ANALYSIS OF ROAD TRAFFIC ACCIDENTS IN LIMPOPO PROVINCE USING GENERALIZED LINEAR MODELLING

by

MODUPI PETER MPHEKGWANA

RESEARCH DISSERTATION

Submitted in fulfillment of the requirements for the degree of

Master of Science

In

Statistics

in the

FACULTY OF SCIENCE AND AGRICULTURE

(School of Mathematical and Computer Sciences)

at the

UNIVERSITY OF LIMPOPO

SUPERVISOR: PROF A TESSERA

CO-SUPERVISOR: MR N YIBAS

2020

DECLARATION

A research project submitted in partial fulfilment of the requirements for the degree of Masters of Science in Statistics by Research Report in the Faculty of Science and Agriculture, University of Limpopo, South Africa, 2019.

I declare that this research is my own, unaided work. It has not been submitted before for any other degree, part of degree or examination at this or any other university.

Mphekgwana MP **29 April 2019**

Surname, Intials (title) **Date**

DEDICATION

I dedicate my dissertation work to my family and many friends. A special feeling of gratitude to my beloved mother Moshibudi Rebecca Kgatle for her love, care and support; you always said work hard now and play later. With all this hard work, I guess it is my time to play. I also dedicate this work to my wife and children; Tania, Mogale and Ezra Mokgophi who has encouraged me all the way and whose encouragement has made sure that I give it all it takes to finish what I have started.

ACKNOWLEDGEMENT

Firstly, I would like to thank God for blessing me with life and for having given me the strength, wisdom, belief, and as well as for guiding me from my infancy to date.

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 supervisors, Prof. A. Tessera and Mr. N. Yibas for their patience, guidance, encouragement and support in shaping the outlook of this thesis. They provided invaluable insights that have guided my thinking and understanding. Thank you once again and may God bless you.

I am grateful to all the staff at the Department of Statistics and the Research Office f or their encouragement and facilitation.

Last but not the least, a special appreciation to my family for having supported me through all the decisions I have taken and allowing me the opportunity to study.

ABSTRACT

Background: Death and economic losses due to road traffic accidents (RTA) are huge global public health and developmental problems and need urgent attention. Each year nearly 1.24 million people die and millions suffer various forms of disability as a result of road accidents. This puts road traffic injuries (RTIs) as the eighth leading cause of death globally and RTIs are set to become the fifth leading cause of death worldwide by the year 2030 unless urgent actions are taken.

Aim: In this paper, we investigate factors that contribute to road traffic deaths (RTDs) in the Limpopo province of South Africa using models such as the generalized linear models (GLM) and zero inflated models.

Methods: The study was based on retrospective data that comprised of reports of 18,029 road traffic accidents and 4,944 road traffic deaths over the years 2009 – 2015. Generalized linear modelling and zero-inflated models were used to identify factors and determine their relationships to RTDs.

Results: The data was split into two categories: deaths that occurred during holidays and those that occurred during non-holiday periods. It was found that the following variables, namely, Monday, human actions, vehicle conditions and vehicle makes, were significant predictors of RTDs during holidays. On the other hand, during non-holiday periods, weekend, Tuesday, Wednesday, national road, provincial road, sedan, LDV, combi and bus were found to be significant predictors of road traffic deaths.

Conclusion: GLM techniques, such as the standard Poisson regression model and the negative binomial (NB) model, did little to explain the zero excess, therefore, zero-inflated models, such as zero-inflated negative binomial (ZINB), were found to be useful in explaining excess zeros.

Recommendation: The study recommends that the government should make more human power available during the festive seasons, such as the December holidays, and over weekends.

Key concepts: Poisson, ZIP, ZINB, NB, accidents, deaths, RTAs, RTDs, zeros.

TABLE OF CONTENTS

LIST OF ACRONYMS AND SYMBOLS

AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
EM	Expectation Maximization
GLM	Generalized Linear Model
MLE	Maximum Likelihood Estimation
NB	Negative Binomial
RTAs	Road Traffic Accidents
RTDs	Road Traffic Deaths
RTIs	Road Traffic Injuries
ZINB	Zero-Inflated Negative Binomial
ZIP	Zero-Inflated Poisson

CHAPTER 1: INTRODUCTION

1.1. INTRODUCTION

Transportation is the heartbeat of South Africa's economic growth and social development and allows both the development of internal and external merchandising. South African transport compromises of general transport, rail, civil aviation, shipping, motor vehicles and freight (Klynsmith, 2015).

Rapid population growth and urbanization has a dramatic effect on the increasing demand for transport. An increase in demand for transport increases the number of road traffic accidents (RTAs) (Zhang et al., 2006). RTAs cause economic loss to company owners, insurance companies and, subsequently, the country as a whole. RTAs might also result in the loss of lives, with some individuals suffering non-fatal injuries, while others may incur disabilities as a result of RTAs.

1.2. BACKGROUND OF THE STUDY

Globally, RTAs are a major cause of death and severe injuries (WHO, 2013). Each year nearly 1.24 million people die and millions suffer various forms of disability as a result of road accidents (Agyemang et al., 2013; WHO, 2013; Subhan, 2017). This puts road traffic injuries (RTIs) as the eighth leading cause of death globally, which is likely to increase to the fifth leading cause of death worldwide by the year 2030 unless urgent action is taken (Masuri et al., 2012; WHO, 2013 and Subhan, 2017)

Death rates due to road accidents are increasing rapidly in lower- and middle-income countries (Sharma, 2008). The social and economic costs of deaths and injuries due to RTAs are considerable. Road accidents in lower- and middle-income countries cost over US$ 100 billion each year (WHO, 2013). Furthermore, it was reported by the WHO (2013) that road traffic death (RTD) rates vary considerably from region to region. Africa, with only 2% of the world's vehicles, is the least motorised region of the world, but accounts for 16% of all global traffic deaths, with Nigeria and South Africa contributing the most to fatality rates in the region (WHO, 2013).

The Department of Roads and Transport (DOT, 2007), reported that South Africa has one of the worst road safety records in the world, recording road accident-related deaths of approximately 120,000 people per annum and injuries in excess of a million

people per annum. RTDs have increased from 25.1 fatalities per 100,000 people in 1994 to 30.3 fatalities per 100,000 people in 2008. Additionally, this annual road carnage costs the South African economy approximately R43 billion. Approximately 60% of these costs include damage to vehicles and other properties (Harris and Olukoga, 2005). People most affected by the consequences of these RTAs are young people, aged between 20 and 44 (Mohamed et al., 2009).

The Road Traffic Management Corporation (RTMC) (2016), reported that the number of road fatalities in South Africa increased by 10% between 2014 and 2015. Within South Africa, the Limpopo Province had the highest crude RTD rate for the period 2001-2006 (Lehohla, 2009). This shows that there is a need to analyse RTAs in the Limpopo Province in order to identify the important factors that contribute to RTDs.

1.3. STUDY SITE

The province of Limpopo is the northernmost province of South Africa. Statistics South Africa's Census 2011, showed that the Limpopo Province comprises 125,755 square kilometres of the country's total land area (StatsSA, 2012). It is the fifth largest of the country's nine provinces, accounting for 10.3% of South Africa's total land area.

Figure 1: **South Africa map (source: Wikipedia image)**

According to the 2011 census report, 5 404 868 people live in Limpopo, constituting 10.4% of South Africa's total population. The majority of the people living in the province were born in the province (91%), while 3% of the people living in the province were born outside of South Africa (StatsSA, 2012). Furthermore, at the time of the

2011 census, 34% of the population in the province were children aged between 0-14 years, 60% were aged between 15-64 years and 6% of the population were elderly people. Black Africans constitute the majority of the population, followed by Whites, Indians or Asians and Coloureds. Sepedi is the dominant language spoken in Limpopo, followed by Xitsonga and Tshivenda.

The province is divided into five district municipalities. The most populated district in the province is Vhembe (1,294,22 population), followed by Capricorn (1,261,463 population), Mopani (1,092,507 population), Greater Sekhukhune (1,076,840 population) and Waterberg (679,336 population), as reported by (StatsSA, 2012). In 2011, there were more females than males across all districts, with the exception of Waterberg.

Figure 2: **Limpopo province map (source: Wikipedia image)**

Limpopo is the second poorest province in South Africa with a poverty rate of 59.1% of the total population (Kyei, 2011). It is a typical developing area, with many rural settlements practising subsistence farming. According to the 2011 census, the unemployment rate in Limpopo was 38.9%. The Greater Sekhukhune district has the highest unemployment rate and the highest unemployment rate among people without education in the province. The Capricorn district had the highest proportion of the people with Grade 12 or Matric and higher education qualifications.

1.3. PURPOSE OF THE STUDY

The goals and objectives of this study are divided into the main aim and the objectives of the study.

1.3.1. Main Aim of the Study

The main aim of this study was to determine factors that contribute to RTD in Limpopo Province.

1.3.2. Objective of the Study

The study focused on the following specific objectives:

i.) To understand the temporal trend of RTAs and RTDs.

ii.) To compare generalized linear models to zero-inflated models.

iii.) To identify and estimate the effect of each factor contributing to RTDs.

1.4. RESEARCH METHODOLOGY

1.4.1. Data Source

The study was based on secondary data on RTAs obtained from the Limpopo Province Department of Roads and Transport. The study comprised of 18,029 RTAs that occurred and were recorded in the Limpopo Province during the period January 2009 to December 2015. The data consisted of the number of people killed, seriously injured and slightly injured, as well as information on where and when the accident occurred, the vehicle type and the cause of the accident.

1.4.2. Data Analysis

Descriptive statistical analyses, including line graphs, bar charts and cross tabulations, were used in the analysis to summarise the dataset. The study proposes alternative models to the standard Poisson regression model. Competing count models were fitted to road accident data to come up with better models for predicting road fatalities.

1.5. SIGNIFICANCE OF THE STUDY

The issue of road accidents is a public health problem both internationally and locally. Although there has been a great deal of research done on this subject, the research is old, was mostly done for international markets, and does not really use prediction models to determine the effect of each contributing factor. For such a contemporary issue, more recent research is necessary in order to contribute to the body of knowledge on road safety and to help the Department of Roads and Transport in Limpopo assess the progress made towards reducing the number of RTAs and RTDs

in the province. The research was not meant to be conclusive, but it was an attempt to serve as a building block for future research to be done on the subject.

1.6. CONCLUSION

Road accidents are a subset of non-natural causes of deaths in South Africa. The study attempts to determine the factors that contribute to RTDs in the province of Limpopo. Chapter 2 focus on the literature on risk factors associated with, and techniques used to analyse data on, road accidents. Chapter 3 describes the methodology used in this study. The model results are presented and discussed in Chapter 4 and Chapter 5. Finally, in Chapter 6 I will present a summary of the study and give recommendations.

CHAPTER 2: LITERATURE REVIEW

2.1. INTRODUCTION

This chapter will provide a review of various relevant literature, both national and international, closely related to the topic. The review begins by exploring the literature focusing on the risk factors that contribute to road traffic accidents (RTAs), road traffic deaths (RTDs), and road traffic injuries (RTIs). Finally, the literature about methodology is also reviewed.

2.2. RISK FACTORS

Many researchers have studied the causes and effects of vehicular accidents in South Africa, and elsewhere, and made a number of recommendations. The study of the cause of road accidents by Vogel and Bester (2005), classified factors contributing to RTAs as human factors, factors of environmental conditions and factors of the vehicle. Factors in the human factors category were negligence, excess speed, dangerous overtaking, pedestrians in the road and inconsiderate driving behaviour. Factors in the vehicle factors category had mostly to do with defective brakes and tyres. Rush-hour traffic and inadequate facilities for pedestrians were factors included in the environmental factors category. It was found that the highest number of road accidents recorded were as a result of human factors.

A study carried out by Li and Bai (2008) further classified RTA data into the following categories: driver at fault, time, accident environmental conditions, road conditions, accident scene information and other contributing factors. The main variables in the driver at fault category were age and gender. The variables in the time category where time, day, month and year. Variables within the accident environment conditions category were lightning, weather and road surface. Variables within road conditions category were surface type, lane number, road class, speed limit, area information, road character and road special features. The main variables in the accident scene information category were accident location, number of cars involved in the collision, vehicle maneuverer before accidents, accident type, vehicle type, traffic control device, driver and pedestrians.

The study conducted by Bener et al. (2013) investigated the gender- and age-related differences in driver behaviour in Qatar. The study was based on face to face

interviews and found that the majority of the male and female drivers were young drivers in the age group 30-39 years. In this study the Student t-test was used to test for significance differences between mean age values of male and female drivers. A significant difference was found in the mean age of male and female drivers. Drivers between the ages of 25 and 44 often caused more RTAs and accidents among male drivers were more common than among female drivers (Li and Bai, 2008).

A cross-sectional study conducted by Burgut et al. (2010), undertaken from February to June 2009, explored RTA patterns among drivers in Qatar and investigated the contributing factors. Face to face interviews were conducted using a questionnaire covering sociodemographic information, driving history, type of vehicle, driver behaviour, details of crashes and accident pattern. Fisher exact and Chi-square tests were used to test differences in the proportions of categorical variables between; marital status, educational level, on holiday and drivers who did or did not have accidents. No significant difference was found between high and low household income. The frequency of RTAs among drivers who were married was higher than those who were not married and the accidents among drivers with a university degree were more common during non-holidays than during holidays. In contrast, single drivers were involved in more accidents than married drivers in the study by Al-Matawah and Jadaan (2010).

Burgut et al's study has shown that drivers with more driving experience (over 5 years) were more frequently involved in RTAs, followed by drivers with 1-3 years of experience (Burgut et al., 2010). This contradicted Al-Matawah and Jadaan's (2010) study which found that the more experienced the driver, the less involved they were in accidents.

A study conducted by Agbonkhes et al. (2013) in Nigeria, investigated possible causes of RTA in Nigeria with the aim of recommending general preventive action. Despite increased enforcement, speeding was found to be leading cause of accidents in Nigeria. In New Zealand in 2012, the Minister reported that speeding contributed to 68 fatal accidents, 307 severe injury accidents and 1, 049 minor injury accidents (Ministry of Transport, 2013). The Minister also reported that these accidents resulted in 85 deaths, with a total social cost of approximately NZ$637 million.

A 1% increase in speed is approximately associated with a 2% increase in the injury accident rate, a 3% increase in the severe accident rate and a 4% increase in the fatal accident rate (Aarts and Van Schagen 2006). A study by Li and Bai (2008) showed that a 51-60 mph (82-97 km/h) speed zone had the highest proportion of both fatal and injury accidents.

Seatbelts are very important in preventing deaths from road traffic accidents, and the study by Ogundele et al. (2013) showed a significantly increased risk of death among road accident victims who did not wear seatbelts. Using seatbelts can reduce the likelihood that drivers and front passengers will be killed. There is a higher proportion of seatbelt use in female drivers than in male drivers and drivers not involved in accidents (Burgut et al.,2010; Afukaar et al., 2010; Clarke et al.,2010). Seatbelt use is 33.2% among users of private cars, 9.0% for taxis, 8.3% for minibus, 13.1% for large buses and 9.7% for trucks (Afukaar et al., 2010). About 85% of fatalities involved people who were not wearing seatbelt and travelling who were in the front passenger seat (Clarke et al.,2010).

A study on the importance of visual perception for safe driving was conducted by Maffioletti et al. (2009). In the study they found that about 59.13% of accidents are associated with poor eyesight. Drivers for whom the eyesight deficiency is corrected with the eyeglass are likely to be involved in less severe accidents (Zhu and Srinivasan, 2011). Most adult drivers aged 65 and above presenting with eye conditions, such as cataracts, are at more risk of being involved in an accident than younger drivers with no cataracts (Desapriya et al., 2010).

The Zhu and Srinivasan (2011) study set out to determine the factors affecting the severity of overall injury resulting from RTAs. In this study, RTAs were found to be less severe on weekdays than on weekends.

However, these findings were inconsistent with findings in a previous study by Li and Bai (2008). This study found that, over a weekend, Sunday frequently recorded the lowest number of injury accidents. These inconsistent results may be as a result of cultural activity differences taking place on weekends and weekdays.

Generally, the highest number of road accidents and death were observed during the month of December and lowest observed in January and February (Lehohla, 2009). However, a recent study by Ishtiaque (2013) showed that most accidents occurred

during the heavy rainy season months of July to September. Weather seasonality differs from region to region, and this might be one of the reasons why there is inconsistent results between these two studies.

A high proportion of road accidents and deaths occur during the night, between 6 pm and 6 am, with peak times from 12 pm to 6 pm (Goswami and Sonowal, 2009; Zhu and Srinivasan, 2011). Some accidents at night are caused by the lack of street lights, particularly during night time driving on the undivided 2-lane, 2-way rural highways (Ishtiaque, 2013). This could lead to a difficulty in distinguishing the lane separation which might cause an accident. The probability of fatality is estimated to rise when dull lighting conditions are present (Lemp et al., 2011). Multi-vehicle accidents commonly occur during the daytime off-peak hours (Li and Bai, 2008).

A study by Cantillo et al. (2016) investigated the factors affecting urban road accidents. A combined GIS-Empirical Bayesian approach was used this study and it was found that the geometry of the road plays an important role in the frequency of road accidents as well as in the level of accident severity. More accidents commonly occurred in roads with two-way traffic, as opposed to single-way roads. The study also found that risk decreases with the width of the road. Moreover, studies have found a link between road accident frequency and risk factors, such as: road segment length, width, number of ramps and bridges, horizontal and vertical curves and shoulder width (Anastasopoulos and Mannering, 2009).

A study conducted by Jung et al. (2010) in south eastern Wisconsin assessed the effects of rainfall on the severity of single-vehicle accidents, taking into account weather-related factors, such as estimated rainfall intensity for 15 minutes before accident occurrence, water film depth, temperature, wind speed/direction, stopping sight distance and car following distance at the time of the crash. This study found that rainfall intensity, wind speed, and horizontal or vertical curve, were all linked to an increasing the likelihood of accident severity in rainy weather.

Distracted driving is a comportment dangerous to drivers and passengers. A report by the National Highway Traffic Safety Administration (NHTSA) (2013) revealed that 10% of fatal accidents in 2011 were reported as distraction-affected crashes. Additionally, 12% of the drivers involved in these accidents were using a cell phone at the time of the crash. The use of mobile phones reduces situation awareness and

increases unsafe behaviour, putting pedestrians at greater risk of accidents and crime victimization (Nasar et al., 2008; Zhu and Srinivasan 2011; Agbonkhese et al., 2013). People who use their cell phone while driving, are four times more likely to be involved in an accident (WHO, 2011).

Driving while under the influence of alcohol and drugs increases the risk of a RTAs and the chances of causing death or serious injury on roads (Burgut et al., 2010; Romana et al., 2014). South Africa has national laws to combat drunken driving. However, more drunken driving-related deaths occur in this country than in anywhere else in the world. South Africa has four out of ten in its ability to implement these laws (WHO, 2015). Approximately 60% to 70% of South African drivers and pedestrians killed in road accidents were found to have a concentration blood alcohol (BA) above 0.08g (WHO, 2015).

A study undertaken by Al-Matawah and Jadaan (2010), involved creating a model of accident prediction related to the frequency of accidents in Kuwait. This study found that the more aggressive the driving, the greater the number of road accidents. Furthermore, drivers who think that enforcement is ineffective experience more road accidents than drivers who perceive enforcement as effective.

A study carried out by Agbonkhese et al. (2013), examined the problems associated with road accidents in Nigeria and found that vehicle factors alone had the greatest influence on the frequency of accidents resulting in fatalities or serious injury. Vehicle parts, such as: tyres, engines, braking system, side mirrors, wipers, the horn and light systems, were also found to be the main contributing factors to RTAs in the country. A South African study showed that the most common defects in minibus taxis were found with braking systems, such as brake pads identified as being cheap imports (Govender and Allopi, 2007).

2.3 METHODOLOGY

A study conducted Lemp et al. (2011) examined the impact of vehicle, occupant, driver and environmental characteristics on accident severity for those involved in truck crashes. In this study, the ordered probit model was used to model road fatalities and it was found that the likelihood of fatalities and serious injury was estimated to increase with the number of truck trailers, but fall with the total length of the truck and the gross weight of the vehicle.

Anowar et al. (2012) compared two different models, the traditional ordered logit model, the latent segmentation based ordered logit model, with two segments and with three segments. They deployed two goodness of fit Bayessian information criterion (BIC) and Ben-Akiwa and Lerman's adjusted likelihood ratio (BL) test, to compare the goodness of fit of three models. The latent segmentation based ordered logit model with two segments was found to outperform other models in identifying the factors that influence injury severity of highway vehicle occupants involved in accidents.

The study by Lemp et al. (2011) examined the impact of environmental factors, and drivers and vehicle factors on the severity of injury resulting from large truck crashes by running two regression models namely the ordered probit (OP) model and the heteroskedastic ordered probit (HETOP) model. The study found that the HOP model performed significantly better than the OP.

In order to better understand the injury severity distributions of accidents on highway segments, and the effect that traffic, highway and weather characteristics have on these distributions, the mixed (random) logit model was used by Milton, et al. (2008) to model road accidents on a highway. The authors found that weather effects, such as snowfall, are best modelled as random parameters, while roadway characteristics, such as the number of horizontal curves, number of grade breaks per mile and pavement friction, are best modelled as fixed parameters. However, the disadvantage of using mixed effect methods is that the results may not be easily transferable to other datasets (Lord and Mannering, 2010).

Stepwise logistic regression analysis was applied by Çelik and Senger (2014) in the case of the Kars Province in Turkey to analyse data and investigate critical factors that contributed significantly to fatal versus non-fatal traffic accidents. They found that the stepwise logistic regression model fitted the RTA data in the Kars Province well.

A study by Jung et al. (2010) compared two predicting models, the ordinal logistic regression model and the sequential logistic regression model, to predict accident severity, that is, a polychotomous response. In the study the data was divided into forward format from lowest injury severity to the highest injury severity, and the backward format, reversing the sequence. The study found that the backward format

sequential logistic regression model outperformed the logistic regression model in predicting accident severity.

The multiple linear regression model was used by Gupta et al. (2017) to try to identify factors that contributed to the cause of accidents, and also to develop an accident prediction model for the road segment. The number of accidents was treated as outcome variable in the model, while the predictor variables were road width, segment length of the road, traffic volume, pedestrian volume and the number of passageways. The results of the study found that the risk of being involved in an accident increased as the traffic volume, pedestrian volume, carriage-way width, segment length and number of passageways.

When investigating the impact of traffic congestion on the frequency of road accidents in England, Poisson-lognormal, Poisson-gamma and Poisson-lognormal with conditional autoregressive prior models were used to account for the effect of both heterogeneity and spatial correlation (Wang et al. 2009). The results of the study showed that there was no link between traffic congestion and road accidents.

Applying linear regression to count data leads to inconsistent standard errors and may produce negative predictions for the dependent variable (Al-Matawah and Jadaan, 2010; Ayati and Abbasi, 2014). Therefore, the Poisson regression model is one of the most widely used statistical models for the analysis of count data.

