STATISTICAL ANALYSIS OF CORRELATES OF NUMBER OF FATALITIES PER TRAFFIC ACCIDENT IN ADDIS ABABA USING COUNT DATA MODELS

AHMED ABDELLA

A thesis submitted to

The Department of statistics

PRESENTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE

DEGREE OF MASTER OF SCIENCE IN STATISTICS

Addis Ababa University

Addis Ababa, Ethiopia

November, 2013

Addis Ababa University
School of Graduate Studies

This is to certify that the thesis prepared by Ahmed Abdella, entitled: statistical analysis of correlates of number of fatalities per traffic accident in Addis Ababa using count data models and submitted in Partial fulfillment of the requirements for the Degree of Master of science in Statistics complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Approved by the Examining Committee:

_____	_____
Advisor	Signature
_____	_____
Examiner	Signature
_____	_____
Examiner	Signature

_____

Chair of Department or Graduate Program Coordinator

Abstract

Ethiopia is a country with a very large number of traffic accidents and fatality rate, and the share of the city of Addis Ababa is quite big (about 60%). Pedestrians and the disabled, children and the aged in particular, are the major victims of these accidents. The aim of this thesis is to analyze fatal traffic accidents in Addis Ababa and identify factors that contribute to the occurrence of road traffic accidents leading to fatalities. This study applies four count models namely Poisson, negative binomial (NB), zero-inflated Poisson (ZIP), and zero-inflated negative binomial (ZINB) models to establish the relationship between the number of fatalities per traffic accident and human, environmental, road and vehicle related factors. Data regarding to number of fatalities per traffic accident were obtained from AATCID for a one year period from April 2012-March 2013. Drivers in the age group 18-30 years, the drivers having less than 5 years of experience and those who are employed were highly associated with the number of fatalities per accident. Among vehicle related factors automobile, taxi-minibuses and vehicles less than five years of service were highly associated with the number of fatalities.

ACKNOWLEDGEMENTS

My first thanks go to the Almighty God, without whose provisions and guidance, my participation in this program of study would have been futile. I would like to express my heartfelt gratitude to my principal advisor, Butte Gotu (PhD), who read, criticized and provided necessary support and encouragement to accomplish this research. My special thanks for Department of Statistics. Also to friends and classmates that contributed in diverse ways to my success in this program, I say a big thank you. Finally, my thanks go to my family, both home and abroad for their continued support, prayer, contributions and bearing with me throughout this program of study

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURE

List of Abbreviations

AATCID	Addis Ababa Traffic Control and Investigation Department
AIC	Akaike information criterion
BIC	Bayesian information criterion
CAIC	Consistent Akaike information criterion
GLM	Generalized linear models
GRSP	Global Road Safety Project
LR	Likelihood Ratio
NB	Negative Binomial
RTA	Road Traffic Accident
ZIP	Zero inflated Poisson
ZINB	Zero inflated Negative Binomial

CHAPTER 1; INTRODUCTION

1.1 Background of the study

A Road Traffic Accident (RTA) is when a road vehicle collides with another vehicle, pedestrian, animal or geographical or architectural obstacle. Globally over 1.2 million people are killed and more than 20 million injured in crashes every year. The global economic losses due to road traffic crashes exceed US$ 500 billion (WHO, 2009). Road traffic accidents pose a significant burden in Ethiopia, as is the case for other developing countries. Currently, developing countries contribute over 90% of the world‘s road traffic fatalities (WHO, 2009) and overall road injury disability-adjusted life year (DALYs) increased by 2.5% between 1990 and 2010, with pedestrian injury DALYs increasing by 12.9%, more than any other category (Murray, et al, 2012). This implies that pedestrian injury on the road is a problem that has increased at a global level and the increase is most likely to be attributable to developing countries. The social and economic impacts of road crashes in developing countries are not well assessed. It is believed that the implications are immense and that road safety needs the attention of researchers, professionals, and politicians. Developing countries have embarked on achieving the United Nations Millennium Development Goals as their main priority, and these do not explicitly include road safety. However, road traffic accident and poverty are linked because family bread winners are highly represented among the fatalities. At least one study has demonstrated that road traffic accident has a negative impact on the achievement of the Millennium Development Goals (Ericson & Kim, 2011). Therefore, the road traffic accident problem in Ethiopia merits investigation both in its own right and because of its links with other development objectives.

