FACTORS AFFECTING THE SEVERITY OF ROAD ACCIDENTS IN SRI LANKA: A LOGISTIC REGRESSION APPROACH

Navarathna Arrachchilage Manoja Rasanjali Senaviratna

158384M

Degree of Master of Science in Operations Research

Department of Mathematics

University of Moratuwa
Sri Lanka

November 2018

FACTORS AFFECTING THE SEVERITY OF ROAD ACCIDENTS IN SRI LANKA: A LOGISTIC REGRESSION APPROACH

Navarathna Arrachchilage Manoja Rasanjali Senaviratna

158384M

Thesis submitted in partial fulfillment of the requirements for the degree Master of Science

Department of Mathematics

University of Moratuwa
Sri Lanka

November 2018

Declaration of the candidate and the supervisor

I declare that this is my own work and this thesis does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my thesis, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: Date: 09.11.2018

The above candidate has carried out research for the Masters thesis under my supervision.

Signature of the supervisor: Date:

Mr.T.M.J.A.Cooray
Senior Lecturer
Department of Mathematics
University of Moratuwa

Abstract

Road accidents have become a leading cause of death and injury as well as property damage worldwide. Ever increasing road accidents and traffic flow is a heavy burden to a developing country like Sri Lanka. In year 2016, 38915 accidents were reported where 7% of them are fatal contributing to 2824 deaths. Therefore, it is urgently needed to find solutions and reduce road accident deaths and injuries. The objective of this study is to identify the significant factors affecting for motorcycle and motor vehicle accidents in Sri Lanka. Secondary data used in this study between the period 2014 to 2016 were acquired from the police traffic headquarters, Colombo in Sri Lanka. A total number of 111457 road accidents where drivers at fault were included in the analysis. Among them 78531were motor vehicle accidents and 32926 were motor cycle accidents. Motorcycle accidents are analyzed separately due to high accident rate of motorcycles.

Factors considered in the study were vehicle type, gender of driver, validity of license, accident cause, alcohol test, time of accident, weekday/weekend, road surface, weather condition, light condition, location and age of driver. Results revealed that male drivers (98%) have greater tendency to be involved in motorcycle and motor vehicle accidents rather than female drivers (2%). High number of motorcycle (75%) and motor vehicle (73%) accidents reported due to aggressive /negligent driving. Highest number of motor vehicle accidents (20.5%) reported by the drivers in between 29 - 34 years old. Highest number of motorcycle accidents (28.5%) reported by the drivers in between 19-24 years old. Majority of the accidents were occurred, while the vehicle was moving on a straight road. Among drivers and motorcyclists (7%) were found to have consumed alcohol. Most of motorcycle and motor vehicle accidents occurred in daytime under daylight on weekdays.

Binary logistic regression is applied motorcycle and motor vehicles accidents separately to evaluate the odds of grievous accidents compared to non-grievous accidents. For motor vehicle accidents vehicle type, validity of license, time, location, alcohol test, accident cause, age of driver and gender have a significant effect on the severity of accidents. Bend or junction location, aggressive/negligent driving, drive by male drivers, drive at daytime, driving light vehicle and drivers who use alcohol below legal limit or no alcohol, have a high chance to be a grievous accident. Moreover, the older drivers have less accident risk. For motorcycle accidents, location type, time, age of driver, accident cause and gender have a significant effect on the severity of accidents. Among them, location type, accident cause and gender have an increasing effect on the probability of a grievous accident. Time and age of driver have a decreasing effect on the probability of a grievous accident. Straight road, aggressive/negligent driving, drive by male motorcyclists, daytime have a high chance to be a grievous accident. Moreover, the older motorcyclists have less accident risk. These findings can aid modifying regulations and laws and establishing preventive and protective approaches and strategies.

Keywords: Road accidents, Logistic Regression, Accident severity, Motorcycle accidents, Motor vehicle accidents

Acknowledgement

I would like to acknowledge the efforts, support, guidance, cooperation and encouragement of numerous people who have made it possible for me to undertake this study.

I wish to express my sincere gratitude to my supervisor, Mr.T.M.J.A.Cooray, for his guidance, encouragement and support in shaping the outlook of this thesis. He provided invaluable insights that have guided my thinking and understanding.

I also would like to express gratitude to the Senior DIG of Police, Police Head Quarters for his permission to supply the data for this study. Besides all the police officers who support to get the data that needed for this research are also greatly appreciated.

I would like to gratefully acknowledge the staff of Department of Mathematics, The Open University of Sri Lanka for their support given to me.

Last but not least I would like to thank my parents for their endless support throughout my journey. Without them I wouldn’t be able to accomplish any of the achievements that I have attained so far.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Page

LIST OF ABBREVIATION

Abbreviation	Description
B	Coefficient of logistic regression model
BJ	Bend/Junction
BL	No alcohol/below legal limit
Cause1	Speeding
Cause2	Aggressive/ negligent driving
Cause3	Influenced by alcohol/drugs
Cause4	Fatigue/fall asleep
Cause5	Others
CI	Confidence Interval
CL	Clear
D	Dry
DD	Dusk/Dawn
Df	Degrees of freedom
DL	Daylight
DT	Day Time
DW	Durbin Watson Test
F	Female
GSL	Night, Good street lighting
HV	Heavy Vehicle
LM	Lagrange Multiplier Test
LV	Light Vehicle
M	Male
ML	Maximum Likelihood
NSL	Night, no street lighting

NT	Night Time
OL	Over Legal Limit
OR	Odds Ratio
Others	OT
RA	Rainy
RD	Road
RTA	Road Traffic Accidents
S.E	Standard error
Sig.	Significant
W	Wet
WD	Weekday
WE	Weekend
WL	With Valid License
WO	Others
WOL	Without Valid License

1. INTRODUCTION

1.1 Overview

This chapter presents the background of the study, statement of the problem, objectives and significance of the study. The organization of the report is presented at the end of the chapter.

1.2 Background of the study

Road accidents are a leading cause for many of deaths around the world. World Health Organization (WHO) has found more than 1.2 million people die each year on the world's roads and most of these deaths are in low and middle-income countries. WHO indicated road traffic injuries are currently estimated to be the 9th leading cause of death across all age groups globally and predicted to become the 7th leading cause of death by 2030. Road accidents are highly influenced to the public health in a country. Furthermore, increasing road accidents evolve social and economic problems due to loss of lives and damage properties.

Ever increasing road accidents and traffic flow is a heavy burden to a developing country like Sri Lanka. The rate of increase in road accidents is 7% per year in Sri Lanka. Increasing in vehicle population is 11% per year. The analysis of past accident data has clearly shown that in Sri Lanka about 50,000 accidents occur annually on average out of which 2000 were fatal accidents and 15,000 were injury accidents. Traffic Police reveals that a Sri Lankan is killed in a road accident every three and half hours and two are critically injured. This is a heavy economic burden to the country. Following Figure 1.1 illustrates that how number of different types of accidents fluctuates during the period 1977-2016.

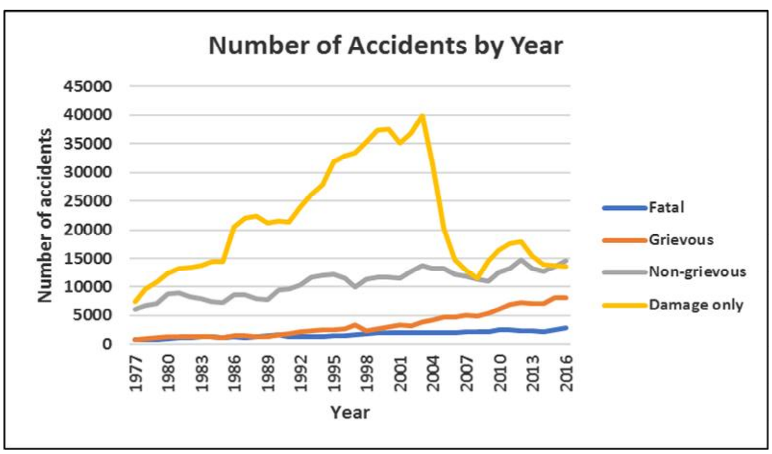

Figure 1.1: Number of accidents by year

It clearly shows that number of fatal, grievous and non-grievous accidents increased during the period. However, number of damage only accidents increased dramatically from 1977 to 2004 and then decreased from 2005 to 2008.

Moreover, Traffic Police in Sri Lanka said that a notable increase has been witnessed in motorcycle accidents during the last 5 years. The daily increase in motorcycle accidents has become a serious issue. According to Police reports, the number of motorcycle which are operative in the island so far has exceeded 2.9 million. It is around 53 percent of the entire number of vehicles operating on roads.
Therefore, researches should be undertaken, because road accidents have negative impacts on social and economical improvements in developing countries like Sri Lanka.

1.3 Statement of the problem

Road accidents are a serious public health problem and one of the leading causes of the death and injuries around the globe with ever rising trend. The magnitude of the problem of road traffic injuries in Sri Lanka significantly increased in the last decade. However, only few studies have done in the past related to road accidents in Sri Lanka. Among them most of studies were only descriptive analysis. Moreover, even though accident rate of motorcycles is high in Sri Lanka, thorough analysis related to motorcycle accidents haven't done so far. Therefore, this study statistically explored the significant factors influencing motor vehicles and motorcycle accidents separately that are occurring in Sri Lanka and attempt to fill the gaps by proposing solutions to the problem.

1.4 Objectives

The objectives of this study are as follows.

- Identifying the significant factors affecting motor vehicle accidents and motorcycle accidents in Sri Lanka.
- Estimating the effect of the statistically significant factors on accident severity.

1.5 Significance of the study

Road accidents can be caused due to different factors namely road characteristics, environmental characteristics, vehicle characteristics, human characteristics, etc. Therefore, identifying these factors is very important to a country. It enables police and policy makers to modify regulations and laws and establishing preventive and protective approaches and strategies. Recommendations can be utilized by public on prevention of road accidents.

1.6 Source of Data

Secondary data used in this study were acquired from the Traffic Police headquarters, Colombo in Sri Lanka. Data is received from 2014 to 2016 time period and those data has collected by the police officers. They have reported the related data according to the questionnaire which was prepared by Traffic Police headquarters.

1.7 Limitations of the study

In this study, data considered only for accidents occurred by drivers at fault. It is not considered accidents occurred by pedestrian fault.

1.8 Outline of the report

This report consists of six chapters. Chapter 01 discusses overview of the study. It consists of background of the study, statement of the problem, objectives and significance of the study. Chapter 02 describes the researches done in the past related to this study in Sri Lanka as well as in other countries. Chapter 03 reviews the methodology used in the study. It also describes the data set and variables used in the study. Chapter 04 presents the results of the analysis of motor vehicles accidents. It explains the most influential factors affecting for the accident severity. Chapter 05 describes findings of the analysis of motor cycle accidents. It presents the risk factors associated to motor cycle accidents in Sri Lanka. Finally, chapter 06 contains the final discussion, conclusions and recommendations.

2. LITERATURE REVIEW

2.1 Overview

This chapter presents a detailed literature review related on factors affecting motor vehicles and motorcycle accidents. This literature review investigated for available published research notes and journal papers on this study. During the literature review, also a task of finding available methodologies for analyzing influential factors for road accidents was undertaken.

2.2 Previous Studies related to Motor Vehicle Accidents in Sri Lanka

This section describes previous studies carried out related to motor vehicle accidents in Sri Lanka. This section categorized into 2 sections namely descriptive analysis and risk factors influencing motor vehicle accidents.

2.2.1 Descriptive Analysis

Somasundaraswaran, 2006 analyzed accident data in Sri Lanka during 1989-2005. The results of this study revealed that the main reason for the rapid increase of traffic accidents is due to the alarming rate of vehicle ownership together with inadequate road network development to support the demand.

Kumarage, et al., 2003 studied the relationship among the various causes of accidents in Sri Lanka by using the accident statistics of the year 1997. They found that speed related accidents be the most contributory of fatal accidents. Vehicle defects, driving on the wrong side and aggressive driving were identified as most important factors for fatal accidents.

Jeepara & Pirasath, 2011 studied etiological factors, type of injuries and treatments of road traffic accidents (RTA) in Eastern Sri Lanka. They found that most victims of RTAs were young, male adults in urban areas. The commonest victims are motorbike riders. The personal characteristics of victims are important contributors to an RTA. Head and limb injuries are the commonest injuries.

2.2.2 Risk Factors Influencing Motor Vehicle Accidents

Renuraj, et al., 2015 conducted a research on factors influencing traffic accidents in Jaffna. In this study, they used 692 accident cases for the analysis based on Jaffna police records during the period 2010 -2013. They have used logistic regression approach for the analysis. Results from this study reveal that the fitted logistic regression model can be used for the safety improvements against the traffic accidents in Jaffna. The conclusion of this research expressed that independent variables “Type of vehicle” and “Age” were identified as more influential variables affecting the accident severity.

Dhananjaya & Alibuhtto, 2016 conducted a study to identify the factors that mainly contribute to accident severity in Sri Lanka during the period 2010-2014. Binomial Logistic Regression was used to analyze the data. Based on the results, they concluded that variables such as light condition, age of the driver, the validity of the license, urban / rural, weather, vehicle type and age of the vehicle have a decreasing effect on the probability of a fatal accidents. Location type, alcohol test and accident cause have an increasing effect on the probability of a fatal accident. Among them, Accident Cause is the most important variable in the model.

Liyanage & Rengarasu, 2015 conducted a study to develop traffic accident prediction models based on traffic police report data within the Galle police division for years 2011, 2012 and 2013. These models relate accident numbers, as a dependent variable, with possible causes of accidents that are related to accident occurrence such as: time, type of the day, road geometry, light condition, year of driver license provision and vehicle type as independent categorical variables. Count data models Poisson and Negative-Binomial models along with a non-parametric decision-tree model were applied. Out of those models, Negative Binomial regression model demonstrated a better fit than the Poisson model. Further, decision tree can also be used to model traffic accident frequencies. Considering the results of Negative Binomial model, the key variables which cause the occurrence of accidents are experience of the driver (year of driver license issue), vehicle type, light condition and time of the accident. Moreover, decision tree results show that, road geometry with straight roads contributing to the highest number of accidents.

Senasinghe, Wirasinghe, & De Barros, 2017 carried out a study of accidents occurred on two major intercity highways (A001 and A004) in Sri Lanka. The study was conducted using 2010 to 2012 records of accidents that occurred within an approximately 50-km length on the each A001(Peliyagoda to Ambepussa) and A004 (Vilasitha Niwasa to Avissawella) highway segments. Negative Binomial (NB) regression model was used to predict the frequency of crashes of a specific severity level, as a function of explanatory variables. It is found that age, gender, protection, light conditions, urbanicity, traffic control, and mode of transport, have a significant and direct impact on the severity level of the accidents that occurred on both highway segments.

It is important to note that any researcher haven't study accident data related to motor cycles separately in Sri Lanka even though accident rate of motor cycles is high in Sri Lanka.

2.3 Previous Studies related to Motor Vehicles Accidents in Other Countries

Many scholars have investigated the factors influencing the road accidents in worldwide using different methods. This section presents studies related to descriptive and risk factors influencing motor vehicle accidents.

2.3.1 Descriptive Analysis

Recently, Baruah and Chaliha (2015) have analyzed the incidence of alcohol consumption among the victims of road traffic incidents brought for autopsy to the Department of Forensic Medicine, Guwahati. They found that males are most exclusively involved in accidents following alcohol consumption with 20-29 age group most affected. Majority of the victims were lowly educated and were pedestrians or riders of two wheelers.

