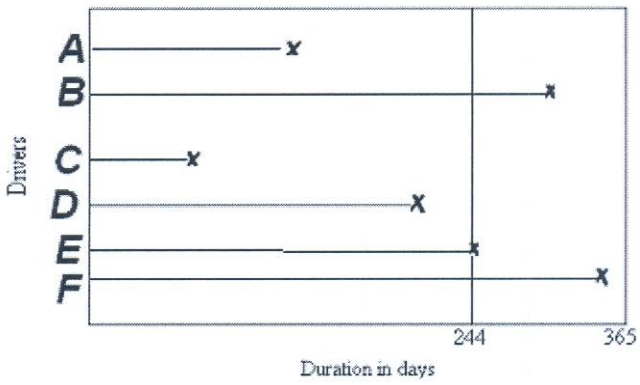


Figure 3.1: Theoretical display of survival time in the accident data

It is assumed that all the drivers entered the study at the first day of the year and are followed for twelve months. The study end was set to (244 days) about eight months. The horizontal axis represents time (Duration in days). Each of the horizontal lines labeled A through F represents a single driver. An X indicates that the accident occurred at the point in time. The vertical line at 2444 is a point at which we stop following the driver. Any accident that occurred at 244 days or earlier was considered uncensored, and hence, those accidents times are censored at time 244 days. Therefore, drivers A, C, D and E have uncensored accident times while drivers B and C have censored accident times.

Individual potential variables were selected and included into the models with the help of Kaplan-Meier estimates directly from the data. Next, the similarities of the survival functions obtained were also tested with the Log Rank test. The Log Rank test score used in the comparisons of the survival functions is asymptotically normally distributed. When there is a distinct differences between the survival functions, the variable's significance in the model will be tested the rest of the variables.

The evaluation of the proportional models' explanatory power and statistical significance was based on the size of the residuals and the DfBeta estimates of coefficients.

Because the variables in survival models can strongly correlate with each other, their impact on the characteristics of the most important models were examined by alternately removing variables from the models.

CHAPTER 4

DATA PRESENTATION AND ANALYSIS

4.1 Motor Traffic and Transport Union (MTTU) Data

4.1.1 Nature of the data

The Motor Traffic and Transport Union (MTTU) of the Ghana Police Service, Tamale gathers extensive and detailed information from each accident they investigate. Details of the information that is always taken can be found in appendix A. the whole data available during the beginning of the study contained information on 398 fatal accidents for the period of 2007 – 2009. Therefore, the interpretation of the findings of the study is limited to the Northern Region.

4.1.2 Driver and accident characteristics

The table 4.1 presents the ages of the involved drivers and the consequences to drivers (every accident had at least one fatality who could be driver or another occupant). Out of the 398 drivers, 91 (22.9%) had survived without injury, 186 (46.7%) had died and 121 (30.4%) had been seriously injured. Most (68%) drivers were 26 – 56 years old, 93 (23%) were younger, and 17 (4%) older.

Only 21 were female drivers.

Table 4.1: Distribution of drivers by injury severity and age according to the accident data

		Severity of injury			Total
		Fatalities	No/minor	Serious	
Age	$\leq$ 25 years	41	19	33	93
	26 -50	128	62	82	272
	50+	17	10	6	33
Total		186	91	121	398
Percentage of Total		46.70%	22.90%	30.40%	100.00%

Table 4.2 presents the yearly and monthly distribution of the accidents.There are clear fluctuations in the number of cases over years and months. It can be seen that out of the total number of 398 cases for the entire period of 3 years, 62 (15.6%) occurred in December alone.

Table 4.2: The number of drivers involved in accident by year and month

		Month												Total
		Jan	Feb	Mar	April	May	June	July	Aug	Sept	Oct	Nov	Dec	
year	2007	10	8	6	12	6	8	5	3	10	8	15	17	108
	2008	12	15	8	17	10	11	9	12	9	13	20	22	158
	2009	10	11	8	10	8	10	2	10	12	11	17	23	132
Total		32	34	22	39	24	29	16	25	31	32	52	62	398
Percentage of Total		8.0	8.5	5.5	9.8	6.0	7.3	4.0	6.3	7.8	8.0	13.1	15.6	100.0

4.2 Preliminary Analysis: Kaplan - Meier estimates of the accident data variables

In any data analysis it is always a great idea to do some univariate analysis before proceeding to more complicated models. In survival analysis it is highly recommended to look at the Kaplan-Meier curves for all the categorical predictors/variables. This will provide insight into the shape of the survival function for each group and give an idea of whether or not the groups are proportional (i.e. the survival functions are approximately parallel). Tests of equality across strata were used to explore whether or not to include the predictor in the final model. A variable was considered for

inclusion in to the model as a predictor if the test has a p-value of 0.05 or less. This elimination scheme criterion is used because all the predictors in the data set are variables that could be relevant to the model. If the predictor has a p-value greater than 0.05 in a univariate analysis it is highly unlikely that it will contribute anything to the model which includes other predictors.

Detailed information on the Kaplan-Meier estimates of all the variables under consideration can be found in appendix C.

Driver Age

Driver age was categorized into three groups: those aged less than or equal to 25 years constituted the first group, the second group was aged between 26 to 50 years and the last group aged above 50 years. However, the log-rank test of equality across these age groups (strata) realized a p-value of 0.0001 indicating that the variable driver age had a significant effect on accident time, thus age was included as potential candidate for the final model. From the graph it can be seen that those drivers aged between 26 and 50 had higher share of accidents, followed by those above 50 years and lastly followed by those up to or less than 25. In general, the pattern of one survivorship function lying above another means that the group defined by the upper curved live longer or had a more favourable survival experience than the group defined by the lower curve.

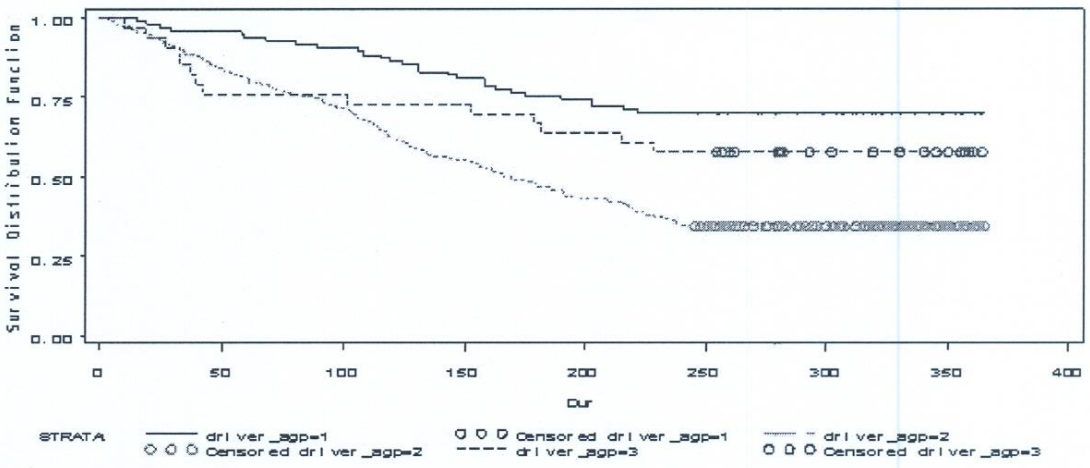


Figure 4.1: Survival distribution of age groups of drivers

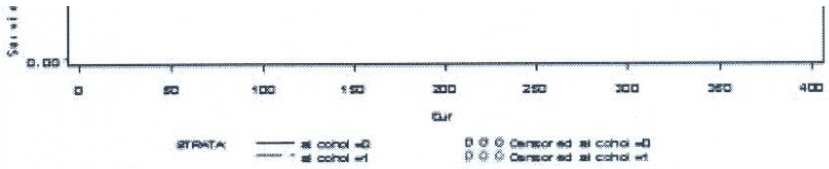


Figure 4.3: Survival distribution of the alcohol status of drivers

Use of safety belt

Drivers who used safety belts were labeled 1 and 0 for those who did not use. The log-rank test of equality across strata for the predictor usebelt realized a p-value of 0.0001 indicating that usebelt had a significant effect on the accident time distribution. Thus usebelt was included as potential candidate for the final model. From the graph it can see that drivers who used safety belts had higher survival experience than those who did not. Aside its significant effect on accidents time it also increases the severity and consequences of the accident.

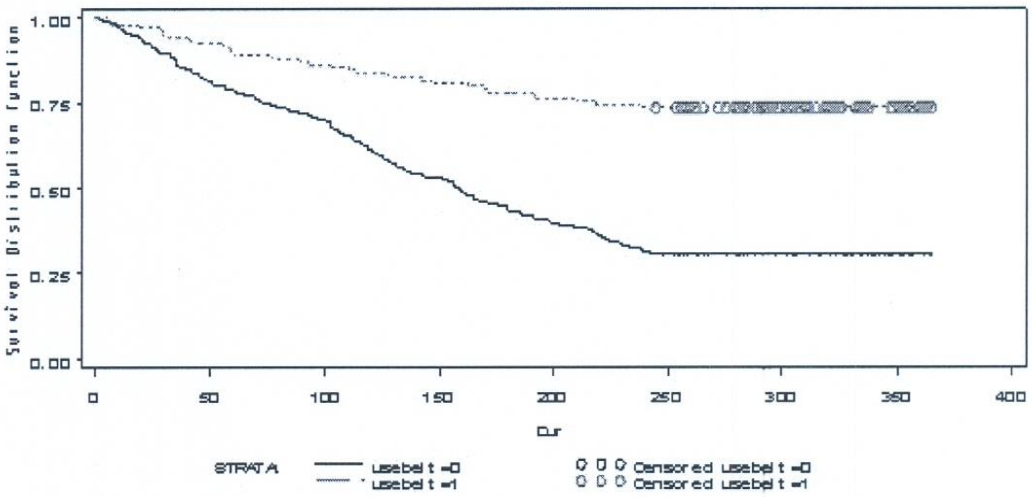


Figure 4.4: Survival distribution of the use of seat belts status of drivers

Annual Vehicle Kilometerage

Annual vehicle kilometerage was used to measure the exposure of drivers to traffic. This variable was categorized in to three groups: less than 5,000km/a, this category of drivers were given 1, 2 for those who traveled between 5,000 to 14,999km/a and 3 for those that traveled at least 15,000km/a. The log-rank test of equality across strata for the predictor annukil had a p-value of 0.4016 indicating that annukil had no significant effect on the accident time distribution. Thus annukil should not be included as a potential candidate for the final model.

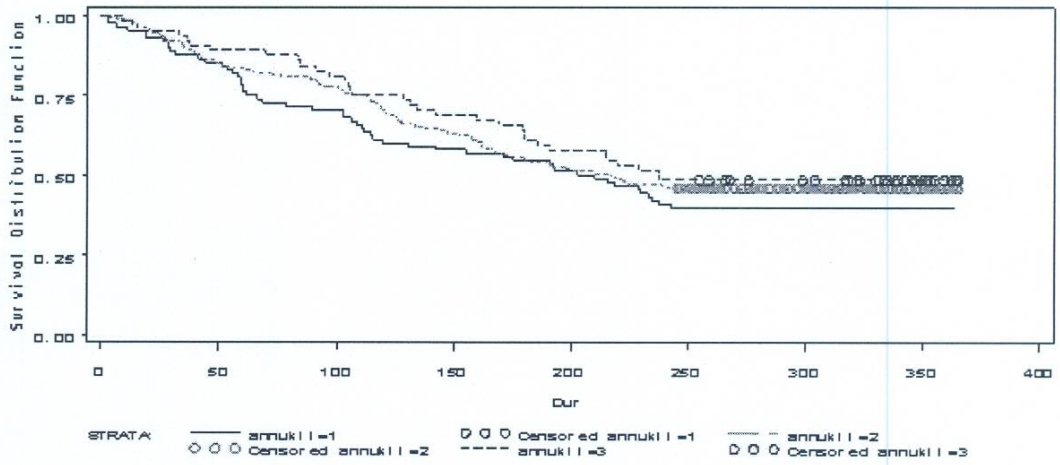


Figure 4.5: Survival distribution of the annual kilometers traveled by drivers

Road Section/Scene of Accidents

Accident locations were separated into links (labeled as 1) and junctions (labeled as 2). The log-rank test of equality across strata for the predictor scene has a p-value of 0.08017 indicating that scene had no significant effect on the accident time distribution. Thus scene was not included as a potential candidate for the final model, since the p-value is greater than the cut-off point.

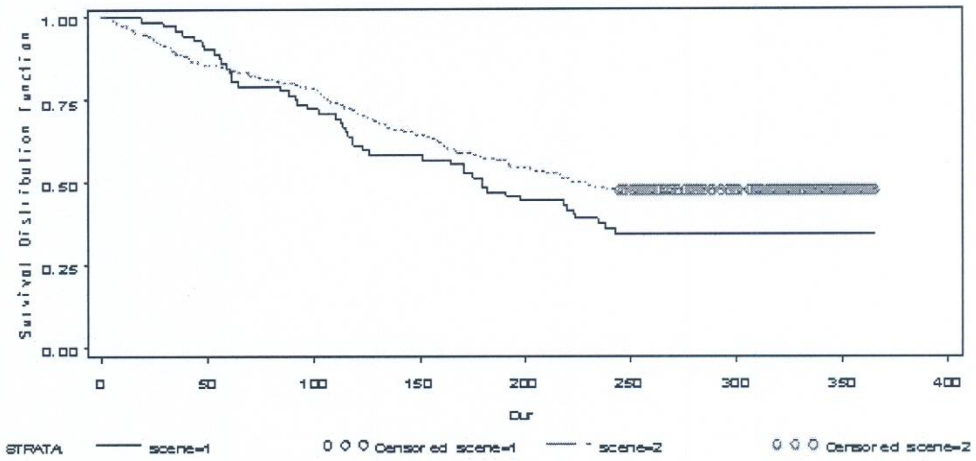


Figure 4.6: Survival distribution of Scene of accidents among drivers

Estimated Speed of Vehicle at the time of accident

Speed was categorized in to two classes: up to 80km/h (which represented 1) and over 80km/h (which represented 2). Speed remained a statistically significant variable (P = 0.0001). Thus speedveh was included as potential candidate for the final model. From the graph it can be seen that drivers whose speeds exceeded 80km/h had lower survival experience than those who did not.