Recently, Ethiopia has become one of the fastest growing non-oil producing economies in the world (AfDB, et al., 2012). Car ownership has grown rapidly at about 8% per annum on average (Akloweg, Hayshi, & Kato, 2011). The construction of roads is one of the major focal areas of the government to fast-track economic growth. Although the vehicle population growth rate per annum is increasing, the number of total vehicles remains low compared to other developing countries. Fatal traffic accidents are a major public health concern. Ethiopia experiences the highest rate of such accidents in Sub-Saharan Africa. Out of all the accidents registered in Ethiopia, Addis Ababa accounts about 60% on average (Bisrat, 2010)

Road traffic injuries are growing as the vehicle use of developing countries rises. By 2020, road traffic accident are expected to be the third leading cause of death and disability worldwide, by some calculations matching the toll of AIDS. Residents of developing countries are at much higher risk of road traffic fatalities than are residents of high-income countries. They are also at greater risk of death when a crash occurs. Developing countries also have inadequate trauma systems and are often unable to care for crash victims. It was indicated that unless action is taken to improve road safety systems, poor countries will continue to bear the heavy toll of road traffic fatalities (Lauren and Hill, 2005).

According to the WHO data published in April 2011, Road Traffic Accidents deaths in Ethiopia reached 22,786 per year (2.77% of total deaths). The age adjusted death rate of 37.83 per 100,000 of population ranks Ethiopia 12th in the world. Road accidents appear to occur regularly at some flash points such as where there are sharp bends, potholes and at bad sections of the highways. At such points over speeding

drivers usually find it difficult to control their vehicles, which then results in fatal traffic accidents, especially at night (Atubi, 2009).

The United Nations has declared this an issue of great concern. In this study, a road traffic accident is defined as an accident which took place on the road between two or more objects, one of which must be any kind of a moving vehicle (Jha et al, 2004). Road traffic accidents are increasing with a rapid pace and presently these are one of the leading causes of death in developing countries. According to world report on traffic injury prevention-2004, road traffic accidents were ranked as the 6th place (was the 9th in 1990) of a major cause of death worldwide, will rise to become the 3rd leading cause of DALYs lost by 2020; the 2nd leading cause of DALYs lost for low and middle income countries; fatalities will increase worldwide from 0.99-2.34 million (representing 3.4% of all deaths); fatalities will increase on average by over 80% in low-income and middle-income countries and decline by almost 30% in high-income countries; DALYs lost will increase worldwide from 34.3-71.2 million (representing 5.1% of the global burden of disease).

The morbidity and mortality burden in developing countries is rising due to a combination of factors, including rapid motorization, poor road and traffic infrastructure as well as the behavior of road users (Nantulya and Reich, 2002). Traffic accidents are a ‗global tragedy' with ever-rising trend in fatalities and injuries in the developing countries.

Accident rates in developing countries are often 10-70 times higher than in developed countries. Whereas traffic accident situation is slowly improving in the industrialized societies (e.g. Australia, USA, UK etc.), most developing countries face a worsening situation. For developing measures aimed at reducing the rate of road traffic accidents

and the consequent injuries and fatalities, there is the need for regular evaluation of the road traffic accidents in terms of developing statistical models.

The annual cost of road crashes is in excess of US $500 billion, and in the developing world the estimated cost is about US $65 billion each year. Due to the scarcity of costing data for African countries, it is difficult to make a precise cost of road crashes in Sub-Saharan Africa. The estimate of costs of crashes in the continent is US$ 3.7 billion per year, of which South Africa alone accounts for 2 billion. However, the estimated cost as a percentage of the Gross National Product (GNP) in most African countries range from about 0.8% in Ethiopia and 1% in South Africa to 2.3% in Zambia and 2.7% in Botswana to almost 5% in Kenya (Kopits and Cropper, 2003)

A Global Road Safety Project (GRSP) study shows that about 10 per cent of global road deaths in 1999 took place in Sub-Saharan Africa where only 4 per cent of global vehicles are registered. Conversely, in the entire developed world, with 60 per cent of all globally registered vehicles, only 14 per cent of road deaths occurred. However, given the widely recognized problem of under-reporting of road deaths in Africa (like the rest of the developing world); the true figures are likely to be much higher, as the police- reported road fatalities represent only the tip of the injury pyramid.