Singh et al. (2014) have done a study on road traffic fatalities in adults of North West India. This study was based on the autopsy records of unnatural deaths occurred in a leading tertiary health care center of North West India. The adult road traffic fatalities constituted of all unnatural deaths with male preponderance throughout the study period. People in the age group 21-30 years particularly from rural areas were most affected. The pedestrians and two wheeler users formed the majority of fatalities. Collision between two wheeler and light motor vehicle was the most common crash pattern and injury to head & neck region was the most common cause of death. Maximum number of accidents occurred between 4pm to 8pm and in the month of November. Unskilled workers, agricultural workers and government employees constituted a larger proportion of fatalities.

Shruthi et al. (2013) have conducted a retrospective observational study in the Department of Forensic Medicine and Toxicology, Kempegowda Institute of Medical Sciences, Bangalore between January 2010 to December 2012. Results of this study revealed that, most of victims were between 21-30 years of age, males constituted 78.22% of the total victims, and four wheeler vehicles were involved in 68.44% RTAs. Maximum RTAs occurred during the daytime, between 6 AM to 12 PM. Head injures constituted 30.22% of the total injuries, followed by injuries involving abdomen, thorax and limb. Hemorrhagic shock caused 63.11% of deaths, while head injury caused death in 30.22% of cases.

Singh et al. (2013) have done the study on Elucidation of risk factors in survivors of road traffic accidents in North India. This study was conducted from 1 March 2012 to 30 May 2012 at the Trauma Centre of King George's Medical University, Lucknow, India and the questions were asked from survivors of road traffic accidents using a pretested questionnaire after they received pre -medical care. At the end of the study, it was found that severe injuries are more likely to be due to over-speeding of vehicles, not using helmets and seat belts.

Singh and Aggarwal (2010) have analyzed the fatal road traffic accidents among young children in Muzaffarnagar. In this study, descriptive statistical analysis was used and it was found that fatal road accidents are a major cause of childhood mortality up to sixteen years of age involving mainly males. Pedestrians and cyclists were the common group injured and majority of the accidents occurred during the winter season.

Komba (2007), have done a case study on Risk factors and road traffic accidents in Tanzania. This study has revealed the pattern and trends of motor traffic accidents in Kibbaha district from 2001 to 2004. It shows that the accident occurrence was increasing every year, passengers and pedestrians are always at highest risk of being injured or killed on the road, young males are highly prone to motor traffic accidents. Males are more involved in road accidents than females, the risk of dying in an accident during the night was significantly higher than during the day, especially when it was raining. Further age, gender, over speeding, reckless driving, being a pedestrian, or a motor cyclist were identified as risk factors to motor vehicle crashes. This study has also identified qualitatively (by interviews) that the technical element of the highway construction, corruption, irresponsibility, poor management, driving while using cell phone, driving without training, failure to respect and obey traffic regulations, bad condition of vehicles, age of the vehicles and poor condition of services as the important risk factors associating to the cause of traffic accidents in Kibaha district.

2.3.2 Risk Factors Influencing Motor Vehicle Accidents

Al-Ghamdi, 2002 conducted a research on using logistic regression to estimate the influence of accident factors on accident severity. Traffic police records are collected in order to examine the contribution of several variables to accident severity in Riyadh. The data set used in this study was derived from a sample of 560 subjects involved in serious accidents reported in traffic police records in Riyadh, the capital of Saudi Arabia. Only accidents occurring on urban roads in Riyadh were examined. The conclusion of this study expressed that location and cause of accident were significantly associated with accident severity among nine variables.

Amarasingha & Dissanayake, 2013 carried out a study to identify the relationships between large truck crashes and traffic and geometric characteristics on limited access highways in state of Kansas during the period 2005 to 2010. Poisson regression model and a negative binomial regression model were applied to analyze data. It is found that from the results of Negative Binomial regression, number of lanes, annual average daily traffic, and large truck percent have a specific impact on large truck crashes. Results of Poisson regression model shows that section length, number of lanes, lane width, horizontal curvature, vertical grade, annual average daily traffic per lane, inside shoulder width, inside rumble strip have significant impact on large truck crashes.

Haadi, 2014 conducted a case study on identification of factors that cause severity of road accidents in Ghana: Northern Region. In this study, the binary logistic regression has applied to a total of 398 accident data from 2007-2009 collected from traffic-police records. The conclusion of the research expressed that overloading and obstruction were the most significantly associated with accident severity.

Wedagama & Dissanayake, 2009 have done a study to investigate the influence of accident related factors on road fatalities using logistic regression technique. Logistic Regression models were separately developed for fatal accidents considering motorcycles and all vehicles including motorcycles in Bali, Indonesia as a case study. Seven predictor variables were employed in the developed models. The study found that probabilities of female motorcyclists and motorists contributed more on motorcycle and motor vehicle fatal accidents than males. In addition, age was also significant to influence all vehicle fatalities.

Chengye & Ranjitkar, 2013 carried out a study to develop accident prediction models that link accident frequencies and factors including traffic conditions, geometric and operational characteristics of road and weather conditions using negative binomial regression model. The study used a sample of accidents occured from 2004 to 2010 on a 74 km long section of Auckland motorway. It is found that segment length, annual average daily traffic per lane and number of lanes have the most profound effects on accident frequency.

Chen, et al., 2016 have conducted a study to identify the main factors affecting serious road traffic crash and particularly serious road traffic crash during the period 2007-2014 in China. They acquired information on 18 risk factors and applied multinomial logistic regression technique. They found that five risk factors namely location, vertical alignment, roadside safety rating, driver distraction and overloading of cargo were significant factors for crash severity. They indicated that intersections were more likely to have side impact on serious road traffic crashes and particularly serious road traffic crashes, especially with poor visibility at night.

Robin, 2014 have done a study to identify factors which affect the severity of crashes in Missouri work zone during the period 2009-2011. Multinomial Logistic Regression technique was applied to analyze the data. Road alignment, road condition and road profile are influential factors for severity of crashes.

Zhang, et al., 2013 have conducted a study to investigate the influential factors to accident severity during the period 2006–2010 in Guangdong Province, China. They applied binary logistic regression technique to analyze the data. Light condition, overloading and gender of driver factors were identified as highly influential factors on accident severity.

Celik & Oktay, 2014 have done a study to determine the risk factors affecting the severity of traffic injuries during the period 2008-2013 in Turkey. Data were classified into three injury severity categories: fatal, injury, and no injury. Based on this classification, a multinomial logistic regression analysis is performed. The estimation results reveal that the drivers over the age of 65, primary educated drivers, accidents occurring on state routes, highways or provincial roads and the presence of pedestrian crosswalks increase the probability of fatal injuries. The results also indicate that accidents involving cars or private vehicles or those occurring during the evening peak, under clear weather conditions, on local city streets or in the presence of traffic lights decrease the probability of fatal injuries.

2.4 Previous Studies related to Motorcycle Accidents in Other Countries

This section describes previous studies related motorcycle accidents in the world. Similarly, it presents 2 sections namely descriptive analysis and risk factors influencing motorcycle accidents using logistic regression analysis.

2.4.1 Descriptive Analysis

Tanaboriboon & Satinnam , 2005 conducted a research on motorcycle accidents in Thailand during the period 2000–2002. They found in this research, motorcycles can be anticipated throughout the country which will result in more road casualties and tremendous economic losses, especially the extra health care costs for the accident victims and therefore, remains a challenging issue to all concerned parties to address this significant social problem and concurrently, to implement all the necessary measures promptly to fight this long and seemingly endless battle.

Jama, et al., 2011 have done a study on characteristics of fatal motorcycle crashes into roadside safety barriers in Australia and New Zealand. Seventy seven motorcycle fatalities involving a roadside barrier in Australia and New Zealand were examined. They found that the fatalities usually involved a single vehicle crash and young men. The roadside barriers predominantly involved were steel W-beams, typically on a bend in the horizontal alignment of the road. A majority of fatalities occurred on a weekend, during daylight hours, on clear days with dry road surface conditions indicating predominantly recreational riding. Speeding and driving with a blood alcohol level higher than the legal limit contributed to a significant number of these fatalities.

Teoh & Campbell, 2010 have conducted a study on role of motorcycle type in fatal motorcycle crashes. They found that strong effects of motorcycle type were observed on driver death rates and on the likelihood of risky driving behaviors such as speeding and alcohol impairment. Although the current study could not completely disentangle the effects of motorcycle type and rider characteristics such as age on driver death rates, the effects of both motorcycle type and rider age on the likelihood of risky driving behaviors were observed among fatally injured motorcycle drivers.

2.4.2 Risk Factors Influencing Motorcycle Accidents Using Logistic Regression

Zhu, 2014 conducted a study to identify influential factors that cause motorcycle-motor vehicle crashes during the period 2008 to 2012 in the State of Ohio. Multinomial logistic regression model was applied to the data. The conclusion of the research expressed that age, time of crash, number of units, vehicle in error, road contour, collision type, alcohol used, posted speed, and helmet used were the influential factors for the crash severity. Moreover, he found that driver of motorcycle or vehicle that uses alcohol increased the chance of a fatality or injury. Crashes that occur on highways or freeways with higher speed limits were more likely to result in injuries and fatalities.

Wedagama & Dissanayake, 2009 have done a study to investigate the influence of accident related factors on motorcycle injuries on two arterial roads in Bali. Multinomial logistic regression analysis is applied considering three severity classes such as slight injury, serious injury and fatal injury as response variables. The results showed that sideswipe accidents, motorcycles collided with other vehicles, motorcyclist failed to yield and motorcycle at fault were the influential factors on motorcycle injuries. Probability analysis showed that a change in 1% of these variables could influence motorcycle injuries between 33% and 34%.

3. METHODOLOGY

3.1 Overview

This chapter presents the description of the data and methods carried out in the analysis.

3.2 Description of Data

The initial database had 111457 accidents. Initially, it has detected important 16 factors influencing in road accidents. However, it is found some issues exist in pedestrian location, road pre-crash factor, vehicle pre-crash factor and accident type factor of the database. Basically, it is performed descriptive statistics and graphical analysis roughly. Then it leads to ascertain these 4 factors recorded more data (more than 100,000) under not known/not applicable level. Therefore, those factors were removed from the analysis. Finally, a database having 12 factors is prepared and used it for further analysis.

3.3 Explanatory Variables

In this study, it is considered only the road accidents involved drivers at fault. Response variable is accident severity which consists of two levels namely grievous and non-grievous. Accidents result in death or critically injured are named as grievous accidents. Accidents result in non-critically injured or damage only accidents are named as non-grievous accidents. The following factors are considered in this study. Age is the only one continuous variable among the variables considered in this study. Dummy variables are used to represent the categorical variables in the analysis.

3.4 Research Methodology

In this study, factors affecting motor vehicle accidents and motorcycle accidents are studied separately. Data analyses are arrayed mainly under preliminary and fundamental analyses. In preliminary analysis will be included univariate analysis and bivariate analysis. Univariate analysis is performed to get a general understanding of the whole dataset and bivariate analysis is functioned to examine the relationships between the variables. In fundamental analysis, one sample proportion test is used to reduce the levels of factors that are influencing the severity of accidents. Pearson Chi-Square test is performed to check the association between each contributory factor and the accident severity. Finally, due to the dichotomous nature of the dependent variable, binary logistic regression analysis is carried out as advanced analysis to investigate the combined effect of the variables. These statistical data analyses are conducted by using MS Excel, EVIEWS and SPSS software.

Table 3.1: Description of Factors

Factor	**Levels**	**Abbreviation**
Vehicle Type	Light vehicle	LV
	Heavy vehicle	HV
Gender	Male	M
	Female	F
Validity of License	With valid license	WL
	Without valid license	WOL
Accident Cause	Speeding	Cause1
	Aggressive/negligent driving	Cause2
	Influenced by alcohol/drugs	Cause3
	Fatigue/fall asleep	Cause4
	Others	Cause5
Alcohol Test	No alcohol/below legal limit	BL
	Over legal limit	OL
Time	Day time	DT
	Night time	NT
Weekday/Weekend	Weekday	WD
	Weekend	WE
Road surface	Dry	D
	Wet	W
	Others	OT
Weather Condition	Clear	CL
	Rainy	RA
	Others	WO
Light condition	Daylight	DL
	Night, Good street lighting	GSL
	Night, no street lighting	NSL
	Dusk/dawn	DD
Location	Bend/Junction	BJ
	Road	RD
Age of Driver at fault		Age

4. ANALYSIS OF MOTOR VEHICLE ACCIDENTS

4.1 Overview

This chapter presents the results obtained by analyzing motor vehicle accidents occurred in Sri Lanka. The first part of this chapter presents the descriptive results based on the motor vehicle accidents and the second part discusses the results of significant factors affecting on accident severity, goodness of fit measures, model diagnostics and predictive accuracy.

4.2 Frequency of accidents

Following Table 4.1 presents the frequencies and percentages of accident during the study period 2014 to 2016.

Table 4.1: Frequency of accidents over 3 years

Year	Number of accidents	Percentage (%)
2014	34657	31.1
2015	37885	34.0
2016	38915	34.9

According to the Table 4.1 the numbers of reported accidents to the police were 34657 in 2014 and it has increased to 38915 in 2016. It is noted that these are figures based on the accidents that have been reported. Most of the non-grievous accidents are not reported to the police and damage only accidents are settled between the parties amicably.

4.3 Severity of accidents by year

Figure 4.1 shows severity of accidents reported during the period 2014 to 2016.

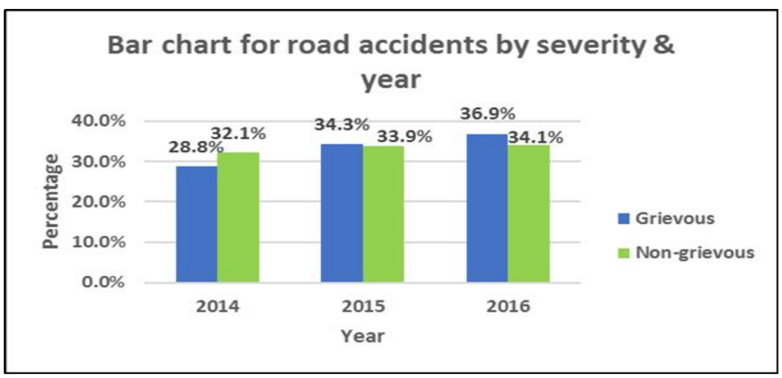

Figure 4.1: Severity of accidents by year

According to the Figure 4.1 the highest numbers of grievous and non-grievous accidents have been recorded in the year 2016. It also indicates that the percentages of grievous and non-grievous accidents increased from 2014 to 2016.

4.4 Road Accidents by Vehicle Types

Figure 4.2 shows the percentages of accidents occurred by different types of vehicles.

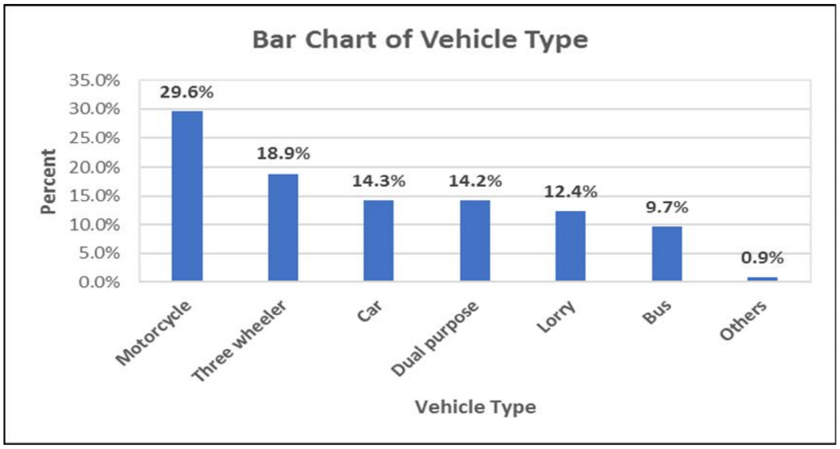

Figure 4.2: Road accidents by vehicle types

Figure 4.2 illustrates that motor cycle accidents were accounted for almost 30% of total accidents. Therefore, it is decided to study the factors affecting motor cycle and motor vehicle accidents separately.