One of the advantages of Poisson regression over a standard linear regression model is that this model includes a skew and restriction of predicted values to non-negative integer values (Ayati and Abbasi, 2014). In most count data sets seen in practice, the Poisson regression model tends to fit the data poorly, as indicated by the deviance. This may be because of the restriction that the conditional variance of the dependent variable is equal to the conditional mean.

In the case when the Poisson model assumption is violated, Ayati and Abbasi (2014) and Oppong (2014) suggest that an alternative approach is to apply the negative binomial regression model as this model relaxes the assumption of equality of the conditional mean and conditional variance by adding a gamma distributed error term.

Anastasopoulos and Mannering (2009), using accident data from rural interstate highways in Indiana collected over a 5-year period (1995-1999), explored the use of random parameter count models as a methodological alternative in analysing

accident frequencies in order to gain new insights into the ways that factors significantly influence accident frequency. They found that the random parameter negative binomial model resulted in the best statistical fit (relative to the random and fixed parameter Poisson models).

Aderson's study (2009), investigated road accident hotspots using data collected by the Metropolitan Police in the United Kingdom over a 4-year period from 1999 to 2003. Geographical information system (GIS) and kernel density estimation (KDE) information were used in this study to model road accidents. The study found that KDE with K-means clustering can be used to identify accident hotspot locations and to predict the impact of the road on the fragmentation of the landscape. However, a study by Thakali et al. (2015) found that the Gaussian process regression method outperformed the KDE method in its ability to detect hotspots. These inconsistent results may be a result of the fact that the Gaussian process regression method allows for interpolated cells to exceed the boundaries of the sample range.

Analysing potential factors that affect the odds of having fatalities in a vehicle collision in Namibia over 3 years (2007-2009), analysis of variance (ANOVA) and the binary logistic regression model were used by Nangombe (2012) to test whether there was any difference in the average number of fatalities between the years and also to calculate the odds of fatalities occurring, respectively. The study found that there was no significant difference in the number of fatalities over the years. The study also found that road users were 1.83 times more likely to die on Sundays than on Fridays and 1.96 times more likely to die when weather conditions were unknown than when weather conditions were clear.

A study by Zong et al. (2013), compared two modelling techniques, namely, the Bayesian neural network and the regression models, by employing them in the analysis of accident severity. Mean absolute percentage error (MAPE) and the hit ratio were used to compare the goodness of fit of these two models. The study found that, based on the goodness of fit, the Bayesian neural network outperformed the regression model in modelling road accident severity. However, the Bayesian neural network model has a disadvantage in that may not have interpretable parameters and complex estimation processes (Lord and Mannering, 2010).

Several RTA-prediction models have been developed and assessed for their predictive ability using different models. Imran and Nasir (2015), determined the trend of road accidents in Pakistan form January 2002-2003 to December 2011-2012. They deployed a set of eleven curve fitting models, namely: linear, quadratic, cubic, logarithmic, inverse, exponential growth model, logistic curve and compound models, for predicting RTAs. The cubic model was found to be the appropriate or convincing model for predicting the annual road accident rate for the total number of accidents, fatal accidents, non-fatal accidents, killed, injured people and the number of vehicles involved.

To understand the pattern of road accidents, the autoregressive integrated moving average (ARIMA) model was used by Sanusi el at. (2016) to determine patterns of RTAs cases along Nigeria's motorway between 1960 and 2013. The ARIMA model was developed and found to perform well in predicting minor cases, serious cases, fatal cases and total cases. Additionally, it was shown that RTA cases were on the increase in Nigeria. However, a study by Quddus (2008) found that real-valued time series models, such as the ARIMA model, and structured time series models may be inappropriate when modelling non-negative integer-valued data, such as road accidents. This is mainly because of the normality assumption of errors in the ARIMA model being violated (Quddus, 2008; Junus and Ismail, 2014)

Road accident models developed in one country might not be suitable for other countries (Mohanty and Gupta, 2016). This makes traffic accident analysis and modelling a task suitable for data mining and machine learning approaches that develop models based on actual real-world data (Kromer et al., 2014).

Ogwueleka et al. (2014), used neural network (NN) model for analysing historical data in Nigeria in order to predict future trends. Input variables were selected by examining the strength of the correlation between the annual number of accidents and related variables. The model was found to be a potentially powerful tool for analysing and forecasting the number of accidents.

A joint probability model was developed by Pei et al. (2011) to evaluate the effect of explanatory factors on accident occurrence and accident severity at signalized intersections in Hong Kong. Twelve neutral independent variables were selected using correlation analysis. The Markov chain Monte Carlo (MCMC) approach full

Bayesian method was applied to estimate the effect of explanatory factors and the deviance information criterion (DIC) and Chi-square test statistics were used to evaluate statistical fit. The authors found that the negative binomial logistic model is superior to the negative binomial truncated Poisson model in analysing accident occurrence.

The generalized Pareto model was deployed for modelling the number of road accidents in Spain between 2003 and 2007, and the discrete Lomax distribution model was applied in order to model number of fatalities (Prieto et al., 2014). Both models were found to outperform the negative binomial model.

The study by Ma et al. (2014), provided an alternative method to analyse the accident risk. The datasets from various sources were first integrated under a GIS platform and then fitted to a quasi-Poisson regression model because of its advantage over the traditional Poisson and negative binomial regression models, since this model does not require a predefined distributional from of the responses and hence may produce more mature and accurate results. The results showed that the model is appropriate for dealing with over-dispersed count data and several key explanatory variables were found to have a significant impact on the estimation of the Accident Hazard Index (AHI).

A study by Pollak et al. (2014), compared four predicting models, namely, the Poisson, the negative binomial, the zero-inflated Poisson and the zero-inflated negative binomial models, and found that the most significant model was negative binomial for modelling RTAs. The models were improved by using the empirical Bayes method, which increased the accuracy of the assessment by considering historical data and correcting the biases.

Prasetijo and Musa (2016) used accident data from south of Peninsular Malaysia collected over 5-year period from 2010 to 2014, and fitted the Poisson regression model with excess zero outcomes on the response variable. The study found that a generalized linear modelling (GLM) technique, such as the Poisson regression model and the negative binomial model, were insignificant in explaining and handling over-dispersion due to the high number of zeros. This suggests that zero-inflated models can be deployed to cater for excess zero outcomes on the response variable.

The goodness of fit of a statistical model describes how well it fits into a set of observations. Its indices summarize the discrepancy between the observed values and the values expected under a statistical model. The goodness of fit, such as: the Akaike information criterion (AIC), the Bayesian information criterion (BIC), the Ben-Akiwa and Lerman's adjusted likelihood ratio test (BL), Chi-squared test, Kolmogorov-Smirnov test (KS) and the deviance information criterion (DIC) can used to assess a model's goodness of fit (Pei et al., 2011; Anowar et al., 2012; Pollak et al., 2014, Prieto et at., 2014).

To estimate the parameters of the generalized linear and zero-inflated models, the Markov chain Monte Carlo (MCMC), the maximum likelihood estimation (MLE) and the expectation maximization (EM) algorithms are commonly used methods to estimate model parameters (Anastasopouos and Mannering,2009; Pei et al., 2011).

2.4. CONCLUSION

The literature on risk factors and models for modelling road accidents and their severity has been considered in this chapter. The study shows the importance of modelling to understanding the factors that contribute to road accidents and their severity. The literature guides the author to assume that models of road traffic developed in one country may not be suitable for application in other countries or in provinces within a country, thus demonstrating a need to understand the pattern of road accidents and associated risk factors, and the need to develop a model for the province of Limpopo, South Africa.

CHAPTER 3: METHODOLOGY

This chapter focusses on the different approaches which I will employ to model road traffic accident (RTA) data in the Limpopo province. I discuss the statistical methodologies used to carry out the analysis of data gathered in this study.

3.1. GENERALIZED LINEAR MODELS

McCullagh and Nelder (1989) developed generalized linear models (GLMs) as flexible generalizations of the ordinary linear model that allow for response variables that have error distribution other than a normal distribution. The GLM approach has the following two advantages:

i). it gives a general framework for the commonly used statistical models.

ii). one general algorithm can be used for estimation, inference and assessing model adequacy for all the models.

GLMs have the following three components:

i). Random component: This refers to the probability distribution of the response variable (Y) that belongs to the exponential family with density function of the form:

$$\ln\big(f(y;\theta,\phi)\big) = \frac{y\theta + b(\theta)}{a(\phi)} + c(y,\phi). \tag{1}$$

It can be shown that the conditional mean and the variance of Y are given by:

$$\mathrm{E}(Y|X) = \mu = b'(\theta) \text{ and } Var(Y|X) = \sigma^2 = b''(\theta)a(\phi).$$

ii). Systematic component: This component specifies the explanatory variables $(X_1, X_2, \cdots, X_k)$ in the model, more specifically, their linear combination:

$$\alpha + \beta_1 X_1 + \cdots + \beta_k X_k, \tag{2}$$

where β is the vector of regression coefficients and X_i are the explanatory variables.

iii). Link function: This component specifies the link between random and systematic components:

$$g(\mu) = \alpha + \beta_1 X_1 + \cdots + \beta_k X_k, \tag{3}$$

where $g(\mu)$ is a known link function which is a one to one continuous differentiable function and monotonic. The link function, $g(\cdot)$, connects the stochastic and systematic components.

The simplest link function is $g(\mu) = \mu$. This models the mean directly and is called the identity link. It specifies a linear model for the mean response:

$$\mu = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k. \tag{4}$$

This is the form of ordinary regression models for continuous responses.

Another link function is $g(\mu) = \log(\mu)$. This models the log of the mean and is called the log link. The log link function applies to positive numbers, so the log link function is appropriate when μ cannot be negative, such as count data. It specifies a linear model for the mean response:

$$\log(\mu) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k. \tag{5}$$

The link function $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$ models the log of the odds. It is appropriate when μ is between 0 and 1, such as a probability. This is called the logit link.

If the link function is $g(\mu) = \theta$ we say we have a canonical link, which transforms the mean to the natural parameter. The link function that uses the natural parameter as $g(\mu)$ in the GLMs is called the canonical link.

In summary, GLMs extend the general linear models in two ways. Firstly, it allows for stochastic components following distributions other than the normal distribution. Secondly it links functions other than the identity function. The Poisson, negative binomial, logistic regression models are special cases of the GLM framework.

3.2. LOGISTIC REGRESSION MODEL

The logistic regression model is one of the special cases of the GLM framework for binary data. This is the most important model for categorical response data. Let Y be a random variable that takes either 0 or 1, defined below as follows:

$$P(Y = 1) = \pi.$$
$$P(Y = 0) = (1 - \pi).$$

This likelihood of $Y = y$ is given as follows:

$$P(Y = y) = \pi^y (1 - \pi)^{1-y} \ = \exp\left(y \log\left(\frac{\pi}{1 - \pi}\right) + \log(1 - \pi)\right)$$

Bernoulli distribution is one of the exponential family represented as:

$$\theta = \log\left(\frac{\pi}{1 - \pi}\right),$$
$$b(\theta) = -\log(1 - \pi)$$
$$a(\phi) = 1,$$
$$c(y; \phi) = y$$

The random component for the outcome (success, failure) has a binomial distribution. The link function for logistic regression model uses the logit link function of π defined as:

$$\pi(x) = \frac{e^{x_i'\boldsymbol{\beta}}}{1 + e^{x_i'\boldsymbol{\beta}}} \tag{6}$$

where parameter $\boldsymbol{\beta}$ represent the rate of increase or decrease of the curve. When $\boldsymbol{\beta} > 0$, both $\pi(x)$ and $\boldsymbol{x}$ increase. When $\boldsymbol{\beta} < 0$, $\pi(x)$ decreases as $\boldsymbol{x}$ increases. When $\boldsymbol{\beta} = 0$, Y is independent of $\boldsymbol{x}$. The π is restricted to the 0-1 range, the logit can be any real number. The linear predictor ($\alpha + \beta_1 x_1 + \cdots + \beta_k x_k$) form the systematic component of a GLM.

3.3. POISSON DISTRIBUTION

Many discrete response variables have counts as possible outcomes. Counts also occur in summarising categorical variables with contingency tables. The simplest GLM for count data assumes a Poisson distribution for the random component. The Poisson distribution is a discrete probability distribution mainly used to model the number of events that occur randomly within a given time interval. Let Y denote a count and $\mu = \mathrm{E}(Y)$. The Poisson probability mass function for Y is defined as follows:

$$f(y;\mu) = \frac{e^{-\mu}\mu^{y}}{y!}, \qquad y \geq 0$$

where y is number of events in a given interval and $\mu(>0)$. Taking logarithm, we get

$$\log\big(f(y;\mu)\big) = \log\left(\frac{e^{-\mu}\mu^{y}}{y!}\right)$$
$$= \log e^{-\mu} + \log \mu^{y} - \log y!$$
$$= -\mu + y\log\mu - \log y!$$
$$= \frac{y\log\mu - \mu}{1} - \log y!.$$

Matching the generic functions and parameters in equation 1:

$$\theta = \log\mu,$$
$$b(\theta) = \mu$$
$$a(\phi) = 1,$$
$$c(y;\phi) = \log y!$$

Thus, the canonical parameter for the Poisson distribution can be written as $\mu = e^{\theta}$, where θ is the canonical parameter for the exponential family and $\log\mu$ is the canonical parameter for the Poisson distribution.

Second differencing function $b(\theta)$ given as:

$$b''(\theta) = e^{\theta} = \mu$$

The Poisson distribution has only a single parameter, $\mu(>0)$, that is, the rate parameter, which is both the mean and variance, so it is described as equi-dispersed given as follows:

$$\mathrm{E}(Y) = Var(Y) = \mu.$$

This shows that an effect on the mean will also affect the variance.

3.4. POISSON REGRESSION MODEL

The Poisson regression model is derived from the Poisson distribution by parameterizing the relationship between the mean parameter $\mu(>0)$ and the linear predictors, given by:

$$\mu = \boldsymbol{x_i'\beta}, \qquad i = 1, \cdots, n.$$

To ensure that $\mu > 0$, the standard assumption is to use the natural logarithm on the mean because it is a strictly monotonically increasing function defined as:

$$\log(\mu) = g(\mu_i),$$

where $g(\mu)$ is the canonical link function. We consider the GLM with link log function resulting in a log-linear relationship between the mean parameter μ and the linear predictor $\boldsymbol{x_i'\beta}$ defined as:

$$\log(\mu) = \boldsymbol{x_i'\beta}, \tag{7}$$

where the regression coefficient $\boldsymbol{\beta}$ represents the effect of a one unit change in the predictor on the log of the mean.

Taking the logarithm in equation (7), we obtain:

$$\mu = \exp(\boldsymbol{x_i'\beta}).$$

The Poisson regression model assumes that variance $[Var(Y) = \mu)]$ is equal to the mean $[\mathrm{E}(Y) = \mu]$, thus the dispersion is fixed at $\phi = 1$. This assumption in most count data seems to be violated in practice.

3.5. NEGATIVE BINOMIAL REGRESSION MODEL

The phenomenon of the data having greater variability than expected for GLM is called over-dispersion. This may be because some of the relevant explanatory variables are not in the model, or this may be due to unobserved heterogeneity. An alternative approach to model an over-dispersed dataset is to use models that are less restrictive, such as the negative binomial regression model.

The negative binomial model is another distribution that concentrates on the nonnegative integers. Suppose the distribution of a random variable Y follows the Poisson with the parameter $\theta\mu$ $\left(Y \sim Poisson(\theta\mu)\right)$. Where θ has a gamma distribution with parameters $\left(\theta \sim \Gamma(\alpha, \beta)\right)$. The corresponding probability density function in the shape rate parametrization is defined as:

$$f(x;\alpha,\beta) = \frac{\beta^{\alpha} x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)}.$$

We assign parameters $\alpha = \beta = \sigma^2$ with $\mathrm{E}(\theta) = 1$ and $Var(\theta) = \alpha$ and where μ is a deterministic function of x. The probability mass function of the negative binomial distribution is given by:

$$P(Y = y) = \frac{\Gamma(\alpha^{-1} + y)}{\Gamma(\alpha^{-1})\Gamma(y+1)} \left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu}\right)^{\frac{1}{\alpha}} \left(\frac{\mu}{\mu + \alpha^{-1}}\right)^{y} \tag{8}$$

The negative binomial belongs to an exponential family. Equation (8) can exponentially be represented as

$$P(Y = y) = \frac{\Gamma(\alpha^{-1} + y)}{\Gamma(\alpha^{-1})\Gamma(y+1)} exp\left(\frac{1}{\alpha} ln\left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu}\right) + y\, ln\left(\frac{\mu}{\mu + \alpha^{-1}}\right)\right)$$

where $\mu > 0$ is the mean of Y, α is the shape parameter and $\Gamma(\cdot)$ is the gamma function. The negative binomial distribution has mean $\mathrm{E}(Y) = \mu$ and variance $Var(Y) = \mu + \alpha\mu$. If $\alpha = 0$ we obtain Poisson variance. If $\alpha > 0$ and $\mu > 0$, therefore, the variance will exceed the mean.

Let μ depend on the explanatory variables through a log-linear model. Then, the negative binomial regression model is given by:

$$\mu = \exp(\boldsymbol{x_i'\beta}).$$

3.7. ZERO-INFLATED MODEL

In practice many count data exhibit zero inflation, therefore the Poisson regression model may not be adequate. One of the extensions is use of the zero-inflated regression model. This model provides one method to explain the excess zeros by modelling the data as a mixture of two separate data generation processes. The first process is a constant distribution that can generate only zero counts, called the structural zeros, and the second process is a Poisson distribution that generates both zero and non-zero counts, called sample zeros (Ridout et al., 1998; Erdman et al., 2008). There are two types of zeros observed in count data, the zeros coming from a Poisson distribution having probability of occurrence $1 - \omega$ and the zeros coming from

a zero generating distribution having probability ω, which is called the zero-inflation probability Equation Lambert (1992).

The mass function of the two-component mixture distribution is given by:

$$P(Y=y)=\begin{cases}\omega+(1-\omega)g(0|\mu) & \text{for } y=0\\ (1-\omega)g(y|\mu) & \text{for } y>0\end{cases}, \tag{9}$$

where $0 \leq \omega \leq 1$, $\lambda \geq 0$ and y is the observed count dataset. The ω is the probability of being a structural zero (i.e. belonging to the first components). The term $g(y|\mu)$ is the probability mass function for belonging to the second component and typically chosen to be either from a Poisson or a negative binomial.

3.7.1. Zero-inflated Poisson Model

The probability mass function of Y can be written as follows:

$$P(Y=0)=\omega+(1-\omega)e^{-\mu}, \tag{10}$$

$$P(Y=y)=(1-\omega)\frac{e^{-\mu}\mu^{y}}{y!}, \quad y=1,2,\cdots \tag{11}$$

where the outcome variable Y has any non-negative integer value, μ is the expected Poisson count for the individual and ω is the of being a structural zero.

The mean and variance are defined below as:

$$\mathrm{E}(Y)=(1-\omega)\mu.$$
$$Var(Y)=(1-\omega)\mu(1+\omega\mu).$$

It can be observed that Equation (11) reduces to the Poisson regression model when $\omega=0$, and also when $\omega>0, P(Y=0)>e^{-\mu}$, which indicates zero-inflation.

3.7.2. Zero-inflated Negative Binomial model

The probability mass function of Y_i can be written as follows:

$$P(Y=0)=\omega+(1-\omega)\left(\frac{\theta}{\theta+\mu}\right)^{\theta}, \tag{12}$$

$$P(Y=y)=(1-\omega)\frac{\Gamma(\theta+y)}{y!\,\Gamma(\theta)}\left(\frac{\mu}{\theta+\mu}\right)^{y}\left(\frac{\theta}{\theta+\mu}\right)^{\theta}, \ y=1,2,\cdots \tag{13}$$

where the outcome variable Y has any non-negative integer value, μ is the expected Poisson count for the individual, θ overdispersion parameter and ω is the of being a structural zero.

The mean and variance are defined below:

$$\mathrm{E}(Y) = (1-\omega)\mu.$$
$$Var(Y) = (1-\omega)\mu(1+(\omega+\theta)\mu).$$

Again, it can be observed that Equation (13) reduces to the negative binomial regression model when $\omega = 0$, and also when $\omega > 0, P(Y=0) > \left(\frac{\theta}{\theta+\mu}\right)^{\theta}$, which indicates zero inflation.

Lambert (1992) suggested that the logit link function can be used to model the probability of being structural zeros ω and the canonical log link function can be used to model the Poisson mean μ, defined as follows:

$$\log(\mu) = \boldsymbol{X\beta} \quad \text{and}$$
$$\log\left(\frac{\omega}{1-\omega}\right) = \boldsymbol{Z\gamma},$$

where X and Z are vectors of covariates, β and γ are $p \times 1$ and $q \times 1$ vectors of regression coefficients. The logit link function enables us to determine the effect of the intercept and each covariate on the structural zeros.

3.8. PARAMETER ESTIMATION

3.8.1. Logistic Regression Model

In the logistic model we have the following expression for the likelihood:

$$L(\beta_0, \beta) = \prod_{i=1}^{n}\left[\left(\pi(x_i)\right)^{y_i}\left(1-\pi(x_i)\right)^{1-y_i}\right].$$

The log-likelihood turns product into sums:

$$\ell(\beta_0, \beta) = \sum_{i=1}^{n} \big(y_i \log\big(\pi(x_i)\big) + (1 - y_i) \log\big(1 - \pi(x_i)\big)\big)$$
$$= \sum_{i=1}^{n} \log\big(1 - \pi(x_i)\big) + \sum_{i=1}^{n} y_i \log\left(\frac{\pi(x_i)}{1 - \pi(x_i)}\right)$$
$$= \sum_{i=1}^{n} \log\big(1 - \pi(x_i)\big) + \sum_{i=1}^{n} y_i(\beta_0 + \beta x_i)$$
$$= \sum_{i=1}^{n} \log\big(1 + e^{\beta_0 + \beta x_i}\big) + \sum_{i=1}^{n} y_i(\beta_0 + \beta x_i).$$

To find the maximum likelihood estimates we differentiate the log-likelihood with respect to the parameters, set the derivatives equal to zero and solve:

$$\frac{\partial \ell}{\partial \beta} = -\sum_{i=1}^{n} \frac{x_i e^{\beta_0 + \beta x_i}}{1 + e^{\beta_0 + \beta x_i}} + \sum_{i=1}^{n} y_i x_i. \tag{15}$$

We can equate Equation (15) to zero, we cannot solve exactly. We can, however, approximately solve it numerically.