According to this GRSP study, the adjusted true estimate of total road deaths for all Sub-Saharan African countries for the year 2000, based on the police department‘s records, ranges between 68,500 and 82,200. However, the estimated fatality figure of 190,191 per year for Sub-Saharan Africa presented in the 2004 World Report, based on health care data, is much higher, and reflects the magnitude of under-reporting in police statistics.

1.1.1 Ethiopian context

Ethiopia is one of those developing countries with low level of income accompanied by the high rate of population growth. As part of the developing world, Ethiopia is predominantly an agrarian country with a low level of urbanization. The economic performance of different sectors of the national economy is very low. This low performance is due to a number of constraints such as low level of investment in different sectors of the national economy. Among these the existing transport could be mentioned as one. Transport is an important sector in facilitating different economic activities in the national economy.

Out of all the accidents registered in Ethiopia, Addis Ababa accounts about 60% on average. This is partly because the city has only five outlets that connect to all regions of the country. In addition to this about 77% of vehicles in Ethiopia are registered here. Thus Addis Ababa, having a great concentration of vehicles and traffic, takes the lion‘s share in vehicle accidents. Statistical data from the Addis Ababa Traffic Control and Investigation Department (AATCID) show that the city is experiencing around 700 accidents per month and the costs of such fatalities and injuries have a great impact on various aspects of the society (Tesema and Tibebe, 2005).

Among the 3100 records which were collected by AATCID within 310 consecutive days from 04/03/06 to 07/01/07, a total of 1141 (37%) accidents involved human injuries (slight, series and death). The remaining 1959 (63%) is considered as accidents with material damage only. According to Tewolde (2007), the highest mean number of injuries per accident took place in residential areas by drivers in the age group of 18-30 who have an elementary school level of education. Pedestrians and the disabled, children and the aged in particular, are the major victims of these accidents.

In addition to human life and bodily harm costs, the severity of the situation in economic terms is also very alarming. It was stated that more than 12 million Birr is lost every year because of traffic accidents (Asfaw, 1999).

1.2 Statement of the problem

Developing countries face many challenges and have many resource needs. Road safety tends not to receive due consideration because not all road accidents and casualties are reported to the police and there is usually no other system of estimating road accidents and the corresponding casualties nationwide. There is also a problem with the perception that road accidents are random, unintentional, or predestined; i.e., unavoidable. Road accidents are too often accepted as inevitable negative side effects of motorization (Murad, 2011). The root causes of road accidents and its effects on human lives and properties have been associated with human errors and superstition.

The rate of traffic accidents in Addis Ababa goes up together with the increase of motor vehicles and population size. The rise in automobile ownership together with the poor condition of the roads has resulted in the high level of traffic safety and congestion problems. The traffic accidents have been categorized by the AATCID as fatal, serious and minor. This classification is based on the extent of damage to human lives and properties.

Over the years, the Addis Ababa Traffic Control and Investigation Department (AATCID) uses descriptive statistics techniques and charts such as bar graphs, histograms and frequency polygons to organize number of fatalities per accident in Addis Ababa city. This statistical approach of analyzing does not inform the department about the estimates of RTA deaths.

In light of this, it is necessary to use statistical analysis such as nonlinear regression model which include Poisson NB, ZIP and ZINB models in order to better describe and model the accident data. It is also important to find out the effect of factors such as human, environmental, road condition and vehicle type on traffic accidents.

1.3 Objective of the study

1.3.1 General objectives

The objective of this study is to identify and analyze correlates of number of fatalities (human death) per accident

1.3.2 Specific objective

- To examine the effects of various factors on the number of fatalities per traffic accident.
- To model road accident fatality in Addis Ababa using count data models

1.4 Significance of the study

The results of this study will be useful for road safety planning in Addis Ababa. The results could also provide a reliable statistical technique for analyzing accident data and identify patterns of road traffic accidents.