4.5 Results of Analysis of Motor Vehicle Accidents

This section presents the descriptive results of motor vehicle accidents.

4.5.1 Severity of Accidents by Motor Vehicles

Table 4.2 indicates frequencies and percentages of accidents occurred by motor vehicles.

Table 4.2: Severity of Accidents by Motor Vehicles

Vehicle Type	Grievous		Non-grievous	
	Count	Percentage	Count	Percentage
Car	2956	15.3	13040	22.0
Dual purpose vehicle	4127	21.4	11671	19.7
Lorry	3619	18.8	10218	17.2
Three wheeler	5316	27.6	15716	26.5
Bus	2846	14.8	7987	13.5
Others	413	2.1	622	1.1
Total	19277	100.0	59254	100.0

According to the Table 4.2 among motor vehicles, three wheels are the most frequently involved vehicle in grievous and non-grievous accidents accounting for 27.6% and 26.5% respectively. For grievous accidents, second highest type is dual purpose vehicle and lorries are the third type of vehicles. For non-grievous accidents, second highest type is cars and dual purpose vehicles are the third type of vehicles. “Other vehicle” category indicated articulated vehicle, prime mover, land vehicle and tractors. For analysis purpose, motor vehicles are classified into 2 categories as light vehicles and heavy vehicles. Car, dual purpose vehicle, three-wheeler are categorized as light vehicles and lorry, bus and other vehicles are categorized as heavy vehicles. Following figure shows the distribution of both types of vehicles with the severity of accidents.

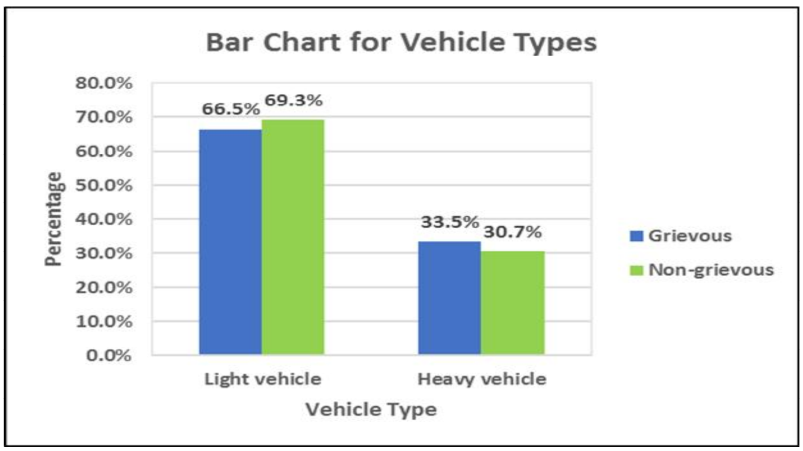

Figure 4.3: Main types of motor vehicles

Figure 4.3 shows the majority of grievous and non-grievous accidents occurred by light vehicles accounting for 66.5% and 69.3% respectively. Percentage of grievous and non-grievous accidents occurred by heavy vehicles contributes less than 35% towards these accidents. It reveals that most of accidents occurred by car, dual purpose vehicle and three-wheelers.

4.5.2 Severity of Accidents by Validity of License

Figure 4.4 shows the percentages of accidents according to the drivers whether they have valid license or not.

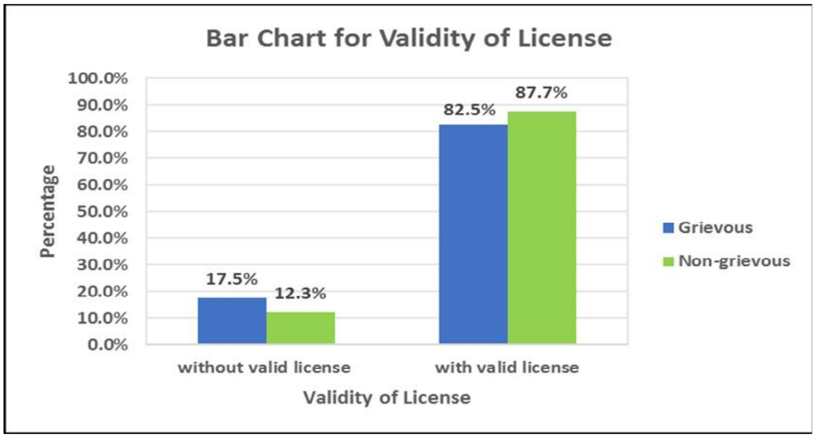

Figure 4.4: Severity of accidents by validity of license

Figure 4.4 illustrates that more than 80% of drivers in the dataset had valid license when they occurred an accident.

4.5.3 Severity of Accidents by Accident Cause

Following Table 4.3 shows the main causes for accidents during the period 2014-2016.

Table 4.3: Severity of accidents by accident cause

Accident Cause	**Grievous**		**Non-grievous**	
	Count	**Percentage (%)**	**Count**	**Percentage (%)**
Speeding	2888	15.0	7465	12.6
Aggressive/negligent driving	14070	73.0	44462	75.0
Influenced by alcohol/drugs	856	4.4	3055	5.2
Fatigue/fall asleep	474	2.5	961	1.6
Others	989	5.1	3311	5.6

Table 4.3 indicates major cause for both grievous and non-grievous accidents is aggressive/negligent driving which contribute more than 70% towards these accidents.

Second major cause is speeding. However, it contributes for grievous and non-grievous accidents 14.8% and 12.5% respectively. Percentages of accidents occurred by influence by alcohol/drugs and fatigue/fall asleep are very less when compare with aggressive/ negligent driving.

4.5.4 Severity of Accidents by Alcohol Test

Figure 4.5 shows the effect of the influence of alcohol used to the percentages of grievous and non-grievous accidents.

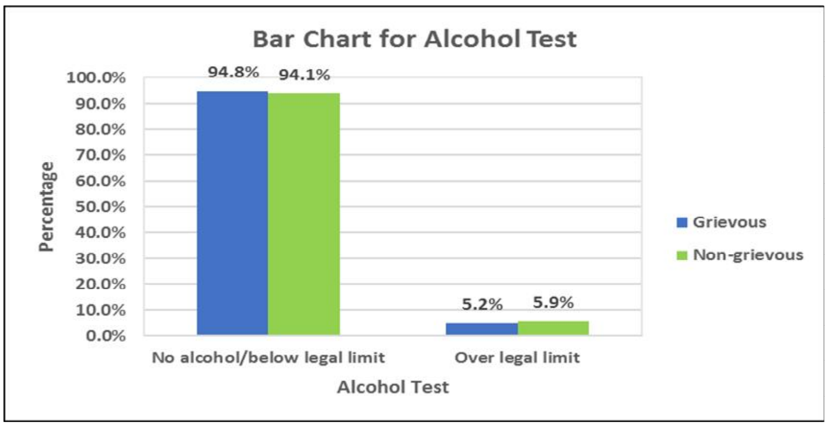

Figure 4.5: Severity of accidents by alcohol test

According to the Figure 4.6, percentage of no alcohol/below legal limit for grievous and non-grievous accidents contributes more than 90% towards these accidents. Percentages of drivers who use alcohol over legal limit for both types of accidents are very less in the dataset.

4.5.5 Severity of accidents by Time of Accident

The following Figure 4.6 indicates percentages of grievous and non-grievous accidents occurred during day and night time of the day.

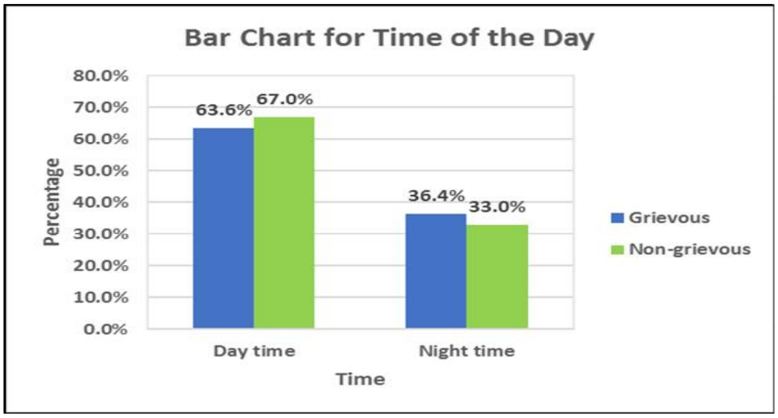

Figure 4.6: Severity of accidents by time of accident

Figure 4.6 shows that most of grievous and non-grievous accidents are occurred in daytime. For the two categories of time, the percentages of accidents occurred in daytime and night time are approximately 65% and 35% respectively. It indicates that the highest numbers of grievous and non-grievous accidents are occurred in daytime of the day.

4.5.6 Severity of accidents by Weekday/ Weekend

Figure 4.7 displays the distribution of road accidents occurred in weekdays and weekend.

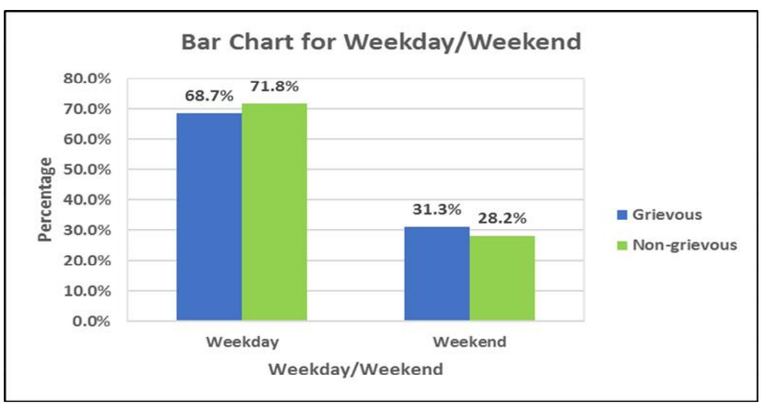

Figure 4.7: Severity of accidents by weekday/ weekend

The day of the week were grouped into two levels, that is, into weekdays or weekends. Percentages of grievous accidents occurred in weekdays and weekends are approximately 68.7% and 31.3% respectively. Percentages of non-grievous accidents occurred in weekdays and weekends are 71.8% and 28.2% respectively. It indicates that majority of accidents occurred in weekdays.

4.5.7 Severity of Accidents by Road Surface

Figure 4.8 shows the distribution of road accidents occurred in different types of road surfaces.

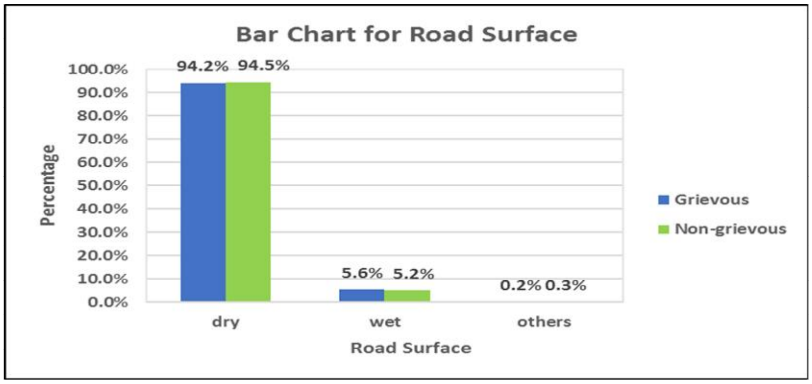

Figure 4.8: Severity of accidents by road surface

According to the figure 4.8, percent of grievous accidents occurred in dry, wet and other surface are 94.2%, 5.6% and 0.2% respectively. Percent of non-grievous accidents occurred in dry, wet and other surface are 94.5%, 5.2% and 0.3% respectively. It indicates that majority of accidents are occurred on dry surfaces than wet and other surfaces.

4.5.8 Severity of Accidents by Weather Condition

Figure 4.9 shows the percent of grievous and non-grievous accidents occurred during different weather conditions.

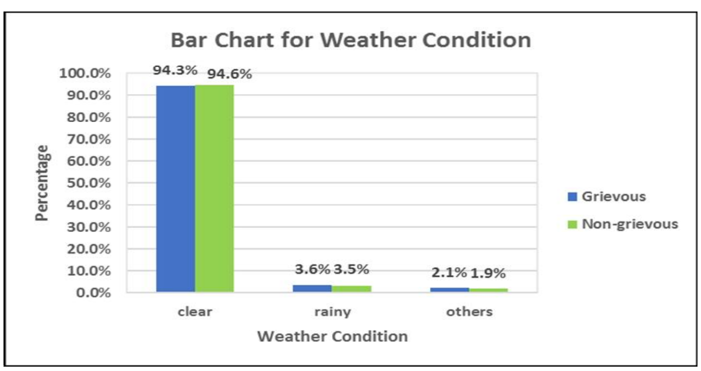

Figure 4.9: Severity of accidents by weather condition

Figure 4.9 illustrates that percent of grievous accidents during clear, rainy and other weather conditions are 94.3%, 3.6% and 2.1% respectively. Percent of non-grievous accidents during clear, rainy and other weather conditions are 94.6%, 3.5% and 1.9% respectively. It reveals that majority of accidents occurred during clear weather for both grievous and non-grievous accidents.

4.5.9 Severity of Accidents by Light Condition

Light condition may have an impact on the severity of accidents due to visibility issues. Table 4.4 shows the frequency and percentages of grievous and non-grievous accidents occurred under different light conditions.

Table 4.4: Severity of accidents by light condition

Light Condition	**Grievous**		**Non-grievous**	
	Count	**Percentage (%)**	**Count**	**Percentage (%)**
Daylight	12069	62.6	39070	65.9
Night, good street lighting	1106	5.7	4594	7.8
Night, no street lighting	5511	28.6	13874	23.4
Dusk, dawn	591	3.1	1716	2.9

Table 4.4 indicates that percentages of grievous and non-grievous accidents occurred under day light are 62.9% and 65.9% respectively. It reveals that majority of grievous and non-grievous accidents occurred under daylight. It seems that nighttime or improper street lighting is not a main reason to occur grievous or non-grievous accident.

4.5.10 Severity of Accidents by Location of the Accident

Percentages of accidents occurred in different locations are displayed in Figure 4.10.

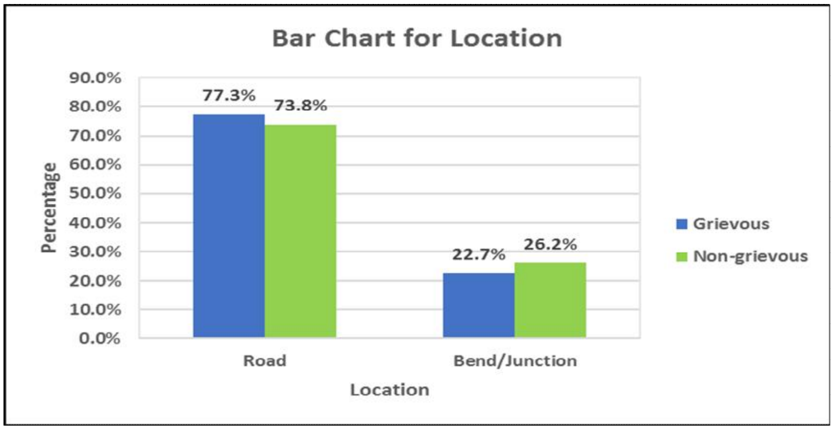

Figure 4.10: Severity of accidents by location of the accident

Figure 4.10 shows that majority of grievous and non-grievous accidents were occurred, while the vehicle was moving on a straight road accounting for 77.3% and 73.8% respectively. Percentages of grievous and non-grievous accidents occurred at bend or junction are less than 30%.