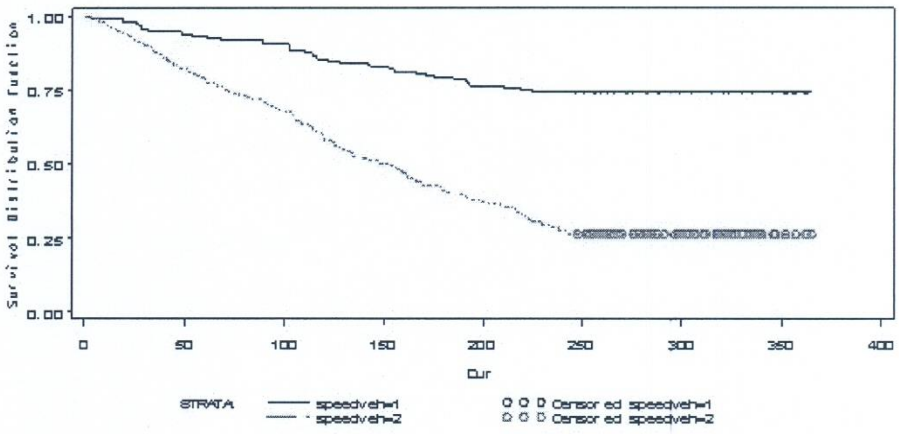


Figure 4.7: Survival distribution of the speed of drivers

Weight of Vehicle

The weight of vehicle was categorized into two groups: less than or equal to 1,000Kg (labeled as 1) and over 1,000Kg (labeled as 2). The log-rank test of equality across strata for the predictor wghtveh had a p-value of 0.4483 indicating that wghtveh had no significant effect on the accident time distribution. Thus wghtveh should not be included as a potential candidate for the final model, since the p-value is greater than our cut-off of 0.05.

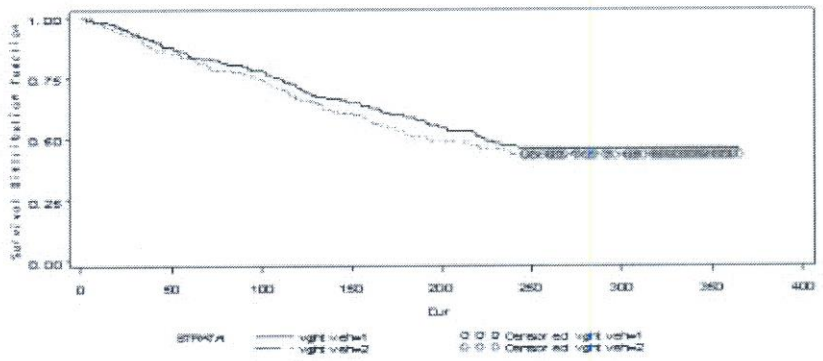


Figure 4.8: Survival distribution of the weight of vehicle

Tyres Condition/tyres tread depth

The tyres tread depth had two levels: 0 for vehicles whose tyres tread were less than or equal to 4mm and 1 for those with over 4mm. the log-rank test of equality across strata for the predictor tyrescon realized a p-value of 0.3897 indicating the tyrescon had no significant effect on the accident time distribution. Thus tyrescon should not be included as a potential candidate for the final model, since the p-value is greater than our cut-off of 0.05.

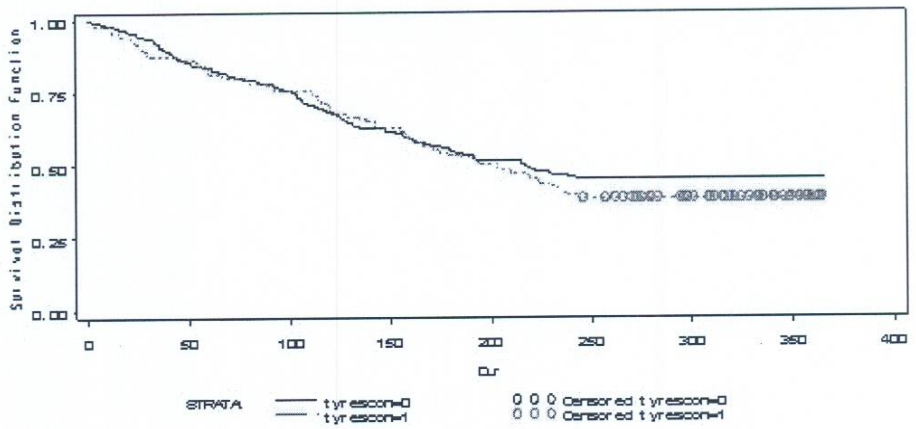


Figure 4.9: Survival distribution of the tyres condition of the vehicles

Driving experience (years since licensing)

The age of driving license had two levels: 1 for those with at least 5 years since licensing and 2 for drivers with under 5 years. The log-rank test of equality across strata for the predictor agelic realized a p-value of 0.0001, indicating that agelic had a significant effect on the accident time distribution. Thus agelic was included as a potential candidate for the final model, since the p-value is less than the cut-off of 0.05.

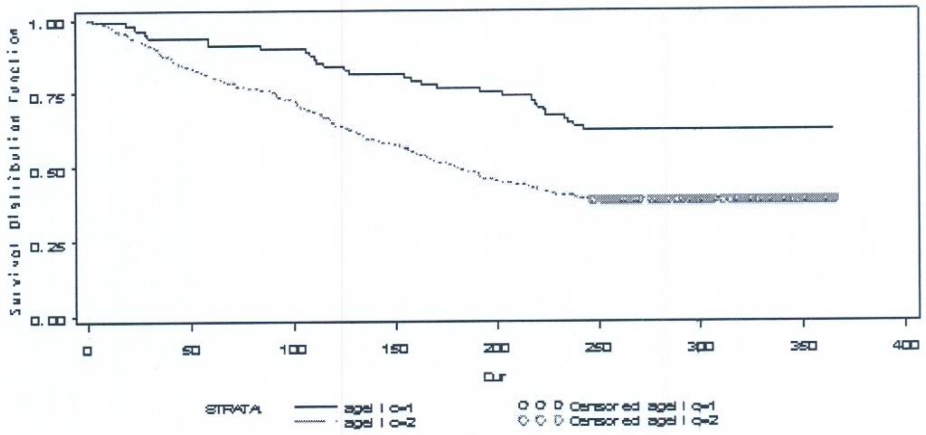


Figure 4.10: Survival distribution of the license duration of drivers

Route familiarity

Route familiarity had two levels: 0 for those drivers who seldom pass the scene of accident more

than once in a month and 1 for those that pass scene of accident at least once a month. The log-rank test of equality across strata for the predictor rutfam had a p-value of 0.8079. Thus rutfam was not included as a potential candidate for the final model, since the p-value is greater than the cut-off of 0.05.

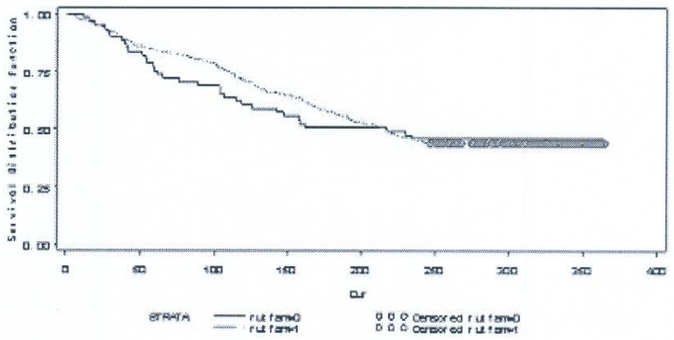


Figure 4.11: Survival distribution of familiarity of route of drivers

Road Surface Condition

Road surface condition had two levels: 1 for dry surface and 0 for wet surface. The log-rank test of equality across strata had a p-value of 0.5470 for the predictor rdsurf, thus rdsurf must not be included in the final model.

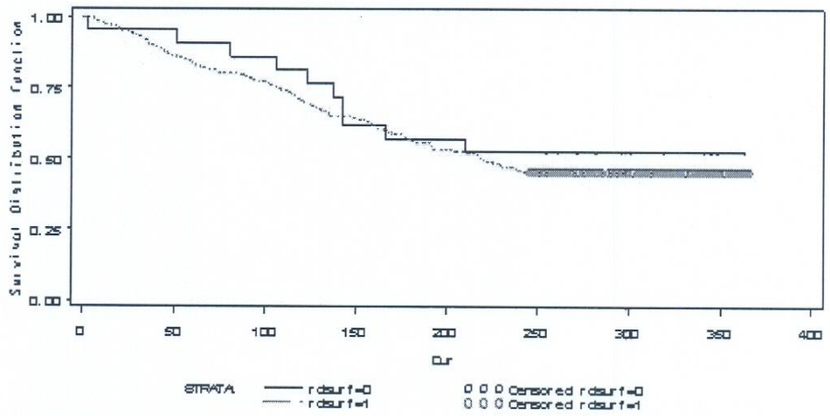


Figure 4.12: Survival distribution of the road surface condition at scene of accident

Ownership

The ownership of the vehicle was categorized into two levels: 1 for own vehicle and 0 for not own vehicle. The log-rank of equality across strata for the predictor owner had a p-value of 0.1280, thus owner was not included as a potential candidate for the final model because this p-value is greater than the cut-off of 0.05. Therefore, ownership of the vehicle did not have a significant effect on drivers' accident time distribution.

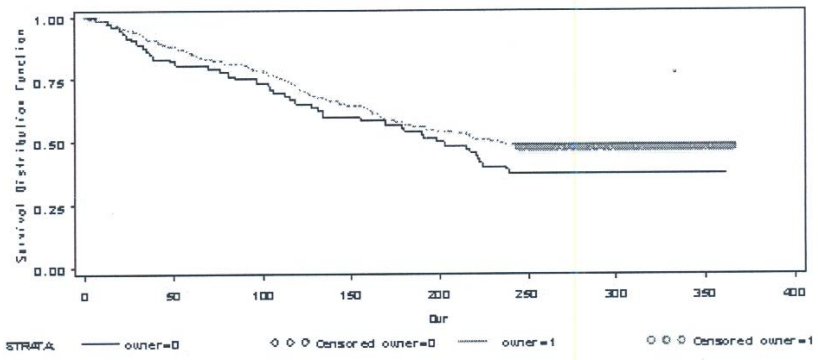


Figure 4.13: Survival distribution of vehicles ownership

Age of vehicle

Age of vehicle was categorized into two classes: up to 10 years. Age of vehicle remained a statistically variable significant variable (P = 0.0001). Thus ageveh was included as potential candidate for the final model. From the graph one can see that drivers whose vehicle aged over 10 years had lower survival experience than those with vehicle less than 10 years.

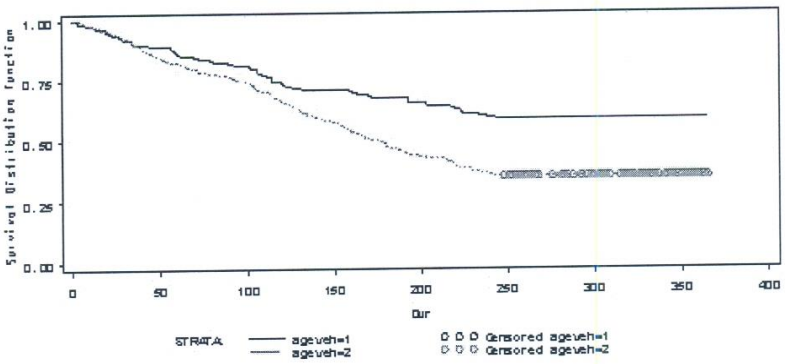


Figure 4.14: Survival distribution of ages of vehicles

Type of vehicle and weather condition

These two variables proved statistically insignificant and hence were excluded from the final model.

Summary of the effects of the variables with the Kaplan-Meier Estimate

Kaplan- Meier estimates identified variables that had major effects on accident time distribution. These included driver age, sex, use of safety belt, use of alcohol, speed of vehicle, age of driving license and age of vehicle.

4.3 Models for the MTTU data

4.3.1 Models and their compilation principles

The Kaplan-Meier suggested which of the many variables should be included in the models (that is predictors that had a p-value of less than 0.05. Some of these variables represented characteristics associated with the driver and the vehicle. The other variables referred specifically to the accident situation.

The variables were first divided into two groups:

1) The most important variables from the point of view of Kaplan-Meier estimate and
2) Other interesting variables though the Kaplan-Meier analysis did not support the inclusion of those variables in the final model, but they are associated with the research objectives. These

Variables include: annual vehicle kilometreage, tyres condition, weight of vehicle and route familiarity.

Model 1: The model was estimated using the first group of variables to obtain model 1 A. the significant variables in model 1 A were selected out and re-run to obtain 1 B. Next, interactions were added to the model 1 B to obtain the final model 1C.

Model 2: The model was estimated on the basis of both the first and second group of variables to obtain model 2A. The significant variables in model 2A were selected and re-run to obtain the final model 2B. There was no interaction effect in this model.

Table 4.3: List of the most important variables, categories, values and reference levels used for the development of the proportional hazards models of the MTTU data

Variables Name	Content	Categorisation	Reference level
annukil	Annual vehicle kilometerage	1. < 5000km/a	< 5000km/a
		2. 5000–14000km/a	
		3. 15000km/a	
driver_agp	Age of driver	1. $\leq$ 25 years	$\leq$ 25 years
		2. 26-50 years	
		3. 50 +	
alcohol	Use of alcohol	0. No	No
		1. Yes	
sex	Sex of driver	1. Male	female
		2. Female	
wghtveh	Vehicle weight	1. > 1000kg	>1000kg
		2. $\geq 1000kg$	
speedveh	Estimated speed of vehicle at time of accident	1. $\leq 80Km/h$	$\geq 80Km/h$
		2. >80Km/h	
ageveh	vehicle Age	1. $\leq$ 10 years	$\leq$ 10 years

Table 4.3: continued

Variables Name	Content	Categorisation	Reference level
		2. >10 years	
tyrescon	Tyres tread depth	1. > 4mm	> 4mm
		2. ≤ 4mm	
rutfam	Familiarity of route	0. Scene of accident passed more seldom than once in a month.	0. scene of accident
		1. scene of accident passed more often than once in a month	
agelic	Age of driving license	1. ≥ 5 years	1. ≥ 5 years
		2. <5 years	

4.3.1.1 Development of Models 1

The categorical predictor driver_agp has three levels and therefore it was included using dummy variables with the group driver_agp=1 as the reference group. The proc phreg yielded the following output;

Table 4.4: Cox model 1A: which includes the most important variables from the point of view of Kaplan Meier estimate

Output 1: Model 1A
The PHREG Procedure
Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr>ChiSq	Hazard Ratio	95% Hazard Ratio Confidence limits	
sex	1	1.27725	0.28114	20.6395	<.0001	3.587	2.067	6.223
usebelt	1	-0.73455	0.23413	9.8426	0.0017	0.480	0.303	0.759
alcohol	1	0.87965	0.20132	19.0921	<.0001	2.410	1.624	3.576
ageveh	1	0.25537	0.20106	1.6132	0.2040	1.291	0.870	1.914
agelic	1	0.33735	0.24374	1.9156	0.1663	1.401	0.869	2.259
speedveh	1	1.02858	0.22639	20.6414	<.0001	2.797	1.795	4.359
driver_agp2	1	0.66861	0.22999	8.4513	0.0036	1.952	1.243	3.063
driver_agp3	1	0.52192	0.39461	1.7493	0.1860	1.685	0.778	3.652

Linear Hypotheses Testing Results

Label	Wald Chi-Square	DF	Pr > ChiSq
driver_agp	8.5020	2	0.0143

Interpretation of the above output (Model 1A)

The output in table 4.4 provides the coefficient estimates and their associated statistics. It is noted that there is no intercept estimate - a characteristic feature of partial likelihood estimation. The column labeled Hazard Ratio is e^{β} . For indicator or (dummy) variables with values 1 and 0, one can interpret the hazard/risk ratio as the ratio of the estimated hazard for those with a value of 1 to the estimated hazard for those with a value of zero (controlling for other covariates). For example, the estimated hazard ratio for the variable usebelt (use of safety belt) is 0.480. This means that the hazard of accident for those drivers who did not use safety belt is only about 48% higher than the hazard for those who did use safety belt (controlling for other covariates).