3.8.2. Poisson Regression Model

In the Poisson model we have the following expression for the likelihood:

$$L(\mu; y) = \prod_{i=1}^{n} \frac{e^{-\mu} \mu^{y_i}}{y_i!}. \tag{16}$$

The log-likelihood of equation (16):

$$l(\mu; y) = \sum_{i=1}^{n} [y_i \log(\mu) - \mu - \log(y_i)] \tag{17}$$

Substitute $\mu = e^{\boldsymbol{x_i'\beta}}$ in equation (17):

$$l(\mu; y) = \sum_{i=1}^{n} \Big[y_i(\boldsymbol{x_i'\beta}) - e^{\boldsymbol{x_i'\beta}} - log(y_i)\Big]. \tag{18}$$

Taking the derivative with respect to $\boldsymbol{\beta}$ we get:

$$\frac{\partial l(\mu; y)}{\partial \beta} = \sum_{i=1}^{n} \left(y_i \boldsymbol{x_i'} - \boldsymbol{x_i'} e^{\boldsymbol{x_i'\beta}} \right)$$

$$\frac{\partial l(\mu; y)}{\partial \beta} = \sum_{i=1}^{n} \boldsymbol{x_i'} \left(y_i - e^{\boldsymbol{x_i'\beta}} \right).$$

To get the maximum likelihood estimator, we have to solve the estimating equations given by

$$\sum_{i=1}^{n} \boldsymbol{x_i'} \left(y_i - e^{\boldsymbol{x_i'\beta}} \right) = 0.$$

This does not have a closed form solution and, because of this, numerical methods, such Newton-Raphson method, are used to get the estimator of $\boldsymbol{\beta}$. The linear predictor is then given by $\hat{\mu} = e^{\boldsymbol{x_i'\beta}}$.

3.8.3. Negative Binomial Model

The likelihood function for the negative-binomial model is defined as follows:

$$L(\mu, \alpha) = \prod_{i=1}^{n} \frac{\Gamma(\alpha^{-1} + y_i)}{\Gamma(\alpha^{-1})\Gamma(y_i + 1)} \left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu} \right)^{\frac{1}{\alpha}} \left(\frac{\mu}{\mu + \alpha^{-1}} \right)^{y_i}.$$

Calculating the log-likelihood function:

$$\ell(\mu, \alpha) = \sum_{i=1}^{n} \left\{ y_i \log \mu + \alpha^{-1} \log \alpha^{-1} - (\alpha^{-1} + y_i) \log(\alpha^{-1} + \mu) + \log \frac{\Gamma(\alpha^{-1} + y_i)}{\Gamma(\alpha^{-1})} - \log y_i! \right\}$$

$$= \sum_{i=1}^{n} \{ y_i \log \mu + \alpha^{-1} \log \alpha^{-1} - (\alpha^{-1} + y_i) \log(\alpha^{-1} + \mu) + dlg(y_i, \alpha^{-1}) - \log y_i! \}.$$

The link function for the negative-binomial is given as:

$$\mu = e^{\boldsymbol{x_i'\beta}}.$$

To find the maximum we take the derivatives with respect to $\boldsymbol{\beta}$ and α:

$$\frac{\partial \ell(\mu,\alpha)}{\partial \beta_j} = \sum_{i=1}^{n} \left\{ \frac{y_i}{\mu} - \frac{\alpha^{-1} + y_i}{\alpha^{-1} + \mu} \right\} \frac{\partial \mu}{\partial \beta_j}$$

$$= \sum_{i=1}^{n} \left\{ \frac{(y_i - \mu)}{\mu\left(1 + \frac{\mu}{\alpha^{-1}}\right)} \frac{1}{\mu} x_i \right\} = 0$$

$$\frac{\partial \ell(\mu,\alpha)}{\partial \alpha^{-1}} = \sum_{i=1}^{n} \left\{ ddg(y_i, \alpha^{-1}) - \log(\mu + \alpha^{-1}) - \frac{\alpha^{-1} + y_i}{\alpha^{-1} + \mu} + \log \alpha^{-1} + 1 \right\} = 0$$

This is in a closed form, the Newton's iterative technique method is used to maximize the parameters $\boldsymbol{\beta}$ and α.

3.8.4. Restricted Maximum Likelihood Estimation

The expectation maximization (EM) algorithm was introduced by Dempster et al. (1977). EM is a convenient tool to use in statistical estimation problems if we encounter missing or hidden data. It is a very general iterative method for parameter estimation by maximum likelihood estimation in statistical models (Borman, 2004; Chang and Kim, 2007). In order to estimate θ, it is typical to introduce the complete log-likelihood function defined as:

$$\ell(\theta; Y, Z) = \log P(Y, Z|\theta).$$

where Z denotes a set of missing or unobserved values and Y represent observed data. We want to estimate parameters θ in a model. The EM consists of two main steps:

1) Expectation (E) step: Determine the conditional expected value of the log-likelihood function defines as:

$$Q^n = \mathrm{E}_{Z|Y,\theta_n}[\log P(Y, Z|\theta)]. \quad (19)$$

2) Maximization (M) step: Maximize Q^n obtained in equation (18) with respect to θ. This is defined as:

$$\theta_{n+1} = \underset{\theta}{argmax}\left(\mathrm{E}_{Z|Y,\theta_n}[\log P(Y, Z|\theta)]\right).$$

3.8.4.1. Zero-inflated Poisson Model

Let us denote $P(y_j; 0) = \frac{\exp\{0\}0^{y_j}}{y_j!}$, and $P(y_j; \mu) = \frac{\exp\{\mu\}\mu^{y_j}}{y_j!}$. Then $P(y_j; 0) = 1$ if $y_j = 0$ and $P(y_j; \mu) = 0$ otherwise. Therefore, the likelihood function of the zero-inflated Poisson model is given by:

$$\ell = \prod_{j=1}^{n} pP(y_j; 0) + (1-p)P(y_j; \mu).$$

Estimation of this model would be trivial if it was known to which process each observation belongs (Ugarte et al., 2004). We consider the labels of the data as unobserved or latent variables. In this case, the result of a Bernoulli trial is used to determine which of the two processes generate an observation. This can be expressed as:

$$f(z; p) = \left[\omega P(y_j; 0)\right]^{1-z_j}\left[(1-\omega)P(y_j; \mu)\right]^{z_j}. \quad (20)$$

Therefore, the likelihood function of equation (20) is given as:

$$\ell_c = \prod_{j=1}^{n}\left[\omega P(y_j; 0)\right]^{1-z_j}\left[(1-\omega)P(y_j; \mu)\right]^{z_j},$$

where $z_j \epsilon \{0,1\}$.

The log-likelihood is then:

$$\ell_c = \sum_{j=1}^{n} z_j \ln(1-\omega) + (1-z_j)\ln\omega + z_j \ln P(y_j, \mu).$$

Maximum likelihood estimates for μ and ω can be estimated via the EM algorithm. In the E-step, using above Equation, the conditional expected value of the log-likelihood function is obtained as follows:

$$Q^j = \mathrm{E}_{z_j|y_j} = \frac{(1-\omega)\frac{\exp\{\mu\}\mu^{y_j}}{y_j!}.}{\omega\frac{\exp\{0\}0^{y_j}}{y_j!} + (1-\omega)\frac{\exp\{\mu\}\mu^{y_j}}{y_j!}.},$$

where $y_j > 0$, then $P(y_j, 0) = 0$ and $Q^j = 1$.

The M-step, we maximize Q^j with respect to μ and ω, leading to:

$$\hat{\mu} = \frac{\sum_{j=1}^{n} Q^j y_j}{\sum_{j=1}^{n} Q^j}.$$

$$\widehat{\omega} = 1 - \frac{\sum_{j=1}^{n} Q^j}{n},$$

where $\mu = \mu_0$ and $\omega = \omega_0$ are initial values and both steps, E-step and M-steps are repeated until convergence is achieved.

3.8.4.2. Zero-inflated Negative Binomial Model

The log-likelihood function for the ZINB regression model (assuming $\theta = 1$) is given by:

Let us denote $P(y_j; 0) = \frac{1}{1+0} = 1$, and $P(y_j; \mu) = \left(\frac{\mu}{1+\mu}\right)^{y_i}\left(\frac{1}{1+\mu}\right)$. Then $P(y_j; 0) = 1$ if $y_j = 0$ and $P(y_j; \mu) = 0$ otherwise. Therefore, the likelihood function of the zero-inflated Poisson model is given by:

$$\ell = \prod_{j=1}^{n} pP(y_j; 0) + (1-p)P(y_j; \mu).$$

We consider the labels of the data as unobserved or latent variables. In this case, the result of a Bernoulli trial is used to determine which of the two processes generate an observation. This can be expressed as:

$$f(z; p) = \left[\ \omega P(y_j; 0)\right]^{1-z_j}\left[(1-\ \omega)P(y_j; \mu)\right]^{z_j}. \tag{21}$$

Therefore, the likelihood function of equation (21) is given as:

$$\ell_c = \prod_{j=1}^{n}\left[\ \omega P(y_j; 0)\right]^{1-z_j}\left[(1-\ \omega)P(y_j; \mu)\right]^{z_j},$$

where $z_j \epsilon \{0,1\}$.

The log-likelihood for the ZINB is given by

$$\ell_c = \sum_{j=1}^{n} z_j \ln(1-\ \omega) + (1-z_j) \ln\ \omega + z_j \ln P(y_j, \mu).$$

Maximum likelihood estimates for μ and p can be estimated via the EM algorithm. In the E-step, using the above Equation, the conditional expected value of the log-likelihood function is obtained as follows:

$$Q^j = \mathrm{E}_{z_j|y_j} = \frac{(1-\ \omega)\left(\frac{\mu}{1+\mu}\right)^{y_i}\left(\frac{1}{1+\mu}\right).}{\omega + (1-\ \omega)\left(\frac{\mu}{1+\mu}\right)^{y_i}\left(\frac{1}{1+\mu}\right).},$$

where $y_j > 0$, then $P(y_j, 0) = 0$ and $Q^j = 1$.

The M-step maximizes:

$$\mathrm{E}_{z_j|y_j} = \sum_{j=1}^{n} z_j \ln(1-\ \omega) + (1 - z_j)\ln\ \omega + z_j \log\left(\left(\frac{\mu}{1+\mu}\right)^{y}\left(\frac{1}{1+\mu}\right)\right),$$

with respect to μ and ω, leading to:

$$\hat{\mu} = \frac{\sum_{j=1}^{n} Q^j y_j}{\sum_{j=1}^{n} Q^j},$$

$$\widehat{\omega} = 1 - \frac{\sum_{j=1}^{n} Q^j}{n},$$

where $\mu = \mu_0$ and $\omega = \omega_0$ are initial values and both steps, the E-step and the M-step, are repeated until convergence is achieved.

3.9. TESTING HYPOTHESES

Testing for two alternative models; one model is saturated and the other model is unsaturated. We test for:

$$H_0 : reduced\ model\ is\ true\ vs.\ H_1 :\ current\ model\ is\ true$$

3.9.1. Wald Test

The test statistic uses the large sample distribution of the maximum likelihood given as follows:

$$\max L\,(\beta, y) = L(\hat{\beta}, y),$$

where $\hat{\beta}$ is multivariate normal denoted as follows:

$$\hat{\beta} \sim N_p(\beta, I(\beta)^{-1}),$$

where $I(\beta)$ is the information matrix, defined as follows:

$$I(\beta) = \frac{(X'WX)}{\phi}.$$

The multivariate normal with mean β and variance covariance matrix $(X'WX)^{-1}\phi$, where X is the model matrix and W is the diagonal matrix of estimation weights.

The test statistic is given as:

$$W_p = \left(\hat{\beta} - \beta_0\right)^T \left[Cov\left(\hat{\beta}\right)\right]^{-1} \left(\hat{\beta} - \beta_0\right). \quad (22)$$

The asymptotic multivariate normal distribution for $\hat{\beta}$ implies an asymptotic Chi-squared distribution for W_p with the rank of $Cov\left(\hat{\beta}\right)$ as a degree of freedom. Under the null hypothesis, Equation (22), the Wald statistic W_p converges in distribution to a Chi-square distribution with k degrees of freedom.

3.9.2. Likelihood Ratio Test

The basic idea is to compare the maximized likelihoods of the two models. Let L_1 be the likelihood of the data with all the parameters unrestricted and maximum likelihood estimates substituted for these parameters. The maximum likelihood of L_1 is given as:

$$\max L_1\,(\theta, y) = L_1\left(\hat{\theta}_{model1}, y\right),$$

where $\hat{\theta}_{model1}$ denotes the maximum likelihood estimator of θ under model 1.

Let L_0 be the maximum value of the likelihood when the parameters are restricted (and reduced in number) based on the assumption. Maximum likelihood of L_0 is given as below:

$$\max L_0\,(\theta, y) = L_0\left(\hat{\theta}_{model0}, y\right),$$

where $\hat{\theta}$ denote the maximum likelihood estimator of θ under model 2.

The likelihood ratio is defined as follows:

$$\lambda = \frac{L_0\left(\hat{\theta}_{model0}, y\right),}{L_1\left(\hat{\theta}_{model1}, y\right),}.$$

This ratio is always between 0 (likelihoods are non-negative) and 1 (the likelihood of the smaller model cannot exceed that of the larger model because it is nested on it)

and the less likely the assumption is, the smaller λ will be. Values close to 1 indicate that the smaller model is almost as good as the larger model, making the data just as likely. Values close to 0 indicate that the smaller model is not acceptable, compared to the larger model, because it would make the observed data very unlikely.

Under certain regularity conditions, multiplying the log-likelihood ratio $\log(\lambda)$ by -2, given bellow:

$$-2\ \log(\lambda) = -2\ \log\left(\frac{L_0(\hat{\theta}_{model0}, y)}{L_1(\hat{\theta}_{model1}, y)}\right)$$

$$= -2\log\left(L_0(\hat{\theta}_{model0}, y)\right) + 2\log\left(L_0(\hat{\theta}_{model1}, y)\right).$$

In large samples, the log of the probability ratio has a Chi-square distribution with degrees of freedom equal to the difference between the two models in the number of parameters. The likelihood ratio test computes X^2 and rejects the assumption if X^2 is larger than a Chi-Square percentile $100(1-\alpha)$ with k degrees of freedom.

3.9.3. Score Test

The score function has an asymptotic normal distribution with mean 0 and variance covariance matrix equal to the information matrix, so that:

$$U(\beta) \sim N\big(0, I(\beta)\big).$$

The quadratic form:

$$Q = U(\beta_0)' I^{-1}(\beta_0) U(\beta_0),$$

has approximately a Chi-squared distribution with k degrees of freedom. The information matrix may be evaluated at the hypothesized value β_0 or at the maximum likelihood estimator of β. Under the null hypothesis, both versions of the test are asymptotically equivalent. One advantage of using β_0 is that calculation of the maximum likelihood estimation may be bypassed.

3.10. GOODNESS OF FIT STATISTICS

After fitting the models, we want to choose the model which best represents the data. The model fit reflects whether the appropriate link function and structural model have been specified.

3.10.1. Deviance

Deviance is a measure of the discrepancy between observed and fitted values. It provides the summary of the adequacy of the fitted model. The goodness of fit of the GLM can be based on the deviance statistic, which is given by:

$$D(y;\hat{\mu}) = 2\sum\left\{y_i \log\left(\frac{y_i}{\hat{\mu}}\right) - (y_i - \hat{\mu})\right\}, \tag{23}$$

where y_i is observations and $\hat{\mu}_i$ is the fitted model mean for i-th observation. The right hand side of Equation (23) is the sum of differences between observed and fitted values. The deviance statistic has an approximate Chi-square distribution with $n - p$ degrees of freedom, where n is the number of observations and p the number of parameters. If our model fits the data well, the Deviance to degree of freedom ratio should be about one.

3.10.2. Pearson's Chi-squared Statistic

The Pearson's Chi-squared is one of the alternative measures of the goodness of fit, denoted as follows:

$$\chi^2 = \sum \frac{(y_i - \hat{\mu})^2}{\hat{\mu}}.$$

The sum is the squared difference between the observed and fitted values y_i and $\hat{\mu}$, divided by the variance of the observed value $\hat{\mu}$.

3.10.3. Akaike Information Criterion (AIC)

AIC is the measure that is used to describe the trade-off between the accuracy and the complexity of the mode. It is also a valid procedure to use to compare non-nested models. AIC is defined as:

$$AIC = -2\log L + 2p,$$

where L is the maximized value of the likelihood function for the estimated model and p is the number of parameters in the statistical model. The AIC penalizes models with large numbers of parameters and selects the model with fewer parameters that best represents the data. The lower the AIC, the better the model.

3.10.4. Bayesian Information Criterion (BIC)

The BIC is closely related to the AIC. It is known as the Schwarz Criterion, after Gideon Scharz. It is normally used for comparing models. It incorporates both estimation uncertainty and parameter uncertainty. The BIC is defined as follows:

$$BIC = 2\log P\left(D|M,\hat{\theta}\right) - d \times \log(n),$$

where D is observed data, M is the model, $\hat{\theta}$ is the MLE, d number of free parameters and D number of data points.

The BIC assumes that one of the models is the true model and that one is trying to find the model most likely to be true in the Bayesian sense. It attempts to mitigate the risk of over-fitting by introducing the penalty term $d * \log(n)$, which grows with the number of parameters. The BIC is an asymptotic result derived from the assumption that the data distribution is an exponential family. A lower BIC score signals a better model.

3.10.5. Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov test is used to verify that a sample comes from a population with some known distribution and also that two populations have the same distribution. It is defined by:

$$H_0 : The\ data\ follow\ a\ specified\ distribution$$
$$H_a : The\ data\ does\ not\ follow\ a\ specified\ distribution$$

The Kolmogorov-Smirnov test statistic is defined as:

$$D = \max_{1 \leq 1 \leq N}\left(F(Y_i) - \frac{i-1}{N}, \frac{i}{N} - F(Y_i)\right),$$

where F is the theoretical cumulative distribution of the distribution being tested.

The hypothesis regarding the distributional form is rejected if the test statistic D is greater than the critical value.

3.10.6. Vuong Test

The Vuong test is mostly used to compare two non-nested models. It is based on Kullback-Leibler information criterion defined by:

$$KLIC = E[\ln h(Y_i|X_i) - E(\ln f(Y_i|X_i|\beta))],$$

where $h(Y_i|X_i)$ is the conditional density of Y_i given X_i and $f(Y_i|X_i|\beta)$ is the model with parameter β. The model which minimizes the $KLIC$ is the one that is closest to the true model.

Considering two models $U_\beta = f(Y_i|X_i|\beta)$ and $U_\theta = f(Y_i|X_i|\theta)$. The null hypothesis of the test is:

$$H_0: E\left(\log\frac{U_\beta}{U_\theta}\right) = 0,$$

which indicates that two models are equally close to the specification. The alternative hypothesis is defined by:

$$H_a: E\left(\log\frac{U_\beta}{U_\theta}\right) > 0, \quad model\ U_\beta\ is\ better$$

$$H_b: E\left(\log\frac{U_\beta}{U_\theta}\right) < 0 \quad model\ U_\theta\ is\ better.$$

3.11. K-MEANS CLUSTERING

The K-means clustering is a popular method for cluster analysis in data mining. It partitions n observations into K clusters in which each observation belongs to the cluster with the nearest mean. Given a set of observations $(x_1, x_2, \cdots, x_n)$, where each observation is a d- dimensional real vector, K-means clustering aims to partition the observation into K $(\leq n)$ sets in order to minimize the within-cluster sum of squares. It is defined by the following steps:

$$\min_{\mu}\min_{C}\sum_{i=1}^{K}\sum_{x\epsilon C_i}|x-\mu_i|^2.$$

Step 1:

Fix μ, optimize C

$$\min_{C}\sum_{i=1}^{K}\sum_{x\epsilon C_i}|x-\mu_i|^2 = \min_{C}\sum_{i=1}^{n}|x-\mu_i|^2$$

Step 2:

Fix C, optimize μ

$$\min_{\mu} \sum_{i=1}^{K} \sum_{x \epsilon C_i} |x - \mu_i|^2$$

take the partial derivatives of μ_i and set to zero, we get

$$\mu_i = \frac{1}{|C_i|} \sum_{x \epsilon C_i} x.$$

The K-means algorithm is a heuristic that requires initial means.

3.12. CHAPTER SUMMARY

In chapter 3 the research methodology to be adopted in this study has been reviewed. In the next chapter the data used in the study will be described, analysed and interpreted.

CHAPTER 4: EXPLORATORY DATA ANALYSIS

4.1. INTRODUCTION

In this chapter we perform exploratory data analysis of the road traffic accidents (RTAs) data in order to identify distributional properties associated with road accidents and associated deaths. We look at yearly total number of RTAs and road traffic deaths (RTDs). We then look at how RTAs and RTDs are distributed monthly, day of the week, hourly, vehicle type, vehicle involved and per district, and categorise contributing factors.

4.2. EXPLORATORY ANALYSIS

4.2.1. The Yearly Distribution of RTAs and RTDs

The data that was used is the daily RTA data in the Limpopo Province from January 2009 to December 2015. It was found that 18,029 RTAs occurred in the province over the 7-year period. Table 1 below shows the distribution of yearly RTAs and RTDs.

Table 1: The yearly distribution of RTAs and RTDs recorded from 2009 to 2015.

Year	No of deaths	%	%Δ	No of accidents	%	%Δ	Rate of death per accident
2009	759	15.35%	-	2416	13.40%	-	0.3141
2010	790	15.98%	4.08%	2240	12.42%	-7.28%	0.3527
2011	680	13.75%	-13.92%	2540	14.09%	13.39%	0.2677
2012	511	10.34%	-24.85%	2409	13.36%	-5.16%	0.2121
2013	556	11.25%	8.81%	2545	14.12%	5.65%	0.2185
2014	799	16.16%	43.71%	3047	16.90%	19.72%	0.2622
2015	849	17.17%	6.26%	2832	15.71%	-7.06%	0.2998
Total	**4944**			**18029**			**0.2742**

It can be seen from the table that 4,944 lives were lost between 2009 and 2015. It can also be seen that 2015 recorded the highest number of deaths, accounting for about 17% of all deaths; while in 2012 the least number of deaths were recorded, that is, 10% of all deaths recorded during the period under review. In 2014, the highest number of accidents were recorded, with approximately 17% of all accidents occurring in that year. The least number of accidents were recorded in 2010. The highest death rate per road accident was recorded in 2010, while 2012 recorded the lowest death rate.

Figure 3 below is a map indicating the district municipalities in Limpopo and shows the percentage contribution to RTDs per district. It can be seen that the Capricorn district recorded the highest number of deaths, followed by the Waterberg and Vhembe districts. Mopani and Greater Sekhukhune recorded the least number of deaths.