4.5.11 Descriptive Statistics of Age of Driver

The descriptive statistics of age of drivers whom responsible for accidents is shown in the following Table 4.5.

Table 4.5: Descriptive statistics of age of driver

	Mean	Std. Deviation	Minimum	Maximum	Q1	Q3
Age of Driver	37.9	11.84	14	89	29	46

Driver's age is the only one continuous variable in this dataset. The above Table shows that average age of driver who is responsible for accident is 37.9 years. Furthermore, even though valid age for issuing a driving license is 18 years, above Table indicates that minimum age of driver who is responsible for accident is 14 years. To get an idea of which age range drivers are more responsible for accidents, histogram is drawn as follows.

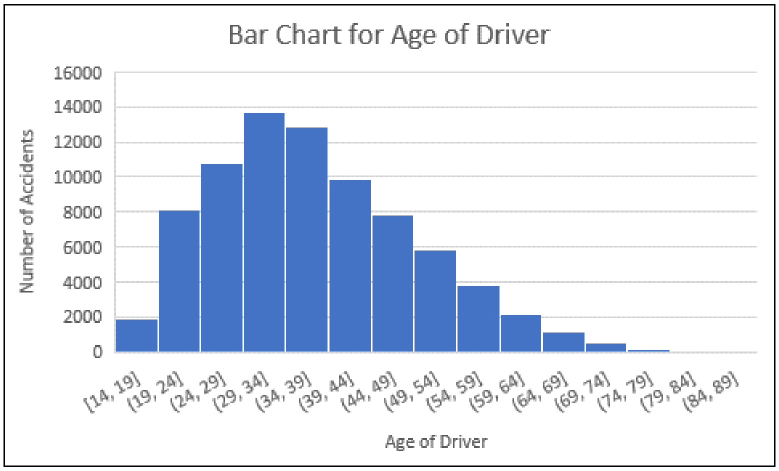

Figure 4.11: Histogram of age of driver

According to the Figure 4.11, the drivers in between the age group 29-34 were more responsible for the highest number of accidents which is approximately 14000 accidents. Furthermore, it indicates that younger drivers whose age less than 40 years were accounted for 60% of total accidents. It should be noted that younger drivers were more responsible for the highest number of accidents.

4.6 Reduction of Levels of Factors

It is better to have as few levels of factors as possible for easy interpretation. Therefore, one sample proportion test is used for factors having more than 2 levels to reduce the levels. The summary statistics of proportion test is listed in the Table 4.6.

Table 4.6: Summary statistics for Proportion Test

Factor	Levels	X	Proportion (N=78531)	95% Confidence Interval	
				Lower	Upper
Vehicle Type	Light vehicle	53861	0.686	0.683	0.689
	Heavy vehicle	24670	0.314	0.311	0.317
Gender	Male	77311	0.984	0.983	0.985
	Female	1220	0.015	0.014	0.016
Validity of License	With valid license	67859	0.864	0.862	0.866
	Without valid license	10672	0.136	0.133	0.138
Accident Cause	Speeding	10353	0.132	0.129	0.134
	Aggressive/negligent driving	58532	0.745	0.742	0.748
	Influenced by alcohol/drugs*	**3911**	**0.049**	**0.048**	**0.051**
	Fatigue/fall asleep*	**1435**	**0.018**	**0.017**	**0.019**
	Others*	**4300**	**0.054**	**0.053**	**0.056**
Alcohol Test	No alcohol/below legal limit	73944	0.941	0.939	0.943
	Over legal limit	4587	0.058	0.056	0.060
Time	Day time	51939	0.661	0.658	0.664
	Night time	26592	0.338	0.335	0.312
Weekday/ Weekend	Weekday	55761	0.710	0.706	0.713
	Weekend	22770	0.289	0.287	0.293
Road surface	Dry	74172	0.944	0.943	0.946
	Wet*	**4161**	**0.053**	**0.051**	**0.054**
	Others*	**198**	**0.0025**	**0.0021**	**0.0028**
Weather Condition	Clear	74242	0.945	0.944	0.947
	Rainy*	**2775**	**0.035**	**0.034**	**0.036**
	Others*	**1514**	**0.019**	**0.018**	**0.020**
Light condition	Daylight	51139	0.651	0.648	0.654
	Night, Good street lighting*	**5700**	**0.072**	**0.071**	**0.074**
	Night, no street lighting	19385	0.247	0.244	0.249
	Dusk/dawn*	**2307**	**0.029**	**0.028**	**0.030**
Location	Bend/Junction	19919	0.254	0.250	0.256
	Road	58612	0.746	0.743	0.749

* indicates status of non-significance

Based on the results of Table 4.6, some factor levels can be neglected or merged because of their small proportions. Proportions of ‘Influenced by alcohol/drugs’, ‘fatigue/fall asleep’ and ‘others’ under accident cause factors are non-significant. Thus, they were merged and generated a new classification called ‘Others’. Moreover, wet and others under road surface were merged and named the new classification as ‘Others’. Rainy and others under weather condition were merged and named the new classification as ‘Others’. Good street lighting and dusk/dawn under light condition were merged and named as ‘Others’.

4.7 Results of Pearson Chi-Square Test

Pearson Chi-square test is performed to test the significant relationship between the independent variables and the dependent variable. According to this analysis, following table describes the association between each factor and the accident severity.

Table 4.7: Results of Chi-square test

Variables	**χ^2 value**	**P value**
Vehicle Type	χ^2 (1) = 53.446	0.000
Gender of Driver	χ^2 (1) = 19.271	0.000
Validity of License	χ^2 (1) = 331.396	0.000
Alcohol Test	χ^2 (1) = 16.812	0.000
Accident Cause	χ^2 (2) = 72.183	0.000
Time	χ^2 (1) = 74.149	0.000
Weekday/Weekend	χ^2 (1) = 66.322	0.000
Road Surface	**χ^2 (1) = 2.896**	**0.089**
Weather	**χ^2 (1) = 3.221**	**0.073**
Light Condition	χ^2 (2) = 231.087	0.000
Location	χ^2 (1) = 81.086	0.000

According to the results of Table 4.7, it can be identified that vehicle type, gender of driver, validity of license, alcohol test, accident cause, time, weekday/weekend, light condition and location type are significantly associated with the accident severity. Only two factors namely road surface and weather, are not significantly associated with the accident severity. Therefore, non-significant factors are removed and continued the analysis.

4.8 Detecting Multicollinearity in Binary Logistic Regression

One of the assumptions in logistic regression is that explanatory variables should not be highly correlated with each other. Therefore, before applying logistic regression, multicollinearity should be checked among explanatory variables.

4.8.1 Correlation Coefficients of Explanatory Variables

The correlation coefficients among the explanatory variables can be used as first step to identify the presence of multicollinearity (Field, 2009). Correlation matrix of highly correlated explanatory variables presented in Table 4.8 and remained presented in the Appendix as large number of dummy variables are exist.

Table 4.8: Pearson Correlation matrix between 2 explanatory variables

Variables		Time		Light condition
		DT	NT	NSL
Light condition	DL	**0.978** (0.000)	**-0.978** (0.000)	**-0.782** (0.000)
	NSL	**-0.800** (0.000)	**0.800** (0.000)	1.000 (0.000)

Cell value: correlation coefficient

p value

It is mentioned earlier about the rule of thumb that if Pearson correlation coefficient is greater than 0.8 or 0.9 then multicollinearity is a serious concern. Results of Table 4.8 indicates that the Pearson correlation coefficients between two variables light condition and time are highly correlated and indicated them as bold in Table 4.8. These high correlation coefficients signify the presence of severe multicollinearity between the explanatory variable light condition and time of accident.

4.8.2 Detection of Multicollinearity based on Collinearity Statistics

Examining the correlation matrix may be helpful but not sufficient. It is quite possible to have data in which no pair of variables has a high correlation, but several variables together may be highly interdependent. Much better diagnostics are produced by tolerance and VIF values. Table 4.9 indicates the collinearity statistics.

Table 4.9: Collinearity statistics

Factor	Model	Collinearity Statistics	
		Tolerance	VIF
Vehicle type	LV	.961	1.029
	HV	.966	1.035
Validity of license	WL	.966	1.030
	WOL	.964	1.038
Alcohol test	BL	.678	1.475
	OL	.680	1.473
Time	**DT**	**.041**	**24.349**
	NT	**.045**	**21.456**
Weekend/Weekday	WE	.998	1.003
	WD	.993	1.007
Location	RD	.992	1.008
	BJ	.995	1.005
Gender	M	.991	1.010
	F	.989	1.008
Accident cause	Cause1	.976	1.025
	Cause2	.965	1.039
	Others	.676	1.479
Light condition	**DL**	**.047**	**21.278**
	NSL	**.058**	**20.146**
	Others	**.110**	**9.115**
Age	Age	.989	1.011

Results of Table 4.9 observes that the high tolerances for the variables vehicle type, gender, validity of license, accident cause, alcohol test, weekday/weekend, location and age of driver but very low tolerances for the variables time and light condition. Similarly, the variance inflation factor corresponding to the explanatory variables vehicle type, gender, validity of license, accident cause, alcohol test, weekday/weekend, location and age of driver are very close to 1, but for variables time and light condition, the VIF are larger than 2.5. Using these collinearity statistics, it can be concluded that the data almost certainly indicates a serious collinearity problem.

4.9 Solutions to Multicollinearity

Once the collinearity between variables has been identified, the next step is to find solutions in order to remedy this problem. Therefore, variables causing multicollinearity need to be dropped from the analysis. Any of the collinear variables could be omitted. There is no statistical ground for omitting one variable over another. Thus, first, time is removed from the data and repeats the analysis. However, collinearity still exists among the levels of light variable. Then time is added and light condition is removed and repeats the analysis. Results are presented in Table 4.10.

Table 4.10: Collinearity statistics for remained variables

Factor	Model	Collinearity Statistics	
		Tolerance	VIF
Vehicle type	LV	.961	1.029
	HV	.967	1.035
Validity of license	WL	.995	1.006
	WOL	.965	1.036
Alcohol test	BL	.678	1.475
	OL	.677	1.471
Time	DT	.970	1.031
	NT	.965	1.029
Weekend/Weekday	WE	.997	1.002
	WD	.993	1.002
Location	RD	.997	1.003
	BJ	.992	1.007
Gender	M	.991	1.009
	F	.990	1.011
Accident cause	Cause1	.976	1.024
	Cause2	.981	1.031
	Others	.676	1.479
Age	Age	.989	1.011

According to Table 4.10, tolerances for all the predictors are very close to 1 and all the VIF values are smaller than 2.5. Therefore, it can be concluded that multicollinearity is not a concern when one of the correlated variable is omitted.

4.10 Selection of Number of Variables

Best subsets regression technique is used to get an idea about suitable number of variables for the further analysis. Best subsets regression fits all possible models based on the explanatory variables selected. Mallows' Cp close to the number of variables plus constant indicates that best fitting models for this process. Results of best subsets regression are shown in the Table 4.11.

Table 4.11: Results of best subsets

Variables	Mallows C-p	Vehicle type	Gender	Age	License	Time	Workday	Location	Alcohol	Cause
1	393.9				×					
1	646.5	×			×			×		
2	311.3									
2	319.8				×			×		
3	236.9	×			×	×				
3	240.3	×			×			×		
4	166.5	×			×	×		×		
4	170.8	×			×	×				×
5	100.9	×			×	×		×		×
5	106.0	×		×	×	×		×		
6	39.9	×		×	×	×		×		×
6	53.0	×		×	×	×		×	×	
7	21.4	×		×	×	×		×	×	×
7	31.1	×	×	×	×	×		×		×
8	**8.2**	**×**	**×**	**×**	**×**	**×**		**×**	**×**	**×**
8	19.7	×		×	×	×	×	×	×	×
9	10.0	×	×	×	×	×	×	×	×	×

Table 4.11 indicates that Cp value (8.2) is less than the number of parameters (9) when there are 8 variables in the model. That is the best fitting model as Cp in other models does not close to number of parameters in the model. It also indicates that variable weekday/weekend is the only variable which should be removed from the model.

4.11 Checking Linearity Assumption

One of the assumptions in logistic regression is that explanatory variables have a linear relationship with the logit of the dependent variable. Age is the only continuous variable in this study. Therefore, the linear relationship between continuous explanatory variable and the logit of the dependent variable is checked by visually inspecting the scatter plot and is shown in the Figure 4.12.

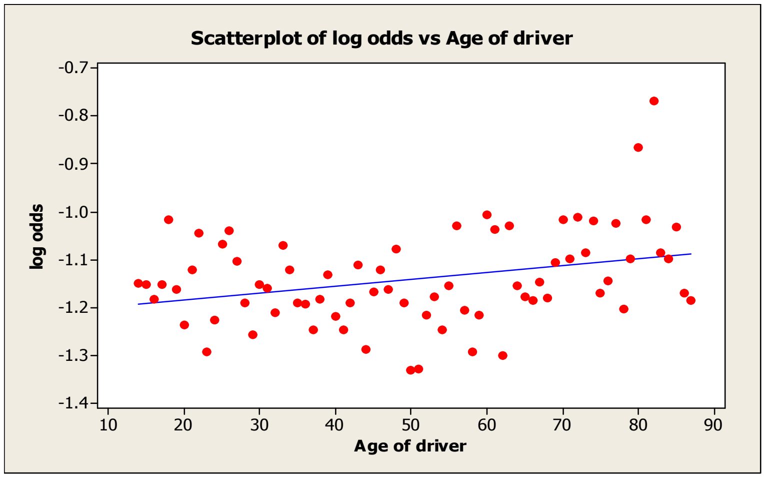

Figure 4.12: Log odds of severity vs. age of driver

Figure 4.12 shows that variable age is quite linearly associated with the severity in logit scale. Thus, binary logistic regression is carried out for further analysis.

4.12 Binary logistic regression analysis

Since the response variable is dichotomous (grievous/non-grievous), the binary logistic regression model is applied to fit the data. The maximum likelihood procedure is used to estimate the parameters of the logistic regression model. Forward (Likelihood Ratio) stepwise selection method was applied under the binomial logistic regression analysis, with variable entry testing based on the significance of the score statistic (the significance level was set at $p<0.05$), and removal testing based on the probability of a likelihood ratio statistic based on the maximum likelihood estimates (the significance level was set at $p>0.10$).

According to the methodology, main vehicle dataset (78531) was divided into two portions; 2/3 of data (52354) was used to develop the model, and the remaining 1/3 of data (26177) was used to validate the model (Dhananjaya & Alibuhtto, 2016).

4.12.1 Baseline Model

Table 4.12 presents the results of the baseline model which is the model with only the constant included before explanatory variables are entered into the model. Logistic regression compares this model with a model including all the significant factors to determine whether the latter model is more appropriate.