For quantitative covariates, a more helpful statistic is obtained by subtracting 1.0 from the hazard ratio and multiplying by 100. This gives the estimated percentage change in the hazard for each one unit increase in the covariate. For example, the variable ageveh, the hazard ratio is 1.291 which yields 100 (1.291 - 1) = 29.1. Therefore for each one-year increase in the age of the vehicle the hazard of accident goes up by an estimated 29.1%.

Also looking at the individual coefficients of the two groups of the variable driver_agp and their associated chi-square statistics, we see that both of the dummy variables are statistically significant. However, each of these coefficients is a comparison with the omitted category of driver_agp1(which

corresponding to those drives aged less than or equal 25 years). More useful is the global test reported for driver_agp at the bottom part of the output, which has a Wald chi-square of 8.5920 with 2 degrees of freedom, again significant at the 0.05 level.

Overall, the highly significant variables include sex, use of safety belt, use of alcohol, speed of vehicle, and age of driver. Therefore, the final model of main effects will include all the aforementioned significant variables. These significant variables were re-runned to obtain the output of model 1B.

Table 4.5: Cox model 1B: which includes the most important variables from Model 1A

Output 2: Model 1B
The PHREG Procedure
Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr>ChiSq	Hazard Ratio	95% Hazard Ratio Confidence limits	
sex	1	1.20582	0.25893	21.6865	<.0001	3.339	2.010	5.547
usebelt	1	-0.70347	0.21491	10.7150	0.0011	0.495	0.325	0.754
alcohol	1	0.90190	0.18941	22.6721	<.0001	2.464	1.700	3.572
speedveh	1	1.18898	0.21875	29.5426	<.0001	3.284	2.139	5.042
driver_agp	1	0.36236	0.15427	5.5174	0.0188	1.437	1.062	1.944

Interactions

All possible interactions of these five significant variables were also conducted to ascertain interactions effects that should be included in the model. All the interactions proved statistically insignificant except the interactions of driver sex and use of safety belt and hence was included to arrive at the final model (Model 1C).

Table 4.6: Cox model 1C: which includes the significant interaction of sex and safety belt use

Output 3: Final Model 1C
The PHREG Procedure
Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr>ChiSq	Hazard Ratio	95% Hazard Ratio Confidence limits	
sex	1	0.82338	0.33798	5.9351	0.0148	2.278	1.175	4.419
usebelt	1	-2.17920	0.68889	10.0068	0.0016	0.113	0.029	0.436
alcohol	1	0.83028	0.19066	18.9651	<.0001	2.294	1.579	3.333
speedveh	1	1.22874	0.22195	30.6496	<.0001	3.417	2.212	5.279
driver_agp	1	0.36533	0.15422	5.6113	0.0178	1.441	1.065	1.950
sexbelt	1	1.25998	0.54301	5.3840	0.0203	3.525	1.216	10.219

4.3.1.2 Development of Models 2

This model is developed using both the first and the second group of variables. All these variables were included (to obtain model 2A) and those that were not statistically significant were eliminated from the model, and the significant ones were re-runned to arrive at the final model 2B. The variables included driver age and sex, use of safety belt, use of alcohol, speed of vehicle, age of driving license, age of vehicle, annual vehicle kilometreage, vehicle tyres condition, weight of vehicle and route familiarity.

Also, the categorical predictor annukil has three levels and therefore the annukil =1 was used as the reference group and hence was created inside the proc phreg. This yielded the output in table 4.7

Overall, there are highly significant effects of driver age, annual vehicle kilometreage, use of alcohol, use of safety belt, speed of vehicle, tyres condition and a marginally significant effect of age

Table 4.7: Cox model 2A: which includes variables used in model 1A and some other interesting variables

Output 4: Model 2A
The PHREG Procedure
Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr>ChiSq	Hazard Ratio	95% Hazard Ratio Confidence limits	
sex	1	0.59775	0.34221	3.0510	0.0807	1.818	0.930	3.555
usebelt	1	-0.85539	0.25954	10.8622	0.0010	0.425	0.256	0.707
alcohol	1	1.10444	0.23274	22.5194	<.0001	3.018	1.912	4.762
ageveh	1	0.21207	0.23830	0.7920	0.3735	1.236	0.775	1.972
agelic	1	0.51500	0.27747	3.4448	0.0635	1.674	0.972	2.883
speedveh	1	1.18762	0.25391	21.8768	<.0001	3.279	1.994	5.394
tyrescon	1	0.48967	0.20705	5.5932	0.0180	1.632	1.087	2.448
wghtveh	1	-0.02177	0.18038	0.0146	0.9039	0.978	0.687	1.393
rutfam	1	-0.44185	0.25370	3.0332	0.0816	0.643	0.391	1.057
driver_agp2	1	0.61749	0.24996	6.1025	0.0135	1.854	1.136	3.027
driver_agp3	1	0.77469	0.42432	3.3332	0.0679	2.170	0.945	4.985
annukil2	1	-0.54953	0.25281	4.7249	0.0297	0.577	0.352	0.947
annukil3	1	-0.80285	0.32834	5.9788	0.0145	0.448	0.235	0.853

Linear Hypotheses Testing Results

Label	Wald Chi-Square	DF	Pr > ChiSq
driver_agp	6.4771	2	0.0392
annukil	6.5258	2	0.0383

of license. The predictor sex is also not significant but from prior research we know that it is a very important variable to have in the final model and therefore it was included in the final model (model 2B). However, of the interactions of these variables proved statistically insignificant.

Table 4.8: Cox model 2B: which includes the most important variables from model 2A

Output 5: Model 2B
The PHREG Procedure
Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr>ChiSq	Hazard Ratio	95% Hazard Ratio Confidence limits	
sex	1	0.60565	0.29601	4.1862	0.0408	1.832	1.026	3.273
usebelt	1	-0.79550	0.22300	12.7254	0.0004	0.451	0.292	0.699
alcohol	1	0.91737	0.19118	23.0260	<.0001	2.503	1.721	3.640
speedveh	1	1.29852	0.22201	34.2100	<.0001	3.664	2.371	5.661
tyrescon	1	0.33968	0.18380	3.4153	0.0646	1.404	0.980	2.014
driver_agp	1	0.38741	0.15863	5.9641	0.0146	1.473	1.079	2.010
annukil	1	-0.43520	0.16507	6.9511	0.0084	0.647	0.468	0.894

4.3.2 Hazard ratios/Relative risks to drivers

All the models in the above outputs displayed the hazard ratios/relative risks along with their 95% confidence interval in the last three columns.

Driver age

Drivers age proved to be a significant accident risk factor in all the models. In model 1A for instance which indicated the individual contribution of each category of the age groups (the age group ≤ 25 serves as reference level). It can be seen that drivers aged between 26 to 50 years had 1.952 times (with 95% confidence interval: 1.243 - 3.063) greater at risk than drivers aged 25 and those who were older than 50 years had 1.685 times greater at risk than those in their early 20s. Also, the final model lC gave the overall contribution of driver age (p=0.0178) with a hazard ratio

of 1.441 which means that for any one year increase in the age of the driver, The hazard of accident increases by 44.1 %.

Driver Sex

According to the Kaplan Meier estimate sex had a strong effect on accidents time, which actually proved to be a significant variable in several of the models. In model 1 C for instance, indicates that male drivers had 2.278 (1.175 - 4.419 with 95% confidence interval) times greater than the risk of female drivers. However when the second group of variables were added, which included annual vehicle kilometreage, route familiarity, tyres condition and weight of the vehicle, the effect of sex had a moderate influence.

Driver's use of alcohol

Driving under the influence of alcohol significantly increased drivers accident risks. Alcohol use was a significant variable in all the models. According to model 2B, the relative risks of drivers under the influence of alcohol, was 2.5031 times (1.721 - 3.640 with 95% confidence interval) greater than that of sober drivers.

Use of Safety Belt

Use of safety belt had a very strong explanatory power in all the models. According to model 2B, the hazard ratio for use of belt is 0.451, indicating that if a driver changes from not use of belt to use of belt, whiles holding other covariates constant the hazard of accident decreases by (100% - 45.1 %) = 54.9%.

Annual Vehicle Kilometreage

Drivers' annual vehicle kilometreage had a strong explanatory power in all the models it featured. In model 2A for instance, it can be seen that drivers that had traveled between 5,000km/a, to 14,000krn/a, had 42.3% lower than those that had travelled less than 5,000krn/a (the reference group). Also those that had travelled for at least 15,000km/a had (100% - 44.8%) = 55.2% lower than those that had travelled less than 5,000km/a. In the final model 2B gave the overall contribution of drivers' annual kilometreage (p=0.0084), the hazard ratio of 0.647 which means that for any one year increased in driver's exposure to traffic, it is associated with (100% - 64.7%) = 35.3% decrease in expected time to accident holding all other covariates constant.

This result indicates that drivers' accident risks decreases as annual kilometerage increases. It

can be seen in the model that annual vehicle kilometerage is negatively related to the hazard of accidents and hence, positively related to survival. But it is generally known that higher exposure to traffic is associated with higher risk, but this result and for that matter this research is in opposition to this believe. This perhaps might be due to accumulated experience on the part of these drivers. That is to say drivers who had a high annual kilometreage had also more driving experience and hence lower involvement in accidents. However this might need further probing.

Route Familiarity

The route familiarity variable was not a significant variable as indicated in model 2A ($p = 0.0816$). The hazard ratio is 0.643 indicating that, the accident risks was about 35.7% lower for drivers who were familiar with the site of the accident compared to other drivers.

Vehicle Weight

Vehicle weight proved to be a discriminating factor although its statistical significance in model 2A was very weak ($p = 0.9039$). According to model 2A, the estimated risk ratio is 0.978. This means that the hazard of accident for drivers using light vehicles had 97.8% of the hazard for those drivers who used the larger weight vehicles.

Vehicle age

Vehicle age had no significant effect on accident risk, in the models though it proved to be a significant variable in the Kaplan-Meier estimate. However, in model 2A, the hazard ratio is 1.236, indicating that vehicles older than 10 years had 1.2 times the risk of vehicles that are less or equal to 10 years.

Tyres condition treads depth

Tyres condition/tread depth proved to be statistically significant ($p = 0.018$) in model 2A. According to this model, drivers with very worn out tyres vehicles had 1.632 times that of drivers who used less worn out tyres (>4mm).

Speed of vehicle

Speed is a statistical significant variable in predicting the hazard of accidents as confirmed in the models. According to model 2A, the hazard of accidents for drivers who drove over 80km/h is 3.279 times (with a 95% confidence interval: 1.994 - 5.394) that of those who drove less than 80km/h.

Age of license

This variable was used as a proxy to assess the level of experience of the driver. Age of license was a moderate significant variable in model 2A. However, the hazard ratio is 1.674, indicating that those drivers with duration of license less than 5 years had 1. 7 times the risk of those with license duration of at least 5 years.

4.4 Evaluation of the models and methods

The goal of statistical model development is to obtain the model which best describes the data. That is to say, the fitted model must provide an adequate summary of the data upon which it is based. Therefore, a complete and thorough examination of the model's fit and adherence to the model's assumption is of great importance and concern. There are several methods for assessment of a fitted proportional hazards model which include;

1. Testing for the assumption of proportionality of the models.
2. Overall summary measures of goodness of fit.
3. Identification of influential and poorly fit subjects/individuals.

4.4.1 Testing for the assumption of proportionality of the developed models

The Cox proportional hazard assumes that the hazard of one individual is proportional to the hazard of any other individual, where the proportionality constant is independent of time. This means that the ratio of the risk of dying (if death is the event of interest) of two individuals is the same no matter how long they survive.

This requires that covariates not be time-dependent. If any of the covariates varies with time, the Proportional hazards assumption is violated. This fact can be used to test the assumption by including a time-covariate interaction terms in the model and testing if the coefficient for interaction is significantly different from zero. The proportional hazards assumption is vital to the interpretation and use of a fitted proportional hazards model.

Though, there are several methods for verifying that a model satisfies the assumption of proportionality, we will, in this project, check proportionality by including time-dependent covariates in the model. In proc phreg it is very easy and convenient to include data step programming inside the procedure. Time dependent covariates are interactions of the predictors with time. In this analysis

the interactions with log (time) was used because this is the most common function of time used in time-dependent covariates but any function of time could be used. If a time-dependent covariate is significant this indicates a violation of the proportionality assumption for that specific predictor. We use a test statement to test all the time dependent covariates together in one collective test. Testing for proportionality of model 1B yields the following output;

Table 4.9: Testing for assumption of proportionality of model 1B using time-dependent covariate interaction

Output 6: Model 1B
The PHREG Procedure
Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr>ChiSq	Hazard Ratio	95% Hazard Ratio Confidence limits	
sex	1	-1.74057	1.62511	1.1471	0.2841	0.175	0.007	4.240
usebelt	1	0.30600	0.97340	0.0988	0.7532	1.358	0.202	9.151
alcohol	1	0.17103	0.89018	0.0369	0.8476	1.187	0.207	6.792
speedveh	1	0.47408	1.01201	0.2194	0.6395	1.607	0.221	11.677
driver_agp	1	1.25640	0.73653	2.9099	0.0880	3.513	0.829	14.879
sext	1	0.67410	0.35283	3.6503	0.0561	1.962	0.983	3.918
usebeltt	1	-0.22095	0.21520	1.0541	0.3046	0.802	0.526	1.222
alcoholt	1	0.17613	0.19710	0.7985	0.3715	1.193	0.810	1.755
speedveht	1	0.16255	0.22184	0.5369	0.4637	1.177	0.762	1.817
driver_agpt	1	-0.20129	0.16140	1.5554	0.2123	0.818	0.596	1.122

Linear Hypotheses Testing Results

Label	Wald Chi-Square	DF	Pr > ChiSq
test_proportionality	7.9568	5	0.1586

It can be seen in the above output of testing of proportionality for model 1B that tests of all the time-dependent variables were not significant either individually or collectively so we do not have enough evidence to reject proportionality and will assume that we have satisfied the assumption of proportionality for this model. Since all the covariates in this model proved to be time independent, we conclude that Cox Proportionality assumption is satisfied for model 1B, and hence best describes the MTTU data.