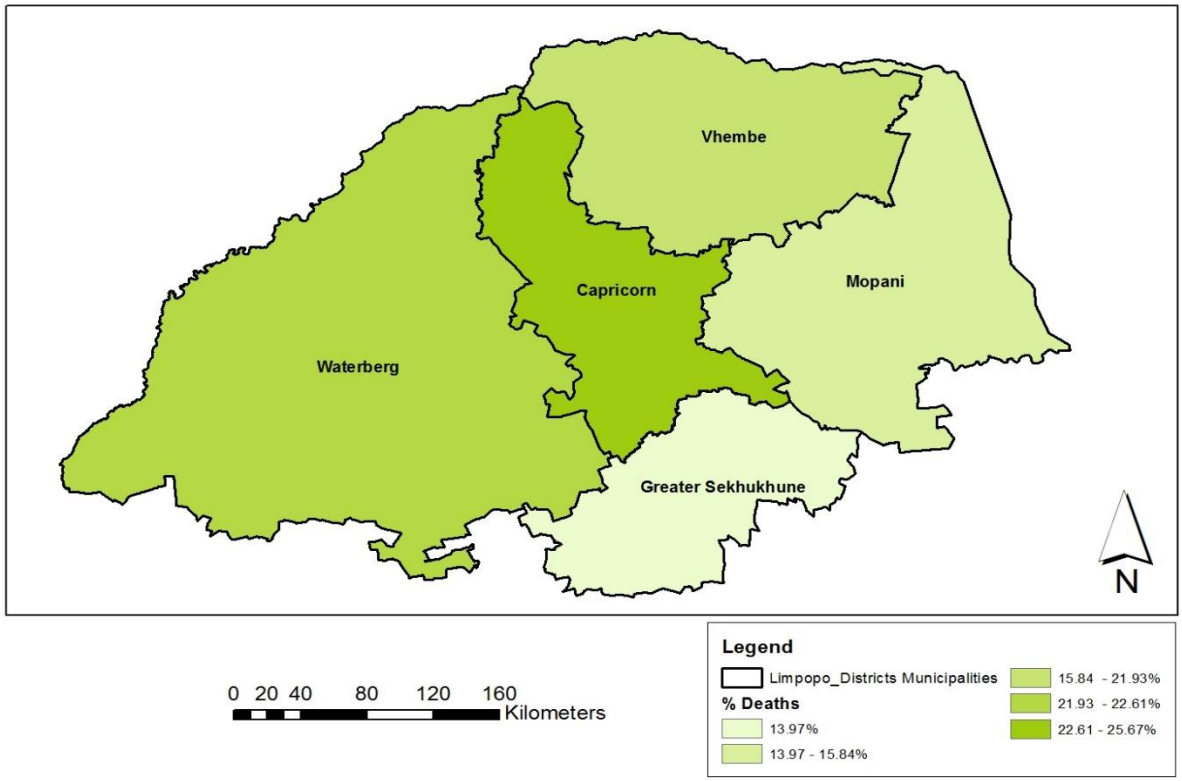

Figure 3: The Road Traffic Deaths distribution per district.

Figure 4 depicts a map of the Limpopo municipal districts and shows the percentage contrition to RTAs per district. It can be seen that the Capricorn district recorded the highest number of accidents, followed by the Mopani and Vhembe districts. Waterberg and Greater Sekhukhune recorded the least number of cases.

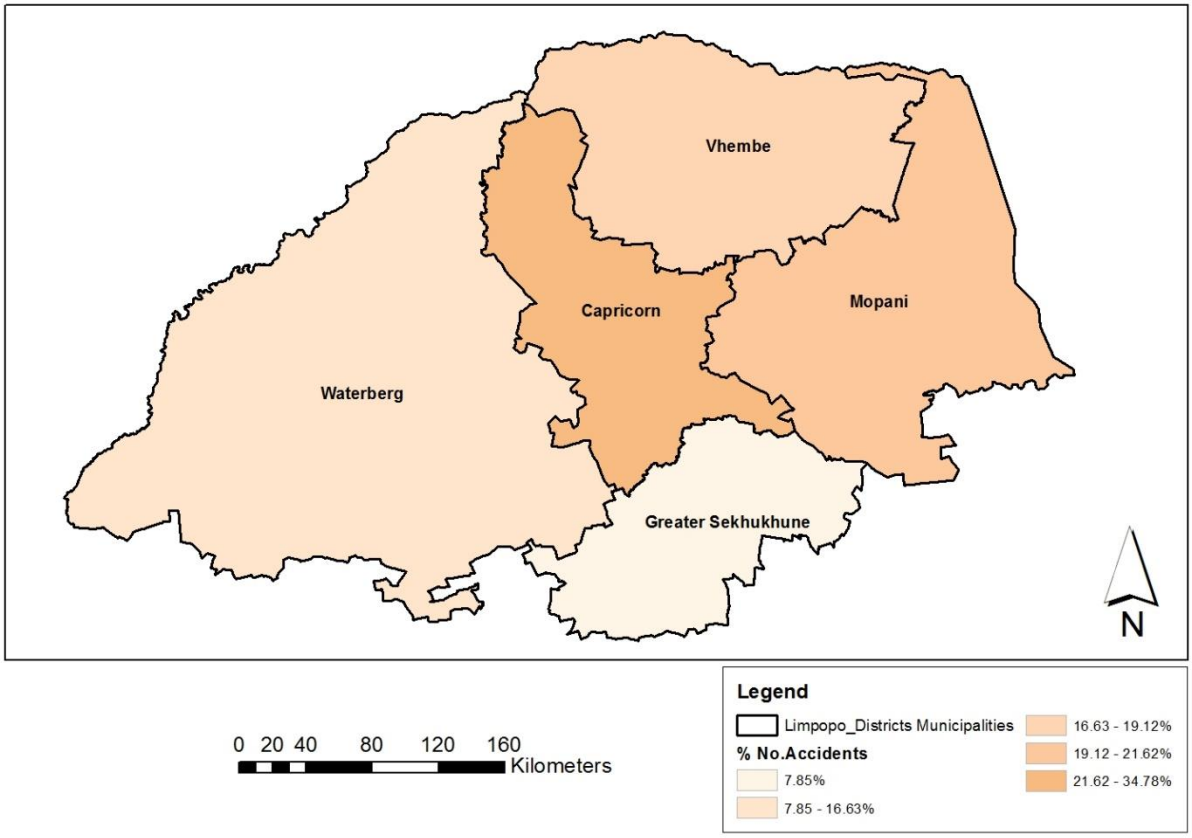

Figure 4: The Road Traffic Accidents distribution per district.

Table 2: The yearly distribution of deaths percentage contribution per district.

District	2009	2010	2011	2012	2013	2014	2015	% Total
Capricorn	35%	28%	23%	28%	26%	19%	23%	26%
Mopani	13%	15%	14%	16%	13%	20%	18%	16%
Sekhukhune	9%	14%	12%	12%	15%	19%	16%	14%
Vhembe	20%	20%	21%	27%	23%	21%	23%	21%
Waterberg	23%	23%	30%	17%	23%	21%	20%	23%

The Capricorn district recorded the highest number of deaths during the years under review, with the exception of 2011 and 2014. The Sekhukhune district recorded the lowest number of deaths during the period under review, as depicted in Table 2. The Capricorn district recorded the highest number of accidents during the period under review, except for 2012. The Sekhukhune district recorded the lowest number of accidents during this period, as shown in Table 3. Overall, the Capricorn district recorded highest number of deaths and accidents between the years 2009 and 2015.

Table 3: The yearly distribution of accidents percentage contribution per district

District	2009	2010	2011	2012	2013	2014	2015	% Total
Capricorn	48%	46%	36%	29%	28%	28%	31%	35%
Mopani	15%	18%	19%	20%	25%	26%	25%	21%
Sekhukhune	6%	8%	8%	8%	9%	8%	9%	8%
Vhembe	6%	7%	20%	31%	25%	23%	20%	19%
Waterberg	25%	21%	17%	12%	13%	15%	15%	17%

Table 4: Number of vehicles involved in accidents.

Vehicle Involved	No of Accidents	%	No of Deaths	%	Rate of death per accident
1	12302	68%	3187	65%	0.2591
2	5525	31%	1645	33%	0.2977
3 or more	202	1%	112	2%	0.5545
Total	**18029**	**100%**	**4944**	**100%**	

Table 4 shows the number of vehicles involved in road accidents. It can be seen from this table that, in approximately 68% of accidents, only one vehicle was involved. Road accidents involving one vehicle accounted for approximately 65%of the total number of deaths. The deaths rate increased significantly when more vehicles are involved in an accident.

4.2.2. The Monthly Distribution of RTAs and RTDs

I investigated how RTAs and RTDs were distributed on a monthly basis and the rate of deaths per accidents within that month. The month of December recorded that the highest number of accidents and deaths, as illustrated in Table 5. This month alone accounted for approximately 14% of all the RTAs and 17% of all RTDs recorded during the period under review. The month of January recorded the lowest number of RTAs and RTDs during this period. If there were 100 accidents in December, there is a risk that such incidents will account for more than a quarter of all deaths reported. If 100 accidents occurred in July, there is a chance that 31 deaths would result.

Table 5: Monthly distribution of Road Traffic Accidents and Road Traffic Deaths from January 2009 to December 2015.

Month	No of deaths	%	No of accidents	%	Rate of death per accident
January	289	5.85%	1127	6.25%	0.2564
February	321	6.49%	1143	6.34%	0.2808
March	369	7.46%	1362	7.55%	0.2709
April	441	8.92%	1700	9.43%	0.2594
May	357	7.22%	1397	7.75%	0.2556
June	385	7.79%	1307	7.25%	0.2946
July	443	8.96%	1447	8.03%	0.3062
August	351	7.10%	1554	8.62%	0.2259
September	379	7.67%	1620	8.99%	0.2340
October	347	7.02%	1447	8.03%	0.2398
November	433	8.76%	1463	8.11%	0.2960
December	829	16.77%	2462	13.66%	0.3367

Table 6: Monthly distribution of Road Traffic Injuries from January 2009 to December 2015.

Month	Minor Injury	Serious Injury	Death	Total	% Casualty
January	1734	949	289	2972	32%
February	1492	1075	321	2888	37%
March	1973	1175	369	3517	33%
April	2200	1411	441	4052	35%
May	1752	1147	357	3256	35%
June	1762	1183	385	3330	36%
July	1828	1295	443	3566	36%
August	1948	1232	351	3531	35%
September	2096	1329	379	3804	35%
October	1745	1165	347	3257	36%
November	1929	1220	433	3582	34%
December	3118	1823	829	5770	32%
Total	**23577**	**15004**	**4944**	**43525**	
%	**54%**	**34%**	**11%**	**100%**	
Average	**1965**	**1250**	**412**	**3627**	

Table 6 shows the monthly variation in the (RTI) between 2009 and 2015. It can be seen from this table that all the months during this period recorded at least a 32% casualty rate $\left(\frac{serious\ injuries}{deaths+serious\ injuries+minor\ injuries}\right)$. In other words, any accident in any of the month had more than a 32% chance of resulting in the loss of life. On average, approximately 281 (1,965/7 years of study period) persons sustained minor injuries in RTAs per month, while approximately 178 persons (1,250/7 years of study period)

persons sustained serious injuries and approximately 59 lives (412/7 years of study period) were lost per month.

4.2.3. The Distribution of RTAs and RTDs by Day of Week

When there are more cars on the road, such as on a particular day of the week or over a weekend, a person's chances of being involved in a road accident increases. Table 7 below shows the death rate per accident distributed per day of week recorded during the seven-year study period.

Table 7: Total number of Road Traffic Accidents and deaths by day of week

Day of week	No of Deaths	%	No of Accidents	%	%Rate of Death per Accident
Sun	1054	21%	3328	18%	32.7%
Mon	470	10%	1805	10%	26.0%
Tue	382	8%	1692	9%	22.6%
Wed	336	7%	1625	9%	20.1%
Thu	554	11%	1873	11%	29.6%
Fri	862	17%	3221	18%	26.8%
Sat	1286	26%	4485	25%	28.7%
Total	**4944**	**100%**	**18029**	**100%**	**27.4%**

The highest number of accidents and deaths were recorded on Saturday, with 25% of all accidents and 26% of all deaths being recorded on this day. The lowest number of accidents and deaths were recorded on Wednesday, followed by Tuesday.

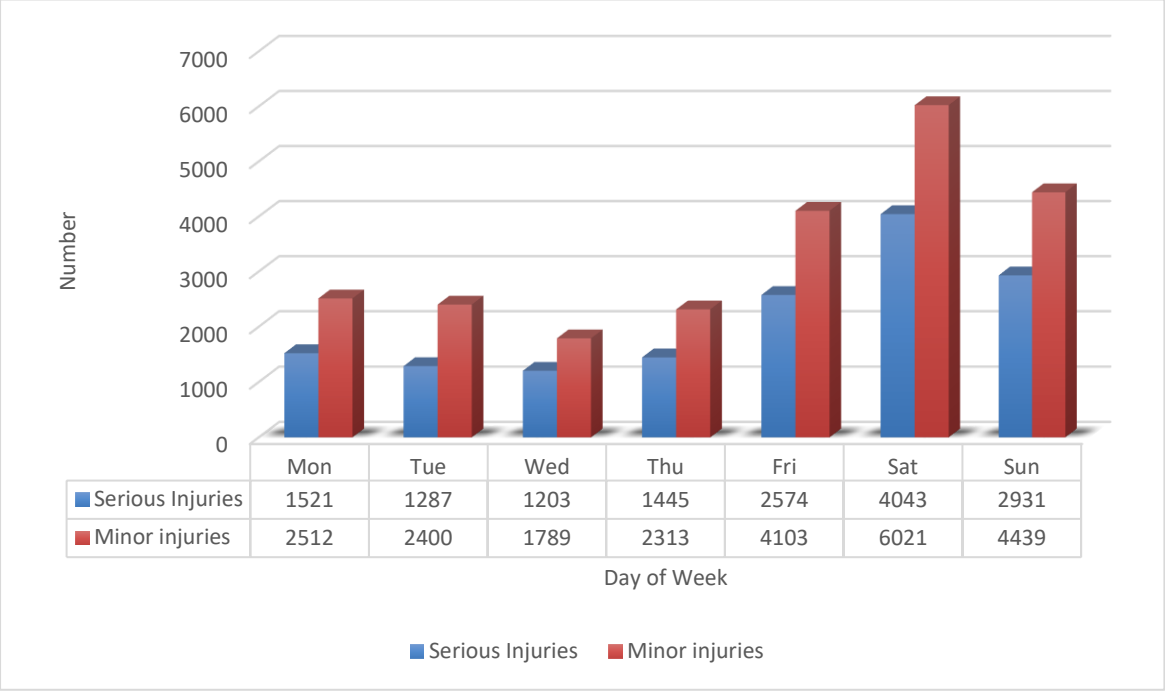

Figure 5: Road Traffic Injuries distributed by day of week.

Figure 5 illustrates road injuries in the Province that occurred during the period under review. Saturday recorded the highest number of serious and minor injuries with 27% of all serious injuries and 26% of all minor injuries recorded on this day. The lowest number for serious and minor injuries were recorded on Wednesday.

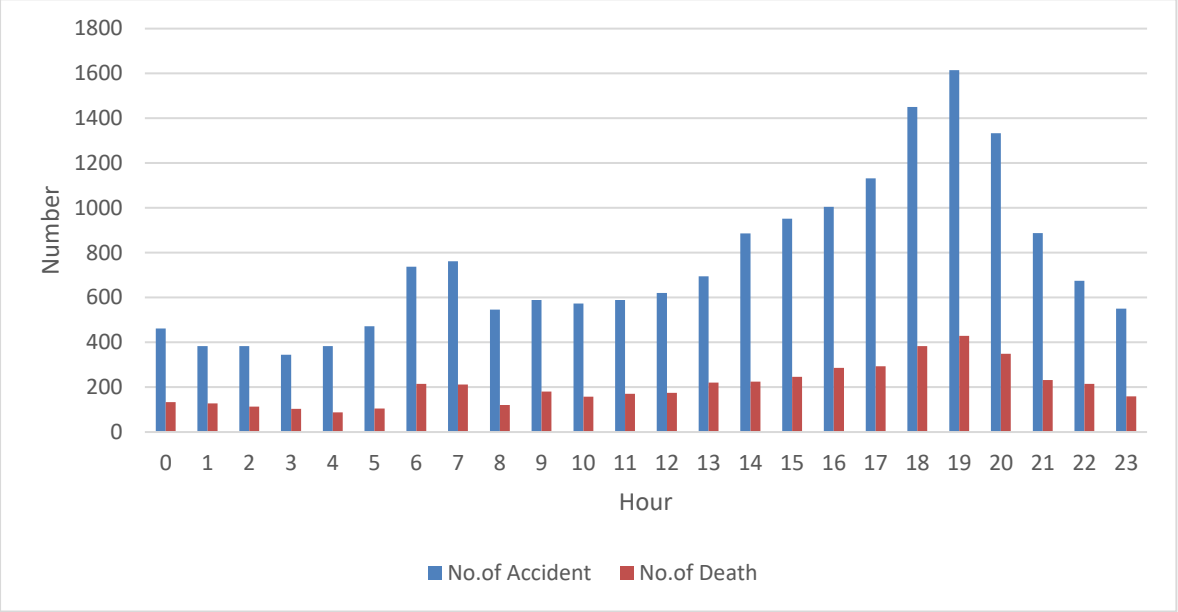

Figure 6: Hourly distribution of Road Traffic Accidents and Road Traffic Deaths, 2009-2015.

The hourly distribution of RTAs and RTDs during the period under review is presented by Figure 6. It can be seen from this figure that highest number of RTAs occurred during rush hour, between 5 pm to 8 pm. The highest number of people lost their lives during that time (5 pm to 8 pm) and also in the morning between 5 am and 7 am.

4.2.4. Distribution of RTAs and RTDs

According to the literature, road accidents occur as the result of one or more of the following factors: human factors, vehicle factors and road and environment factors. Human factors are described as factors directly attributable to the operator of the vehicle or to people involved in an accident. Human factors include the following: speeding, traffic violation, alcohol, drugs, negligence, driver error and fatigue. Road and environment factors refer to all aspects of road design, weather conditions, road conditions, traffic signs and lights. Vehicle factors refer to vehicle condition, maintenance and mechanical faults in the vehicle. These classifications aim to assist in the conceptualization of the problem.

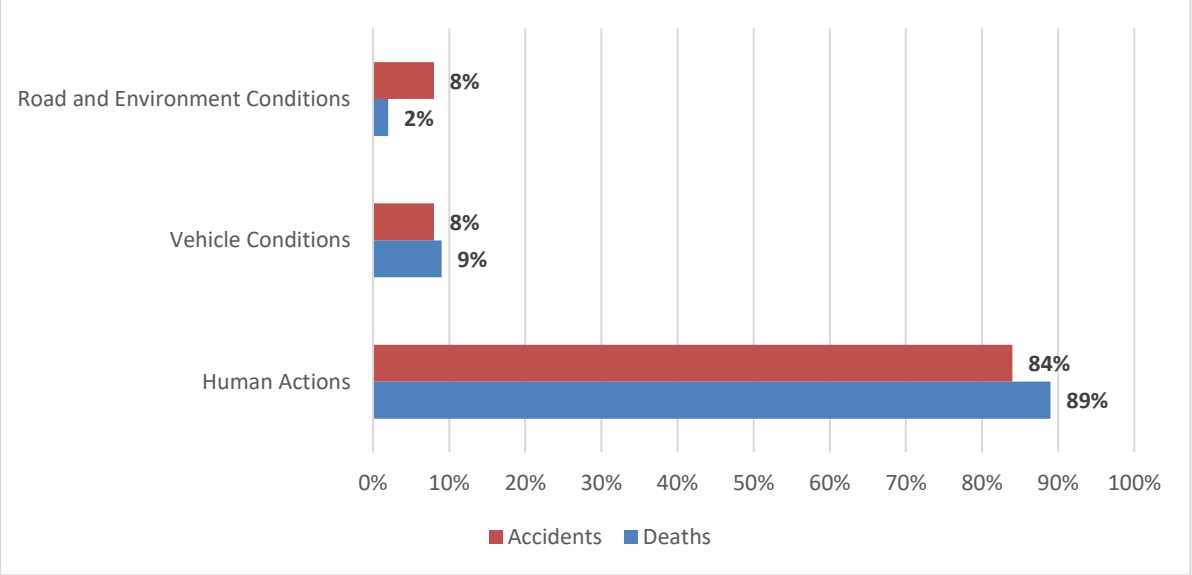

Figure 7: Categorised contributing factors.

The three factors that contribute to road accidents and deaths are represented by Figure 7. It can be seen in the Figure 7 that human factors contributed to the highest number of deaths during the period under review, with 89% of all deaths said to be as a result of human factors, with speeding and pedestrian carelessness being the primary contributors. The number of deaths as a result of human factors was followed by vehicle condition factors, accounting for 9% of deaths during the period under review, with tyre bursts being the primary contributor as shown in Table 8. It can also be seen in Figure 7 that that road and environment condition factors resulted in the lowest number of deaths, at 2%, with animals on the roadway being the highest contributor. Human factors accounted for 84% of the total number of accidents, while vehicle and road and environment conditions both accounted for 8% of the total number of accidents.

An analysis of the data found that human factors accounted for the highest number of serious and minor injuries, followed by vehicle condition factors. Road and environment condition factors accounted for the lowest number of serious and minor injuries between January 2009 and December 2015.

4.3. CHAPTER SUMMARY

Descriptive statistics analyses was conducted in this chapter 4, the data was summarised as bar charts, cross tabulations and line graphs.

Table 8: Contributing factors to road accidents deaths and injuries in the Limpopo Province.

Contributing Factors	Deaths		Serious Injuries		Minor Injuries	
	Number	Percent	Number	Percent	Number	Percent
Environmental Condition	**107**	**2%**	**661**	**4%**	**1355**	**8%**
Animal in roadway	73	1%	486	3%	1099	7%
Deposit on road	6	0%	42	0%	19	0%
Multi vehicle pile up	1	0%	7	0%	4	0%
Defective road surface	2	0%	23	0%	42	0%
Potholes	6	0%	22	0%	22	0%
Rain	6	0%	49	0%	100	1%
Road layout	0	0%	1	0%	7	0%
Slippery Road	4	0%	17	0%	31	0%
Parked vehicle	9	0%	14	0%	31	0%
Human Action	**4404**	**89%**	**12441**	**82%**	**13824**	**83%**
Change lane	10	0%	92	1%	71	0%
Road marking	98	2%	553	4%	675	4%
Crossing the road unsafe	0	0%	0	0%	3	0%
Cyclist in roadway	28	1%	29	0%	59	0%
Disobeyed stop sign	2	0%	17	0%	24	0%
Driving into an obstacle	0	0%	1	0%	10	0%
Entering the road unsafe	21	0%	142	1%	151	1%
Entering the road unsafe	3	0%	48	0%	57	0%
Following too close	165	3%	1055	7%	2133	13%
Head On	28	1%	72	0%	58	0%
Head rear end	4	0%	7	0%	16	0%
Overtaking	436	9%	1372	9%	911	5%
Overturned	0	0%	0	0%	1	0%
Reckless driving	332	7%	1483	10%	1797	11%
Sideswipe	0	0%	18	0%	32	0%
Speeding	1655	33%	5358	36%	4931	30%
Concentration	0	0%	0	0%	2	0%
Driver distraction	8	0%	16	0%	22	0%
Drunken driving	79	2%	342	2%	297	2%
Fatigue	260	5%	515	3%	463	3%
Lost control	68	1%	243	2%	215	1%
Passenger fell	27	1%	16	0%	27	0%
Pedestrian careless	1082	22%	785	5%	1603	10%
Vehicle Condition	**433**	**9%**	**1892**	**13%**	**1445**	**9%**
Brakes failure	50	1%	188	1%	219	1%
Defective lights	11	0%	68	0%	42	0%
Mechanically fault	6	0%	40	0%	44	0%
Overloaded or poorly loaded	0	0%	3	0%	3	0%
Overloaded	22	0%	36	0%	52	0%
Tyre burst	337	7%	1550	10%	1026	6%
Vehicle burned	7	0%	7	0%	58	0%
Visor or widescreen dirty	0	0%	0	0%	1	0%
Overall Total	**4944**	**100%**	**14994**	**100%**	**16624**	**100%**

CHAPTER 5: MODEL FITTING

In this chapter, will first fit the logistic regression model to occurrence of death due to accidents. We then present the models and extensions of the Poisson and the negative binomial regression models. Model diagnosis will also be conducted in each section and the model goodness of fit will be measured. R statistical software version 3.5.2 (Venables and Smith, 2003) was used to fit the models.