Table 4.12: Results of baseline model

Baseline Model	B	S.E.	Wald	df	Sig.	Exp(B)
Constant	1.162	.010	12828.477	1	.000	3.196

Initial -2 Log Likelihood: 57499.931

Table 4.12 indicates that the coefficient of constant is 1.162 and standard error of the coefficient for the constant is 0.01. Wald Chi-Square tests the null hypothesis that the constant equals 0. Results of the above table shows that constant is statistically significant to the model. Initial log likelihood value of the baseline model is 57499.931. This value is used to select an optimal model. Predictive power of the baseline model is 69.5%, which indicates the overall percentage of correctly classified cases when there are no explanatory variables in the model.

4.12.2 Developed model

As said earlier, Forward selection method is used to develop the binary logistic model. First, all variables were entered into the model according to their -2 log likelihood ratio. Variable which has minimum -2 log likelihood ratio is added first. Model iteration occurred up to eight steps. The analysis was performed on p value = 0.05 significance level to formulate the model.

4.12.3 Developed model interpretation

The interpretation of any fitted model requires the ability to draw practical inferences from the estimated coefficients. Comparing the difference in impact or risk among the levels of each variable by looking at the regression coefficients. The interpretation of the estimated parameter coefficients is that, for a one unit change in the predictor variable, the difference in log-odds for a positive outcome is expected to change by the respective coefficient, given the other variables in the model are held constant. Accordingly, those predictors with positive coefficients cause an increasing tendency to result into fatalities. Similarly, negative coefficients indicate decreasing tendency for those significant predictors. Table 4.13 explains the variables in the developed model used to predict the severity of accidents.

Table 4.13: Variables in the model

Variables	B	S.E.	Wald	df	Sig.	Exp(B)
HV	-.183	.022	68.737	1	.000	.833
WOL	-.444	.029	15.826	1	.000	.641
NT	-.143	.022	43.081	1	.000	.866
BJ	.175	.024	51.573	1	.000	1.191
BL	.301	.056	28.445	1	.000	1.351
Cause2	.120	.030	236.856	1	.000	1.128
Age	-.138	.023	37.564	1	.000	.871
M	.201	.094	4.514	1	.034	1.222
Constant	1.219	.032	1479.642	1	.000	3.382

When exploring results of the Table 4.13, validity of license, vehicle type, location type, time, age of driver, alcohol test, accident cause and gender have a significant effect on the severity of accidents. Wald chi-square tests whether each of the predictors included make a significant contribution to the model while controlling other predictors.

Exp(B) is an odds ratio. Odds ratio of vehicle type indicates that a grievous accident occurred by heavy vehicles is 83% less likely to be a grievous accident occurred by light vehicles. Similarly, according to the result by the odds ratio, the odds of a grievous accident occurred by drivers who, without a valid license are 64% less likely to be odds of a grievous accident occurred by drivers who having a valid license. Odds of a grievous accident occurred in night time is 87% less likely to be a grievous accident occurred in day time. Odds of a grievous accident occurred in bend/junction is 19% more likely to be a grievous accident occurred in road. Odds of a grievous accident occurred by drivers who used alcohol below legal limit or no alcohol 35% more likely to be a grievous accident occurred by drivers who used alcohol over legal limit. Odds of a grievous accident occurred by aggressive/negligent driving is 13% more likely to be a grievous accident occurred by speeding. The odds ratio of age is 0.871. It indicates that for every one unit increase in age (one additional year of living), the odds of occurring a grievous accident decreases which implies the older the driver, the less the accident risk. Odds of a grievous accident occurred by male drivers 22% more likely to be a grievous accident occurred by female.

4.12.4 Logit Model

From the analysis, the logit model with the significant variables is:

$$\text{logit}(p) = 1.219 - 0.183HV - 0.444WOL - 0.143NT + 0.175BJ + 0.301BL + 0.12Cause2 - 0.138Age + 0.201M$$

4.12.5 Variables not in the Developed Model

Table 4.14 indicates that variables which are not in the model.

Table 4.14: Variables not in the model

Variable	Score	df	Sig.
Weekday/Weekend	3.067	2	.216
WD	.249	1	.618
HD	2.877	1	.090
Overall Statistics	3.067	2	.216

According to the results of Table 4.14, weekday/weekend variable is not significantly associated with the severity of accidents ($p = 0.216 > 0.05$). Thus, weekday/weekend variable is removed from the model.

4.12.6 Importance of Variables in the Model

Table 4.15 presents the information how the model is affected if an explanatory variable is added to the model. In other words, which variable is important for the model. Results of following table are used to examine the importance of a variable in the model.

Table 4.15: Importance of variables in the model

Step	Improvement			Model			Variable
	Chi-square	df	Sig.	Chi-square	df	Sig.	
1	201.658	1	.000	201.658	1	.000	IN: Accident cause
2	65.982	1	.000	267.640	2	.000	IN: Vehicle Type
3	55.460	1	.000	323.100	3	.000	IN: Location
4	39.785	1	.000	362.885	4	.000	IN: Time
5	36.441	1	.000	399.326	5	.000	IN: Age
6	32.678	1	.000	432.004	6	.000	IN: Alcohol
7	16.706	2	.000	448.710	8	.000	IN: License
8	4.221	1	.030	452.931	9	.000	IN: Gender

Table 4.15 indicates that adding the variable accident cause to the model makes the biggest change in the model's log likelihood value. Therefore, accident cause is the most important variable in this model. It is followed by the vehicle type, location type, time, age of driver, alcohol test, validity of license and gender respectively.

4.13 Measures of Goodness of Fit

Once a logistic model is fitted to the data it is essential to check that the assumed model is actually a valid model. Various measures are used to test the goodness of fit of the model.

4.13.1 Test of Model coefficients

The test of model coefficients is used to check that the new model (with explanatory variables included) is an improvement over the baseline model. It uses chi-square test to see if there is a significant difference between the Log-likelihoods (specifically the -2LLs) of the baseline model and the new model. If the new model has a significantly reduced -2LL compared to the baseline then it suggests that the new model is explaining more of the variance in the outcome and is an improvement.

Table 4.16: Test of developed model coefficients

	Chi-square	df	p value
Step 8	4.221	1	0.028
Block	452.931	9	0.000
Model	452.931	9	0.000

Results of Table 4.6 indicates that the chi-square is highly significant (chi-square=452.931, $p<0.000$ with df =9). Thus, it can be concluded that the developed model is significantly better than the baseline model. That means the accuracy of the model improved when added the explanatory variables.

4.13.2 Model Summary

Under the model summary section, developed model is checked whether it is an improvement over the baseline model. Results are shown in the Table 4.17.

Table 4.17: Model Summary

Step	-2 Log likelihood	Cox & Snell R^2	Nagelkerke R^2
8	57047.735	.604	.651

According to the results of the Table 4.17, the developed model has a significantly reduced log likelihood value (57047.735) compared to the baseline model. It is revealed that the developed model is explaining more of the variance in the outcome and it is an improvement over the baseline model. Thus, it can be concluded that the developed model is better at predicting the severity of the accidents than the baseline model where no predictor variables were added.

In addition to that, Table 4.17 contains the Cox & Snell R^2 and Nagelkerke R^2 values, which are used to calculate the explained variation. These values are sometimes referred to as pseudo R^2 values. According to these both values, the explained variation in the dependent variable based on the model are 60.4% and 65.1% respectively.

4.13.3 Predictive Accuracy of Developed Model

Classification table shows how many of the cases where the observed values of the dependent variable have been correctly predicted. Results of the classification table are shown in the Table 4.18.

Table 4.18: Classification Table

Observed	Predicted		
	Grievous	Non-grievous	Percentage Correct
Grievous	9123	3355	73.11
Non-grievous	8953	30911	77.54
Overall Percentage			76.47

Table 4.18 indicates that 73.1% were correctly classified for grievous accidents and 77.5% for non-grievous accidents. Overall 76.5% were correctly classified. It can be seen that the developed model is correctly classifying the outcome for 76.5% of the cases compared to 69.5% in the null model.

4.13.4 Hosmer and Lemeshow test

Hosmer and Lemeshow test is used to indicate which extent to the estimated model provides a better fit to the data than the null model.

Table 4.19: Results of Hosmer and Lemeshow test

Chi-square	df	Sig.
7.755	8	0.458

As the results shown in the Table 4.19, Hosmer & Lemeshow test of the goodness of fit suggests the model is a good fit to the data as $p=0.458$ (>0.05).

4.13.5 Receiver Operating Characteristic Curve

Receiver operating characteristic (ROC) curve is used to evaluate the fit of a logistic regression model. ROC curve is shown in the Figure 4.13 and the area under curve is presented in Table 4.20.

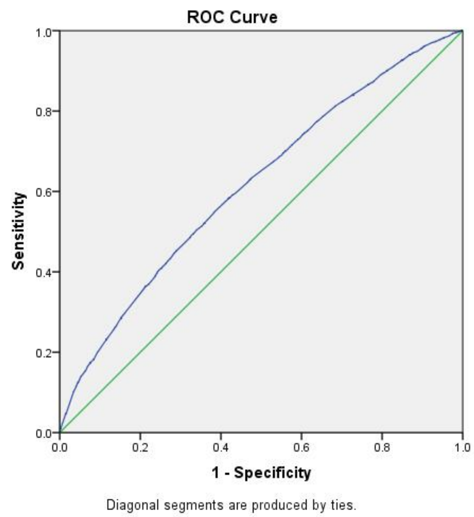

Figure 4.13: ROC Curve

Table 4.20: Area Under the Curve

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.692	.003	.000	.687	.697

According to the above Table and Figure, the area under the curve is 0.692 with 95% confidence interval (0.687, 0.697). Moreover, the area under the curve is significantly different from 0.05 since the p =0.000 <0.05. That means, the logistic regression classifies the group significantly better than by chance.

4.14 Model Diagnostics

Two assumptions of binary logistic regression, linearity and multicollinearity are checked earlier. Assumption of independent errors and influential observations are checked in this section.

4.14.1 Testing Residuals

Table 4.21 provides two measures that can be used to assess how well the model fits the data, as shown below.

Table 4.21: Testing Residuals

Method	Chi-Square	df	Sig.
Pearson	294.983	269	0.133
Deviance	306.894	269	0.065

Pearson and Deviance residuals present the Pearson and Deviance chi-square statistics. Table 4.21 indicates that p values of both tests are greater than 0.05 which implies that residuals are not statistically significant. Based on these measures, the model fits the data well.

4.14.2 Checking Normality in Residuals

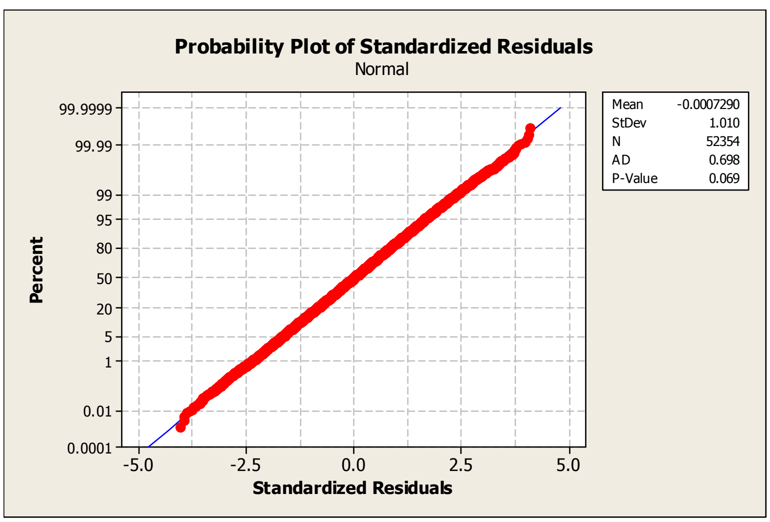

Figure 4.14: Normal probability plot of residuals

In logistic regression, the errors are not assumed to have a normal distribution. Instead, it is assumed that the distribution of the errors follows a binomial distribution, which approximates a normal distribution only for large samples. If the residuals are normally distributed for a large sample, then it can be more confident that inferential statistics are correct, because normal and binomial (the assumed distribution) distributions are about the same for large samples. Therefore, residuals are checked whether normally distributed and displayed in Figure 4.14. It confirms that residuals are normally distributed since p=0.069>0.05.

4.14.3 Testing Heteroskedasticity and Correlation in Residuals

Heteroskedasticity in residuals is tested from Lagrange Multiplier (LM) test. Correlation in residuals is tested using Durbin Watson (DW) test. Results of both tests are displayed in Table 4.22.

Table 4.22: DW and LM test of residuals

Test	Test Statistic	p value
DW Test	1.7512	0.08517
LM Test	14.815	0.0628

Results of Table 4.22 illustrates that residuals have constant variance as p = 0.0628> 0.05. Moreover, residuals are uncorrelated as p=0.085>0.05.

4.14.4 Detecting Influential Observations

Plot of Cook's distance is used to detect influential observations and displayed in Figure 4.15.

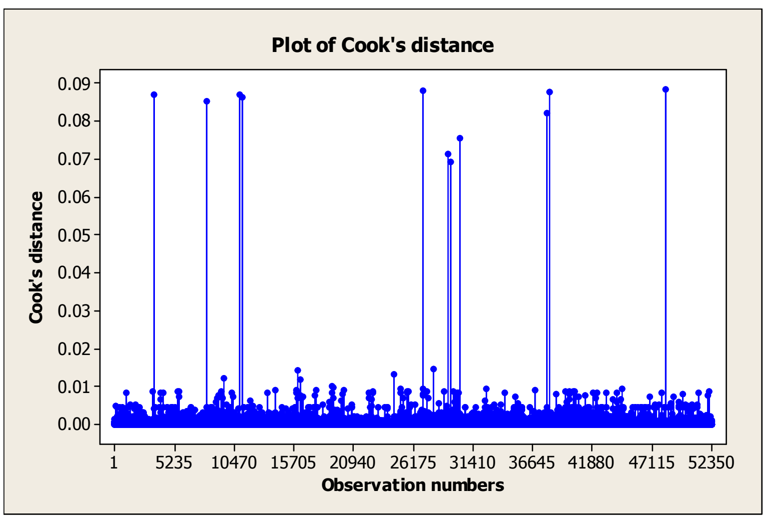

Figure 4.15: Plot of Cook's Distance

It can be seen that all observations are less than 1 and even less than 0.5. Therefore, it can be said that there are no outliers nor influential observations in this data set.

4.15 Validation of the Model

Developed logistic regression model is used to validate the model. One third of data is used for validation and results are shown in the following Table 4.23.

Table 4.23: Category prediction

Observed	Predicted		
	Grievous	Non-grievous	Percentage Correct
Grievous	5026	1773	73.9
Non-grievous	3482	15896	82.0
Overall Percentage			79.9

Table 4.23 indicates that the model correctly predicted 79.9 % of the validation data which is greater than to the predictive power of the baseline model 69.8%. That means the developed model more accurately predicts the severity of accidents than the prediction in baseline model.

4.16 Summary

This section summarizes the results obtained by analyzing motor vehicle accidents occurred in Sri Lanka. According to the logistic regression model, vehicle type, validity of license, time, location, alcohol test, accident cause, age of driver and gender have a significant effect on the severity of accidents. Bend or junction location, aggressive/negligent driving, drive by male drivers, drive at daytime, driving light vehicle and drivers who use alcohol below legal limit or no alcohol, have a high chance to be a grievous accident. Moreover, the older drivers have less accident risk.