However, for model 1C which included the interaction term between sex and belt, the variable sex proved not to satisfy the proportional assumption. The output for the testing of proportionality of model 1C is seen as follows;

Table 4.10: Testing for assumption of proportionality of model 1C using time-dependent covariate interaction

Output 7
The PHREG Procedure
Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standardd Error	Chi-Square	Pr>ChiSq	Hazard Ratio
sex	1	-2.67039	1.72167	2.4057	0.1209	0.069
usebelt	1	-1.22318	1.18350	1.0682	0.3014	0.294
alcohol	1	0.20301	0.88497	0.0526	0.8186	1.225
speedveh	1	0.34893	1.01451	0.1183	0.7309	1.418
driver_agp	1	1.22159	0.73613	2.7538	0.0970	3.393
sexbelt	1	1.38475	0.56009	6.1126	0.0134	3.994
sext	1	0.78382	0.36146	4.7024	0.0301	2.190
usebeltt	1	-0.23834	0.22487	1.1233	0.2892	0.788
alcoholt	1	0.14947	0.19572	0.5833	0.4450	1.161
speedveht	1	0.20414	0.22373	0.8325	0.3615	1.226
driver_agpt	1	-0.19142	0.16130	1.4083	0.2353	0.826

Linear Hypotheses Testing Results

Label	Wald Chi-Square	DF	Pr > ChiSq
test_proportionality	8.5163	5	0.1300

The conclusion is that all of the time-dependent variables are not significant supporting the assumption of proportional hazard except sex when the interaction term was included in the model.

Testing for assumption of proportionality of model 2B with time-varying covariate.

Table 4.11: Testing for assumption of proportionality of model 2B using time-dependent covariate interaction

Output 8
The PHREG Procedure
Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr>ChiSq	Hazard Ratio
sex	1	-3.35264	1.73038	3.7540	0.0527	0.035
usebelt	1	-0.34508	0.99697	0.1198	0.7292	0.708
alcohol	1	0.27916	0.88813	0.0988	0.7533	1.322
speedveh	1	0.98160	1.00662	0.9509	0.3295	2.669
tyrescon	1	0.22133	0.85419	0.0671	0.7955	1.248
driver_agp	1	1.47413	0.77195	3.6467	0.0562	4.367
annukil	1	-1.77499	0.76301	5.4117	0.0200	0.169
sext	1	0.91063	0.37918	5.7676	0.0163	2.486
usebeltt	1	-0.09454	0.22115	0.1828	0.6690	0.910
alcoholt	1	0.15504	0.19716	0.6184	0.4316	1.168
speedveht	1	0.07510	0.22145	0.1150	0.7345	1.078
tyrescont	1	0.03108	0.19239	0.0261	0.8717	1.032
driver_agpt	1	-0.24352	0.16889	2.0792	0.1493	0.784
annukilt	1	0.30722	0.17027	3.2556	0.0712	1.360

Linear Hypotheses Testing Results

Label	Wald Chi-Square	DF	Pr > ChiSq
test_proportionality	10.8699	7	0.1444

It can be seen in the above output of testing of proportionality for model 2B that tests of all the time-dependent variables were not significant either individually (except sex) or collectively so we do not have enough evidence to reject proportionality and will assume that we have satisfied the assumption of proportionality for this model.

4.4.2 Assessing Goodness of Fit of the models using Residuals

There are several graphical methods available for assessing the goodness of fit of a proportional hazards model. These graphical methods are based on residuals and are often used as diagnostic tools. In multiple regression methods, residuals are referred to as the difference between the

observed and the predicted values (based on the regression model) of the dependent variable. However, when censored observations are present and only a partial likelihood function is used in the proportional hazards model, the usual concept of residuals is not applicable. In the following we introduce three different types of individual residuals: Cox-Snell, martingale and deviance residuals. Martingale residuals are obtained by transforming Cox-Snell residuals, and deviance residuals are a further transformation of martingale residuals. However, these three different residual statistics are computed for each individual in the sample.

These individual residuals can be plotted versus the survival time or a covariate. The pattern of the graph provides some information about the appropriateness of the proportional hazards model. It also provides information about outliers and other patterns. Similar to other graphical methods, interpretation of the residual plots may be subjective.

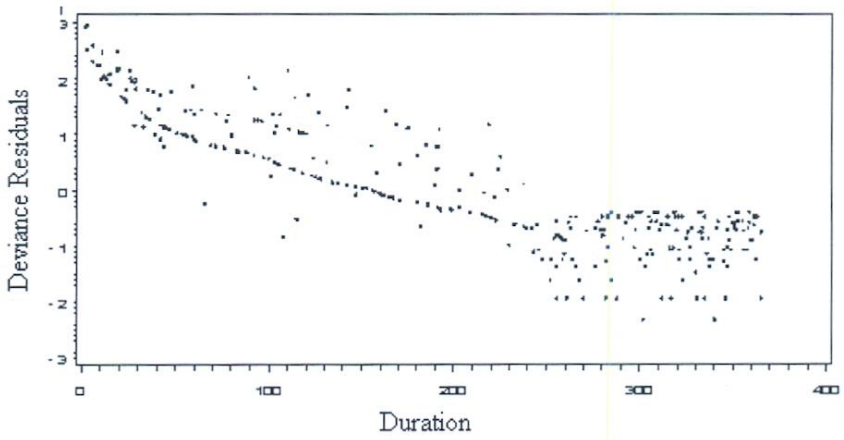
While Cox-Snell residual method is useful in assessing the goodness of fit of parametric models, they are not/ so desirable for a Cox proportional hazards model where a partial likelihood function is used and the survivorship function is estimated by nonparametric methods. The martingale residuals hare a skewed distribution with mean zero.

The deviance residuals behave much like residuals from Ordinary Least Square regression: They have a mean of zero and an approximated standard deviation of 1.0, for that matter they are symmetrically distributed about zero when the fitted model is adequate. They are negative for observations that have longer survival times than expected and positive for observations with survival times that are smaller than expected. In other words deviance residuals are positive for persons who survivor for shorter time than expected and negative for those who survive longer. The deviance often used in assessing the goodness of fit of a proportional hazards model. They can be used like residuals from OLS. Very high or very low values suggest that the observation may be an outlier in need of special attention.

The graph of deviance residuals against survival time or a covariate can be used to check the ad-equacy of the proportional hazards model. The presence of certain patterns in these graphs may indicate departures from the proportional hazards assumption, while extreme departures from the main cluster indicate possible outliers or potential stability problems of the model.
The plot of deviance residuals against the survival time (Dur) using model 1C yields the following

figure;

Figure 4.15: Plot of Deviance residuals of Model 1C against the survival time



The figure above plots the deviance residuals against survival time (Dur). Roughly speaking, the residuals are distributed symmetrically around zero between -3 and 3 with peculiar patterns. Larger positive (negative) residuals are associated with smaller (larger) t values. The deviance residuals suggest that the proportional hazards model provides a reasonable fit to the data.

Covariate-wise residuals

Apart from these individuals residuals mentioned above, we also have covariate-wise residuals statistics which include Schoenfeld residuals, weighted Schoenfeld residuals and Score residuals. All these three share a rather unusual property: instead of a single residual for each individual, as it is in the case of the individual residuals category, there is a separate residual for each covariate for each individual. They sum to 0 (approximately) in the sample i.e. the sum of say, the Schoenfeld residuals for a covariate is zero A major difference among them is that the Score residuals are defined for all observations, while the Schoenfeld residuals (both weighted and unweighted) are not defined for censored observations (they are missing in the output data set). The examinations of graphs of these three residuals behave much the same. Both graphical and statistical testing approach of Schoenfeld residuals were used to evaluate the models.

The main function of residuals is to detect possible departures from the proportional hazards assumption. Since Schoenfeld residuals are, in principle, independent of time, a plot that shows a

18	40	1	1	.	1	1	40	.	.	.	.	.	.	.	.
19	42	1	1	0	0	2	31	0	-0.12739	-0.13650	-0.82560	0.10645	-0.4740	-0.19461
20	47	1	1	0	1	2	30	0	-0.11624	-0.12584	0.18106	0.11220	-1.5086	-0.16905
21	60	1	1	1	1	2	27	1	-0.10951	0.87820	0.17716	0.11699	-4.6030	0.84024

Table 4.13: Assessing the goodness of fit of Model 2B using Schoenfeld residuals data set

Output 10

obs	Dur	Status	sex	usebelt	alcohol	Speedveh	agedriver	annukil	tyrescon	schsex	schusebelt	schalcohol	schspeedveh	schagedriver	schannukil	schtyrescon
1	35	1	1	0	1	.	30	2	0	.	.	.	.	.	.	.
2	36	1	1	0	1	2	29	3	0	-0.13536	-0.12693	0.18455	0.11165	-2.6661	1.14877	-0.31796
3	51	1	1	0	1	2	34	2	0	-0.13179	-0.12277	0.18949	0.11761	2.4818	0.15471	-0.33021
4	83	1	1	0	1	2	29	3	0	-0.10834	-0.12416	0.18860	0.13791	-2.2659	1.09494	-0.31387
5	102	1	1	0	1	1	27	2	0	-0.11578	-0.12931	0.19307	-0.85599	-4.2235	0.11232	-0.32039
6	105	1	1	.	1	2	29	3	0	.	.	.	.	.	.	.
7	112	1	1	0	1	2	26	2	0	-0.10167	-0.13454	0.19557	0.14404	-5.3065	0.07835	-0.33756
8	118	1	1	0	1	2	27	2	1	-0.07443	-0.14215	0.20518	0.15231	-4.5654	0.04952	0.67514
9	6	1	1	0	1	2	26	2	0	-0.12590	-0.13373	0.17160	0.10627	-5.9759	0.15007	-0.33374
10	12	1	1	0	1	2	28	2	1	-0.12884	-0.13335	0.17090	0.10875	-3.8851	0.14894	0.66317
11	15	1	1	.	0	2	30	3	0	.	.	.	.	.	.	.
12	19	1	1	1	1	2	45	1	1	-0.13195	0.86344	0.17502	0.11137	12.9813	-0.84380	0.66998
13	21	1	1	0	1	2	40	2	0	-0.13358	-0.13256	0.17490	0.10839	8.1068	0.14963	-0.32840
14	23	1	1	0	1	2	28	2	1	-0.13544	-0.13441	0.17734	0.10991	-3.8740	0.15172	0.67486
15	26	1	1	0	1	1	29	1	1	-0.13803	-0.13698	0.18073	-0.88799	-2.9978	-0.84538	0.67663
16	31	1	1	.	1	1	33	1	0	.	.	.	.	.	.	.
17	34	1	1	0	1	2	42	2	0	-0.13331	-0.12501	0.18410	0.10996	10.2779	0.14652	-0.31315
18	40	1	1	.	1	1	40	2	0	.	.	.	.	.	.	.
19	42	1	1	0	0	2	31	2	0	-0.13484	-0.12627	-0.81760	0.11351	-0.5364	0.15258	-0.32326
20	47	1	1	0	1	2	30	2	0	-0.13067	-0.12173	0.18788	0.11852	-1.5111	0.15339	-0.32741
21	60	1	1	1	1	2	27	1	1	-0.12638	0.88112	0.18503	0.12507	-4.5926	-0.86747	0.67508

The outputs in tables 4.12 and 4.13 display 21 observations for each of the two models with the lowest accident time in the data sets.

Consider a driver who was involved in accident on the 60th day (according to table 4.12), was a male of 27 years old which according to the Schoenfeld residuals was 4.6 years younger than the model predicts. He had used safety belt, although the model predicts a probability of only 0.22 that a driver involved in an accident on the 60th day will be using safety belt, (schbelt=0.88), he had used alcohol and the model predicts a probability of 0.92 that a driver involved in accident in

60 days would be using alcohol (schalco=0.18), his speed was over 80km/h and the model actually predicts a probability of 0.87 that a driver involved in accident at that time would have been moving at a speed that is over 80km/h (schspd=0.13).

Almost a similar result can be seen on that same individual predicted by model 2B in table 4.13. The driver had used a vehicle whose tyre's tread depth was above 4mm, although the model predicts a probability of only 0.32 that a driver involved in accident in 60 days would have been using a vehicle whose tyres tread will be over 4mm (schtyrescon= 0.68).

Next step is to plot the residuals of each covariate against the survival time (Dur), and if the plot shows a relationship with time, it is evidence against the assumption of proportionality. The underlying idea behind the statistical test and the graphical approach is that if the PH assumption holds for a particular covariate then the Schoenfeld residuals for that covariate will not be related to survival time. The SAS codes produced the graphs in figures 4.16 and 4.17.