5.1. LOGISTIC REGRESSION MODEL

5.1.1. Model Fitting

I fitted a logistic regression model to occurrence of death, given that an accident had occurred, as a function of vehicle type, time of day, region, holiday, day of week, road type and categorised contributing factors. The K-means clustering was used to group time by hour intervals. The model goodness of fit is based on the Akaike information criterion (AIC). In Table 9, I fitted the logistic regression model to each of the explanatory variables and calculated the AIC values as a way of coming up with the optimal model.

Table 9: Logistic regression models with one and all combined explanatory variables

Models	AIC
1. $\text{logit}\,(\pi(deaths)) = \alpha_2 + \beta_k * Vehicle\ type, \quad k = 1,2,\cdots,6$	17248
2. $\text{logit}\,(\pi(deaths)) = \alpha_2 + \beta_k * Day\ of\ Week, \quad k = 1,2,\cdots,7$	16438
3. $\text{logit}(\pi(deaths)) = \alpha_2 + \beta * Holidays$	16433
4. $\text{logit}(\pi(deaths)) = \alpha_2 + \beta_k * Hour\ Interval, \quad k = 1,2,3$	16484
5. $\text{logit}(\pi(deaths)) = \alpha_2 + \beta_k * Road\ Type, \quad k = 1,2,\cdots,5$	16469
6. $\text{logit}(\pi(deaths)) = \alpha_2 + \beta_k * Contributing\ Factors, \quad k = 1,2,\cdots,6$	16721
7. $\text{logit}(\pi(deaths)) = \alpha_2 + \beta_k * Region, \quad k = 1,2,\cdots,5$	16761
8. $\text{logit}(\pi(deaths)) = \alpha_2 + \beta(vehicle + Day\ of\ Week + Holidays + Hour\ Interval + Road\ Type + Contributing\ Factors)$	16449

Table 9 shows that the model with all the combined variables that had the smallest AIC value of 16449. The coefficient estimates of the selected logistic regression model (Model 8) with only significant variables being selected by the bidirectional elimination stepwise regression method based on the AIC is shown in Table 10.

Table 10: Parameter estimates for logistic regression model, using maximum likelihood estimation.

					CI	
Variables	**Coefficient (β)**	**SE**	**P-value**	**Exp(β)**	**2.5**	**97.5**
Intercept	-1.1285	0.0987	0.00***	0.3235	0.2662	0.3920
Holiday	0.1590	0.0594	0.007**	1.1723	1.0426	1.3162
Day of week						
Friday	0.1457	0.0673	0.0305*	1.1568	1.0137	1.3202
Monday	0.1481	0.0790	0.0610	1.1596	0.9925	1.3534
Tuesday	0.3748	0.6873	0.0684	1.4547	0.4535	1.0280
Saturday	0.1851	0.0627	0.003**	1.2034	1.0645	1.3612
Sunday	0.2856	0.0663	0.00***	1.3305	1.1685	1.5156
Thursday	0.2056	0.0766	0.007**	1.2283	1.0565	1.426
Road types						
National road	0.1503	0.0531	0.005**	1.1622	1.0469	1.2894
Others	0.3202	0.0571	0.00***	1.3774	1.2308	1.5398
Provincial road	-0.2866	0.1112	0.009**	0.7507	0.6013	0.9302
District						
Capricorn	-0.5526	0.0490	0.00***	0.5754	0.5226	0.6333
Sekhukhune	0.5579	0.0674	0.00***	0.5781	0.5155	0.6479
Mopani	-0.5478	0.0583	0.00***	1.7470	1.5302	1.9932
Time interval						
00-05	0.2730	0.0555	0.00***	1.3139	1.1777	1.4645
06-13	-0.2077	0.0466	0.00***	0.8123	0.7411	0.8898
Contributing factors						
Human Actions	0.5069	0.0763	0.00***	1.6601	1.4319	1.9320
Environment Conditions	-1.1310	0.1411	0.00***	0.3226	0.2433	0.4233
Vehicle types						
Sedan	-0.5606	0.0347	0.00***	0.5708	0.5330	0.6108
LDV	-0.6150	0.0402	0.00***	0.5406	0.4993	0.5847
Combi	-0.3678	0.0562	0.00***	0.6922	0.6193	0.7722
Truck	-0.1766	0.0537	0.001**	0.8380	0.7536	0.9304
Significant Codes						
0.001 '***'	0.01 '**'		0.05 '*'		0.1 '.'	

This is the fitted logistic regression

$$x'\beta = -1.13 + 0.16 * Holiday + 0.15 * (Friday + Monday) + 0.38 * Tuesday + 0.18 * Saturday + 0.29 * Sunday + 0.21 * Thursday + 0.15 * National\ roads + 0.32 * other\ roads - 0.29 * Provincial\ roads - 0.55 * Capricorn + 0.56 * Sekhukhune - 0.55 * Mopani + 0.27 * Time\ interval\ [00 - 05] - 0.21 * Time\ interval\ [06 - 13] + 0.51 * human\ actions - 1.13 * environment\ conditions - 0.56 * sedan - 0.62 * LDV - 0.37 * combi - 0.18 * truck$$

$$\pi = \frac{e^{x'\beta}}{1 + e^{x'\beta}}$$

The logistic regression coefficients of the exponential can be interpreted as follows. The odds are conditional on the occurrence of accidents:

- The odds of death occurring during holidays is 1.17 times the odds of death occurring during non-holidays.
- The odds of death occurring on a Sunday is 1.33 times the odds of death occurring on a Wednesday.
- The odds of death occurring on other roads is 1.38 times the odds of death occurring on districts roads.

5.1.2. Model Diagnostics

After fitting a regression model, it is important to determine whether there are any assumption violations of the logistic regression model. Therefore, I performed appropriate model diagnostics. The model diagnosis involved a graphical plot of the residual against the predicted, and the observed against the expected.

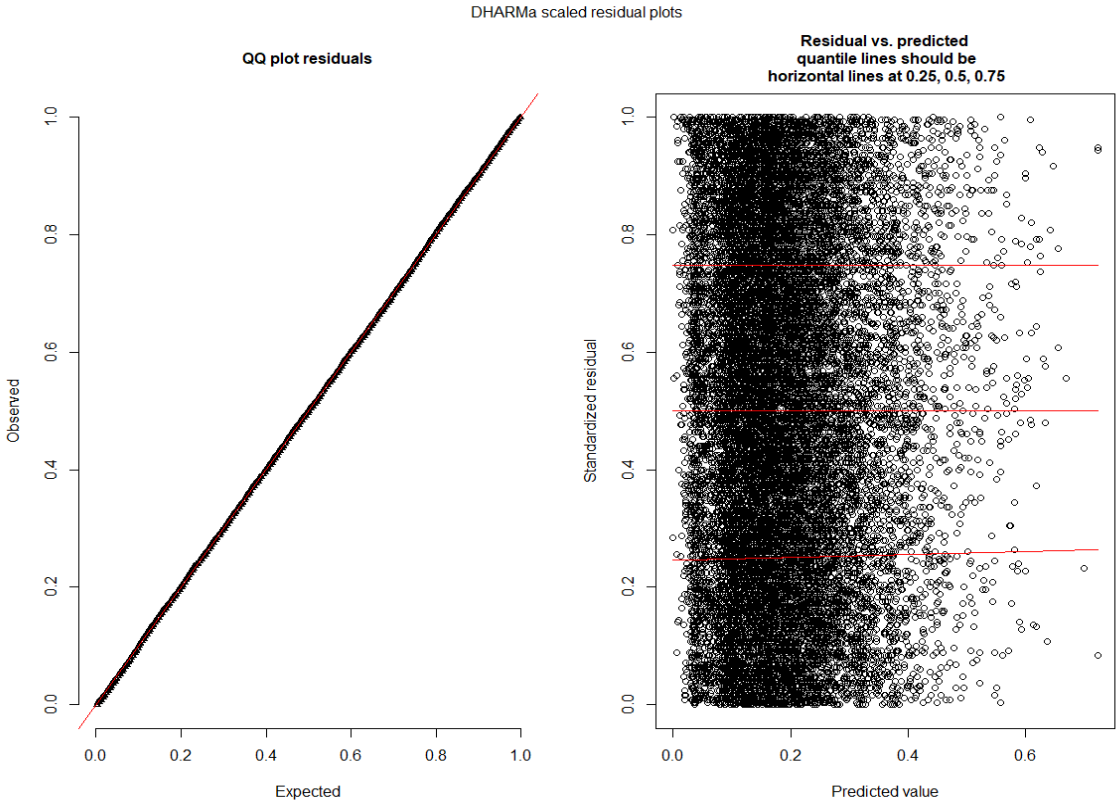

Figure 8: Logistic regression model diagnostic, expected against predicted.

In the observed against the predicted plots, the points seem to fall along a straight line, indicating that there is a good relationship between the predicted and the observed values. In the residuals against the predicted plot, it can be seen that the red dashed lines look straight and horizontal, suggesting that the residuals are spread equally along the ranges of the predictors.

The receiver operating characteristic (ROC) curve plot is generated by plotting sensitivity (probability of correctly detecting a death) against specificity (probability of correctly detecting a non-death), as shown in Figure 9. The diagonal line, from (0, 0) to (1, 1), is indicative of an independent variable that discriminates no difference of sensitivity against specificity. The area under the ROC curve illustrates the likelihood that the proposed model will determine deaths with higher probability than non-deaths. A model with no discrimination will have no area under the curve, which would produce a straight line.

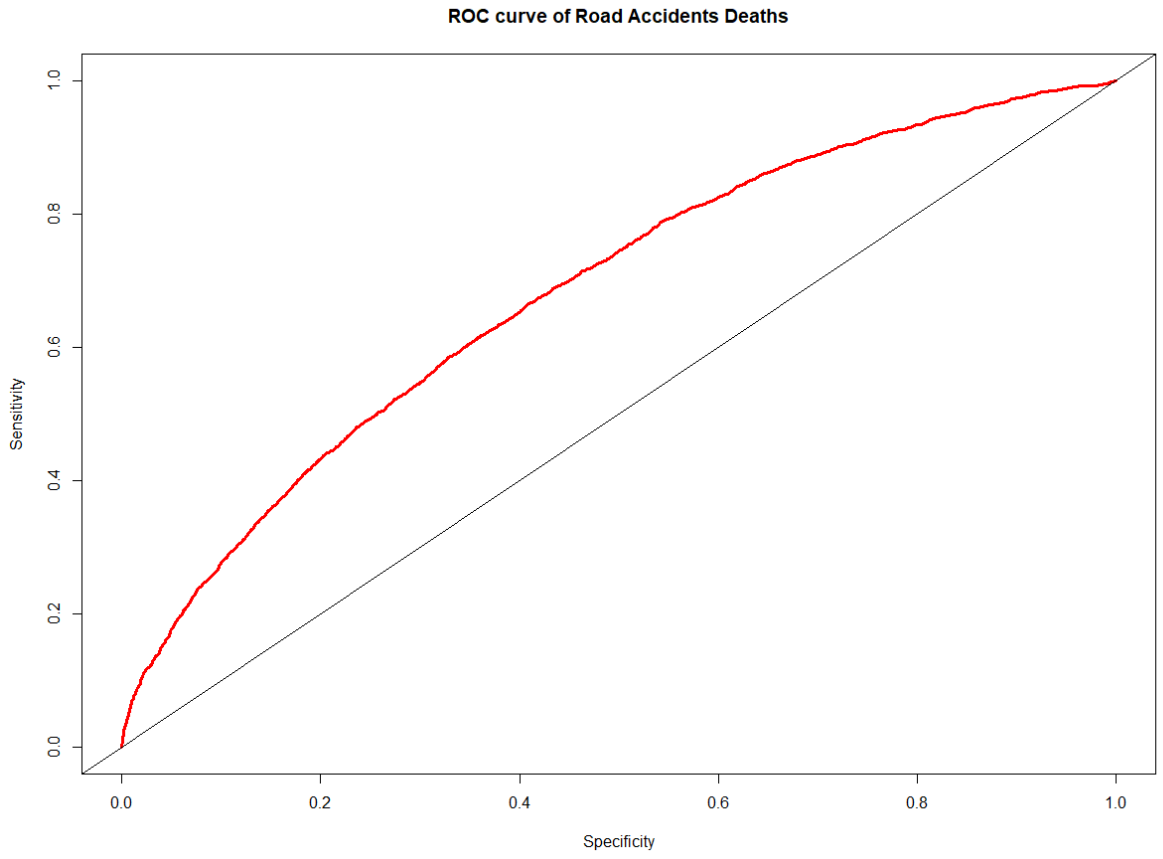

Figure 9: ROC curve for logistic regression model.

The area under the ROC curve for the model had a value of 0.68 above the diagonal line, suggesting that the logistic regression model would be considered to be fair at

separating road deaths from non-deaths. After undertaking all the diagnostics, I can safely conclude that the logistic regression model represents a better fitting model in predicting the probability of the occurrence of death given that an accident has occurred.

5.2. POISSON REGRESSION MODEL

The data used is the daily RTAs data gathered from January 2009 to December 2015. It was found, as depicted in Table 11, which the majority of deaths occurred during holidays when compared to non-holidays, with the highest number of deaths recorded on a Saturday, as a result of human actions on national roads.

Table 11: Frequency of death distributed by holidays and no-holidays

	Non-Holiday	Holiday
Variables	**N(%)**	**N(%)**
Day of week		
Sunday	978 (23%)	76 (12%)
Monday	411 (10%)	59 (9
Tuesday	339 (8%)	43 (7%)
Wednesday	283 (7%)	53 (8%)
Thursday	438 (10%)	116 (18%)
Friday	699 (16%)	163 (25%)
Saturday	1148 (27%)	138 (21%)
Contributing Factors		
Environment Conditions	101 (2%)	6 (0.93%)
Human Actions	3815 (89%)	589 (90.90%)
Vehicle Conditions	380 (9%)	53 (8.17%)
Road Types		
District Road	390 (9%)	61 (9%)
National Road	1119 (26%)	169 (26%)
Others Road	612 (14%)	113 (18%)
Provincial Road	104 (3%)	31 (5%)
Regional Road	2071 (48%)	274 (42%)

The patterns of road deaths differ depending on whether or not it was a holiday. Again, it was found that both the explanatory variables and the response variable (number of deaths) have an effect on the variable holiday. Therefore, the data was split into two: deaths that occurred during holidays and those that occurred during non-holidays. Holidays were New Year's Day, Human Rights Day, Good Friday, Family Day, Freedom Day, Labour Day, Public Day, Youth Day, National Women's Day, Heritage Day, Day of Reconciliation, Christmas Day and Day of Good Will.

Using Chi-square test statistics to test for independence between the number of deaths and the explanatory variables, only variables: region, type of vehicle, yearly quarters, categorised contributing factors, day of week and types of road were found to be statistically associated with the reported number of deaths (see Table 22(A)). These variables are used to fit the standard regression model for Poisson.

5.2.1. Deaths During Holidays

The standard Poisson regression model coefficient estimates for death during holidays is displayed in the table below.

Table 12: Coefficient estimates for the standard Poisson model for deaths during the holidays.

			C.I			
Variables	**Estimate**	**Std. Error**	**2.5%**	**97.5%**	**P-value**	
Intercept	-3.3907	0.4422	-4.3655	-2.6037	<0.0001	***
Sunday	-0.2250	0.1395	-0.5034	0.0443	0.1067	
Monday	-0.4755	0.1530	-0.7829	-0.1817	0.0018	**
Tuesday	-0.4436	0.1723	-0.7932	-0.1158	0.0100	*
Wednesday	-0.0229	0.1591	-0.3436	0.2815	0.8852	
Thursday	0.1915	0.1217	-0.0490	0.4289	0.1158	
Saturday	0.0207	0.1165	-0.2085	0.2486	0.8586	
Human Actions	2.0878	0.4114	1.3737	3.0153	<0.0001	***
Vehicle Conditions	1.8445	0.4326	1.0768	2.8030	<0.0001	***
National road	0.0957	0.1533	-0.1991	0.4031	0.5323	
Others road	0.1974	0.1607	-0.1135	0.5179	0.2193	
Provincial road	-0.1794	0.2226	-0.6275	0.2487	0.4204	
Regional road	-0.0278	0.1435	-0.3016	0.2619	0.8462	
Sedan	-0.2636	0.0553	-0.3736	-0.1567	<0.0001	***
LDV	-0.1740	0.0675	-0.3082	-0.0433	0.0099	**
Combi	0.1639	0.0735	0.0165	0.3049	0.0257	*
Bus	0.1103	0.1885	-0.2866	0.4560	0.5585	
Motorcycle	0.6740	0.2382	0.1731	1.1123	0.0046	**
Likelihood						
Log-likelihood	-1403.338		**Model *df***		19	
Significant Codes						
0.001 '***'	0.01 '**'		0.05 '*'		0.1 '.'	

When modelling of death occurring due to accidents, the number of accidents was treated as an offset variable in the model. It can be observed that not all the predictors in the model were highly statistically significant. In the table above, highly statistically significant variables at the 0.001 significant level are indicated with three asterisks, two asterisks indicate highly statistically significant at the 0.01 significant level, while

one asterisk indicates statistical significance at the 0.05 significant level and a full-stop it shows statistical significant at the 0.1 significant level.

The regression coefficients are interpreted as any other unstandardized coefficients from a standard Poisson regression model. The regression coefficient associated with the human actions factor is 2.09, meaning that for each one-unit increase in the number of accidents due to human actions, log mean death increases by 2.09 units. The mean deaths that occurred on a Monday is 0.62 ($e^{-0.48} = 0.62$) times less than the mean deaths that occurred on a Friday.

Testing for model goodness of fit based on the Chi-square test with residual deviance 1833.4 and 1720 degrees of freedom, the p-value was 0.028, less than the 0.05 significance level, and, therefore, the null hypothesis cannot be accepted and one must conclude that the standard Poisson regression model, as a whole, does not fit the data significantly better than baseline model with interception alone. The standard Poisson regression model diagnostic is shown in Figure 10. The plots compare two distributions, the observed and the fitted distribution using their quartiles.

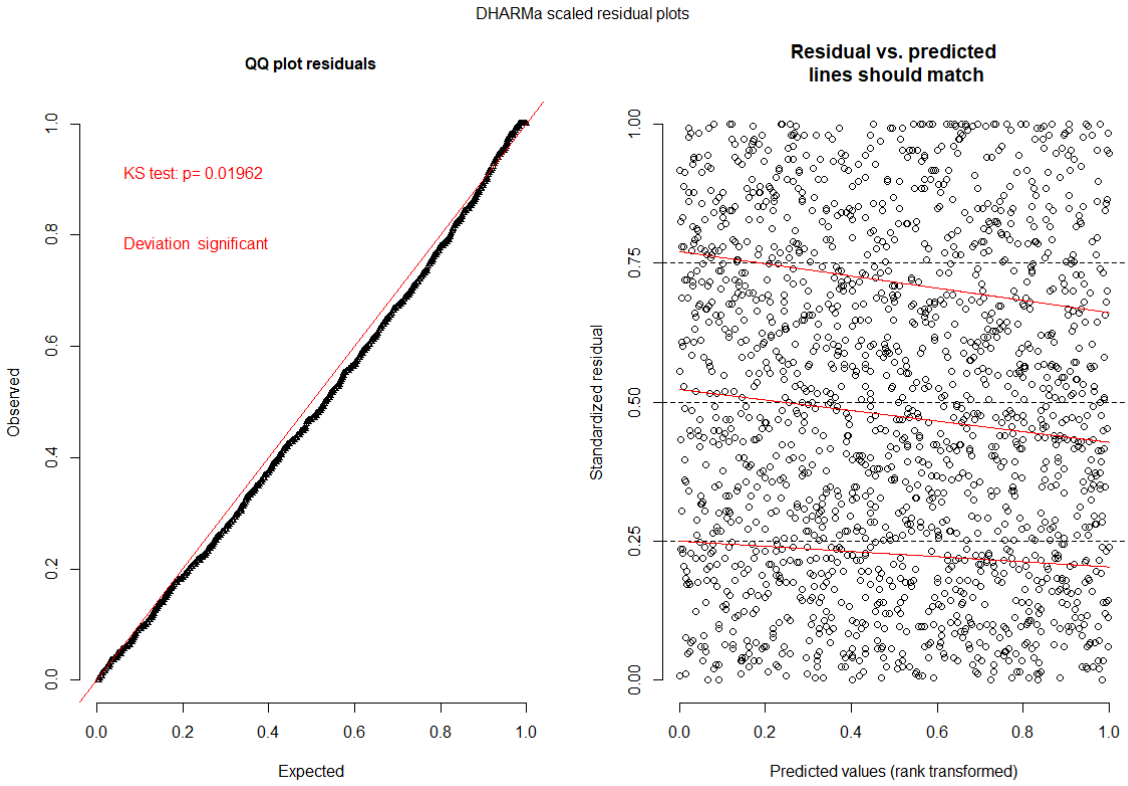

Figure 10: Standard Poisson model diagnostic, observed against predicted values.

The standard Poisson regression model diagnostic is shown in Figure 10. It can be observed that the red dashed lines are not straight and horizontal at y-values of 25%, 50% and 75% quartiles. This suggests that there is no agreement between the expected and the observed values.

To examine over-dispersion in our model, we used a non-parametric over-dispersion test from the R package, called DHARMa (Harting, 2016), to test for over-dispersion or under-dispersion.

Table 13: Testing for over-dispersion or under-dispersion in the model.

Parameter	Estimate	Z-value	P-value
Dispersion	1.7942	3.9559	<0.0001

The test statistic and p-values, respectively, are shown in Table 13, testing the null hypothesis that the true dispersion is equal to one. It was found that the p-value is less than the 0.05, and the null hypothesis is rejected, leading to the conclusion that the dispersion parameter is not equal to 1, instead it is greater than one, suggesting over-dispersion relative to the standard Poisson regression model. This implies that the conditional variance is greater than the conditional mean. This suggest that the model is mis-specified and that the explanatory variables may not well explain the number of deaths.

One common cause for over-dispersion is zero inflation. This is a phenomenon found in data where there are more zeros than expected. To factor this phenomenon in, we first tested for zero inflation using testZeroInflation function in R software. This function compares the observed number of zeros to the zeros expected from simulations. The ratio of observed against expected was found to be 1.0596, with a p-value less than 0.0001. I, therefore, reject the null hypothesis stating that the expected zeros and observed zeros are equal and conclude that the data poses zero inflation.

5.2.2. Death During Non-Holidays

When modelling of deaths occurring due to accidents, the number of accidents was treated as an offset variable in the mode. Based on the Chi-square test with residual deviance 13812 and 13475 degrees of freedom, the p-value was found to be less than 0.05. We, therefore, failed to reject the null hypothesis and conclude that the

standard Poisson regression model, as a whole, does not fit significantly better than a model with only the intercept. The standard Poisson regression model coefficient estimates for death during non-holidays is displayed in the table below.