5. ANALYSIS OF MOTOR CYCLE ACCIDENTS

5.1 Overview

This chapter presents the results obtained by analyzing motorcycle accidents occurred in Sri Lanka. The first part of this chapter presents the descriptive results based on the motorcycle accidents and the second part discusses the results of significant factors affecting on severity of accidents, goodness of fit measures, model diagnostics and predictive accuracy.

5.2 Severity of Motorcycle Accidents

Motorcycles are the vehicles which offer a greater sense of freedom and openness than any other vehicle. Being more fuel efficient than other vehicles, motorcycles are more preferred by people as a cost effective mode of transportation in developing countries like Sri Lanka. Motorcycle accidents are most commonly serious and hazardous that can not only cause serious damage in injuries but also can take away the life. Therefore, it is important to find the risk factors associated with motorcycle accidents.

Here, also dependent variable is dichotomous and categories are grievous and non-grievous. Following Figure 5.1 presents grievous and non-grievous accidents occurred during the period 2014-2016.

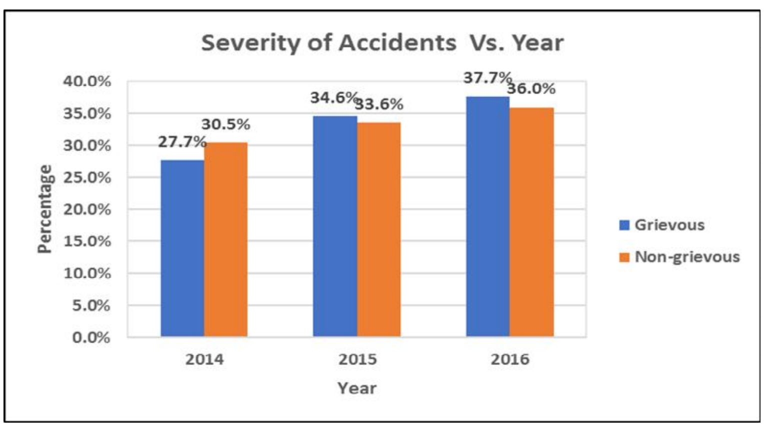

Figure 5.1: Severity of accidents vs. year

Figure 5.1 clearly shows that percentage of grievous and non-grievous accidents increased from 2014 to 2016. The important thing is the increment of grievous accident percentage. It is increased by 10% more in 2016 than in 2014.

5.1 Factors Associated with Severity of Accidents

Table 5.1: Frequency Table for accident severity and risk factors

Factor	Levels	Grievous		Non-grievous	
		Count	Percentage	Count	Percentage
Gender	Male	13691	97.9	18450	97.3
	Female	282	2.1	503	2.7
Validity of License	With valid license	7586	54.3	12263	64.7
	Without valid license	6387	45.7	6690	35.3
Accident Cause	Speeding	2551	18.3	2663	14.1
	Aggressive/negligent driving	9741	69.7	13938	73.5
	Influenced by alcohol/drugs	750	5.4	1047	5.5
	Fatigue/fall asleep	80	0.6	43	0.2
	Others	851	6.0	1262	6.7
Alcohol Test	No alcohol/below legal limit	13083	93.6	17731	93.5
	Over legal limit	890	6.4	1222	6.5
Time	Day time	8294	59.4	12705	67.1
	Night time	5679	40.6	6248	32.9
Weekday/ Weekend	Weekday	9541	68.3	13212	69.7
	Weekend	4432	31.7	5741	30.3
Road surface	Dry	13287	95.1	18135	95.7
	Wet	642	4.6	759	4.0
	Others	44	0.3	59	0.3
Weather Condition	Clear	13302	95.2	18157	95.8
	Rainy	393	2.8	488	2.6
	Others	278	2.0	308	1.6
Light condition	Daylight	8102	57.9	12440	65.6
	Night, Good street lighting	632	45	908	4.8
	Night, no street lighting	4759	34.1	5024	26.5
	Dusk/dawn	480	3.4	581	3.1
Location	Road	10956	78.4	14059	74.2
	Bend/Junction	3017	21.6	4894	25.8

Table 5.1 indicates that factors associated with accident severity. It shows that major cause for both grievous and non-grievous accidents is aggressive/negligent driving which contribute more than 70% towards these accidents. Second major cause is speeding. However, it contributes for grievous and non-grievous accidents 18.3% and 14.1% respectively. Percentages of influence by alcohol/drugs and fatigue/fall asleep are very less when compare with aggressive/ negligent driving.

Majority of accidents occurred in weekdays and daytime. More accidents are occurred on dry surfaces than clear and other surfaces. It also reveals that the percentage of grievous and non-grievous accidents occurred in clear weather is higher than that of rainy and other conditions. Majority of grievous and non-grievous accidents occurred under daylight. It also indicates that majority of grievous and non-grievous accidents were occurred, while the vehicle was moving on a straight road.

5.2 Distribution of Age of Driver

Driver's age is the only one continuous variable in this dataset. The distribution and descriptive statistics of age of drivers at fault is shown in the following Figure 5.2 and Table 5.2 respectively.

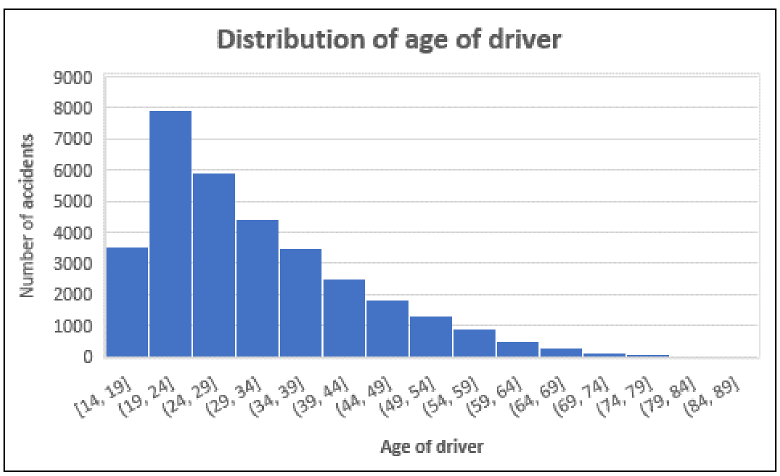

Figure 5.2: Distribution of age of driver

Table 5.2: Descriptive statistics of age of driver

	Mean	Std. Deviation	Minimum	Maximum	Q1	Q3
Age of Driver	31.7	11.95	14	88	22	38

Table 5.2 shows that average age of driver at fault for accident is 31.7 years. Furthermore, even though valid age for issuing a driving license is 18 years, above Table indicates that minimum age of driver at fault for accident is 14 years. To get an idea of which age range of drivers are more responsible for accidents, histogram is drawn as above.

According to the Figure 5.2, the drivers in between the age group 19-24 were more responsible for the highest number of accidents which is approximately 8000 accidents. Furthermore, it indicates that younger drivers whose age less than 30 years were accounted for 60% of total accidents. it should be noted that younger drivers were more responsible for the highest number of motorcycle accidents.

5.3 Reduction of Levels of Factors

It is better to have as few levels of factors as possible for easy interpretation. Therefore, one sample proportion test is used to reduce the levels of factors that are influencing the severity of accidents. The summary statistics of proportion test is listed in the Table 5.3.

Table 5.3: Summary statistics for Proportion Test

Factor	**Levels**	proportion	95% Confidence Interval	
			Lower	Lower
Gender	Male	0.974	0.973	0.977
	Female	0.024	0.022	0.025
Validity of License	With valid license	0.603	0.601	0.608
	Without valid license	0.397	0.391	0.401
Accident Cause	Speeding	0.158	0.154	0.162
	Aggressive/negligent driving	0.719	0.716	0.723
	Influenced by alcohol/drugs*	**0.054**	**0.052**	**0.057**
	Fatigue/fall asleep*	**0.004**	**0.003**	**0.004**
	Others*	**0.064**	**0.061**	**0.066**
Alcohol Test	No alcohol/below legal limit	0.936	0.933	0.938
	Over legal limit	0.064	0.061	0.066
Time	Day time	0.638	0.634	0.642
	Night time	0.362	0.359	0.365
Weekday/ Weekend	Weekday	0.691	0.688	0.694
	Weekend	0.309	0.305	0.312
Road surface	Dry	0.954	0.952	0.956
	Wet*	**0.042**	**0.040**	**0.045**
	Others*	**0.003**	**0.0026**	**0.0037**
Weather Condition	Clear	0.955	0.953	0.958
	Rainy*	**0.027**	**0.025**	**0.028**
	Others*	**0.018**	**0.016**	**0.019**
Light condition	Daylight	0.624	0.620	0.628
	Night, Good street lighting*	**0.047**	**0.044**	**0.049**
	Night, no street lighting	0.297	0.294	0.301
	Dusk/dawn*	**0.032**	**0.030**	**0.034**
Location	Bend/Junction	0.239	0.235	0.244
	Road	0.759	0.755	0.764

Based on the results of Table 5.3, some factor levels can be neglected or merged because of their small proportion. Influenced by alcohol/drugs, fatigue/fall asleep, others under accident cause were merged and generated a new classification called ‘Others’. Moreover, wet and others under road surface, were merged and named as ‘Others’. Rainy and others under weather condition were merged and named as ‘Others’. Good street lighting and dusk/dawn under light condition were merged and named as ‘Others’.

5.4 Results of Pearson Chi-Square Test

Pearson Chi-square test, which performed to check whether there exist or not a significant relationship between the independent variables and a dependent variable. According to this analysis, following table describes the association between each variable and the severity.

Table 5.4: Results of Chi-square test

Variables	**χ^2 value**	**P value**
Gender of Driver	$\chi^2(1) = 13.969$	0.000
Validity of License	$\chi^2(1) = 364.172$	0.000
Alcohol Test	$\boldsymbol{\chi^2(1) = 0.082}$	**0.775**
Accident Cause	$\chi^2(2) = 107.181$	0.000
Time	$\chi^2(1) = 205.187$	0.000
Weekday/Weekend	$\chi^2(1) = 7.678$	0.006
Road Surface	$\chi^2(1) = 6.5$	0.011
Weather	$\chi^2(1) = 6.853$	0.009
Light Condition	$\chi^2(2) = 229.953$	0.000
Location	$\chi^2(1) = 78.846$	0.000

According to the results of Table 5.4, it can be identified that gender of driver, validity of license, accident cause, time, weekday/weekend, road surface, weather, light condition and location type are significantly associated with the severity. Only one variable namely alcohol test is not significantly associated with the severity. Therefore, non-significant variable is not considered for binary logistic regression analysis.

5.7 Detecting Multicollinearity in Binary Logistic Regression

In this section, multicollinearity among explanatory variables is checked using Pearson's correlation coefficient, tolerance and VIF values.

5.7.1 Correlation Coefficients of Explanatory Variables

The correlation coefficients among the explanatory variables can be used as first step to identify the presence of multicollinearity (Field, 2009). Correlation matrix of highly correlated explanatory variables presented in Table 5.5 and remained presented in the Appendix as large number of dummy variables are exist.

Table 5.5: Pearson Correlation matrix between 2 explanatory variables

Variables		Time		Weather Condition		Light Condition
		DT	NT	CL	RA	GSL
Light Condition	DL	**0.971** (0.000)	**-0.971** (0.000)	0.095 (0.124)	-0.095 (0.124)	**-0.837** **(0.000)**
	GSL	**-0.862** (0.000)	**0.862** (0.000)	-0.092 (0.247)	0.092 (0.247)	1.000 (0.000)
Road Surface	D	0.088 (0.164)	-0.088 (0.164)	**0.966** (0.000)	**-0.966** (0.000)	-0.085 (0.321)
	W	-0.088 (0.164)	0.088 (0.164)	**-0.966** (0.000)	**0.966** (0.000)	0.085 (0.326)

Cell value: correlation coefficient

p value

Table 5.5 shows that the correlation coefficients between variables light and time as well as road surface and weather are highly correlated with each other and indicated them as bold. These high correlation coefficients signify the presence of severe multicollinearity between the explanatory variables light condition and time of accident as well as road surface and weather condition.

5.7.2 Detection of Multicollinearity based on Collinearity Statistics

Table 5.6 indicates the collinearity statistics for each levels of factors.

Table 5.6: Collinearity statistics

Factor	Model	Collinearity Statistics	
		Tolerance	VIF
Validity of license	WL	.945	1.023
	WOL	.985	1.016
Time	**DT**	**.045**	**19.456**
	NT	**.050**	**20.181**
Weekend/Weekday	WE	.993	1.102
	WD	.997	1.003
Location	RD	.997	1.004
	BJ	.994	1.006
Gender	M	.994	1.006
	F	.993	1.004
Accident cause	Cause1	.971	1.030
	Cause2	.965	1.023
	Others	.959	1.043
Road surface	**D**	**.059**	**15.457**
	Others	**.067**	**14.927**
Weather	**CL**	**.061**	**15.451**
	Others	**.067**	**14.944**
Light condition	**DL**	**.052**	**19.314**
	NSL	**.057**	**18.654**
	Others	.186	5.364
Age	Age	.984	1.017

Table 5.6 observes the high tolerances for the variables gender, validity of license, accident cause, weekday/weekend, location and age of driver but very low tolerances for the variables time, light condition, road surface and weather condition. Similarly, the variance inflation factor (VIF) corresponding to the explanatory variables gender, validity of license, accident cause, weekday/weekend, location and age of driver are very close to 1, but for variables time, light condition, road surface and weather condition, the VIF are larger than 2.5. Using these collinearity statistics, it can be concluded that the data almost certainly indicates a serious collinearity problem.

5.8 Solutions to Multicollinearity

Once the collinearity between variables has been identified, the next step is to find solutions in order to remedy this problem. Thus, first, variables time and road surface are removed from the data and repeat the analysis. However, collinearity still exists among the levels of light variable and weather condition. Then time and road surface variables are added and light condition and weather are removed from the analysis and repeat the analysis. Results are presented in Table 5.7.

Table 5.7: Collinearity statistics of remained variables

Factor	Model	Collinearity Statistics	
		Tolerance	VIF
Validity of license	WL	.978	1.010
	WOL	.981	1.020
Time	DT	.980	1.028
	NT	.975	1.026
Weekend/Weekday	WE	.998	1.002
	WD	.997	1.003
Location	RD	.998	1.004
	BJ	.997	1.003
Gender	M	.993	1.007
	F	.991	1.005
Accident cause	Cause1	.639	1.565
	Cause2	.633	1.580
	Others	.638	1.560
Road surface	D	.993	1.004
	Others	.992	1.008
Age	Age	.984	1.017

According to Table 5.7, tolerances for all the predictors are very close to 1 and all the VIF values are smaller than 2.5. Therefore, it can be concluded that multicollinearity is not a concern anymore.

5.9 Checking Linearity Assumption

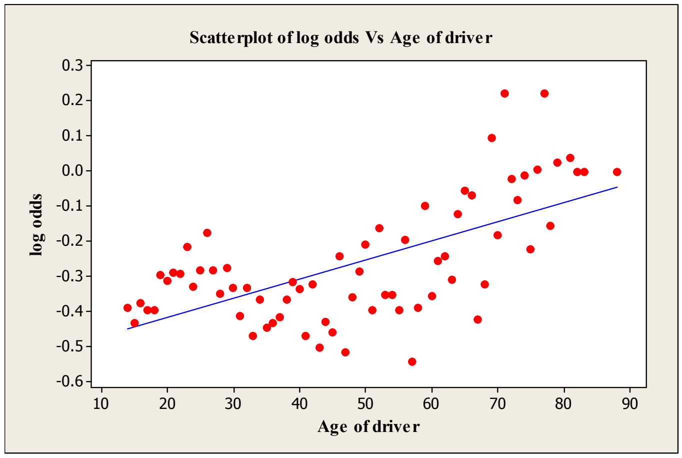

Figure 5.3: Log odds of severity Vs. Age of driver

The linear relationship between continuous explanatory variable age and the logit of the dependent variable is checked by visually inspecting the scatter plot and is shown in the Figure 5.3. It shows that variable age is linearly associated with the severity in logit scale. Thus, binary logistic regression is carried out for further analysis.