Figure 4.16: Plots of Schoenfeld residuals for each of the predictors in model 1C against survival time

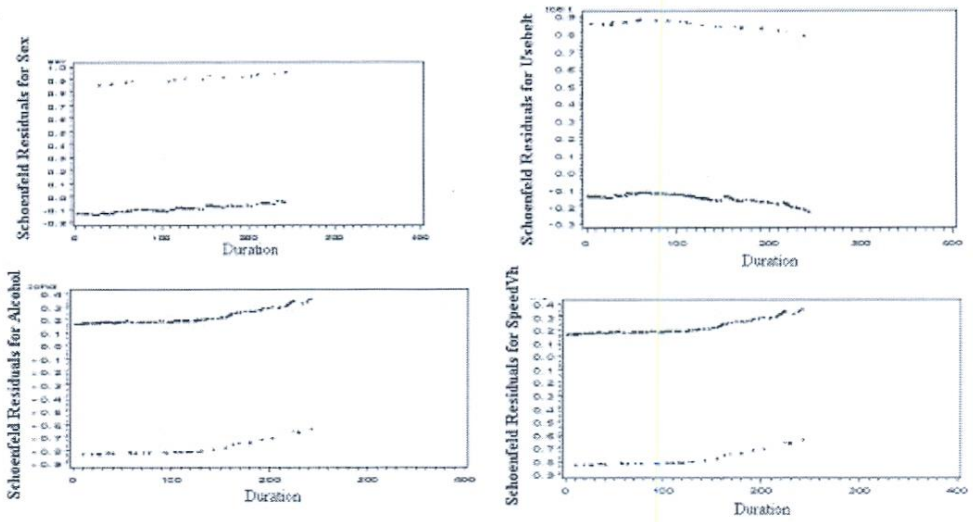
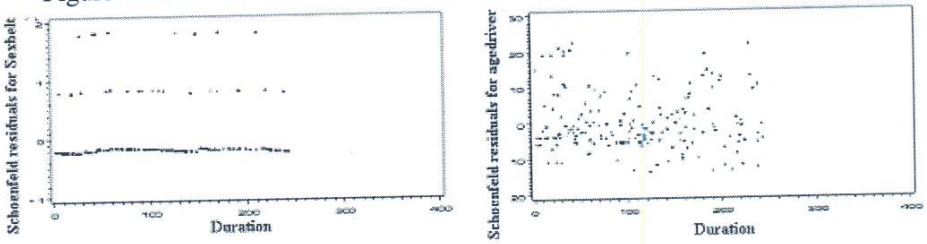


Figure 4.17: Plots of Schoenfeld residuals for model 1C against survival time



The graphs for the Sex, alcohol, usebelt, sexbelt and speed residuals are not very informative, which is typical of graphs for dichotomous covariates. For driver age, the residuals have a fairly random scatter. It can be concluded therefore that there is no evidence against the assumption of proportionality and hence the model adequately fits the data since the plots do not show any relationship with time.

Now, for the implementation of the Schoenfeld statistical test approach can be thought of as a three-step process.

Step 1. Obtain Schoenfeld residuals for each predictor variable in the model as seen above

Step 2. Create a variable that ranks the order of failures. The subject who had the first (earliest) event gets a value of 1; the next gets a value of 2, and so on.

Step 3. Test the correlation between the variables created in the first and second steps. The null hypothesis is that the correlation between the Schoenfeld residuals and ranked failure time is zero. Rejection of the null hypothesis leads to a conclusion that the PH assumption is violated.

The statistical test approach offers a more objective approach for assessing the PH assumption compared to the subjectivity of the graphical approach. However, the graphical approach enables one to detect specific kinds of departures from the PH assumption; the researcher can see what is going on from the graph. Consequently, it is recommended that when assessing the PH assumption, the investigator use both graphical procedures and statistical testing before making a final decision.

The correlation output for model 1C is as follows;

Table 4.14: Assessing the proportional hazard assumption of model 1C using Schoenfeld statistical test approach

Output 11
Pearson Correlation Coefficients
Prob $>|r|$ under $H_0 : \rho = 0$
Number of Observations

	schsex	schusebelt	schalcohol	schspeedveh	schagedriver	schsexbelt
timerank	0.15307	-0.07405	0.10811	0.08157	-0.09264	-0.02832
Rank for Dur	0.0386	0.3191	0.1452	0.2723	0.2123	0.7035
	183	183	183	183	183	183

The sample correlations with their corresponding p-values printed underneath are shown above. The p-values for Sex, usebelt, alcohol, speedveh, agedriver and sexbelt are 0.0386, 0.3191, 0.1452, 0.2723, 0.2123 and 0.7035 respectively, suggesting that the PH assumption is violated for Sex, but reasonable for usebelt, alcohol, speedveh, agedriver and sexbelt. This p-value is used for evaluating the PH assumption for the covariates. A nonsignificant (i.e., large) p-value, say greater than 0.10, suggest that the PH assumption is reasonable, whereas a small p-value, say less than 0.05, suggests that the variable being tested does not satisfy this assumption.

For model 2B, the correlation output is as follows;

Table 4.15: Assessing the proportional hazard assumption of model 2B using Schoenfeld statistical test approach

Output 12
Pearson Correlation Coefficients
Prob $>|r|$ under $H_0 : \rho = 0$
Number of Observations

	schsex	schusebelt	schalcohol	schspeedveh	schagedriver	schannukil	schtyrescon
timerank	0.15216	-0.09784	0.10813	0.07124	-0.10632	0.04124	0.06193
Rank for Variable Dur	0.0444	0.1977	0.1543	0.3488	0.1614	0.5879	0.4156
	175	175	175	175	175	175	175

The sample correlations with their corresponding p-values printed underneath are shown above. The p-values for Sex, usebelt, alcohol, speedveh, agedriver, annukil and tyrescon are 0.0444, 0.1977, 0.1543, 0.3488, 0.1614, 0.5879 and 0.4156 respectively, suggesting that the PH assumption is violated for Sex again in this model, but reasonable for usebelt, alcohol, speedveh, agedriver, annukil and tyrescon. This is the reason why the effect of sex on accident time is less conclusive.

4.4.3 Identification of influential and poorly fit subjects

Another important aspect of evaluation is a thorough examination of regression diagnostic statistics to identify which if any subjects: have an unusual configuration of covariates, exert an undue influence on the estimates of the parameters and/or have an undue influence on the fit of the model. Leverage is a diagnostic statistic that measures how "unusual" the values of the covariates are for an individual. In other words, it measures how far an independent variable deviates from its mean.

These leverage points can have an effect on the estimate of regression coefficient. Therefore how high leverage contributes to a measure of the influence that a covariate value has on the estimate of a coefficient is of concern in proportional hazards model.

An observation is said to be influential if removing the observation substantially changes the estimate of the coefficients. Most OLS regression packages can compute various influence statistics that measure how much would change if a particular observation is removed from the analysis. Such statistics can also be computed for Cox regression models. The Likelihood Displacement (LD) statistic measures influence on the model as a whole. This statistic tells one (approximately) how much the log-likelihood (multiplied by 2) will change if the individual is removed from the sample.

Table 4.16: Identification ofobservations that have leverage in model 2B using the DfBeta coefficient estimates

Output 13. Data set C: Influence statistics of model 2B

obs	Dur	Status	sex	usebelt	alcohol	Speedveh	tyrescon	driver_agp	annukil	dsex	dusebelt	dalcohol	dspeedveh	dtyrescon	ddriver_agp	dannukil
1	35	1	1	0	1	.	0	2	2	.	.	.	.	.	.	.
2	36	1	1	0	1	2	0	2	3	0.012141	0.001596	0.003132	-0.001888	0.001891	-0.000773	0.020836
3	51	1	1	0	1	2	0	2	2	-0.003374	-0.002846	0.002970	0.001920	-0.005749	0.000558	-0.002157
4	83	1	1	0	1	2	0	2	3	0.011250	0.001679	0.002897	0.000025	0.001193	0.000154	0.016048
5	102	1	1	0	1	1	0	2	2	-0.002860	-0.011235	0.007426	-0.037988	-0.009936	-0.002045	0.000163
6	105	1	1	.	1	2	0	2	3	.	.	.	.	.	.	.
7	112	1	1	0	1	2	0	2	2	-0.000401	-0.001800	0.001805	0.002220	-0.003833	0.001072	-0.002574
8	118	1	1	0	1	2	1	2	2	0.002781	-0.000429	0.000673	0.001172	-0.001904	0.000436	-0.001596
9	6	1	1	0	1	2	0	2	2	-0.005225	-0.005328	0.003429	0.001740	-0.008330	-0.000824	-0.003104
10	12	1	1	0	1	2	1	2	2	-0.010517	-0.002869	0.004401	0.005014	0.023395	0.001371	0.006249
11	15	1	1	.	0	2	0	2	3	.	.	.	.	.	.	.
12	19	1	1	1	1	2	1	2	1	-0.022978	0.036582	0.008396	0.017079	0.016261	0.006785	-0.012530
13	21	1	1	0	1	2	0	2	2	-0.005470	-0.004847	0.003208	0.001747	-0.007522	-0.000584	-0.003031
14	23	1	1	0	1	2	1	2	2	-0.010135	-0.002646	0.004186	0.004614	0.021784	0.001229	0.005722
15	26	1	1	0	1	1	1	2	1	-0.028411	-0.017256	0.009033	-0.036386	0.011288	-0.001437	-0.014849
16	31	1	1	.	1	1	0	2	1	.	.	.	.	.	.	.
17	34	1	1	0	1	2	0	2	2	-0.004707	-0.003824	0.003259	0.001760	-0.006226	-0.000107	-0.002565
18	40	1	1	.	1	1	0	2	2	.	.	.	.	.	.	.
19	42	1	1	0	0	2	0	2	2	-0.013473	-0.007866	-0.029537	0.006110	-0.008439	0.001386	-0.002555
20	47	1	1	0	1	2	0	2	2	-0.003473	-0.002912	0.003039	0.002061	-0.005899	0.000574	-0.002224
21	60	1	1	1	1	2	1	2	1	-0.015668	0.026717	0.006300	0.012945	0.011399	0.005719	-0.009193
22	72	1	1	0	1	2	0	2	.	.	.	.	.	.	.	.
23	77	1	1	1	1	2	1	2	.	.	.	.	.	.	.	.
24	69	1	1	0	1	2	0	2	.	.	.	.	.	.	.	.
25	91	1	1	.	1	2	0	2	2	.	.	.	.	.	.	.
26	94	1	1	0	0	2	0	2	2	-0.010835	-0.006641	-0.025588	0.006906	-0.007427	0.001802	-0.003015
27	103	1	1	0	1	2	0	2	3	0.009523	0.001238	0.002647	-0.000100	0.000871	0.000503	0.014004
28	106	1	1	0	1	2	0	2	1	-0.000606	-0.000313	0.000549	0.001205	-0.001088	0.000891	-0.001811
29	114	1	1	.	1	1	1	2	2	.	.	.	.	.	.	.
30	10	1	1	0	1	2	0	3	2	-0.000924	-0.000017	0.002038	0.005150	-0.006219	0.022966	-0.003713
31	28	1	2	1	1	2	1	2	1	0.045208	0.033748	0.013548	0.012590	0.009208	0.008729	0.004491
32	59	1	2	0	0	2	1	1	1	0.026815	-0.006971	-0.012306	0.006716	0.003847	-0.009713	0.000049
33	108	1	2	0	1	2	0	2	1	-0.048687	0.003980	-0.010258	-0.003753	0.018194	-0.003365	0.007634
34	119	1	1	.	1	2	1	2	1	.	.	.	.	.	.	.
35	120	1	1	0	0	2	0	2	2	-0.007739	-0.006798	-0.021860	0.006877	-0.007676	0.001811	-0.003833
36	92	1	1	1	1	1	1	2	2	-0.002184	0.035521	0.012881	-0.030291	0.022223	0.004101	0.015555
37	124	1	1	0	1	2	0	1	2	-0.002394	-0.005544	0.003366	0.001340	-0.005900	-0.011137	-0.002885
38	125	1	1	0	1	2	0	2	2	0.000930	-0.001771	0.001582	0.002152	-0.002671	0.000458	-0.002164
39	131	1	1	0	0	2	0	1	3	0.003272	-0.007238	-0.025665	0.001328	-0.002797	-0.020053	0.019385
40	134	1	1	0	1	2	0	2	.	.	.	.	.	.	.	.
41	135	1	1	0	1	2	0	2	.	.	.	.	.	.	.	.
42	142	1	1	1	1	1	1	2	1	-0.015493	0.022896	0.011892	-0.020449	0.010138	0.003974	-0.008564

The DfBeta statistics is what is considered in this research and it tells one how much each coefficient will change by removal of a single observation. The DfBeta statistics was used to fit model 2B. Model 2B was used because it includes all the significant variables in final model 1C.

Data set C is displayed in the output shown above for 42 cases; the signs of the DfBeta statistics are the reverse of what one might expect - a negative sign means that the coefficient increases when the observation is removed. Almost all the observations have small values for the influence statistics. For example, the estimated coefficients for the covariates of sex and alcohol in model 2B were respectively 0.60565 and 0.91737. However, in Data set C, the value 0.012141 for dsex indicates that if observation 2 is removed, the sex coefficient will decrease to approximately 0.60565-0.012141 = 0.593509, a decrease of 2%. Also for the value 0.0013132 for dalcohol indicates that if observation 2 is removed, the alcohol coefficient will decrease to approximately 0.91737-0.003132 = 0.914238, a decrease of 0.3%. Further, it can be seen in the set the value -0.015493 for dsex indicates that if observation 42 is removed, the sex coefficient will increase to approximately 0.60565+0.015493 = 0.621143, an increase of 1.5%. Overall, it can be seen in the Data set C that none of the observations did exert an undue influence on the estimated coefficients.

Based on the above discussion in trying to find observations that had leverage by using the DfBeta coefficient (estimates the influence of each observation on the resultant coefficients of model variables). It was realized that no deviant observations were found in any of the variables. That is to say that, largest changes were not found in the overall estimated coefficients and hence would not much affect the effects of the coefficients. This means that no observation had exerted an undue influence on the estimates of the parameters and hence the fit of the model.

In summary, the evaluation based on the residuals and the DfBeta indicated that the models structural supposition and proportionality, was acceptable. Elimination of deviant observations from the models will cause minor changes to the coefficients of the variables in the models. The practical conclusion is that removal of variables from the models will result in no or minor changes in the overall coefficients of all the covariates considered and hence have not distorted the models.

In conclusion, the sensitivity testing of the MTTU data models, suggested that it was not necessary to eliminate observations or alter the structure of the models. The calculated relative risk will still remain within the 95% confidence intervals.

However, the explanatory power of the models is not very high, for example consider a driver who was involved in accident on the 60th day (according to table 4.12) of model lC, was a male of 27

years old which according to the Schoenfeld residuals was 4.6 years younger than the model predicts. He had used safety belt, although the model predicts a probability of only 0.22 that a driver involved in an accident on the 60th day will be using safety belt (schbelt=0.88). This suggests a need to further improve the modeling methods.

4.4.4 Methodological Issues considered in developing the accident Models

The following issues to be considered in the process of developing the survival models for the data in this study.

Issue of Multi-Co-linearity

The variables in the models were correlated with each other to various degrees. The so called multi-co-linearity produced by correlations can cause problems in the estimation and interpretation of the models. The central problem of multi-co-linearity is that parameter estimates or variable coefficients cannot be estimated exactly, because the correlations increase the variable's standard errors. The variables' effects may be distorted and their coefficients may reflect the effects of other variables.

Some of the estimation problems observed in the study could be attributed, in part, to multi-co-linearity. Small changes in the set of observations (following DfBeta-tests for deviant observations) caused some insignificant changes in parameter estimates, although the model as a whole was statistical significant since none of the coefficients showed unreasonably large.

Missing Data

Missing data was a general problem. The amount of missing data varies from one variable to another but because of the small number of observations (particularly accident cases) it was not justified to remove all cases with some missing data. This resulted in different number of observations in each model. The models included all the observations that did not have missing data concerning the variables included in the model.