Table 14: Coefficient estimates for standard Poisson model for death during non-holidays.

			C.I			
Variables	**Estimate**	**Std. Error**	**2.5%**	**97.5%**	**P-value**	
Intercept	-2.7908	0.1209	-3.0327	2.5583	<0.0001	***
Sunday	0.2326	0.0496	0.1355	0.3301	<0.0001	***
Monday	-0.0085	0.0623	-0.1314	0.1131	0.8914	
Tuesday	-0.2039	0.0664	-0.3351	-0.0746	0.0021	**
Wednesday	-0.3429	0.0706	-0.4827	-0.2058	<0.0001	***
Thursday	0.0215	0.0610	-0.0986	0.1407	0.7235	
Saturday	0.1334	0.0482	0.0392	0.2283	0.0056	**
Human Actions	1.2882	0.1011	1.0958	1.4929	<0.0001	***
Vehicle Conditions	1.0692	0.1126	0.8528	1.2948	<0.0001	***
National road	0.2869	0.0594	0.1713	0.4045	<0.0001	***
Others road	0.1806	0.0652	0.0532	0.3091	0.0056	**
Provincial road	-0.5026	0.1107	-0.7245	-0.2898	<0.0001	***
Regional road	0.0377	0.0558	-0.0704	0.1486	0.4988	
Sedan	-0.2020	0.0219	-0.2453	-0.1591	<0.0001	***
LDV	-0.2624	0.0259	-0.3135	-0.2118	<0.0001	***
Combi	0.1760	0.0305	0.1156	0.2355	<0.0001	***
Bus	0.6354	0.0627	0.5101	0.7560	<0.0001	***
Motorcycle	0.1262	0.1153	-0.1089	0.3437	0.2738	
Likelihood						
Log-likelihood	-10193.67		**Model *df***		19	
Significant Codes						
0.001 '***'	0.01 '**'		0.05 '*'		0.1 '.'	

The regression coefficient associated with Sunday is 0.23, meaning that for each one-unit increase in the number of accidents on Sunday, the mean death increases by 0.23 units. The mean deaths involving busses is 1.90 ($e^{0.64} = 1.90$) times greater than other vehicle types.

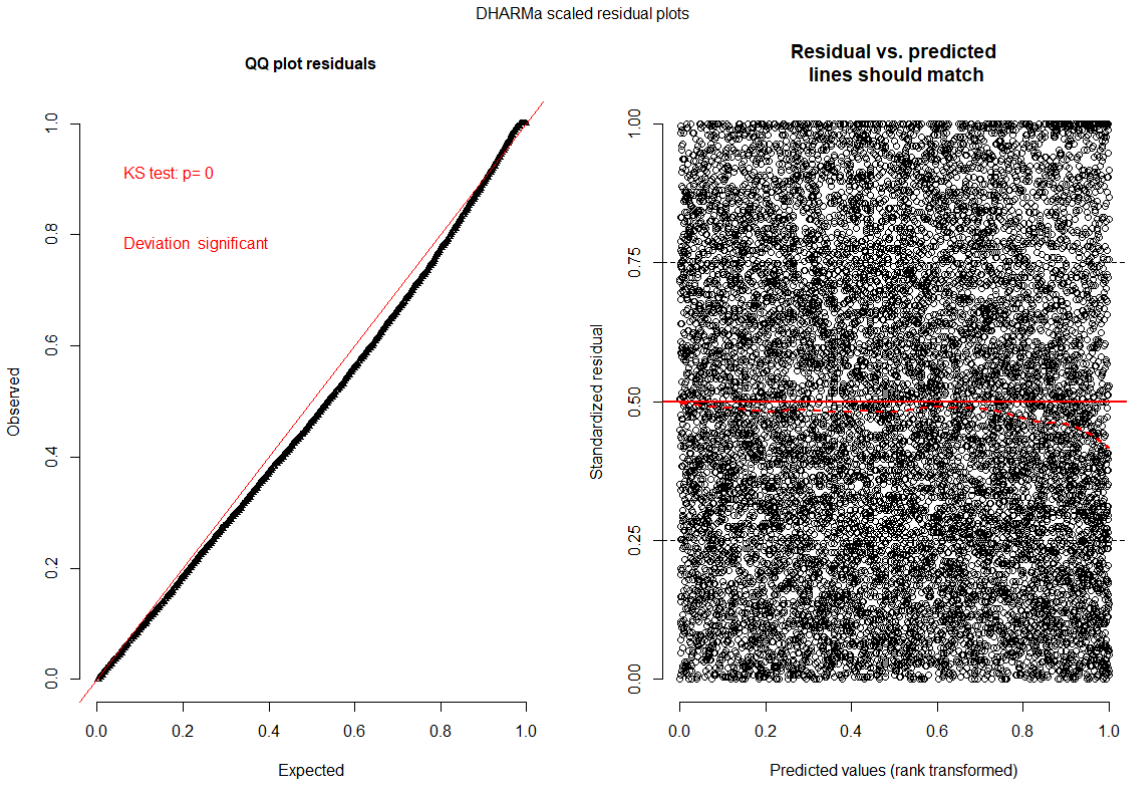

Figure 11: Standard Poisson model diagnostic for deaths occurred during non-holidays.

Figure 11 displays the plot of observed against expected distribution of the number of deaths occurred during non-holidays. Based on the Kolmogorov-Smirnov test statistics with a p-value of 0.000, I reject the null hypothesis and conclude that the data follows the Poisson distribution. This suggests that the standard Poisson regression model does not fit the dataset. Again, it can be observed from the quantile plot that medians for observed and expected are not the same.

Table 15: Testing for over-dispersion or under-dispersion in the model for deaths occurred during holidays.

Parameter	Estimate	Z-value	P-value
Dispersion	1.7942	3.9559	<0.0001

The dispersion parameter is greater than one, suggesting over-dispersion relative to the standard Poisson regression model. Testing for zero inflation, it was found that the ratio of the observed against the expected is 1.1196 with a p-value 0.0000. I reject the null hypothesis stating that the expected zeros and observed zeros are equal. This suggests that the data poses excess of zeros.

5.3. MODEL EXTENSION TO MODEL POISSON

The data, an alternative approach is to use the negative binomial (NB), the zero inflated Poisson (ZIP) and the zero inflated negative binomial (ZINB) regression models. These models can be considered as an extension of the standard Poisson regression model, and considered as flexible regression models that addresses excess zeros and provide flexibility in data dispersion modelling. Tables 19 and 20 show the regression coefficient estimates for the competing count models.

5.3.1. Negative Binomial Regression Model

A limitation of the standard Poisson regression model is the equality of its mean and variance. It was observed in the data from the period under review that the conditional variance is larger than the conditional mean. This renders the assumption of a standard Poisson regression model for the error process untenable. Under the circumstances a reasonable alternative is the NB regression model. This model allows the variance to differ from the mean. The regression coefficients for the two models during holidays and non-holidays, the NB and the standard Poisson regression model are very close. However, the standard errors are larger for the NB regression model. The coefficients for the NB model can be interpreted in the same way as was done previously for the Poisson model.

Comparison of standard regression models Poisson and NB based on AIC and BIC values showed that the NB model had the smallest AIC and BIC values than those for the standard Poisson regression model, indicating that the NB regression model fits the data significantly better than the standard Poisson regression model.

Table 16: Regression coefficient estimates for death during holidays.

Variables	NB	ZIP		ZINB	
		Count	Logistic	Count	Logistic
Coefficients					
Intercept	-3.4692	-2.9185	-0.8615	-2.5600	-2.190
Sunday	-0.1725	-0.1618	-0.0157	-0.1665	-0.0956
Monday	-0.5024	-0.3427	0.3247	-0.4081	0. 4502
Tuesday	-0.3748	-0.0880	0.6786	-0.2665	0.5993
Wednesday	0.0019	-0.0141	-0.0787	0.0685	0.5446
Thursday	0.1767	0.0193	-0.5831	0.0463	-2.366
Saturday	0.0688	0.2784	0.5259	0.2334	0.8944
Human Actions	2.1214	1.4116	-1.4701	0.9882	8.153
Vehicle Conditions	1.8577	1.3547	-0.9731	-	-
National road	0.1118	0.7517	2.2041	0.5663	1.267
Others road	0.2116	0.1982	-0.2158	0.2239	0.9347
Provincial road	-0.1464	0.7652	2.6131	0.6244	1.331
Regional road	-0.0019	0.4119	1.5574	0.1618	1.105
Sedan	-0.2648	-0.1200	-0.0108	-0.1940	-0.0849
LDV	-0.1834	0.1223	0.4262	0.0246	0.8666
Combi	0.1967	0.2229	-0.1887	0.3778	0.6099
Bus	0.0335	-0.1740	-1.5302	0.0438	-1.083
Motorcycle	0.7571	0.2976	16.661	0.5641	-2.444
Standard Errors					
Intercept	0.4774	1.1851	2.5405	0.3020	<0.0001
Sunday	0.1717	0.2003	0.4511	0.1931	0.8247
Monday	0.1877	0.2499	0.5423	0.2255	0.9151
Tuesday	0.2057	0.2477	0.4738	0.2463	0.8797
Wednesday	0.2012	0.2274	0.5580	0.2271	0.9539
Thursday	0.1613	0.1737	0.5304	0.1876	3.466
Saturday	0.1487	0.1686	0.3730	0.1709	0.6992
Human Actions	0.4300	1.0843	1.6963	0.1771	-
Vehicle Conditions	0.4604	1.0873	1.7120	-	-
National road	0.1949	0.2439	1.2209	0.2295	2.319
Others road	0.2047	0.2380	1.4015	0.1926	
Provincial road	0.2753	0.3227	1.2413	0.3564	2.319
Regional road	0.1802	0.2390	1.2736	0.1824	2.319
Sedan	0.0723	0.0665	0.1483	0.0822	0.2263
LDV	0.0865	0.0943	0.2024	0.1059	0.3858
Combi	0.1015	0.1120	0.2326	0.1534	0.5304
Bus	0.2764	0.2535	0.9182	0.2668	1.466
Motorcycle	0.3434	0.2495	2871.2	0.3273	1.068
Likelihood					
Log-likelihood	-1316.369	-1329.164		-1327.987	
Model *df*	20	38		20	

Table 17: Regression coefficient estimates for death during non-holidays.

Variables	NB	ZIP		ZINB	
		Count	Logistic	Count	Logistic
Coefficients					
Intercept	-2.8270	-0.2523	2.994	-1.2045	3.2609
Sunday	0.2405	0.2700	0.0691	0.2497	0.1344
Monday	-0.0066	-0.1560	-0.3139	-0.0743	-0.9271
Tuesday	-0.1927	-0.1728	0.0217	-0.2066	-0.4149
Wednesday	-0.3166	-0.2329	0.1643	-0.3323	-0.3956
Thursday	0.0049	0.0094	0.0071	-0.0123	-0.4374
Saturday	0.1494	0.1729	0.0648	0.1544	0.0826
Human Actions	1.2928	-0.3619	-2.3680	-0.2561	-19.121
Vehicle Conditions	1.0609	0.1416	-1.1870	0.1794	-1.6595
National road	0.2807	0.3774	0.1784	0.3029	0.3411
Others road	0.1731	-0.0721	-0.5678	0.1486	-0.6380
Provincial road	-0.4998	-0.9952	-1.5820	-0.4858	0.1100
Regional road	0.0391	-0.0509	-0.2222	0.0236	-0.4135
Sedan	-0.1958	-0.1865	-0.2509	-0.2011	-0.6939
LDV	-0.2537	-0.1913	-0.0984	-0.2672	-0.7275
Combi	0.2025	0.1003	-0.3410	0.1458	-1.3340
Bus	0.6250	0.7132	0.1078	0.6551	0.2965
Motorcycle	0.2145	-0.5775	-1.4780	0.0987	-13.451
Standard Errors					
Intercept	0.1399	0.1987	0.2808	0.2404	0.6385
Sunday	0.0651	0.0766	0.1426	0.0662	0.3362
Monday	0.0803	0.1014	0.2075	0.0824	0.4322
Tuesday	0.0833	0.1110	0.2052	0.0852	0.4458
Wednesday	0.0867	0.1207	0.2107	0.0889	0.4639
Thursday	0.0790	0.0972	0.1846	0.0804	0.4238
Saturday	0.0627	0.0764	0.1433	0.0637	0.3409
Human Actions	0.1107	0.1562	0.1835	0.2204	876.45
Vehicle Conditions	0.1275	0.1688	0.2015	0.2427	0.3451
National road	0.0768	0.0996	0.1760	0.0777	0.4826
Others road	0.0838	0.1107	0.2143	0.0841	0.5610
Provincial road	0.1333	0.1830	0.7495	0.1333	1.2137
Regional road	0.0709	0.0968	0.1750	0.0714	0.4593
Sedan	0.0299	0.0296	0.0556	0.0297	0.2655
LDV	0.0342	0.0402	0.0739	0.0346	0.2818
Combi	0.0428	0.0454	0.0920	0.0441	0.3209
Bus	0.0977	0.0819	0.1685	0.0981	0.5999
Motorcycle	0.1636	0.1183	1.0770	0.1613	161.52
Likelihood					
Log-likelihood	-9354.226	-9643.509		-9285.377	
Model *df*	20	38		39	

5.3.2. Zero-inflated Regression Model

Using testZeroInfation function in R software, the data showed that it contains excess of zeros, thus limiting description using the standard Poisson regression model. Zero-inflated regression models have been used to better describe a random variable containing excess of zeros. The results from the zero-inflated models are shown in Table 16 and 17. All complete models are shown in the Appendix. Each zero-inflated model has two sets of regression coefficients, count regression model and logistic regression model.

The regression coefficient for the count part can be interpreted as in the same way as the standard Poisson regression model. The coefficient of regression can be interpreted as the expected number of deaths during holidays on provincial roads, which is 2.16 ($e^{0.77} = 2.16$) times greater than the expected number of deaths on district roads during the holidays. The regression coefficients for the logistic regression can be transformed and interpreted as the odd ratios. The interpretation of the coefficient can be expressed as the odds of getting excessive zeros given that an accident has occurred on a provincial road over the odds of getting excessive zeros given that an accident has occurred on a district road is 0.36, while holding other variables in the model constant.

5.4. MODEL COMPARISON

5.4.1. Competing Count Models for Holidays

I looked at model diagnostics for the competitive models. A plot of residuals against predicted values is shown in Figure 12. There should be no relationship or pattern between the residual and the predicted values, so the red dashed line should be horizontal and close to zero.

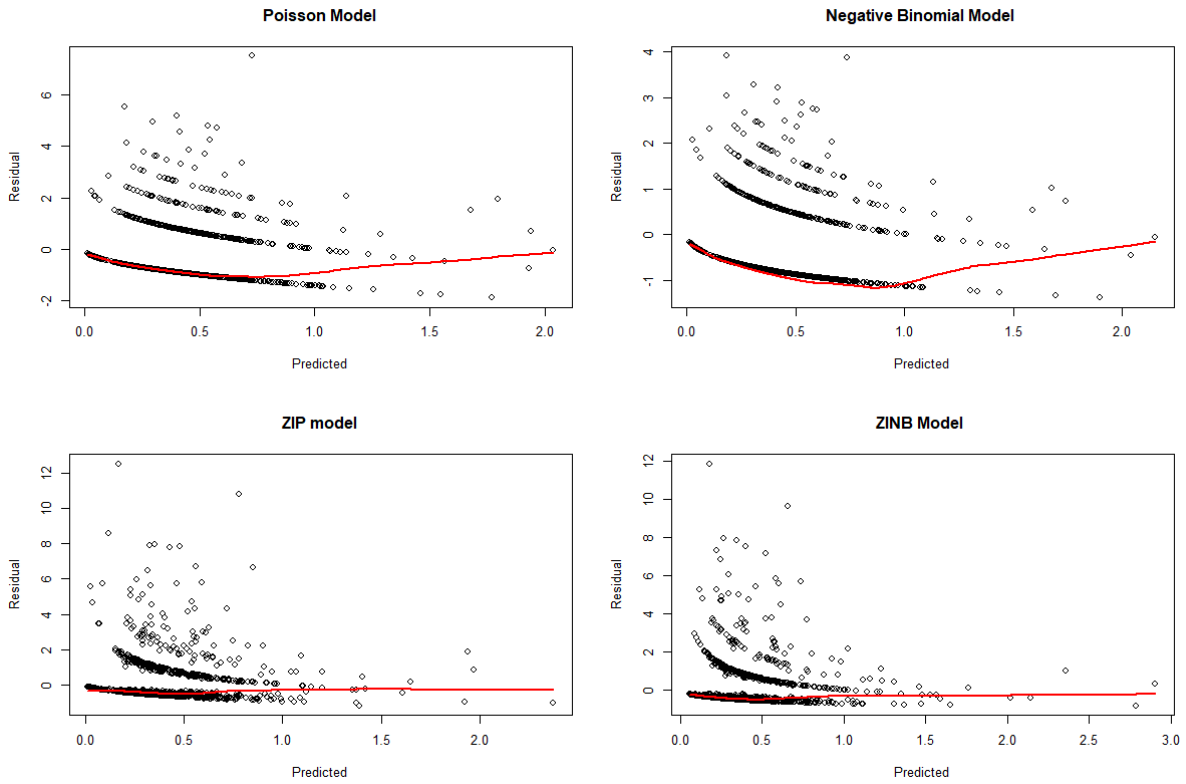

Figure 12: Predicted values against residual plot with LOWESS line.

Graphically examining the fit across all four of the models, the ZIP and ZINB regression models show that there is no relationship between the residuals and the predicted values since the red dashed line is straight, horizontal and close to zero. This favours the ZIP and ZINB regression models over the other two models.

Figure 13 shows a comparison of actual and predicted values total number of deaths. The actuals versus the predicted plots clearly show that there is little agreement between the actual and predicted values for the standard Poisson regression model. The model over-predicts and underpredicts all the count frequencies. The NB and ZINB seem to be doing a better job in predicting the actuals than other two models.

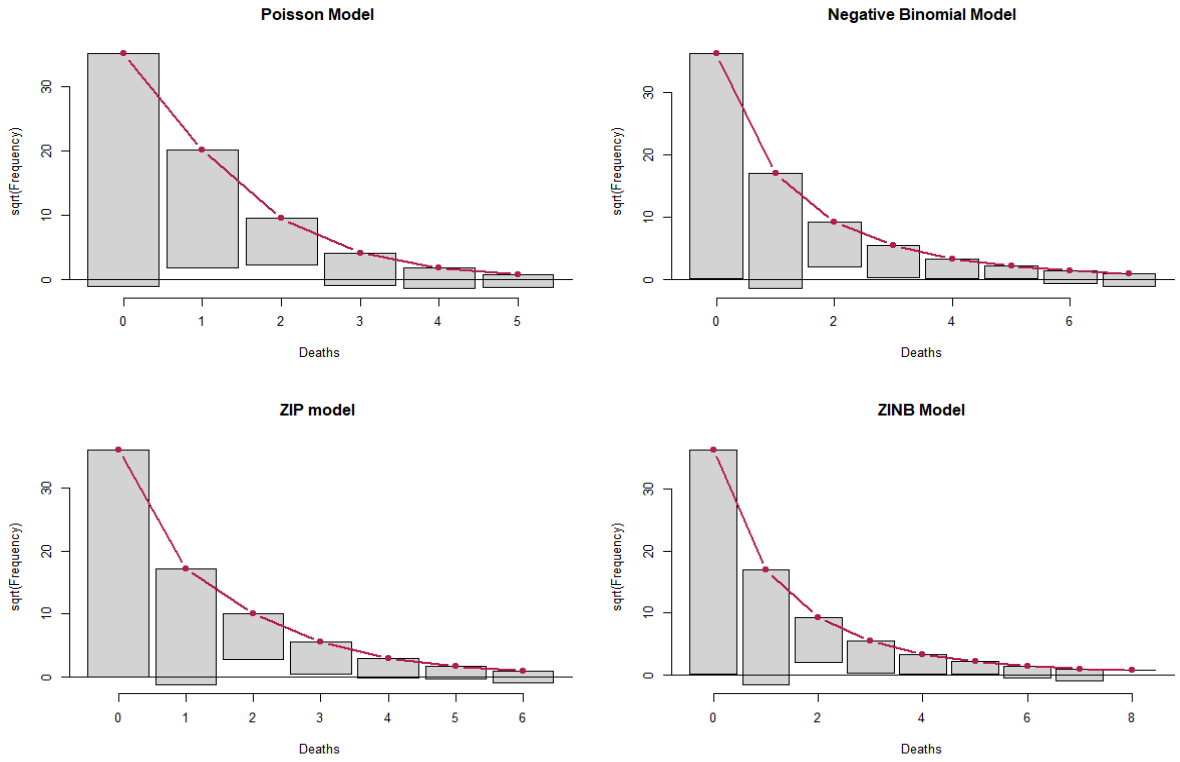

Figure 13: Comparison of actuals and predicted deaths frequency.

Comparing the four count models, the expected number of zero counts based on the ZIP regression model is closer to the observed zeros than for the other three models shown in Table 18. However, the AIC and BIC values for the NB regression model is smaller than those for the other three models, indicating that the NB regression model fits the data somewhat better than the other models do.

Table 18: The observed zero counts compared to the expected number of zeros.

	Observed	PR	NB	ZIP	ZINB
Zero Counts	1299	1226	1224	1301	1316
AIC		2844.677	2672.737	2734.327	2693.68
BIC		2948.437	2781.959	2941.848	2895.74

5.4.2. Competing Count Models for Non-Holidays

A plot of residuals against predicted values is shown in Figure 14.