5.10 Binary logistic regression analysis

Since the response variable is dichotomous (grievous/non-grievous), the binary logistic regression model is applied to fit the data. According to the methodology, main vehicle dataset (32926) was divided into two portions; 2/3 of data (21950) was used to develop the model, and the remaining 1/3 of data (10976) was used to validate the model (Dhananjaya & Alibuhtto, 2016).

5.10.1 Baseline Model

Table 5.8 presents the results of the baseline model which is the model with only the constant included before any explanatory variables are entered into the model. Logistic regression compares this model with a model including all the predictors to determine whether the latter model is more appropriate.

Table 5.8: Results of baseline model

Baseline Model	B	S.E.	Wald	df	Sig.	Exp(B)
Constant	.325	.014	563.373	1	.000	1.384

Initial -2 Log Likelihood: 28858.35

Table 5.8 indicates that the coefficient of constant is 0.325 and standard error of the coefficient for the constant is 0.014. Wald Chi-Square tests the null hypothesis that the constant equals 0. Results of the above table show that constant is statistically significant to the model. Initial log likelihood value of the baseline model is 28858.35. This value is used to select an optimal model. Predictive power of the baseline model is 56.6%, which indicates the overall percentage of correctly classified cases when there are no explanatory variables in the model.

5.10.2 Developed model

Forward selection method is used to develop the binary logistic model. First, all variables were entered into the model according to their -2 log likelihood ratio. Variable which has minimum -2 log likelihood ratio is added first. Model iteration occurred up to five steps. The analysis was performed on p value = 0.05 significance level to formulate the model.

5.10.3 Developed model interpretation

Table 5.9: Variables in the model

Variables	B	S.E.	Wald	df	Sig.	Exp(B)	95% CI.for EXP(B)	
							Lower	Upper
Cause2	.426	.028	226.060	1	.000	1.531	1.492	1.621
NT	-.315	.029	119.342	1	.000	.730	.690	.772
Age	-.059	.030	3.914	1	.048	.942	.889	.999
BJ	-.255	.033	60.402	1	.000	.775	.726	.826
M	.392	.055	60.394	1	.000	1.480	1.329	1.647
Constant	.592	.047	161.535	1	.000	1.807		

Table 5.9 explains the variables in the developed model used to predict the severity of accidents. When exploring results of this table, location type, time, age of driver, accident

cause and gender have a significant effect on the severity of accidents. Wald chi-square tests the effect of individual predictor while controlling other predictors.

Odds of a grievous accident occurred by aggressive/negligent driving is 53% more likely to be a grievous accident occurred by speeding. Odds of a grievous accident occurred in night time is 73% less likely to be a grievous accident occurred in day time. Odds of a grievous accident occurred in bend or junction is 77% less likely to be a grievous accident occurred in road. Odds of a grievous accident occurred by male motorcyclist is 48% more likely to be a grievous accident occurred by female motorcyclist. The odds ratio of age is 0.942. It indicates that for every one unit increase in age (one additional year of living), the odds of occurring a grievous accident decreases which implies the older the motorcyclist, the less the accident risk.

5.10.4 Logit Model

From the analysis, the logit model with the significant variables is:

$\text{logit}(p) = 0.592 + 0.426\text{Cause2} - 0.315\text{NT} - 0.059\text{Age} - 0.255\text{BJ} + 0.392\text{M}$

5.10.5 Variables not in the Developed Model

Table 5.10: Variables not in the Model

Variables	Score	df	Sig.
Weekday/Weekend	1.520	1	.218
Validity of license	.616	1	.433
Road Surface	3.142	1	.076
Overall Statistics	5.334	3	.149

Table 5.10 indicates that variables which are not in the model. According to the Table, it describes that weekday/weekend, validity of license and road surface factors are not significantly associated with the severity of accidents. (p =0.218, 0.433, 0.076 >0.05 respectively).

5.10.6 Importance of Variables in the Model

Table 5.11: Importance of variables in the model

Step	Improvement			Model			Variable
	Chi-square	df	Sig.	Chi-square	df	Sig.	
1	254.230	1	.000	254.230	1	.000	IN: Accident Cause
2	119.363	1	.000	373.593	2	.000	IN: Time
3	62.783	1	.000	436.376	3	.000	IN: Location
4	60.250	2	.000	496.627	5	.000	IN: gender
5	3.910	1	.048	500.537	6	.000	IN: Age

Table 5.11 presents the information how the model is affected if an explanatory variable is added to the model. In other words, which variable is important for the model. Results of following table are used to examine the importance of a variable in the model. According to the results, adding the variable accident cause to the model makes the biggest change in the model's log likelihood value. Therefore, accident cause is the most important variable in this model. It is followed by the age of driver, location type, time and gender respectively.

5.11 Measures of Goodness of Fit

This section presents measures of goodness of fit of the model. Under the measures, test of model coefficients, model summary and predictive accuracy of the developed model are discussed.

5.11.1 Test of Model coefficients

The test of model coefficients is used to check that the new model (with explanatory variables included) is an improvement over the baseline model. It uses chi-square test to see if there is a significant difference between the Log-likelihoods (specifically the -2LLs) of the baseline model and the new model. If the new model has a significantly reduced -2LL compared to the baseline then it suggests that the new model is explaining more of the variance in the outcome and is an improvement.

Table 5.12: Test of developed model coefficients

	Chi-square	df	Sig.
Step 5	3.910	1	.048
Block	500.537	6	.000
Model	500.537	6	.000

Table 5.12 indicates that chi-square of the model is highly significant (chi-square=500.537, p=0.000 with df =6). Thus, it can be concluded that the developed model is significantly better than the baseline model. That means the accuracy of the model improved when added the explanatory variables.

5.11.2 Model Summary

Under the model summary section, developed model will be checked whether it is an improvement over the baseline model. Results are shown in the Table 5.13.

Table 5.13: Model Summary

Step	-2 Log likelihood	Cox & Snell R^2	Nagelkerke R^2
5	28357.808	.621	.643

According to the results of the Table 5.13, the developed model has a significantly reduced log likelihood value (28357.808) compared to the baseline model. It is revealed that the developed model is explaining more of the variance in the outcome and it is an improvement over the baseline model. Thus, it can be concluded that the developed model is better at predicting the severity of the accidents than the baseline model where no predictor variables were added.

In addition to that, Table 5.13 contains the Cox & Snell R^2 and Nagelkerke R^2 values, which are used to calculate the explained variation. According to these both values, the explained variation in the dependent variable based on the model is 62.1% and 64.3% respectively.

5.11.3 Predictive Accuracy of Developed Model

Classification table shows how many of the cases where the observed values of the dependent variable have been correctly predicted. Results of the classification table are shown in the Table 5.14.

Table 5.14: Classification Table

Observed	Predicted		
	Grievous	Non-grievous	Percentage Correct
Grievous	5269	3940	57.21
Non-grievous	1692	11049	86.72
Overall Percentage			74.34

Table 5.14 indicates that 57.2% were correctly classified for grievous accidents and 86.7% for non-grievous accidents. Overall 74.3% were correctly classified. It can be seen that the developed model is correctly classifying the outcome for 74.3% of the cases compared to 56.6% in the null model.

5.11.4 Hosmer and Lemeshow test

Hosmer and Lemeshow test is used to indicate which extent to the estimated model provides a better fit to the data than the null model.

Table 5.15: Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
5	8.602	8	.377

As the results shown in the Table 5.15, Hosmer & Lemeshow test of the goodness of fit suggests the model is a good fit to the data as $p=0.377$ (>0.05).

5.11.5 Receiver Operating Characteristic Curve

Receiver operating characteristic (ROC) curve is used to evaluate the fit of a logistic regression model. ROC curve is shown in the Figure 5.4 and the area under curve is presented in Table 5.16.

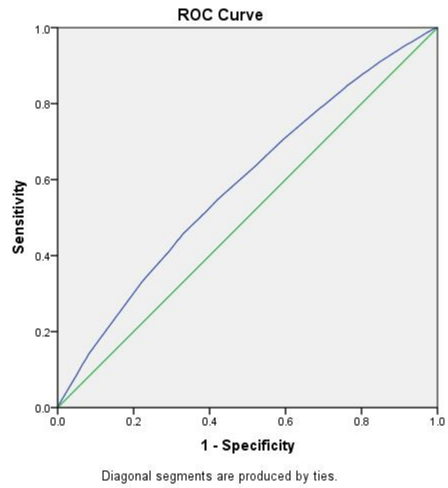

Figure 5.4: ROC Curve

Table 5.16: Area Under the Curve

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.587	.004	.000	.579	.594

According to the above Table and Figure, the area under the curve is 0.587 with 95% confidence interval (0.579, 0.594). Moreover, the area under the curve is significantly different from 0.05 since the p =0.000 <0.05. That means, the logistic regression classifies the group significantly better than by chance.

5.12 Model Diagnostics

Two assumptions of binary logistic regression, linearity and multicollinearity were checked earlier. Assumptions of independent errors and influential observations are checked under this section.

5.12.1 Testing Residuals

Table 5.17 provides two measures that can be used to assess how well the model fits the data, as shown below.

Table 5.17: Testing Residuals

Method	Chi-Square	df	Sig.
Pearson	3101.57	3012	0.125
Deviance	3251.312	3012	0.084

Table 5.17 indicates that p values of both tests are greater than 0.05 which implies that residuals are not statistically significant. Based on these measures, the model fits the data well.

5.12.2 Checking Normality in Residuals

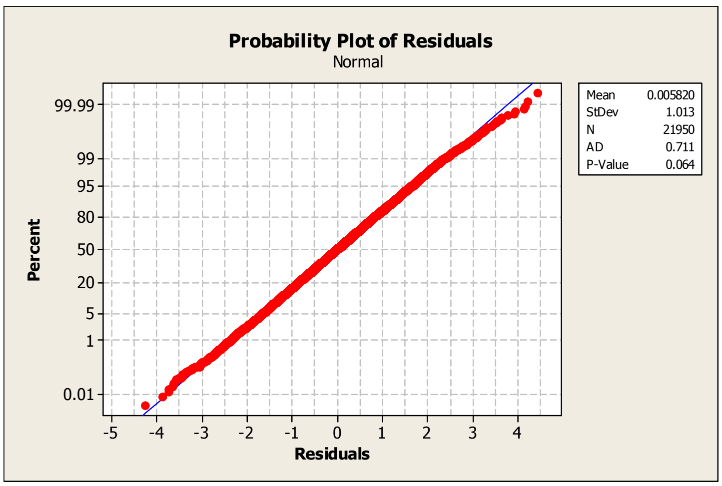

Figure 5.5: Normal probability plot of residuals

Then, residuals are checked whether normally distributed and displayed in Figure 5.5. It indicates that residuals are normally distributed since p=0.064>0.05.

5.12.3 Testing Heteroskedasticity and Correlation in Residuals

Heteroskedasticity in residuals is tested from Lagrange Multiplier (LM) test. Correlation in residuals is tested using Durbin Watson (DW) test. Results of both tests are displayed in Table 5.18.

Table 5.18: DW and LM test of residuals

Test	Test Statistic	p value
DW Test	1.724	0.073
LM Test	12.54	0.076

Results of Table 5.18 illustrates that residuals have constant variance as p = 0.076> 0.05. Moreover, residuals are uncorrelated as p=0.073>0.05.

5.12.4 Detecting Influential Observations

Plot of Cook’s distance is used to detect influential observations and displayed in Figure 5.6.

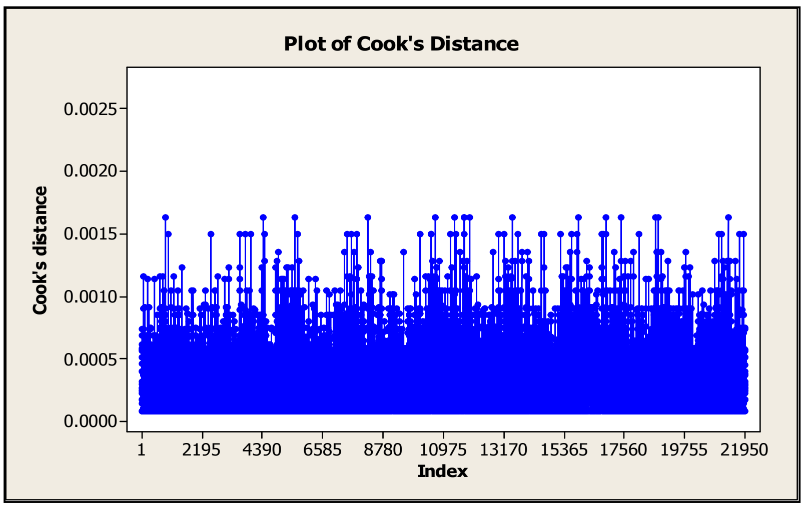

Figure 5.6: Plot of Cook's Distance

According to the Figure 5.5, it can be seen that all observations are less than 1 and even less than 0.5. Therefore, it can be said that there are neither outliers nor influential observations in this data set.

5.12.5 Validation of the Model

Developed logistic regression model is used to validate the model. One third of data is used for validation and results are shown in the following Table 5.19.

Table 5.19: Category prediction

Observed	Predicted		
	Grievous	Non-grievous	Percentage Correct
Grievous	3513	1251	73.74
Non-grievous	990	5222	84.06
Overall Percentage			79.58

Table 5.19 clearly indicates that the model correctly predicted 79.6% of the validation data which is greater than to the predictive power of the baseline model 56.6%. That means the developed model more accurately predicts the severity of accidents than the prediction in baseline model.

5.13 Summary

This section summarizes the results obtained by analyzing motorcycle accidents occurred in Sri Lanka. According to the logistic regression model, location type, time, age of driver, accident cause and gender have a significant effect on the severity of accidents. Among them, location type, accident cause and gender have an increasing effect on the probability of a grievous accident. Time and age of driver have a decreasing effect on the probability of a grievous accident. Straight road, aggressive/negligent driving, drive by male motorcyclists, daytime have a high chance to be a grievous accident. Moreover, the older motorcyclists have less accident risk.

CONCLUSIONS AND RECOMMENDATIONS

6.1 Overview

This chapter highlights the summary of findings of the study. It also presents recommendations for future studies.

6.2 Conclusions and Discussion

The main objective of this study was to identify the significant factors affecting for motorcycle and motor vehicle accidents in Sri Lanka. Based on this main objective, three years accident data during the period 2014 to 2016 obtained from the Traffic Police Head Quarters in Sri Lanka. The conclusions achieve from this study are summarized as below.

Motorcycles are found to have a higher probability of causing road accidents in Sri Lanka. Therefore, motorcycle accidents are analyzed and identified significant factors separately. Then motor vehicles are classified into 2 categories as light vehicles and heavy vehicles. Majority of grievous and non- grievous motor vehicle accidents reported by light vehicles. When motorcycles data excluded from vehicle accidents, highest accident percentages recorded by three wheelers and cars.