Specific causes of missing data were drivers who had died in the accidents and could not be captured in the registry. Some killed drivers were omitted from the MTTU data due to large missing information about them.

4.5 Solving the main questions of the research

The approach for solving the four main questions for this research was based on the following research methods and data sources.

1. Previous research results.
2. Analysis with Kaplan-Meier Method.
3. The compiled accident models

The four specified main question are;

- Can a drive involvement in accident be examined with survival models on the basis of exposure over time?
- Do age and Sex capture differences in accident risk at a general level, as background factors?
- Do vehicle characteristics contribute to accident risks?
- Do unworn out tyres reduces drivers' accident risk?

Hypothesis testing

Statistical hypothesis testing was used in defining solutions to the four questions of the study. The statistical decisions were based on the estimated parameters of the developed survival models. The H_0 is the null hypothesis and H_1 is the alternative or research hypothesis. The purpose of the testing is to decide whether the evidence tends to refute the null hypothesis. Since the research hypothesis is H_1, it is hoped that evidence leads to reject H_0 and thereby accept H_1.

When values of the test statistics, here models and their estimated parameters, are in the critical region specified by α, the chosen level of significance, we reject H_0. We have specified σ to be about 0.05, meaning that there is about 5% risk to reject H_0 when it is actually true (Type I error). It is possible to make another type of error (Type II error), to fail to reject H_0 when the alternative H_1 is true.

Generally, we test at the significance level $\alpha = 0.05$ as follows;

$H_0 : \beta_1 = \beta_2 = \ldots = \beta_n = 0$. e.g. all parameters of the models are 0, i.e. none of the variables is significant

$H_1 : \beta \neq 0$ for at least one i, $i = = 1, 2, \ldots..k$.

Therefore, the estimated hazard (accident) models for the research are model IC and model 2B is shown below;

The Cox PH model is usually written in terms of the hazard model formula shown below;

$h(t, X) = h_0(t) \times exp(X\beta)$

Where $h(x, X)$ is the hazard function, $h_0(t)$ is the base level of the hazard function and $exp(X\beta) = exp(\beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n)$· The β_1 is the coefficients of the variables x_1 (variables)

Model 1C:

$$h(t, X) = ho(t)e^{0.82sex - 2.18usebelt + 1.23speedveh + 0.37driverage + 1.26sexbelt}$$

Model2B:

$$h(t, X) = ho(t)e^{0.16sex - 0.8usebelt + 130speedveh + 0.34typcron + 0.39driverage - 0.44annhlkil}$$

In the next section we first presented a summary of driver's accident risk factors and relative risks based on the analysis and survival models discussed. Then hypothesis is set for each of the research questions and finally a summary is presented on either acceptance or rejection of the hypothesis based on the developed models.

4.5.1 Driver risk factors and relative risks

The probability of drivers to be involved in an accident was influenced by driver characteristics (age, gender, experience), their behaviour in traffic (speed, use of alcohol, use of safety belt), the nature of exposure (annual kilometreage, road surface condition) and by vehicle characteristics (vehicle age, weight, tyre condition). The use of safety belt, which was only supposed to influence the seriousness of injuries, proved to be also a strong risk factor.

The developed models estimated the coefficients of the variables which enabled calculation of relative risks associated with each factor. The time in the models referred to the number of days counted from the beginning of the period under consideration until the day of the accident. The models of the MTTU data gave many of the risks factors. Table 4.17 gives the interesting and most important factors.

In the MTTU accident data, survival models compare "primary parties" and "other involved parties". The obtained results do not necessarily describe the accident risk, the probability of being involved in a fatal accident, but the probability of being an accident primary involved party on the condition that a fatal accident has occurred.

Table 4.17: Interesting and important variables in the survival models of the MTTU data.

Variable	Remark
Driver Age	The relative risk was high for both the young and and the middle aged and lowest for the old
Driver Sex	Male drives had a higher relative risk than female drivers.
Use of alcohol	Significantly increased drivers' relative risk
Not using the safety belt	Significantly increased drivers' relative risk
Speed	Higher relative risk for drivers with speed over 80km/h
Familiarity of route	Drivers' familiar with route had a lower relative risk
Annual vehicle kilomreage	Relative risk decreased with increase in kilometerage
Vehicle Age	Users of new vehicles had somewhat lower risk, (though not statistically significant variable).
Vehicle weight	Users of light vehicles had a higher relative risk, (though not statistically significant variable).
Tyres tread depth	Small tread depth increases relative risk (though not statistically significant variable)
Age of license	Drivers with lessthan 5 years of license duration had higher relative risk.

4.5.2 Can driver involvement in accidents be examined with Survival Models on the basis of exposure over time?

The first main question was that can driver involvement in accident be examined on the basis of exposure over time using survival models and is amount and nature of exposure to traffic a major risk factor?

In the preliminary analysis (Kaplan-Meier estimates) we have seen that there are many factors having a statistically significant effect on the accident probability. Here, we first test if we can find any significant predictor variables by using the survival models.

In order to get the answer to the question concerning the amount of traffic exposure as a major risk factor, we set the following hypothesis to be tested at the significance level $\alpha = 0.05$;

From model 2B, we can write

$H_0 : \beta_7 = 0$

$H_1 : \beta_7 \neq 0$

In other words, it can be stated as:

H_0: The coefficients of driver annual kilometerage does not differ from zero in the model. i.e, the

variable does not significantly contribute to accident occurrence H_1: The coefficients of driver annual kilometerage differ from zero. i.e. The variable does not significantly contribute to accident occurrence.

The p-value for this variable in model 2B is 0.0084, which is less than the pre-assigned significant level of 0.05 indicating that the variable is statistically significant risk factor. Therefore, the hypothesis H_0 is rejected.

In the MTTU data, exposure was described by the annual vehicle kilomreage. According to the MTTU data models, the probability to be a primary party in a fatal accident decreased as the annual vehicle kilomerage increased, which is in contrary to what other researchers finding that the probability of accident increased with increasing kilometreage. This conflicting result might be due to the fact that, in the MTTU data, all drivers were already involved in accident and therefore, annual vehicle kilometerage represented driving experience. It distinguished between those with little experience who got involved as "primary parties" and the more experienced whose involvement was as "others". It is likely that more detailed and accurate information on actual traffic exposure might improve the models.

One can conclude that driver involvement in accident can be examined with survival models based on exposure over time. The amounts of travel are specific risk factors. With better exposure data more accurate models can be specified.

4.5.3 Are driver age and sex major risk factors?

The second main question of the study was that are driver age and sex major background risk factors for differences in accident probability. Several other factors are expressed through age and sex.

We test the nature of driver age and sex a major risk factor at the significant level $\sigma = 0.05$ as follows;

H_o: The coefficients of driver age do not differ from zero in the models 1C and 2B.

H_1: the coefficients of driver age differ from the zero in models 1C and 2B.

The p-value for this covariate in each of the models 1C and 2B are respectively 0.0178 and 0.0146, both values are less than the specified significance level. This indicates that the coefficients of driver age differ from zero in both models. Therefore, H0 is rejected; driver age is a statistically

significant risk factor.

We can also test the sex variable as follows;

H_0: the coefficients of driver sex do not differ from zero in the two models

H_1: the coefficients of driver sex differ from zero in the two models.

Similarly, the p-values for this covariate in each of the models lC and 2B are respectively 0.0148 and 0.0408, both values are less than the specified significance level. This indicates that the coefficients of driver sex differ from zero in both models. Therefore Ho is rejected; driver sex is a statistically significant risk factor.

4.5.4 Do vehicle characteristics contribute to accident risks?

The third main question of the study was that specific characteristics of vehicles can contribute to accident risk probabilities. In the MTTU data, vehicle age and weight were the main characteristics.

We therefore, test the nature of vehicle age and weight as major risk factors at the significance level $\alpha = 0.05$ as follows;

Ho: The coefficients of vehicle age do not differ from zero in the models.

H_1: The coefficients of vehicle age differ from zero in the models.

The p-value for this covariate in model 2A is 0.3735 which is greater than the specified significance level. This indicates that the coefficient of vehicle age do not differ from zero. We conclude that Ho is failed to be rejected; vehicle age is not a statistically significant accident risk factor.

Then for vehicle weight, we test it as follows;

H_0: The coefficient of vehicles weight do not differ from zero in the models.

H_1 : The coefficient of vehicles weight differ from zero in the models.

In the model 2A, the p-value for this covariate is 0.9039 which is greater than the specified significance level. This indicates that the coefficient of vehicles weight do not differ from zero. We conclude that H_0 is failed to be rejected; vehicles weight is not a statistically significant accident risk factor.

4.5.5 Do unworn out Tyres Reduces Drivers' Accident Risk?

The fourth and final main question of the study was that does the use of good tyres reduce drivers accident risk?

The p-values for this covariate in models 2A and 2B are respectively 0.0180 and 0.0646. It can be seen in model 2A that the tyres condition differs from zero, rejecting the H_0, indicating that worst tyre's depth is a statistically significant risk factor. This same variable did not have a large contribution to the occurrence of accident in model 2B after vehicle age and other variables were excluded in the same model. Nevertheless, tyre condition appears to be a better explanatory factor than vehicle age. Poor tyre condition is more likely, of course, with old vehicles.

CHAPTER 5

CONCLUSION AND SUGGESTIONS FOR FUTURE WORK

5.1 Conclusion

The following conclusions can be drawn from the analysis of the Northern regional MITU accident data;

In the models, it indicated that young and middle aged drivers are at the higher risk. This may reflect lack of experience (and perhaps riskier driving style) of young drivers, for the old drivers it can be attributed to reduced capabilities on their part. Traffic conditions set greater demands on all drivers; the young and very old may have a harder time meeting the greater demands.

It is usually difficult to separate the effects of driver age and driving experience due to strong correlation between them. The results suggest that as the driver age and driving experience increase, the risk of mild accident decreases, but the risk factors for the most serious accidents accumulate with both young and old drivers (Elvik and Vaa 1990). In the models, dealing with fatal accidents, experience was better represented by annual kilometerage and less linked to age.

Driver sex did not have a clear impact on accident probability, once exposure and experience were taken into account. There were significant differences between female and male drivers in age distribution, mobility, driving experience, the type of cars used, and in driving related behaviour. Females generally drove less than males and those females involved in accidents lacked driving experience.

Females however, had a higher relative risk across all age groups. Previous studies demonstrated that differences in risks between sexes could be explained by differences in mobility.

Also, female drivers drove fewer kilometers, were less often under the influence of alcohol, used

safety belt more often than male drivers. Female drivers had fewer accidents and committed fewer offences.

To sum up, the statistical testing and other variable information support the conclusion that driver age is a major risk factor. The role of sex as a risk factor is less conclusive, as in model 2A the effect of driver sex was significantly minimised when other factors correlated with it were considered in the model. These included amount of mobility, specific driving experience, access to different types of vehicles and certain behaviours. The sex of drivers may remain a practically useful explanatory variable.

Not use of safety belt, which was only supposed to influence the seriousness of injuries, proved to be also a strong risk factor.

The analysis of the MTTU data indicated that driving under the influence of alcohol doubled drivers' relative risk. According to other previous studies, driving under the influence of alcohol has a significant effect on drivers' accidents risk (Elvik and Vaa 1990). Drivers under the influence of alcohol are 3 times greater at risk than that of non-alcohol users. The use of alcohol is often associated with young drivers' lifestyle and other risk factor related to it.

According to the models, Speed proved to be a statistically significant variable in predicting the hazard of accidents. The hazard of accidents for drivers who drove over 80km/h is 3 times that of those who drove less than 80km/h.

Annual vehicle kilometerage was used to measure the exposure of drivers to traffic, it was however indicated in the models that for any one year increased in driver's exposure to traffic, it is associated with 35.3% decrease in expected time to accident (hazard). This result is indicating that drivers' accident risks decreases as annual kilometerage increases. This perhaps might be due to accumulated experience on the part of these drivers. However this might need further probing. Age of license of drivers was used as a proxy to assess the level of experience of the driver. Age of license was a moderate significant variable in model 2A with p =0.0635. However, the hazard ratio indicated that drivers with duration of license less than 5 years had 1.7 times the risk of those with license duration of at least 5 years.

Route familiarity of drivers was not a significant variable, it however, did indicate that the accident risks was lower for drivers who were familiar with the site of the accident compared to other

drivers.

In the models, vehicles age was not a clear risk factor, perhaps because of the strong correlations with alcohol and safety belt use. Users of old vehicles included many alcohol users, young drivers, and non-user of belt.

Vehicle weight was related inversely to risk, and so was vehicle age. However, these variables were strongly related to driver and exposure factors, such that the unique effect of these vehicle factors was rather small and it depended on the exact combinations of other factors. For example, although it was generally safer for large and older vehicles, the most risky combination was of young drivers using old vehicles. The Models based on the fatal accidents database included a strong influence of driver behaviour - use of alcohol, non-use of safety belt, and excessive speed all of which added to the probability of causing an accident.

It was also realised in the MTTU data that young drivers tended to drive older vehicles. More of drivers of newer vehicles were involved in speeding over 80km/h prior to the accident. Light weight vehicles were more likely to be primary party in a fatal accident. Vehicle characteristics were highly correlated with driver age and sex with vehicle kilometreage. The analysis of the third main question of the study pointed out that the features of a vehicle, it accessories, the manner of driving and the nature of exposure are all interrelated. However, some vehicle characteristics can be separated into their own individual risk factors but they seemed not to have strong explanatory power in these survival models.

In addition, tyre condition proved to be as important factor in determining the overall risk associated with tyres. Norwegian accident studied in the early 90's proved that the condition of tyres, especially their tread depth, is at least as important as having studs. A similar conclusion was derived here from the survival models of the MTTU data. The Norwegian study estimated that a 1mm decrease in the tyre's tread depth increased the accident probability by about 4%, which is totally consistent with the result obtained in the present study (3%-6%). In summary, according to all the estimated models, driving with much worn out tyres was more consistently risky than driving with unworn out tyres.

In conclusion, the application of survival models to the accident data appeared to be a promising approach. The models apply well to the examination of risk factors. These models can however

Be improved further. Improving the models will require better information on drivers' exposure times. Information on the date the driver got his driving license to his first involvement of accident can help improve the models.