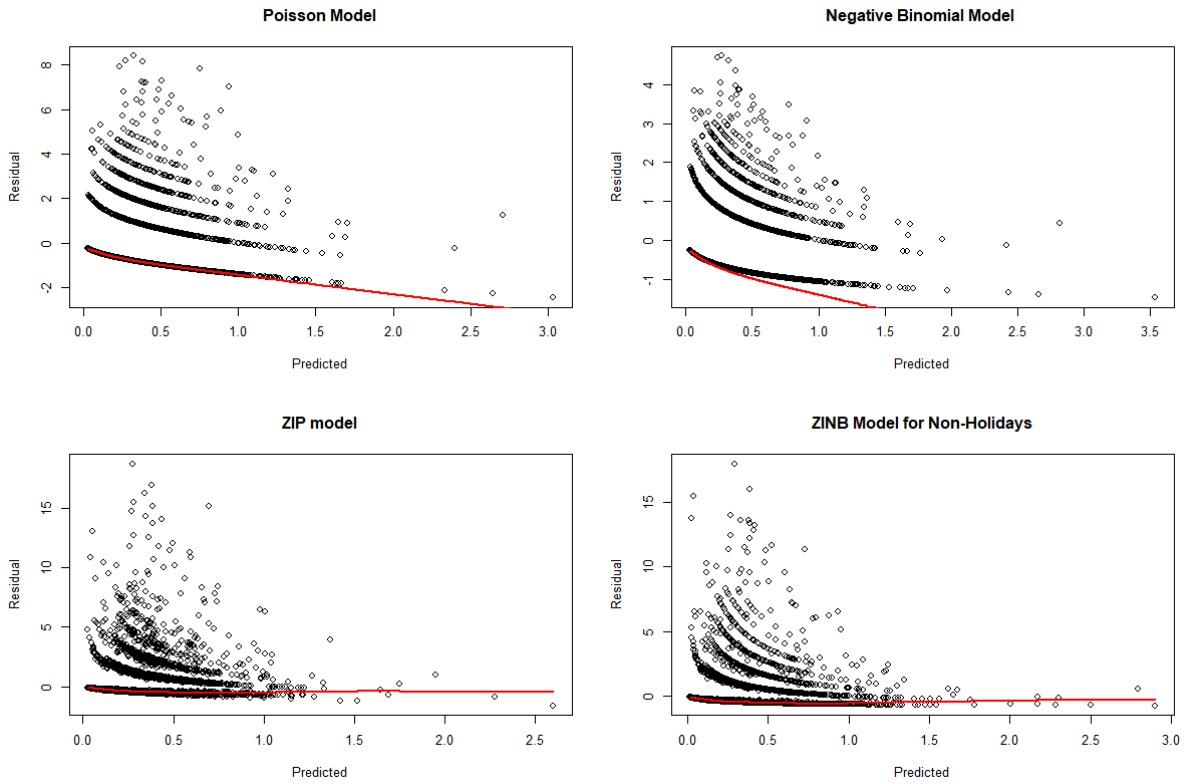

Figure 14: Predicted against residual plot with LOWESS line, death during non-holidays

Examining fit across all four of the models, the zero-inflated models represent a better performance than the other two generalized linear models framework do, since the zero-inflated models show that there was no agreement between the residuals and the predicted values. Figure 15 compares the actual and predicted values, the NB and ZINB regression models seem to be doing a better job of capturing the death count, as the standard Poisson and the ZIP regression models over-predict and under-predict all the count frequencies

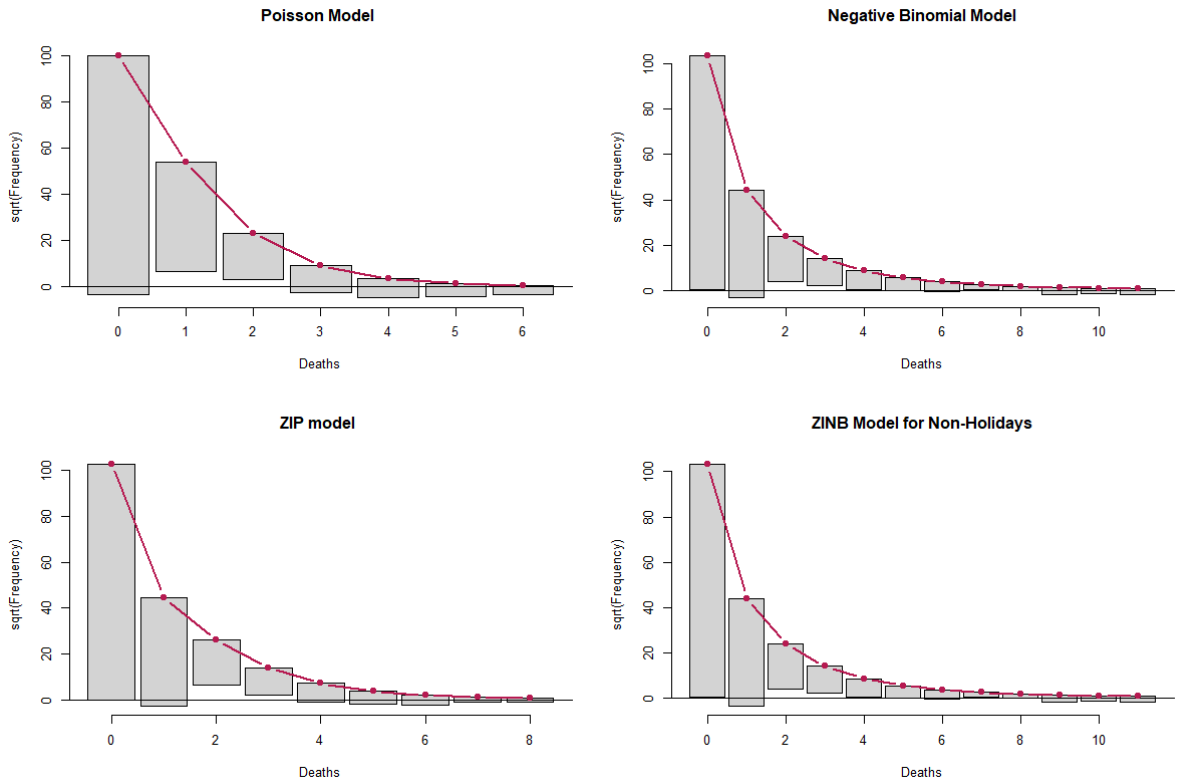

Figure 15: Comparison of actuals and predicted deaths frequency, death during non-holidays.

Comparing the four death count models during non-holidays, the expected number of zero counts based on the ZIP regression model is closer to the observed zeros than in the other three models. However, the AIC and BIC values for the NB regression models are smaller than those for the other three models, indicating that the NB regression model fits the data somewhat better than the other models do.

Table 19: The observed zero counts compared to the expected number of zeros for non-holidays.

	Observed	PR	NB	ZIP	ZINB
Zero Counts	10591	9961	9945	10564	10670
AIC		20425.33	18748.45	19363.02	18648.75
BIC		20568.03	18898.66	19648.41	18941.66

5.5. FINAL COUNT MODELS

5.5.1. Deaths During Holidays

Table 20: The Negative Binomial regression coefficient estimates using Maximum Likelihood Estimate.

		C.I for exp (β)			
Variables	β	Exp(β)	2.5%	97.5%	P-value
Count					
Intercept	-3.4692	0.0311	0.0111	0.0740	<0.0001***
Sunday	-0.1725	0.8414	0.5980	1.1783	0.3148
Monday	-0.5024	0.6050	0.4174	0.8688	0.0074**
Tuesday	-0.3748	0.6873	0.4535	1.0280	0.0684.
Wednesday	0.0019	1.0019	0.6714	1.4844	0.9921
Thursday	0.1767	1.1932	0.8710	1.6335	0.2735
Saturday	0.0688	1.0712	0.7978	1.4380	0.6433
Human Actions	2.1214	8.3430	3.9342	21.572	<0.0001***
Vehicle Conditions	1.8577	6.4091	2.8147	17.336	<0.0001***
National road	0.1118	1.1183	0.7641	1.6461	0.5661
Others road	0.2116	1.2357	0.8305	1.8477	0.3013
Provincial road	-0.1464	0.8637	0.4986	1.4808	0.5948
Regional road	-0.0019	0.9980	0.7037	1.4272	0.9913
Sedan	-0.2648	0.7673	0.6659	0.8834	0.0002***
LDV	-0.1834	0.8323	0.6997	0.9889	0.0341*
Combi	0.1967	1.2174	0.9892	1.4993	0.0526.
Bus	0.0335	1.0340	0.6112	1.7500	0.9034
Motorcycle	0.7571	2.1322	1.0709	4.2826	0.0274*
Significant Codes					
0.001 '***'		0.01 '**'		0.05 '*'	0.1 '.'

This is the fitted NB regression model

$$\log(death) = -3.47 * Intercept - 0.17 * Sunday - 0.50 * Monday - 0.38 * Tuesday + 0.00 * Wednesday + 0.18 * Thursday + 0.07 * Saturday + 2.12 * human\ actions + 1.86 * vehicle\ conditions + 0.11 * national\ roads + 0.21 * other\ roads - 0.15 * Provincial\ Roads - 0.00 * Regional\ Roads - 0.27 * sedan - 0.18 * LDV + 0.20 * combi + 0.03 * bus + 0.76 * motorcycle.$$

The regression model represents only explanatory variables that were found to be statistically significant at 0.05 significant level. The model coefficients are can be interpreted as any other unstandardized coefficients. The coefficient associated with Monday is -0.50. The negative sign indicate that the expected log number of death occurred on Monday is smaller than for those that occurred on Friday. The expected

log death for aaccidents attributed to human actions is 2.12 higher than when the accidents cannot be attributed to human actions.

5.5.2. Deaths During Non-Holidays

Table 21: The ZINB regression coefficient estimates using restricted maximum likelihood estimate.

Variables	β	C.I for exp (β)			
		Exp(β)	2.5%	97.5%	P-value
			Count Model		
Intercept	-1.2045	0.2998	0.1871	0.4803	<0.0001***
Sunday	0.2497	1.2836	1.1273	1.4617	0.0001***
Monday	-0.0743	0.9283	0.7899	1.0911	0.3671
Tuesday	-0.2066	0.8133	0.6881	0.9613	0.0154*
Wednesday	-0.3323	0.7172	0.6024	0.8538	0.0001***
Thursday	-0.0123	0.9877	0.8437	1.1563	0.8779
Saturday	0.1544	1.1670	1.0299	1.3224	0.0154*
Human Actions	-0.2561	0.7740	0.5025	1.1923	0.2452
Vehicle Conditions	0.1794	1.1965	0.7436	1.9254	0.4596
National road	0.3029	1.3538	1.1624	1.5767	<0.0001***
Others road	0.1486	1.1603	0.9839	1.3682	0.0771.
Provincial road	-0.4858	0.6151	0.4737	0.7989	0.0002***
Regional road	0.0236	1.0239	0.8901	1.1777	0.7406
Sedan	-0.2011	0.8177	0.7714	0.8668	<0.0001***
LDV	-0.2672	0.7654	0.7152	0.8192	<0.0001***
Combi	0.1458	1.1570	1.0610	1.2616	0.0009***
Bus	0.6551	1.9254	1.5884	2.3340	<0.0001***
Motorcycle	0.0987	1.1038	0.8046	1.5142	0.5402
			Logistic Model		
Intercept	3.2609	26.073	7.4584	9.1145	<0.0001***
Sunday	0.1344	1.1438	0.5917	2.2111	0.6893
Monday	-0.9271	3.9566	0.1695	0.9230	0.0319*
Tuesday	-0.4149	0.6603	0.2756	1.5821	0.3519
Wednesday	-0.3956	0.6732	0.2711	1.6714	0.3937
Thursday	-0.4374	0.6457	0.2813	1.4817	0.3020
Saturday	0.0826	1.0861	0.5567	2.1189	0.8085
Human Actions	-19.121	<0.0001	0.0000	INF	0.9825
Vehicle Conditions	-1.6595	0.1902	0.0967	0.3741	<0.0001***
National road	0.3411	1.4065	0.5461	3.6223	0.4796
Others road	-0.6380	0.5283	0.1759	1.5867	0.2554
Provincial road	0.1100	1.1163	0.1034	1.2048	0.9277
Regional road	-0.4135	0.6613	0.2687	1.6271	0.3680
Sedan	-0.6939	0.4996	0.2968	0.8407	0.0089**
LDV	-0.7275	0.4830	0.2780	0.8394	0.0098**
Combi	-1.3340	0.2633	0.1404	0.4940	<0.0001***
Bus	0.2965	1.3451	0.4150	4.3594	0.6211
Motorcycle	-13.451	<0.0001	<0.0001	4.4345	0.9336
Significant code	0.001 '***'	0.01 '**'	0.05 '*'	0.1 '.'	

This is the fitted ZINB regression model

$$\log(death) = -1.21 * Intercept + 0.25 * Sunday - 0.07 * Monday - 0.21 * Tuesday - 0.33 * Wednesday - 0.01 * Thursday + 0.15 * Saturday - 0.26 * human\ actions + 0.18 * vehicle\ conditions + 0.30 * national\ roads + 0.15 * other\ roads - 0.49 * provincial\ roads + 0.02 * regional\ roads - 0.20 * sedan - 0.26 * LDV + 0.15 * combi + 0.66 * bus + 0.10 * motorcycle.$$

The regression model represents only explanatory variables that were statistically significant at 0.05 significant level. The model coefficients can be interpreted as the expected log number of death that occurred on Sunday is 0.25 times greater than the death that occurred on Friday. The expected log number of death that occurred on Tuesday is 0.21 times less than the expected log number of deaths that occurred on Friday.

The logistic model part of this fitted model:

$$\pi = 3.26 * Intercept + 0.13 * Sunday - 0.93 * Monday - 0.42 * Tuesday - 0.40 * Wednesday - 0.45 * Thursday + 0.08 * Saturday - 19.12 * human\ actions - 10.66 * vehicle\ conditions + 0.34 * national\ roads - 0.64 * other\ roads + 0.11 * provincial\ roads - 0.41 * regional\ roads - 0.69 * sedan - 0.73 * LDV - 1.33 * combi + 0.30 * bus - 13.45 * motorcycle.$$

$$logit(\omega) = \frac{\pi}{1+\pi}.$$

The log odds of being an excessive zero would decrease by 0.92 for every additional accident on Monday as compared to Friday. The more accidents on Monday the less likely that zero would be due to no death. The log odds of being an excessive zero would decrease by 1.66 for every additional accident caused by vehicle conditions, indicating that the higher the number of accidents caused by vehicle conditions, the higher the likelihood of death from the accidents.

The results from the NB and ZINB regression models are summarised in Tables 20 and 21. Holiday's road accidents caused by human actions and vehicle conditions on Monday, driving in sedan, LDV vehicle types and motorcycles have a significantly positive effect on road deaths. On the other hand, the factors Sunday, Tuesday, Wednesday, Saturday, national roads, provincial roads, sedan, LDV, combi and bus have a significantly positive effect on road deaths during non-holidays, whilst Monday, Thursday, human actions, vehicle conditions, other roads, regional roads and motorcycle have significantly negative effect on road deaths.

CHAPTER 6: DISCUSSION AND CONCLUSION

6.1. INTRODUCTION

This chapter covers the discussion, conclusion, recommendation, further research area and study limitations. The purpose of this study was to analyse road deaths in the Limpopo province in order to determine the factors causing death due to road accidents. The final step was to compare the generalized linear models (GLM) with the zero-inflated models.

6.2. MAIN FINDINGS

The study examined factors that contribute to deaths due to road accidents between 2009 and 2015. There were 18,029 RTAs recorded during this study period under review, resulting in 4,944 deaths. The year 2015 recorded the highest number of incidents or cases. Most accidents and deaths took place in December. This could be due to the Christmas season, where the roads are busier as a result of making last minutes trips to the shops or going on long trips. This month, most people in the country are on leave and schools are closed, resulting in a lot of traffic congestion. More accidents occur on Saturdays (25%), while 18% of all car accidents occur between 5 p.m. and 8 p.m. on Sunday. Nearly half (43%) of all accidents occur on weekends. This could be due to the fact that, on weekends, more people go to church, attend weddings and engage in many other activities. Tuesday and Wednesday are the safest days to drive, accounting for only 9% respectively of all accidents.

Most deaths (26%) due to road accidents occur on Saturday, while 21% of all deaths occur on Sundays between 5 pm and 8 pm and, again, between 5 am and 7 am. Nearly half (47%) of all road deaths occur on weekends. Wednesday is the safest day to drive, accounting for only 7% of all deaths due to road accidents.

More than 82% of all RTAs, RTDs and RTIs in the Province occur as results of human actions, such as speeding, pedestrian carelessness or recklessness, following too close to the vehicle in front, reckless driving, contravention of traffic signs, fatigue and drunken driving.

6.3. LOGISTIC REGRESSION MODEL FINDINGS

Of the 33 variables that were considered for fitting to the model, 20 variables were significant in predicting the occurrence of death given the fact that an accident has occurred. Among the explanatory variables that were significant, it was found that the variables Friday, Monday, Saturday, Sunday and Thursday were significant predictors of road deaths.

Our study detected that human actions and environment conditions were important explanatory variables that can be used in predicting the likelihood of death. These results were similar to the results of previous studies (Siskind, et al., 2011; Zhang, et al., 2013). The study by Siskind et al., (2011), found that human actions, such as speed, were considered by police to be a contributing factor in 18% of fatal accidents, compared to 10% in non-fatal accidents. This study further shows that vehicle type, such as sedan, LDV, combi and truck, was found to be significant in predicting the odds of death occurring as a result of an accident.

The model showed a strong relationship between the observed and predicted values, and the residuals were equally spread along the range of predictors. The area under the curve (AUC) value was 68%, indicating that our model has the ability to predict the probability of death given the fact that an accident has occurred. The model was considered to be valid since our AUC is above 50%.

6.4. COMPETING COUNT MODELS FINDINGS

The standard Poisson regression model was found to be over-dispersed and zero-inflated. An alternative approach to deal with this over-dispersion and zero-inflated was to use the negative binomial (NB) and zero-inflated models. Fitting four competing count models to aggregated data by day, the study found that the NB model performed better than the three other models did in modelling the number of deaths that occurred during the holidays. The model showed no relationship between the residual and the predicted values and the excess of zeros were better captured by the zero-inflated Poisson (ZIP) model than by the NB model.

The aggregated death data for non-holidays contained an excess of zeros, thus limiting description using the standard Poisson and NB regression models. The zero-inflated models were used to better describe such a random variable containing excess of zeros. Based on the AIC and BIC criterion to select the best model, the

zero-inflated negative binomial (ZINB) model had the smallest values when compared to the values for ZIP model. Again, the ZINB model diagnosis showed no relationship between the predicted and residual values. The model captured the zero counts better than standard Poisson and NB regression models. These results were similar to those in a previous study by Prasetijo and Musa (2016).

Among the explanatory variables, it was found that the variables Monday, human actions, vehicle conditions, sedan, LDV and motorcycle were significant predictors of RTDs during holidays. On the other hand, during non-holidays the variables weekend, Tuesday, Wednesday, national road, provincial road, sedan, LDV, combi and bus were found to be significant predictors of RTDs.

The study succeeded in addressing all the objectives that it set out to address. From both the literature review as well as the study, it is clear that the variables human actions, vehicle type, road type and day of week are the main determinants of RTDs in the Limpopo province.

6.5. CONCLUSION

Generalized linear modelling (GLM) techniques, such as the standard Poisson regression model and NB model, did little to explain and handle zero excesses, thus, zero-inflated models, such as ZINB, were found to be effective in catering for, and explaining, excess zeros.

6.6. RECOMMENDATION

- Government investment in the maintenance of district and rural roads should be recommended, as most roads have potholes and road signs are no longer visible.
- During festive seasons, such as December and over weekends, it is recommended that the government provide more manpower for law enforcement.
- Finally, it is recommended that the Limpopo Province Department of Road and Transport consider adding colour of the vehicle, gender, age, alcohol concentration, car roadworthiness and the wearing of seatbelts by the driver when capturing incident information. Literature national wide showed that these are some of determinants for RTAs.

6.7. AREAS FOR FURTHER RESEARCH

The study could be extended to other parts of the Province and designed to investigate variables such as colour of the vehicle, gender of the driver, age of the driver, alcohol concentration, marital status, educational level and car roadworthiness, as determinants of RTDs within the South African context.

6.8. STRENGTH AND LIMITATIONS

The present study has both strengths and limitations. Among the limitations I acknowledge the fact that the study only included data from Limpopo and the data from other provinces should be collected and analysed. This limits the external validity of the study in that results cannot be generalized to include the whole of South Africa. Due to the fact that data analysis involved the use of secondary data, there was no control over what data were collected, or how the data were collected or managed.

REFERENCES

Aarts, L. and Van Schagen, I., 2006. Driving speed and the risk of road crashes: A review. *Accident Analysis & Prevention*, *38*(2), pp.215-224.

Afukaar, F.K., Damsere-Derry, J. and Ackaah, W., 2010. Observed seat belt use in Kumasi Metropolis, Ghana. *Journal of Prevention & Intervention in the Community*, *38*(4), pp.280-289.

Agbonkhese, O., Yisa, G.L., Agbonkhese, E.G., Akanbi, D.O., Aka, E.O. and

Agyemang, B., Abledu, G. K. & Semevho, R., 2013. Regression Analysis of Road Traffic Accidents and Population Growth in Ghana. *International Journal of Business and Social Research.*

Al-Matawah, J. and Jadaan, K., 2010. Application of Prediction Techniques to Road Safety in Developing Countries. *International Journal of Applied Science and Engineering*, *8*(1), pp.11-17.

Anastasopoulos, P.C. and Mannering, F.L., 2009. A note on modeling vehicle accident frequencies with random-parameters count models. *Accident Analysis & Prevention*, *41*(1), pp.153-159.

Anowar, S., Yasmin, S., Eluru, N. and Miranda-Moreno, L., 2012. Analyzing car ownership in two Quebec metropolitan regions: a comparison of latent ordered and unordered response models. *Technical Paper, Department of Civil Engineering and Applied Mechanics, McGill University*.

Arrive Alive (AA), 2009. Texting and distracted driving. *Retrieved October*, *16*, p.2013.

Ayati, E. and Abbasi, E., 2014. Modeling accidents on Mashhad urban highways. *Open Journal of Safety Science and Technology*, *4*(01), p.22.

Bener, A., Dafeeah, E.E., Verjee, M., Yousafzai, M.T., Al-Khatib, H., Nema, N., Mari, S., Choi, M.K., Özkan, T. and Lajunen, T., 2013. Gender and age differences in risk taking behaviour in road traffic crashes. *Advances in Transportation Studies*, *31*, pp.53-62.

Borman, S., 2004. The expectation maximization algorithm-a short tutorial. *Submitted for publication*, *41*.

Burgut, H. R. et al., 2010. Risk factors contributing to road traffic crashes in a fast-developing country: the neglected health problem. *Turkish Journal of Trauma & Emergency Surgery,* 16(6), pp. 497-502.

Cantillo, V., Garcés, P. and Márquez, L., 2016. Factors influencing the occurrence of traffic accidents in urban roads: A combined GIS-Empirical Bayesian approach. *Dyna*, *83*(195), pp.21-28.

Çelik, A.K. and Senger, O., 2014. Risk factors affecting fatal versus non-fatal road traffic accidents: the case of Kars province, Turkey. *International Journal for Traffic and Transport Engineering*, *4*(3), pp.339-351.

Chang, S.C. and Kim, H.J., 2007. Em algorithm.

Clarke, D.D., Ward, P., Bartle, C. and Truman, W., 2010. Killer crashes: fatal road traffic accidents in the UK. *Accident Analysis & Prevention*, *42*(2), pp.764-770.

Dempster, A.P., Laird, N.M. and Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), pp.1-22.

Department of Road and Transport (DOT), 2007. *Department of Road and Transport.* [Online] Available at: http://www.safiri.co.za/lpfdb/roads-traffic-management.html. [Accessed 15 May 2017].

Desapriya, E., Subzwari, S., Scime-Beltrano, G., Samayawardhena, L.A. and Pike, I., 2010. Vision improvement and reduction in falls after expedited cataract surgery: Systematic review and metaanalysis. *Journal of Cataract & Refractive Surgery*, *36*(1), pp.13-19.

Erdman, D., Jackson, L. and Sinko, A., 2008, March. Zero-inflated Poisson and zero-inflated negative binomial models using the COUNTREG procedure. In *SAS Global Forum* (Vol. 2008, pp. 322-2008).

Goswami, A. and Sonowal, R., 2009. A statistical analysis of road traffic accidents in Dibrugarh City, Assam, India. *Division of Epidemiology and Nutrition, Regional Medical Research Centre*.

Govender, R. and Allopi, D.R., 2007. Analysis of the scientific aspects related to minibus taxi collisions. *SATC 2007*.