One of important exposure in descriptive statistics is the highest number of motorcycle and motor vehicle accidents reported due to aggressive /negligent driving. Thus, this is a great teaser of drivers in Sri Lanka. Similarly, highest number of motor vehicle accidents reported by the drivers in between 29 - 39 years old. Highest number of motorcycle accidents reported by the drivers in between 19-24 years old. It is convinced that young motorcyclists are more influential in road accidents. It may be due to most of young motorcyclists have not satisfactory experiencing in driving, lack of relevant knowledge and lack of tolerance. Most of motorcycle and motor vehicle accidents occurred in daytime under daylight on weekdays. In addition to that, majority of motorcycle and motor vehicle accidents were occurring in dry road surface under clear weather condition on a straight road.

Then, according to the Pearson chi-square test, it was concluded that vehicle type, gender of driver, validity of license, alcohol test, accident cause, time, weekday/weekend, light condition and location type are significantly associated with the severity of motor vehicle accidents. Only two variables such as road surface and weather, are not significantly associated with the severity. Furthermore, for motorcycle accidents it was concluded that gender of driver, validity of license, accident cause, time, weekday/weekend, road surface, weather, light condition and location type are significantly associated with the severity. Only one variable namely alcohol test is not significantly associated with the severity.

For motor vehicle accidents, light condition is removed from the analysis due to presence of multicollinearity and for motorcycle accidents, light and weather condition factors are removed.

Based on the binary logistic regression analysis results, weekday/weekend variable is not significantly associated with the severity of motor vehicle accidents. Thus, remaining factors namely vehicle type, validity of license, time of accident, location type, alcohol test, accident cause, age of the driver and gender have found a significant effect on the severity of accidents. Moreover, bend or junction location, aggressive/negligent driving, drive at daytime, driving light vehicle and drivers who use alcohol below legal limit or no alcohol, have a high chance to be a grievous accident. Also, the older drivers have less accident risk. Finally, it is concluded that accident cause is the most important variable in the model. This is an issue which needs high level attention from drivers and high commitment by traffic police. Thus, not only the government of Sri Lanka, but also the drivers is reflected a great responsibility to reduce road accidents, and control this ambience.

For motorcycle accidents, road surface, validity of license and weekday/weekend factors are not significantly associated with the severity of accidents. Accident cause, age of driver, location, time and gender factors have found a significant effect on the severity of accidents. Drive on straight road, aggressive/negligent driving, drive at daytime have a high chance to be a grievous accident. Moreover, the older motorcyclists have less accident risk. There is a tendency for more accidents to take place at the hands of young drivers who are inexperienced or poorly trained. Therefore, drivers must be well trained before they are allowed to drive.

6.3 Recommendations

There are many factors that contributed to this kind of escalation of road accidents such as inadequate safety precautions, increasing number of vehicles on the roads, failure of observing and following road signs, lack of knowledge and experience, irresponsible driving, inadequate safety standards in vehicles, etc., Therefore, it is time to take actions to reduce road accidents. Following recommendations are presented based on this study.

- Educate drivers especially young drivers and young motorists on traffic safety, rules and regulations.
- Renew driving license frequently (once in 2 years) and the knowledge on rules and regulations, health and driving skills can be checked during the renewal of the license.
- Current preventive strategies review again and modify them according to current situation.
- Conduct road safety audit on every road to identify road accidents in terms of where, when and why.

This study considered factors influencing by drivers only. It is better to perform a research by considering the whole database into distinct models such as "Factors influencing by pedestrians", "Factors influencing by drivers", "Factors influencing by vehicles" and so on.

Furthermore, priority is always given to the accommodation of more vehicles and widening roads, less importance is placed on people's safety. Similarly, insurance should not focus only on the vehicle but should also take care of the human factor. It needs to change the current culture and place more value on human life. The public, insurance companies and law authorities should collaborate and play a more responsible role to minimize the number of road accidents and learn to place more value on human life which is an invaluable gift.

REFERENCE LIST

Al-Ghamdi, A., (2002). Using logistic regression to estimate the influence of accident factors on accident severity. *Accident Analysis and Prevention*, Volume 34, pp. 729-741.

Allison, P., (2014). Measures of fit for logistic regression. *SAS Global Forum* Statistical Horizons LLC, pp 60-72.

Amarasingha, N., & Dissanayake, S (2013). Modelling frequency of truck crashes on limited access highways. Journal of the Transportation Research Forum, 123-139.

Baruah, A. & Chaliha, R., (2015). A study of incidence of alcohol use in fatal road traffic accidents. *Journal of Indian Academy of Forensic Medicine,* 37(1), pp. 12-15.

Bursac, Z., Gauss, C., Williams, D. & Hosmer, D., (2008). Purposeful selection of variables in logistic regression. *Source Code for Biology and Medicine,* pp. 3-17.

Celik, A. & Oktay, E., (2014). A multinomial logit analysis of risk factors influencing road traffic injury severities in the Erzurum and Kars Provinces of Turkey. *Accident Analysis and Prevention,* pp. 66-77.

Chen, Y. et al., (2016). Differences in factors affecting various crash types with high numbers of fatalities and injuries in China. *PLOS One,* pp. 1-12.

Chengye, P., & Ranjitkar, P. (2013). Modelling motorway accidents using Negative Binomial Regression. Journal of Eastern Asia Society for Transportation Studies,10,1946-1963.

Dhananjaya, S. & Alibuhtto, M., (2016). Factors Influencing Road Accidens in Sri Lanka.: A Logistic Regression Approach. *Annual Science Research Sessions-South Eastern University of Sri Lanka*, pp. 157-173.

Dissanayaka, M. & Kulatunga, D., (2012). Analysis of fatalities in road accidents considering Peliyagoda police area in Gampaha. *Annual Science Research Sessions-University of Kelaniya*, p. 45.

Giancristofaro, R. & Salmaso, L., (2003). Model performance analysis and model validationnin logistic regression. *STATISTICA anno LXIII,* pp. 375-396.

Haadi, A., (2014). Identification of Factors that Cause Severity of Road Accidents in Ghana: a Case Study of the Northern Region. *International Journal of Applied Science and Technology*, May, pp. 242-249.

Hosmer, D., Hosmer, T., Lemeshow , S. & Cessie , L., (1997). A comparison of goodness of fit tests for the logistic regression model. *Statistics in medicine,* pp. 965-980.

Jeepara, P. & Pirasath, S., (2011). Road traffic accidents in Eastern Sri Lanka: An analysis of admissions and outcome. *Sri Lanka Journal of Surgery,* 29(2), pp. 72-76.

Komba, D. D., (2006). Risk factors and road traffic accidents in Tanzania: A case study of Kibaha District*, Trondheim: Norwegian University of Science and Technology,* pp 24-33.

Kumarage, A. S., Wickramasingha, S. M. & Jayaratne, M. D., (2003). Analysis of road accidents in Sri Lanka*. Supreme Group of Companies*, pp 56-64.

Liyanage, T. & Rengarasu, T., (2015). Traffic Accident Analysis and Development of Accident Prediction Model*.The Institution of Engineers, Sri Lanka*, pp. 105-109.

Lund, B. & MI, D., (2015). Fitting and evaluating logistic regression models*. Marketing Associates, LLC, pp* 105-115.

Midi, H., Sarkar, S. & Rana, S., (2013). Collinearity diagnostics of binary logistic regression model. *Journal of Interdisciplinary Mathematics,* pp. 253-267.

Mohamed Aslam, A., (2015). An Econometric Analysis of Trafffic Accident in Sri Lanka. pp. 1077-1084.

Rana, S., Midi, H. & Sarkar , S.(2005). Validation and Performance Analysis of Binary Logistic Regression Model*. Proceedings of the WSEAS International Conference on ENVIRONMENT, MEDICINE and HEALTH SCIENCES,* pp. 51-55.

Renuraj, S., Varathan, N. & Satkunananthan, N., (2015). Factors Influencing Traffic Accidents in Jaffna. *Sri Lankan Journal of Applied Statistics,* September, 16(2), pp. 117-133.

Robin, P., (2014). Use on multinomial logistic regression in work zone crash analysis for Missouri work zones*. Missouri University of Science and Technology*, pp 45-61.

Sarkar, S., Midi, H. & Rana, S., (2011). Detection of outliers and influential observations in binary logistic regression: An empirical study. *Journal of Applied Sciences,* pp. 26-35.
Seiss, M., (2009). Logistic and Poisson Regression: Modeling Binary and Count Data. *VirginiaTech.*

Senadeera, W. A., (2016). Analyzing traffic accidents in Gampaha district. *Academic Sessions- University of Sri Jayewardenepura,* pp 67-70.

Senasinghe, A.P., Wirasinghe, S.C., & De Barros, A.G (2017). Road safety issues on two major intercity highways in Sri Lanka (A001 and A004)- Preliminary analysis. R4TLI Conference Proceedings, pp.201-206.

Shruthi, P. et al., (2013). Analysis of fatal road traffic accidents in a Metropolitan City of South India. *Journal of Indian Academy of Forensic Medicine,* 35(4), pp. 317-320.

Singh, D., Moorthi, K., Singh, S. & Goel, S., (2014). Profile of road traffic fatalities in adults: A 40 year study in Chandigarh Zone of North West India. *Journal of Indian Academy of Forensic Medicine,* 36(1), pp. 47-51.

Singh, H. & Aggarwal, A. D., (2010). Fatal road traffic accidents among young children. *Journal of Indian Academy of Forensic Medicine,* 32(4), pp. 286-288.

Singh, R. et al., (2013). Elucidation of risk factors in survivors of road traffic accidents in North India. *Hard Tissue,* 2(1), pp. 1-6.

Somasundaraswaran , A. K., (2006). Accident statistics in Sri Lanka. *International Association of Traffic and Safety Sciences Research,* pp. 115-117.

Tanaboriboon , Y. & Satinnam , T., (2005). Traffic accidents in Thailand. *IATSS Research,* 29(1), pp. 88-100.

Wedagama, D. & Dissanayake, D., (2009). Analysing motorcycle injuries on Arterial roads in Bali using a multinomial logit model. *Journal of the Eastern Asia Society for Transportation Studies,* pp. 55-68.

Wedagama, D. & Dissanayake, D., (2009). The Influence of Accident Related Factors on Road Fatalities Considering Bali Province in Indonesia as a Case Study. *Journal of the Eastern Asia Society for Transportation Studies*, Volume 8.

Yan, X., Radwan, E. & Abdel-Aty, M., (2005). Characteristics of rear-end accidents at signalized intersections using multiple logistic regression model. *Accident Analysis and Prevention,* pp. 983-995.

Zhang, G., Yau, K. & Chen, G., (2013). Risk factors associated with traffic violations and accident severity in China. *Accident Analysis and Prevention,* pp. 18-25.

Zhu, D., (2014). Analysis of factors affecting motorcycle-motor vehicle crash characteristics*.* *Ohio University of Dayton*, pp 123-143.

APPENDIX

Results of Motor Vehicle Accidents

Correlation among the factors related to motor vehicle accidents

	veh type1	veh type2	gender1	gender2	license1	license2
veh type2	-1.000					
gender1	-0.078	0.078				
gender2	0.078	-0.078	-1.000			
license1	-0.099	0.099	0.007	-0.007		
license2	0.099	-0.099	-0.007	0.007	-1.000	
alcohol1	-0.095	0.095	-0.025	0.025	0.144	-0.144
alcohol2	0.095	-0.095	0.025	-0.025	-0.144	0.144
time1	-0.088	0.088	-0.034	0.034	0.040	-0.040
time2	0.088	-0.088	0.034	-0.034	-0.040	0.040
workday1	-0.058	0.058	-0.005	0.005	0.022	-0.022
workday2	0.058	-0.058	0.005	-0.005	-0.022	0.022
location1	-0.018	0.018	0.025	-0.025	-0.017	0.017
location2	0.018	-0.018	-0.025	0.025	0.017	-0.017
cause1	0.011	-0.011	-0.001	0.001	0.008	-0.008
cause2	-0.064	0.064	-0.008	0.008	0.084	-0.084
cause3	0.074	-0.074	0.012	-0.012	-0.120	0.120

	alcohol1	alcohol2	time1	time2	workday1	workday2
alcohol2	-1.000					
time1	0.129	-0.129				
time2	-0.129	0.129	-1.000			
workday1	0.040	-0.040	0.042	-0.042		
workday2	-0.040	0.040	-0.042	0.042	-1.000	
location1	-0.003	0.003	-0.003	0.003	-0.011	0.011
location2	0.003	-0.003	0.003	-0.003	0.011	-0.011
cause1	0.067	-0.067	-0.014	0.014	-0.013	0.013
cause2	0.366	-0.366	0.090	-0.090	0.033	-0.033
cause3	-0.555	0.555	-0.105	0.105	-0.030	0.030

	location1	location2	cause1	cause2
location2	-1.000			
cause1	0.017	-0.017		
cause2	-0.022	0.022	-0.667	
cause3	0.012	-0.012	-0.146	-0.640

Cell Contents: Pearson correlation

Results of Motorcycle Accidents

Correlation among the factors related to motorcycle accidents

	license1	license2	time1	time2	workday1	workday2
license2	-1.000					
time1	0.066	-0.066				
time2	-0.066	0.066	-1.000			
workday1	0.026	-0.026	0.039	-0.039		
workday2	-0.026	0.026	-0.039	0.039	-1.000	
location1	-0.034	0.034	-0.024	0.024	-0.009	0.009
location2	0.034	-0.034	0.024	-0.024	0.009	-0.009
gender1	-0.002	0.002	-0.069	0.069	-0.014	0.014
gender2	0.002	-0.002	0.069	-0.069	0.014	-0.014
cause1	-0.005	0.005	-0.003	0.003	-0.010	0.010
cause2	0.069	-0.069	0.065	-0.065	0.024	-0.024
cause3	-0.089	0.089	-0.085	0.085	-0.021	0.021
surface1	0.006	-0.006	0.088	-0.088	0.004	-0.004
surface2	-0.006	0.006	-0.088	0.088	-0.004	0.004
weather1	0.006	-0.006	0.094	-0.094	0.004	-0.004
weather2	-0.006	0.006	-0.094	0.094	-0.004	0.004
light1	0.069	-0.069	0.971	-0.971	0.041	-0.041
light2	-0.079	0.079	-0.862	0.862	-0.036	0.036
light3	0.010	-0.010	-0.282	0.282	-0.011	0.011

	location1	location2	gender1	gender2	cause1	cause2
location2	-1.000					
gender1	0.027	-0.027				
gender2	-0.027	0.027	-1.000			
cause1	0.015	-0.015	0.025	-0.025		
cause2	-0.015	0.015	-0.019	0.019	-0.694	
cause3	0.004	-0.004	-0.001	0.001	-0.162	-0.598
surface1	-0.015	0.015	-0.007	0.007	-0.018	0.024
surface2	0.015	-0.015	0.007	-0.007	0.018	-0.024
weather1	-0.017	0.017	-0.008	0.008	-0.012	0.020
weather2	0.017	-0.017	0.008	-0.008	0.012	-0.020
light1	-0.025	0.025	-0.068	0.068	-0.002	0.063
light2	0.048	-0.048	0.063	-0.063	-0.000	-0.058
light3	-0.035	0.035	0.015	-0.015	0.004	-0.016

	cause3	surface1	surface2	weather1	weather2	light1
surface1	-0.014					
surface2	0.014	-1.000				
weather1	-0.014	0.966	-0.966			
weather2	0.014	-0.966	0.966	-1.000		
light1	-0.084	0.088	-0.088	0.095	-0.095	
light2	0.079	-0.085	0.085	-0.092	0.092	-0.837
light3	0.017	-0.015	0.015	-0.015	0.015	-0.377

	light2
light3	-0.190

Cell Contents: Pearson correlation