5.2 Suggestions for Future Work

The accident survival models is still in progress and it is necessary to test it further and examine it more thoroughly in future work. The following suggestions are made on how this research can be developed further and strengthen the obtained results;

1. The major limitation of this study is that it focused on only one year driving, meanwhile, the driver might have been driving for so many years before the accident occurred. It is therefore strongly suggested that any further work on this should be focused on the date the drivers got their driving license to his first involvement of accident. Results obtained from such a data would be more reliable.
2. The new edition of accident survival models may seek more comprehensive data for a larger number of predictors. There are several covariates such as the date when vehicle was first taken into use, how long the driver had been on the trip when accident occurred, employment status of the driver, criminal records of the driver, driver's history of involvement of accident, medical condition and health of drivers, marital status and family situation and economic status of drivers. This information may play an essential role to the development of the models, but unfortunately these indicators are currently not available. In the future it is hoped to develop a set of covariates that will be widely useful, regularly updated, available and can cover future data needs.
3. The model developed in this study is only limited to Northern region and not the whole country of Ghana. Therefore in order to obtain a generalised model for the entire country, there is the need to have a comprehensive database of accidents in the country. This database should contain detailed information about the involved drivers, which should include all the variables considered in this study and those that have been proposed. The information in the database should be able to solve the shortcomings encountered in this study.

4. It is necessary to improve the accuracy and reduce the uncertainty of the survival model results by taking into consideration the above mentioned recommendations. There is also the need to investigate new applications and alternatives to the methods whether they can provide better results. This may include: better assessment and selection of variables and data. Extended Cox survival approach can be used to take care of time - varying covariates, since the values of the covariates of drivers may change over the follow- up period. Also, other parametric regression models such as log-normal, the exponential model, Weibull model and log-logistic can also be carried out to be able to make comparisons of the all model types and hence determine the best model.

Finally it can be said that the outcomes from this study are very encouraging. The Survival accident models is very promising and worth further applications to all regions of the country. This survival models developed for the northern region has the potential to become a major measure of road safety situation in the Northern region in particular and the country at large and hence enable stakeholders in road safety management to put up better measures to reduce the occurrence of accidents in the region. However, more extensive analysis and applications need to be conducted in the future before making any strong conclusions for now. This makes me conclude this thesis work which concentrates on the concept of "Survival accident model development" is a broad and complex concept. This requires deeper processes, integrated programs and much more cooperation between all the key national and international bodies. I will therefore suggest for more improvements of these models in future projects.

REFERENCES

Afukaar, F. K., Antwi, P., and Ofosu-Amaah, S. (2003). Pattern of Road Traffic Injuries in Ghana: Implications for Control. *Injury Control and Safety Promotion,* 10(12): 69-76.

Adams, J., (1987). *Smeed's law: some further thoughts.* Traffic Engineering and Control 10 (7), pp. 70-73. Bester, C. J., (2001). *Explaining national road fatalities.* Accident Analysis and Prevention. Vol. 33, pp. 663-672.

Broughton, J., (1988). *Predictive Models of Road Accident Fatalities.* Traffic Engineering and Control, May 1988, ISSN: 0041-0683, pp. 296-300.

Brown, S. and Bohnert, P. (1968) *Alcohol Safety Study: Drivers Who Die.* Waco: Baylor University College of Medicine.

Clark D. E. (2003). Effect of Population Density on Mortality after Motor Vehicle Collisions. *Accident Analysis and Prevention,* 35: 965-71.

Clarke, D., Ward, P., Bartle, C. and Truman, W. (2006). Young Driver Accidents in the UK: The Influence of Age, Experience, and Time of Day. *Accident Analysis and Prevention,* 38 (5):871-8. Dewer, R. E. (2002) Individual differences. Chapter 5 in Dewer, R. E. and Olson, P. L., *Human Factors in Traffic Safety,* Lawyers and Judges Publishing Co., Tucson, AZ, ISBN 0-913875-47-3. Elvik, R. (1996). A Meta analysis of Studies Concerning the Safety Effects of Daytime Running Lights on Cars. *Accident Analysis and Prevention,* 28: 68594.

Elvik, R. and Vaa, T. (1990). Human factors, road accident data and information technology. Oslo:

Institute of Transport Economics. pp 155

Eby, D. W., Kostyniuk, L. P., and Vivoda, J. M. (2003). Risky Driving: Relationship between Cellular Phone and Safety Belt Use. In Transportation Research Record: *Journal of the Transportation Research Board,* No. 1843, TRB, National Research Council, Washington, DC.

Elvik, R., Vaa, T., (2004). *The Handbook of Road Safety Measures.* Elsevier Amsterdam, ISBN: 0-08-044091-6, pp. 66 and pp. 676-803.

Evans, L., (1991). *Traffic Safety and the Driver,* New York: Van Nostrand Reinhold, P25-43, pp. 60-95.

Finch, J. R. and Smith, J. P. (1970). *Psychiatric and legal aspects of automo-bile fatalities.* Spring?eld, IL, Charles C. Thomas Publisher.

Fieldwick, R. and Brown R.J. (1987). *The effect of speed limits on road casualties.* Traffic Engineering and Control, Vol. 28, pp 635-640

Forjuoh, S. N. (2003). Traffic-related Injury Prevention Interventions for Low-Income Countries. *Injury Control and Safety Promotion*, 10.1-2: 109-18.

Garber N. J., Gadiraju, R., (1988). *Speed Variance and its Influence on Accidents.* Foundation for Traffic Safety, Washington, DC.

Hanowski, R. J., Wierwille, W.W., Garness, S. A., and Dingus, T. A. (2000). *Impact of Local/Short Haul Operations on Driver Fatigue.* Final Report No. DOT-MC-00-203. Washington, DC, U.S. Department of Transportation, Federal Motor Carriers Safety Administration.

Hijar, M., Arredondo, A., Carrillo, C., and Solorzano, L. (2004). Road Traffic Injuries in an Urban Area in Mexico: An Epidemiological and Cost Analysis. *Accident Analysis and Prevention*, 36: 37-42.

Hakim, S., Shefer, D., Hakkert, A. S., Hocherman, I. (1991). *A Critical Review of Macro Models for Road Accidents.* Accident Analysis and Prevention, Vol. 23, No. 5, pp. 379 400

Hakkert, A. S., and Braimaister, L., (2002). *The uses of exposure and risk in road safety studies.* SWOV Institute for Road Safety Research, the Netherlands.

Hakkanen H. and Summala, H. (1978). Fatal Traffic Accidents among Trailer Truck Drivers and Accident Causes as Viewed by Other Truck Drivers.

Jacobs, G. D., Fouracre, P. R., (1977). *Further research on road accident rate in developing countries.* TRRL report LR 270. Transport and Road Research Laboratory, Crowthome, Berkshire. Jacobs, G. D., Hutchinson, P., (1973). *A study of accident rates in developing countries.* TRRL report LR 546. Transport and Road Research Laboratory, Crowthome, Berkshire.

Koomstra, M., Lynam, D., Nilsson, G., Noordzij, P., Petterson, H-E., Wegman, F., Wouters, P., (2002). *SUNftower: a comparative study of the development of road safety in Sweden, the United Kingdom, and the Netherlands.* SWOV Institute for Road Safety Research, Leidschendam, the Netherlands, pp 67-127.

Koomstra, M. J., (1992). *The evolution of road safety and mobility.* IA.TSS (International Association of Traffic and Safety Sciences), Research, Vol.16, No.2, pp 129-148.

Kulmala, R. (1995). Safety at rural three-and four- arm junctions. Development and application of

accident prediction models. 42-104p.

Kulmala, R. and Peltola, H. (1985). Traffic Safety in the dark on Public roads in Finland. 14p to 51p.

Laapotti, S., Keskinen, E. and Rajalin, S. (2003). Comparison of Young Male and Female Drivers' Attitude and Self-reported Traffic Behaviour in Finland in 1978 and 2001. *Journal of Safety Research,* 34 (5):579.

Lancaster, R. and Ward, R(2002). *The Contribution of Individual Factors to Driving Behaviour: Implications for Managing Work-Related Road Safety.* Entec UK Limited, Health and Safety Executive, Research Report 020, United Kingdom.

Leaf, W. A., and Preusser, D. F, (1999). *Literature Review on Vehicle Travel Speeds and Pedestrian Injuries.* US National Highway, Traffic Safety Administration, *http://www.nhtsa.gov/people/injury/research/ pub/hs809012.html* (last February 15, 2010).

Lynam, D. and Twisk, D. (1995). Car driver training and licensing systems in Europe.

Mayou, R., and Bryant, B. (2003). Consequences of Road Traffic Accidents for Different Types of Road User. *Injury,* 34: 197-202

Meadows, M., and Stradling, S. (1999). Are Women Better Drivers Than Men? In J. Hartley, and A. Branthwaite (Eds.), The Applied Psychologist (2nd ed.). Buckingham: Open University Press. Mayhew, D. R. and Simpson, H. M. (2003) Graduated Driver Licensing: Safety Program Proves Effective in Reducing Crashes. TR News, No.

Michael S. *et al*, (2004), "Individual differences and the high risks of commercial driver"

McCartt A. T., Rohrbaugh J. W., Hammer M. C., Fuller S. Z. (2000) "Factors Associated with Falling Asleep at the Wheel among Long-Distance Truck Drivers." *Accident Analysis and Prevention.* 32(4), pp. 493-504.

Ghana National Road Safety Commission (2005) annual report.

Ghana National Road Safety Commission (2008) annual report.

National Highway Traffic Safety Administration. (2004). Safety Belt Use in 2003 Demographic Characteristics. National Center for Statistics and Analysis, May DOT HS 809 729.

NHTSA. (2007). Graduated Driver Licensing System. [Cited Dec. 31, 2009].

Available *from:http:/lwww.nhtsa.dot.gov/people/injuryffSFLaws!PDFs/810727W.pdf*

Reason, J(1990). Human Error. Cambridge University Press, Cambridge, UK.

Rimmo, P. A.(2002) "Aberrant Driving Behavior: Homogeneity of a Four-Factor Structure in Samples Differing in Age and Gender." *Ergonomics*, Vol. 45, No. 8, pp. 569-582

Sagberg, F.(1999) "Road Accidents Caused by Drivers' Falling Asleep." *Accident Analysis and Prevention,* 31, pp. 639-649.

Smeed, R.J ., (1949). *Some statistical aspects of road safety research.* J oumal of Royal Statistical Society Series A 112, pp. 1-34.

Singh, S.(March 2003) *Driver Attributes and Rear-End Crash Involvement Propensity.* NHTSA Report No. DOT HS 809 540.

Segui-Gomez, M. et al. (2007). Self-Reported Drinking and Driving Amongst Educated Adults in Spain: The Seguimiento Universidad de Navarra (SUN) Cohort Findings. *BMC Public Health,* Silvak, M., (1983). *Society's aggression level as a predictor of traffic fatality rate.* Journal of Safety Research 14, pp. 93-99.

Thouez, J.P. et al. (1991). Geographical Variations of Motor-Vehicle Injuries in Quebec, 1983- 1988. *Social Science and Medicine* 33.4 (1991): 415-21.

Van den Bossche, F. and Wets, G., (2003). *Macro Models in Traffic Safety and the DRAG Family:*

Literature Review. Report RA-2003-08, Diepenbeek, Belgium.

VALT. (1980 - 1999). Accident Statistics in Finland From the Years of 1980- 99. Helsinki, Finland: The Traffic Safety Committee of Insurance Companies

Wells-Parker, E., Popkin, C.L., and Ashley, M. (1996). Drinking and Driving Among Women:

Gender Trends, Gender Differences. In: Howard, J.M .. ; Martin, S.E.; Mail, P.O.; Hilton, M.E.; Taylor, E.D.(1996)., editors. Women and Alcohol: Issues for Prevention Research. Government Printing Office; Washington. NIAAA Research Monograph No. 3Z, NIH Publication No. 96- 3817.

WHO, World Health Organisation, (2004). World Report on Road Traffic Injury Prevention, Chapter 3 "Risk Factors", Geneva.

http://www. who.intlworld-healthday/2004/infomaterials/world_report/en/ (last visited November 10, 2009).

World Bank, (2003). *Traffic Fatalities and Economic Growth.* Policy Research Working Paper.

NHTSA. (2005). Alcohol Related Crashes and Fatalities. [Cited Sept. 18, 2009]. Available from: *http://www-nrd.nhtsa.dot.gov/Pubs/810616.PDF*

NHTSA. (2008). Traffic Safety Facts: Rural/Urban Comparison. [Cited Jan. 11, 2010]. Available from: *http://www-nrd.nhtsa.dot.gov/Pubs/810812.PDF*

NHTSA(2000). *Traffic Safety Facts:* Older Population. DOT HS 809 328. National Highway Traffic Safety Administration, Washington, DC.

NHTSA. (2008). Bicycle Helmet Use Laws. [Cited Jan. 11, 2010].Available from: *http://www.nhtsa.dot.gov/portallnhtsa_static_file_downloader.jsp?file=lstaticfiles IDOTINHTSAI Communication%20and%20Consumer%20Information/Articles/Associated%20Files/810886.pdf* Nyberg, A., and Gregersen, N. P. (2007). Practicing for and Performance on Drivers License Tests in Relation to Gender Differences in Crash Involvement Among Novice Drivers. *Journal of Safety Research,* 38 (1): 71-80.

Noland, R. B., (2003). *Medical Treatment and Traffic Fatality Reductions in Industrialised Countries. Accident Analysis and Prevention,* 35 (6), pp. 877-883.

Oppe, S., (1989). *Macroscopic models for traffic and traffic safety.* Accident Analysis and Prevention, Vol. 21, pp. 225-232.

Over, M., Ellis, P., Huber, J., and Solon, O. (1992). The Consequences Of Adult Ill Health. In: Feachem RGA, Kjellstrom T, Murray CJL, Over M, Phillips M (eds.). The Health of Adults in the Developing World. New York: Oxford University Press, 161-207.

Paul, D. Allison, (2008). Survival Analysis using SAS, A practical guide.12th edition.

Proctor, S., M. Belcher, P. Cook., (2001). *Practical Road. Safety Auditing,* Thomas Telford Publishing, London, United Kingdom.

Page,Y., (2001). *A statistical model to compare road mortality in OECD countries.* Accident Analysis and Prevention, 33, pp. 371-385.

Peden, M., Scurfield, R., and Sleet, D., *et al*. (2004). World Report on Road Traffic Injury Prevention. Geneva: World Health Organization, 2004. [Cited September 18, 2009]. Available *from:http://www.who.int/world-health day/2004/infomaterials/world_report/en/summary _en_rev.pdf* Rivara, F.P., Koepsell, T. D., Grossman, D.C., and Mock, C. (2000). Effectiveness of Automatic Shoulder Belt Systems in Motor Vehicle Crashes. *JAMA*, 283: 28262828.

Series 3035.

Williams, A. F. (2003). Teenage Drivers: Patterns of Risk. Journal of Safety Research, 34(1): 5, 15.