Gupta, M., Solanki, V.K. and Singh, V.K., 2017. Analysis of Datamining Technique for Traffic Accident Severity Problem: A Review. In *Proceedings of the Second International Conference on Research in Intelligent and Computing in Engineering* (pp. 197-199).

Harris, G. T. & Olukoga, I. A., 2005. A cost benefit analysis of an enhanced seat belt enforcement program in South Africa. *Injury Prevention,* pp. 102-105.

Hartig, F., 2016. Package "DHARMa": residual diagnostics for hierarchical (multi-level/mixed) regression models. *An R package*.

Imran, M. and Nasir, J.A., 2015. ROAD TRAFFIC ACCIDENTS; PREDICTION IN PAKISTAN. *Professional Medical Journal*, *22*(6).

Ishtiaque, A., 2013. Road infrastructure and road safety. *Transport and Communications Bulletin for Asia and the Pacific,* p. 80.

Jung, S., Qin, X. & Noyce, D. A., 2010. Rainfall effect on single-vehicle crash severities using polychotomous response models. *Accident Analysis & Prevention,* Volume 42, pp. 213-224.

Junus, N.W.M. and Ismail, M.T., 2014, July. Predicting road accidents: Structural time series approach. In *AIP conference proceedings* (Vol. 1605, No. 1, pp. 816-821). AIP.

Klynsmith, A., 2015. *Laaimylorrie.* [Online] Available at: http://andreklynsmith.blogspot.com/2015/07/transport-is-heartbeat-of-south-africas.html. [Accessed 15 May 2017].

Kromer, P., Beshah, T., Ejigu, D., Snasel, V., Platos, J. and Abraham, A., 2013, April. Mining traffic accident features by evolutionary fuzzy rules. In *Computational Intelligence in Vehicles and Transportation Systems (CIVTS), 2013 IEEE Symposium on* (pp. 38-43). IEEE.

Kyei, A. K., 2011. Some Socio- Economic Indicator from Vhembe District in Limpopo Province in South Africa. *Journal of Emerging Trend in Economic and Management Sciences,* 2(5), pp. 364-371.

Lambert, D., 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, *34*(1), pp.1-14.

Lehohla, P., 2009. South African Statistics, 2009-Statistics South Africa. [online] Available at: http://www.statssa.gov.za/publications/SAStatistics. [Accessed 25 May 2016].

Lemp, J. D., Kockelman, K. M. & Unnikrishnan, A., 2011. Analysis of large truck crash severity using heteroskedastic ordered probit models. *Accident Analysis & Prevention,* Volume 43, pp. 370-380.

Li, Y. and Bai, Y., 2008. Comparison of characteristics between fatal and injury accidents in the highway construction zones. *Safety Science*, *46*(4), pp.646-660.

Lord, D. and Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice*, *44*(5), pp.291-305.

Ma, L., Yan, X. and Qiao, W., 2014. A quasi-Poisson approach on modeling accident hazard index for urban road segments. *Discrete Dynamics in Nature and Society*, *2014*.

Maffioletti, S., Pocaterra, R. and Tavazzi, S., The importance of precise sight correction for safe driving, Undergraduate Degree Course in Optics and Optometry, January 2009.

Masuri, M.G., Isa, K.A.M. and Tahir, M.P.M., 2012. Children, youth and road environment: Road traffic accident. *Procedia-Social and Behavioral Sciences*, *38*, pp.213-218.

McCullagh, P. and Nelder, J.A., 1983. 1989. *Generalized linear models*, *37*.

Milton, J. C., Shankar, V. N. & Mannering, F. L., 2008. Highway accidents severities and mixed logit model: An exploratory empirical analysis. *Accident Analysis & Prevention,* Volume 40, pp. 260-266.

Ministry of Transport, 2013. *Ministry of Transport.* [Online] Available at: http://www.transport.govt.nz/news/land/slow-down-to-survive/. [Accessed 25 May 2016].

Mohamed, S. et al., 2009. Violence and injuries in South Africa: prioritising an agenda for prevention. *Health in South Africa,* p. 978.

Mohanty, M. and Gupta, A., 2016. Investigation of adolescent accident predictive variables in hilly regions. *International Journal of Injury Control and Safety Promotion*, *23*(3), pp.291-301.

Nangombe, A.N., 2012. Statistical analysis of road traffic fatalities in Namibia from 2007 to 2009.

Nasar, J., Hecht, P. & Wener, R., 2008. Mobile telephones, distracted attention, and pedestrain safety. *Accident Analysis & Prevention 40,* pp. 69-75.

National Highway Traffic Safety Administration (NHTSA), 2013. Motor Vehicle Crashes: Overview. *National Highway Traffic Safety Administration*.

Ogwueleka, F.N., Misra, S., Ogwueleka, T.C. and Fernandez-Sanz, L., 2014. An artificial neural network model for road accident prediction: a case study of a developing country. *Acta Polytechnica Hungarica*, *11*(5), pp.177-197.

Oppong, R.A. and Asiedu-Addo, S.A., 2014. Analysis of vehicular type as a risk factor of road accidents fatality in Ghana. *International Journal of Modern Science and Engineering Technology*, *1*(5), pp.106-114.

Ogundele, O.J., Ifesanya, A.O., Adeyanju, S.A. and Ogunlade, S.O., 2013. The impact of seat-belts in limiting the severity of injuries in patients presenting to a university hospital in the developing world. *Nigerian medical journal: journal of the Nigeria Medical Association*, *54*(1), p.17.

Pei, X., Wong, S.C. and Sze, N.N., 2011. A joint-probability model for crash occurrence and crash severity at signalized intersection. In *World Congress of the International Traffic Medicine Association*. The International Traffic Medicine Association.

Pollak, K., Peled, A. and Hakkert, S., 2014. Geo-based statistical models for vulnerability prediction of highway network segments. *ISPRS International Journal of Geo-Information*, *3*(2), pp.619-637.

Prasetijo, J. and Musa, W.Z., 2016. Modeling Zero–Inflated Regression of Road Accidents at Johor Federal Road F001. In *MATEC web of conferences* (Vol. 47, p. 03001). EDP Sciences.

Prieto, F., Gómez-Déniz, E. and Sarabia, J.M., 2014. Modelling road accident blackspots data with the discrete generalized Pareto distribution. *Accident Analysis & Prevention*, *71*, pp.38-49.

Quddus, M.A., 2008. Time series count data models: an empirical application to traffic accidents. *Accident Analysis & Prevention*, *40*(5), pp.1732-1741.

Ridout, M., Demétrio, C.G. and Hinde, J., 1998, December. Models for count data with many zeros. In *Proceedings of the XIXth international biometric conference* (Vol. 19, pp. 179-192). Cape Town: The International Biometric Society.

Road Traffic Management Corporation (RTMC), 2016. *Road Traffic Report Calendar: 1 January - 31 December 2016,* South Africa: Department of Transport.

Romano, E., Torres-Saavedra, P., Voas, R.B. and Lacey, J.H., 2014. Drugs and alcohol: their relative crash risk. *Journal of Studies on Alcohol and Drugs*, *75*(1), pp.56-64.

Sanusi, R.A., Adebola, F.B. and Adegoke, N.A., 2016. Cases of road traffic accident in Nigeria: a time series approach. *Mediterranean Journal of Social Sciences*, *7*(2 S1), p.542.

Sharma, B.R., 2008. Road traffic injuries: A major global public health crisis. *Public health*, *122*(12), pp.1399-1406.

Statistics South Africa (StatsSA), 2012. *Census 2011 Municipal report Limpopo.,* Pretoria: Statistics South Africa.

Subhan, F., Kanwal, H., Sulaiman, M., Naeem, M.M., Shafiq, M.M.I., Sajjad, U., Ajwad, A. and Aqdas, A., 2017. National Road Crash Injuries An Estimation and Comparison with Previous National Studies. *Nucleus*, *54*(4), pp.210-213.

Thakali, L., Kwon, T.J. and Fu, L., 2015. Identification of crash hotspots using kernel density estimation and kriging methods: a comparison. *Journal of Modern Transportation*, *23*(2), pp.93-106.

Ugarte, M.D., Ibáñez, B. and Militino, A.F., 2004. Testing for Poisson zero inflation in disease mapping. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, *46*(5), pp.526-539.

Venables, W.N. and Smith, D.M., 2003. The R development core team. *An Introduction to R, Version*, *1*(0)

Vogel, L. and Bester, C.J., 2005, July. A relationship between accident types and causes. In *Proceedings of the 24th Southern African Transport Conference (SATC 2005)* (Vol. 11, p. 13).

Wang, C., Quddus, M.A. and Ison, S.G., 2009. Impact of traffic congestion on road accidents: A spatial analysis of the M25 motorway in England. *Accident Analysis & Prevention*, *41*(4), pp.798-808.

World Health Organization, 2011. Mobile phone use: a growing problem of driver distraction.

World Health Organization (WHO), 2013. Violence, Injury Prevention and World Health Organization, 2013. *Global status report on road safety 2013: supporting a decade of action*. World Health Organization.

World Health Organization (WHO), 2015. *Global status report on road safety,* Geneva, Switzerland: World Health Organization.

Zhang, D., Jin, W. and Wang, D.Y., 2006. Analysis on the traffic problems and research on the traffic strategy in group urban development. *SATC 2006*.

Zhu, X. and Srinivasan, S., 2011. A comprehensive analysis of factors influencing the injury severity of large-truck crashes. *Accident Analysis & Prevention*, *43*(1), pp.49-57.

Zong, F., Xu, H. and Zhang, H., 2013. Prediction for traffic accident severity: comparing the Bayesian network and regression models. *Mathematical Problems in Engineering*, *2013*.

APPENDIX

Table 22(A): Chi-square test to test for association of variables

Variables	Chi-square	Df	P-value
Week of Day	125.50	90	0.0080
Contributing Factors	2204.50	75	<0.0001
Road Types	132.82	60	<0.0001
Vehicle types			
Sedan	229.67	75	<0.0001
Motorcycle	58.258	30	0.0015
Truck	168.38	45	<0.0001
Bus	244.34	30	<0.0001
Combi	120.51	45	<0.0001
LDV	108.77	60	0.0001
Hour Intervals	30.163	30	0.4573
Season	67.45	45	0.0167

Table 23(A): Negatibve binomial model for deaths during the holidays.

Variables	Estimate	Std. Error	Z-value	P-value	
Intercept	-3.4692	0.4774	-7.266	<0.0001	***
Sunday	-0.1725	0.1717	-1.005	0.3148	
Monday	-0.5024	0.1877	-2.677	0.0074	**
Tuesday	-0.3748	0.2057	-1.822	0.0684	.
Wednesday	0.0019	0.2012	0.010	0.9921	
Thursday	0.1767	0.1613	1.095	0.2735	
Saturday	0.0688	0.1487	0.463	0.6433	
Human Actions	2.1214	0.4300	4.932	<0.0001	***
Vehicle Conditions	1.8577	0.4604	4.035	<0.0001	***
National road	0.1118	0.1949	0.574	0.5661	
Others road	0.2116	0.2047	1.034	0.3013	
Provincial road	-0.1464	0.2753	-0.532	0.5948	
Regional road	-0.0019	0.1802	-0.011	0.9913	
Sedan	-0.2648	0.0723	-3.661	0.0002	***
LDV	-0.1834	0.0865	-2.119	0.0341	*
Combi	0.1967	0.1015	1.938	0.0526	.
Bus	0.0335	0.2764	0.121	0.9034	
Motorcycle	0.7571	0.3434	2.205	0.0274	*
Goodness of Fit					
Residual deviance	1163.8		**Model *df***	1720	

Table 24 (A): Zero inflated Poisson model for deaths during the holidays.

Variables	Estimate	Std. Error	Z-value	P-value	
Intercept	-2.9185	1.1851	-2.463	0.0137	*
Sunday	-0.1618	0.2003	-0.808	0.4189	
Monday	-0.3427	0.2499	-1.371	0.1703	
Tuesday	-0.0880	0.2477	-0.355	0.7224	
Wednesday	-0.0141	0.2274	-0.062	0.9504	
Thursday	0.0193	0.1737	0.112	0.9111	
Saturday	0.2784	0.1686	1.651	0.0988	.
Human Actions	1.4116	1.0843	1.302	0.1929	
Vehicle Conditions	1.3547	1.0873	1.246	0.2127	
National road	0.7517	0.2439	3.082	0.0020	**
Others road	0.1982	0.2380	0.833	0.4048	
Provincial road	0.7652	0.3227	2.371	0.0177	*
Regional road	0.4119	0.2390	1.723	0.0849	.
Sedan	-0.1200	0.0665	-1.805	0.0710	.
LDV	0.1223	0.0943	1.296	0.1949	
Combi	0.2229	0.1120	1.990	0.0466	*
Bus	-0.1740	0.2535	-0.686	0.4924	
Motorcycle	0.2976	0.2495	1.193	0.2328	
Variables	**Estimate**	**Std. Error**	**Z-value**	**P-value**	
Intercept	-0.8615	2.5405	-0.339	0.7345	
Sunday	-0.0157	0.4511	-0.035	0.9721	
Monday	0.3247	0.5423	0.599	0.5493	
Tuesday	0.6786	0.4738	1.432	0.1521	
Wednesday	-0.0787	0.5580	-0.141	0.8878	
Thursday	-0.5831	0.5304	-1.099	0.2717	
Saturday	0.5259	0.3730	1.410	0.1585	
Human Actions	-1.4701	1.6963	-0.867	0.3861	
Vehicle Conditions	-0.9731	1.7120	-0.568	0.5698	
National road	2.2041	1.2209	1.805	0.0710	.
Others road	-0.2158	1.4015	-0.154	0.8776	
Provincial road	2.6131	1.2413	2.105	0.0353	*
Regional road	1.5574	1.2736	1.223	0.2214	
Sedan	-0.0108	0.1483	-0.073	0.9420	
LDV	0.4262	0.2024	2.106	0.0352	*
Combi	-0.1887	0.2326	-0.811	0.4173	
Bus	-1.5302	0.9182	-1.667	0.0956	.
Motorcycle	16.661	2871.2	-0.006	0.9954	
		Likelihood			
Residual deviance	-1329		**Model *df***	38	

Table 25 (A): Zero inflated negative binomial model for deaths during the holidays.

Variables	Estimate	Std. Error	Z-value	P-value	
Intercept	-2.5600	0.3020	-6.411	<0.0001	***
Sunday	-0.1665	0.1931	-1.226	0.2202	
Monday	-0.4081	0.2255	-3.428	0.0006	***
Tuesday	-0.2665	0.2463	-1.654	0.0981	.
Wednesday	0.0685	0.2271	0.433	0.6652	
Thursday	0.0463	0.1876	1.022	0.3067	
Saturday	0.2334	0.1709	0.702	0.4827	
Human Actions	0.9882	0.1771	5.300	<0.0001	***
Vehicle Conditions	-	-	-	-	
National road	0.5663	0.2295	2.142	0.0322	*
Others road	0.2239	0.1926	0.576	0.5645	
Provincial road	0.6244	0.3564	-0.100	0.9201	
Regional road	0.1618	0.1824	0.020	0.9838	
Sedan	-0.1940	0.0822	-3.976	<0.0001	***
LDV	0.0246	0.1059	-2.741	0.0061	**
Combi	0.3778	0.1534	0.318	0.7506	
Bus	0.0438	0.2668	0.005	0.9963	
Motorcycle	0.5641	0.3273	2.084	0.0371	*
Variables	**Estimate**	**Std. Error**	**Z-value**	**P-value**	
Intercept	-2.190	<0.0001	0.000	1.000	
Sunday	-0.0956	0.8247	-0.221	0.825	
Monday	0. 4502	0.9151	-0.014	0.989	
Tuesday	0.5993	0.8797	-0.222	0.824	
Wednesday	0.5446	0.9539			
Thursday	-2.366	3.466	-0.381	0.703	
Saturday	0.8944	0.6992	0.512	0.609	
Human Actions	-	-	-	-	
Vehicle Conditions	-	-	-	-	
National road	1.267	2.319	0.397	0.691	
Others road	0.9347		0.000	1.000	
Provincial road	1.331	2.319	0.000	1.000	
Regional road	1.105	2.319	0.000	1.000	
Sedan	-0.0849	0.2263	-1.007	0.314	
LDV	0.8666	0.3858	-0.974	0.330	
Combi	0.6099	0.5304	-0.004	0.996	
Bus	-1.083	1.466	-1.206	0.228	
Motorcycle	-2.444	1.068	0.391	0.696	
		Likelihood			
Residual deviance	-1315		**Model *df***	37	

Table 26 (A): Negative binomial model for deaths during the non-holidays.

Variables	Estimate	Std. Error	Z-value	P-value	
Intercept	-2.8270	0.1399	-20.204		
Sunday	0.2405	0.0651	3.692	0.0002	***
Monday	-0.0066	0.0803	-0.083	0.9341	
Tuesday	-0.1927	0.0833	-2.312	0.0207	*
Wednesday	-0.3166	0.0867	-3.651	0.0002	***
Thursday	0.0049	0.0790	0.063	0.9498	
Saturday	0.1494	0.0627	2.382	0.0172	*
Human Actions	1.2928	0.1107	11.670	<0.0001	***
Vehicle Conditions	1.0609	0.1275	8.320	<0.0001	***
National road	0.2807	0.0768	3.653	0.0002	***
Others road	0.1731	0.0838	2.066	0.0388	*
Provincial road	-0.4998	0.1333	-3.748	0.0001	***
Regional road	0.0391	0.0709	0.552	0.5806	
Sedan	-0.1958	0.0299	-6.539	<0.0001	***
LDV	-0.2537	0.0342	-7.401	<0.0001	***
Combi	0.2025	0.0428	4.731	<0.0001	***
Bus	0.6250	0.0977	6.393	<0.0001	***
Motorcycle	0.2145	0.1636	1.311	0.1898	
		Likelihood			
Residual deviance	-9354.226		**Model *df***	20	

Table 27 (A): Zero inflated Poisson model for deaths during the non- holidays.

Variables	Estimate	Std. Error	Z-value	P-value	
Intercept	-0.2523	0.1987	-1.270	0.2041	
Sunday	0.2700	0.0766	3.524	0.0004	***
Monday	-0.1560	0.1014	-1.538	0.1239	
Tuesday	-0.1728	0.1110	-1.556	0.1196	
Wednesday	-0.2329	0.1207	-1.930	0.0536	.
Thursday	0.0094	0.0972	0.097	0.9224	
Saturday	0.1729	0.0764	2.261	0.0237	*
Human Actions	-0.3619	0.1562	-2.316	0.0205	*
Vehicle Conditions	0.1416	0.1688	0.839	0.4013	
National road	0.3774	0.0996	3.786	0.0001	***
Others road	-0.0721	0.1107	-0.652	0.5146	
Provincial road	-0.9952	0.1830	-5.437	<0.0001	***
Regional road	-0.0509	0.0968	-0.527	0.5984	
Sedan	-0.1865	0.0296	-6.284	<0.0001	***
LDV	-0.1913	0.0402	-4.755	<0.0001	***
Combi	0.1003	0.0454	2.209	0.0271	*
Bus	0.7132	0.0819	8.701	<0.0001	***
Motorcycle	-0.5775	0.1987	-4.879	<0.0001	***
Variables	**Estimate**	**Std. Error**	**Z-value**	**P-value**	
Intercept	2.994	0.2808	10.666	<0.0001	***
Sunday	0.0691	0.1426	0.485	0.6276	
Monday	-0.3139	0.2075	-1.513	0.1304	
Tuesday	0.0217	0.2052	0.106	0.9157	
Wednesday	0.1643	0.2107	0.780	0.4356	
Thursday	0.0071	0.1846	0.039	0.9692	
Saturday	0.0648	0.1433	0.453	0.6507	
Human Actions	-2.3680	0.1835	-12.905	<0.0001	***
Vehicle Conditions	-1.1870	0.2015	-5.890	<0.0001	***
National road	0.1784	0.1760	1.013	0.3110	
Others road	-0.5678	0.2143	-2.650	0.0080	**
Provincial road	-1.5820	0.7495	-2.111	0.0347	*
Regional road	-0.2222	0.1750	-1.270	0.2040	
Sedan	-0.2509	0.0556	-4.508	<0.0001	***
LDV	-0.0984	0.0739	-1.332	0.1829	
Combi	-0.3410	0.0920	-3.706	0.0002	***
Bus	0.1078	0.1685	0.640	0.5224	
Motorcycle	-1.4780	1.0770	-0.014	0.9890	
		Likelihood			
Residual deviance	-9644		**Model *df***	38	

Table 28 (A): Zero inflated negative binomial model for deaths during the non-holidays.

Variables	Estimate	Std. Error	Z-value	P-value	
Intercept	-1.2045	0.2404	-5.010	<0.0001	***
Sunday	0.2497	0.0662	3.769	0.0001	
Monday	-0.0743	0.0824	-0.902	0.3671	
Tuesday	-0.2066	0.0852	-2.422	0.0154	*
Wednesday	-0.3323	0.0889	-3.735	0.0001	***
Thursday	-0.0123	0.0804	-0.154	0.8779	
Saturday	0.1544	0.0637	2.422	0.0154	***
Human Actions	-0.2561	0.2204	1.162	0.2452	
Vehicle Conditions	0.1794	0.2427	0.739	0.4596	
National road	0.3029	0.0777	3.895	<0.0001	***
Others road	0.1486	0.0841	1.768	0.0771	.
Provincial road	-0.4858	0.1333	-3.643	0.0002	***
Regional road	0.0236	0.0714	0.331	0.7406	
Sedan	-0.2011	0.0297	-6.759	<0.0001	***
LDV	-0.2672	0.0346	-7.720	<0.0001	***
Combi	0.1458	0.0441	3.302	0.0009	***
Bus	0.6551	0.0981	6.674	<0.0001	***
Motorcycle	0.0987	0.1613	0.612	0.5402	
Variables	**Estimate**	**Std. Error**	**Z-value**	**P-value**	
Intercept	3.2609	0.6385	5.107	<0.0001	***
Sunday	0.1344	0.3362	0.400	0.6893	
Monday	-0.9271	0.4322	-2.145	0.0319	*
Tuesday	-0.4149	0.4458	-0.931	0.3519	
Wednesday	-0.3956	0.4639	-0.853	0.3937	
Thursday	-0.4374	0.4238	-1.032	0.3020	
Saturday	0.0826	0.3409	0.242	0.8085	
Human Actions	-19.121	876.45	-0.022	0.9825	
Vehicle Conditions	-1.6595	0.3451	-4.808	<0.0001	***
National road	0.3411	0.4826	0.707	0.4796	
Others road	-0.6380	0.5610	-1.137	0.2554	
Provincial road	0.1100	1.2137	0.091	0.9277	
Regional road	-0.4135	0.4593	-0.900	0.3680	
Sedan	-0.6939	0.2655	-2.613	0.0089	
LDV	-0.7275	0.2818	-2.581	0.0098	
Combi	-1.3340	0.3209	-4.157		
Bus	0.2965	0.5999	0.494	0.6211	
Motorcycle	-13.451	161.52	-0.083	0.9336	
Likelihood					
Residual deviance	-9285		**Model *df***	39	