Zewerling,C., *et. al*. (2005). Fatal Motor Vehicle Crashes in Rural and Urban Areas:

APPENDIX A

FATALITY FORM OF THE UNIT

The following information is what is required of any accident involved driver;

1. Sex of driver
2. Age of driver————
3. Name of driver————
4. Date of accident————
5. Use of alcohol when accident occurred. a) Yes b) No
6. Use of safety belt when accident occurred. a) Yes b) No
7. Vehicle ownership————
8. Type of vehicle————
9. Age of vehicle ————
10. Weight of vehicle————
11. Vehicles tyres condition

    a) New or good as new b) Somewhat worn d) Quite worn e) Very worn
12. Estimated Speed of vehicle when accident occurred
13. Posted speed limit at the site of accident
14. Estimated annual vehicle kilometers driven————
15. Consequences of the accident

    a) Number of minor injuries————b) Number of serious injuries ———— c) Number of fatalities ————
16. Road surface condition when accident occurred

17. Weather condition when accident occurred
18. Traffic lighting condition when accident occurred
19. Drivers familiarity of route. a) Pass scene of accident at least once a month. b) more seldom pass than once in a month
20. Type of driving license————
21. Age of driving license————
22. Scene of accident. a) Junction b) link c) other
23. cause of accident
    a) Excessive speeding
    b) Inattention, confusion of lack of judgment of driver
    c) Drivers careless at road junction and cutting corners
    d) Improperly overtaking or cutting in
    e) Inexperience of driver
    f) Intoxication
    g) Other recklessness or negligence by drivers
    h) Over loading
    i) Mechanical defects
    j) Defective lights
    k) Dazzling lights
    l) Skid and road surface defects
    m) Other road defects
    n) Obstructions
    o) Level erosions
    p) Children
    q) Adults crossing road carelessly
    r) Adults boarding or alighting from vehicles
    s) Other pedestrian faults
    t) Passengers faults
    u) Animals not under control

v) Recklessness or negligence caused by peddle cyclist

w) Recklessness or negligence caused by drivers of horse - driven vehicles

APPENDIX B

SAS CODES FOR ALL THE FIGURES AND OUTPUTS

B.1 DEMONSTRATING PROC LIFETEST TO OBTAIN KAPLAN-MEIER AND LIFE TABLE SURVIVAL ESTIMATES AND PLOTS

PROC LIFETEST produces Kaplan-Meier survival estimates with the METHOD=KM option. The PLOTS=(S) option plots the estimated survival function. The TIME statement defines the time-to-event variable (Dur) and the value for censorship (STATUS = 0). The STRATA statement is used to compare survival estimates for different groups (e.g., strata speedveh). The STRATA statement also provides the log rank test and Wilcoxon test statistics. The SYMBOL statements are optional, it is to distinguish the two survival curves by colour.

The codes are as follows;

For figure 4.1

```
proc lifetest data=sasuser.thesisdata14 plots=(s) graphics;
   time Dur*status(0);
   strata driver_agp;
   symbol1 v=none color=black line=1;
   symbol2 v=none color=red line=2;
```

```
   symbol3 v=none color=blue line=3;
run;
```

For figure 4.2

```
proc lifetest data=sasuser.thesisdata14 plots=(s) graphics;
   time Dur*status(0);
   strata sex;
   symbol1 v=none color=black line=1;
   symbol2 v=none color=red line=2;
run;
```

For Figure 4.3

```
proc lifetest data=sasuser.thesisdata14 plots=(s) graphics;
   time Dur*status(0);
   strata alcohol;
   symbol1 v=none color=black line=1;
   symbol2 v=none color=red line=2;
run;
```

Figure 4.4

```
proc lifetest data=sasuser.thesisdata14 plots=(s) graphics;
   time Dur*status(0);
   strata usebelt;
   symbol1 v=none color=black line=1;
   symbol2 v=none color=red line=2;
run;
```

For Figure 4.5

```
proc lifetest data=sasuser.thesisdata14 plots=(s) graphics;
   time Dur*status(0);
   strata annukil;
   symbol1 v=none color=black line=1;
   symbol2 v=none color=red line=2;
   symbol3 v=none color=blue line=3;
run;
```

For figure 4.6

```
proc lifetest data=sasuser.thesisdata14 plots=(s) graphics;
   time Dur*status(0);
   strata scene;
   symbol1 v=none color=black line=1;
   symbol2 v=none color=red line=2;
run;
```

For figure 4.7

```
proc lifetest data=sasuser.thesisdata14 plots=(s) graphics;
   time Dur*status(0);
   strata speedveh;
   symbol1 v=none color=black line=1;
   symbol2 v=none color=red line=2;
run;
```

For figure 4.8

```
proc lifetest data=sasuser.thesisdata14 plots=(s) graphics;
   time Dur*status(0);
```

```
    strata wghtveh;
    symbol1 v=none color=black line=1;
    symbol2 v=none color=red line=2;
run;
```

For figure 4.9

```
proc lifetest data=sasuser.thesisdata14 plots=(s) graphics;
    time Dur*status(0);
    strata tyrescon;
    symbol1 v=none color=black line=1;
    symbol2 v=none color=red line=2;
run;
```

For figure 4.10

```
proc lifetest data=sasuser.thesisdata14 plots=(s) graphics;
    time Dur*status(0);
    strata agelic;
    symbol1 v=none color=black line=1;
    symbol2 v=none color=red line=2;
run;
```

For figure 4.11

```
proc lifetest data=sasuser.thesisdata14 plots=(s) graphics;
    time Dur*status(0);
    strata rutfam;
    symbol1 v=none color=black line=1;
    symbol2 v=none color=red line=2;
```

```
    symbol1 v=none color=black line=1;
    symbol2 v=none color=red line=2;
run;
```

For Figure 4.13

```
proc lifetest data=sasuser.thesisdata14 plots=(s) graphics;
    time Dur*status(0);
    strata owner;
    symbol1 v=none color=black line=1;
    symbol2 v=none color=red line=2;
run;
```

For figure 4.14

```
proc lifetest data=sasuser.thesisdata14 plots=(s) graphics;
    time Dur*status(0);
    strata ageveh;
    symbol1 v=none color=black line=1;
    symbol2 v=none color=red line=2;
run;
```


B.2 RUNNING A COX PROPORTIONAL HAZARD MODEL WITH PROC PHREG

PROC PHREG is used to request a Cox proportional hazards model. The statement Dur*status(0) in the MODEL statement specifies the time-to-event variable (Dur) and the value for censorship (STATUS = 0). The predictors are then included in the model. The option RL in the MODEL statement provides 95% confidence intervals for the hazard ratio estimates. When we wish to assess interaction effects between variables, we define two interaction terms in a new temporary SAS dataset and then run a model containing those terms. The codes are as follows;

```
Output 1
proc phreg data=sasuser.thesisdata14;
model Dur*status(0)=sex usebelt alcohol ageveh agelic speedveh...
 driver_agp2 driver_agp3/rl;
driver_agp2=(driver_agp=2);
driver_agp3=(driver_agp=3);
driver_agp:test driver_agp2,driver_agp3;
run;

Output 2
proc phreg data=sasuser.thesisdata14;
model Dur*status(0)=sex usebelt alcohol speedveh driver_agp /rl;
run;

Output 3
proc phreg data=sasuser.thesisdata14;
```

```
model Dur*status(0)=sex usebelt alcohol speedveh agedriver sexbelt/rl;
sexbelt=sex*usebelt;
run;

Output 4
proc phreg data=sasuser.thesisdata14;
model Dur*status(0)=sex usebelt alcohol ageveh agelic speedveh tyrescon..
 wghtveh rutfam driver_agp2 driver_agp3 annukil2   annukil3/rl;
driver_agp2=(driver_agp=2);
driver_agp3=(driver_agp=3);
driver_agp:test driver_agp2,driver_agp3;
annukil2=(annukil=2);
annukil3=(annukil=3);
annukil:test annukil2,annukil3;
run;

Output 5
proc phreg data=sasuser.thesisdata14;
model Dur*status(0)=sex usebelt alcohol speedveh tyrescon...
 driver_agp annukil/rl;
run;

Output 6
proc phreg data=sasuser.thesisdata14;
```


```
model Dur*status(0)=sex usebelt alcohol speedveh driver_agp sext usebeltt
 alcoholt speedveht driver_agpt;
sext=sex*log(Dur);
usebeltt=usebelt*log(Dur);
alcoholt=alcohol*log(Dur);
speedveht=speedveh*log(Dur);
driver_agpt=driver_agp*log(Dur);
test_proportionality:test sext,usebeltt,alcoholt,speedveht,driver_agpt;
run;
```

Output 7

```
proc phreg data=sasuser.thesisdata14;
model Dur*status(0)=sex usebelt alcohol speedveh driver_agp sexbelt sext.
 usebeltt alcoholt speedveht driver_agpt;
sexbelt=sex*usebelt;
sext=sex*log(Dur);
usebeltt=usebelt*log(Dur);
alcoholt=alcohol*log(Dur);
speedveht=speedveh*log(Dur);
driver_agpt=driver_agp*log(Dur);
test_proportionality:test sext,usebeltt,alcoholt,speedveht,driver_agpt;
run;
```

Output 8

```
proc phreg data=sasuser.thesisdata14;
model Dur*status(0)=sex usebelt alcohol speedveh tyrescon driver_agp..
annukil  sext usebeltt alcoholt speedveht tyrescont driver_agpt annukilt;
sext=sex*log(Dur);
usebeltt=usebelt*log(Dur);
alcoholt=alcohol*log(Dur);
speedveht=speedveh*log(Dur);
tyrescont=tyrescon*log(Dur);
driver_agpt=driver_agp*log(Dur);
annukilt=annukil*log(Dur);
test_proportionality:test sext,usebeltt,alcoholt,speedveht,tyrescont,...
driver_agpt,annukilt;
run;
```

For figure 4.15

```
proc phreg data=sasuser.thesisdata14;
model Dur*status(0)=sex usebelt alcohol speedveh driver_agp sexbelt;
sexbelt=sex*usebelt;
output out=c resdev=dev;
run;

proc gplot data=c;
symbol1 value=dot h=.2;
plot dev*Dur;
run;
```


```
Output 9
proc phreg data=sasuser.thesisdata14;
model Dur*status(0)=sex usebelt alcohol speedveh agedriver/rl;
output out=b ressch=schsex schusebelt schalcohol schspeedveh schagedriver
run;

proc print data=b;
run;

Output 10
proc phreg data=sasuser.thesisdata14;
model Dur*status(0)=annukil alcohol usebelt tyrescon sex...
 agedriver/ties=efron;
output out=b ressch=schannukil schalcohol schusebelt schtyrescon...
 schsex schagedriver;
run;

proc print data=b;
run;
```

For figures 4.17

```
proc gplot data=b;
plot schsex*Dur schusebelt*Dur schalcohol*Dur schspeedveh*Dur...
 schagedriver*Dur;
```

```
symbol1 value=dot h=.2;

run;
```

B.3 ASSESSING THE PH ASSUMPTION WITH A STATISTICAL TEST

This is accomplished by finding the correlation between the Schoenfeld residuals for a particular covariate and the ranking of individual failure times. The p-value for testing this correlation can be obtained from PROC CORR. The Schoenfeld residuals for a given model can be saved in a dataset using PROC PHREG. The ranking of events by failure time is saved in a dataset using PROC RANKED. First the full model is runned. The output statement creates a SAS dataset, the OUT=option defines an output dataset, and the RESSCH=statement is followed by the variables names so that the output dataset contains the Schoenfeld residuals of these variables. The order of the names corresponds to the order of the independent variables in the model statement. The actual variable names are arbitrary which should contain the Schoenfeld residuals for all the variables.

Next, create a dataset that deletes censored observations (i.e., only contains observations that fail). PROC RANK is used to create a dataset containing a variable that ranks the order of failure times. The variable to be ranked is Dur (the survival time variable). The RANKS statement precedes a defined variable name for the rankings of failure times called TIMERANK. PROC CORR is used to get the correlations between the ranked failure time variable (called TIMERANK) and the variables containing the Schoenfeld residuals. The NOSIMPLE option suppresses the printing of summary statistics. If the proportional hazard assumption is met for a particular covariate, then the correlation should be near zero. The p-value obtained from PROC CORR which tests whether this correlation is zero is the p-value for testing the proportional hazard assumption. The code follows;

```
Output 11

proc phreg data=sasuser.thesisdata14;

model Dur*status(0)=sex usebelt alcohol speedveh agedriver sexbelt;

sexbelt=sex*usebelt;
```

```
output out=b ressch=schsex schusebelt schalcohol schspeedveh...
 schagedriver schsexbelt;
run;

proc print data=b;
run;

data state;
set b;

if status=1;
run;

proc rank data=state out=ranked ties=mean;
var Dur;
ranks timerank;
run;

proc print data=ranked;
run;

proc corr data=ranked nosimple;
var schsex schusebelt schalcohol schspeedveh schagedriver schsexbelt;
with timerank;
```

```
run;

Output 12
proc corr data=ranked nosimple;
var schsex schusebelt schalcohol schspeedveh schagedriver...
 schannukil schtyrescon;
with timerank;
run;

Output 13
proc phreg data=sasuser.thesisdata14;
model Dur*status(0)=sex usebelt alcohol speedveh tyrescon...
 driver_agp annukil;
output out=c dfbeta=dsex dusebelt dalcohol dspeedveh dtyrescon...
 ddriver_agp dannukil;
run;
proc print data=c;
run;
```


APPENDIX C

Table C.1: Detailed description of the Kaplan-Meier Estimates of variables captured in the accident data

Variable	Categories	Total	Uncensored	p-value
Driver age	≤ 25 years	93	28	<0.0001
	26 Ű 50 years	272	179	
	50 +	33	14	
Driver sex	Male	377	203	0.0049
	Female	21	18	
Use of alcohol	No	160	50	<0.0001
	Yes	230	168	
Use of safety belt	No	233	162	<0.0001
	Yes	109	29	
Vehicle annual kilometereage	< 5000km/a	88	53	0.4016
	5000 Ű 14,000km/a	237	128	
	≥ 15000km/a	64	33	
Road section/scene of accident	Link	304	159	0.0807
	Junction	72	47	
Speed when accident occurred	≤ 80km/h	143	37	<0.0001
	>80km/h	241	178	
Vehicle weight	≤ 1,000kg	212	115	0.4483
	>1,000kg	174	99	
Tyres tread depth	≤ 4mm	278	152	0.3897
	>4mm	111	68	
Age of driving license	≥ 5years	295	180	<0.0001
	<5years	80	30	
Route familiarity	Seldom pass scene	61	33	0.8079
	frequent	292	163	
Road surface condition	Dry	324	177	0.547
	Wet	21	10	
Vehicle ownership	Own	297	154	0.128
	Not own	72	45	
Vehicle age	≤ 10 years	142	58	<0.0001
	>10 years	248	160	