CHAPTER 2; LITRETURE REVIEW

This aspect of the study reviews the various literatures related to the topic under consideration in order to uncover critical facts and findings which have already been identified by previous researchers and numerous studies in and around the causes, effects and economic implications of road accidents with particular reference to the number of deaths per accident. As countries develop death rates usually fall, especially for diseases that affect the young and result in substantial life-years lost. Deaths due to traffic accidents are a notable exception: the growth in motor vehicles that accompanies economic growth usually brings an increase in road traffic accidents. Indeed, the World Health Organization has predicted that traffic fatalities will be the sixth leading cause of death worldwide and the second leading cause of disability-adjusted life-years lost in developing countries by the year 2020 (World Bank 2003).

Accident type analysis showed that hitting pedestrian is the dominant accident type both in urban and rural areas with 45% involvement in fatal accidents. Another common accident types are as rearing end (16.5%), heading on (13.2%) and overturning vehicle (9.3%). These four types account for nearly 85% of the fatal accidents. Indeed the running-off-road accident has the highest rate of about 19% fatalities per accident (*Hoque et al., 2003*). The predominant type of fatal collision on any type of road and in either an urban or rural locality is a vehicle hitting a pedestrian. The half of all fatal accidents being collisions where a vehicle hit a pedestrian is evident in all divisions and cities. Minor changes are noticed showing improvement compared to the previous year (55%) in total hit pedestrians. The range is 42%-75%. Dhaka city (75%) and Rajshahi city and Khulna city are at the upper end of the range. The incidence of hit pedestrian is the highest on national highways.

More than half of fatal victims of traffic accidents are pedestrians. Physical works which either provide a safer environment for pedestrian (e.g., speed limit zones) or provide safer pedestrian facilities (e.g., footpaths, full width shoulders, pedestrian over-bridges) are the types of measures which can address this safety problem. Most importantly, reckless behaviour of drivers of buses and trucks will need to be controlled. Following hitting a pedestrian, holding on, rearing end and overturning are other major collision types in rural localities that account for a high proportion of fatal accidents. Dangerous driving, reckless overtaking and over-speeding are some of the factors responsible for these types of accidents. Of course, badly designed and constructed roads are also a contributing factor. Dhaka city and Rajshahi city have comparatively higher fatal accident rates (more than twice that of any other city) with a predominance of pedestrian accidents and rear end accidents (*BRTA, 2007*).

Hoque *et,al, (2003)* found that over the involvement of truck and buses are particularly prevalent in Bangladesh. Earlier studies of police reported accidents on a section of Asian highways revealed that trucks and buses (including minibuses) accounted for about 72% (track 34%, buses 20% and minibuses 18%) of the vehicles involved in fatal accidents. This group of vehicles is particularly over involved in pedestrian accidents accounting for about 79% (trucks 37%, buses 20% and minibuses 22%). At some locations, trucks involvement was found in 43-50 %. *Bose* (2007) found that fatal accidents occur needlessly and take the lives of innocent people, mainly through the reckless behavior of the drivers of buses and trucks.

Ongoing advances in crash count modeling and prediction stem from several issues common to such data. For example, zero-inflated Poisson and negative binomial models were developed as a remedy for the preponderance of zeroes in crash data a

phenomenon particularly common for fatal crash counts. Lord and Mannering (2010) argue that a high share of zero counts (which lead to rather low sample mean values) can create biased estimators, as seen in Lord‘s (2006) smalls ample estimate of the negative binomial model‘s dispersion parameter. The incorrect estimation of dispersion parameters also negatively affects parameter-based inferences. In a study involving all collisions (including collisions involving fatalities), it was found that 57% of fatal crashes were due solely to driver factors, 27% to combined roadway and driver factors, 6% to combined vehicle and driver factors, 3% solely to roadway factors, 3% to combined roadway, driver, and vehicle factors, 2% solely to vehicle factors and 1% to combined roadway and vehicle factors (Lum & Reagan, 1995).

In theory, it seems that there are many variables to look at. A study, which used multivariate logistic regression, revealed that the odds ratio (OR) of a fatal accident increased with age. Gender also seems to be a characteristic that has significance. In the same research study, the majority of fatalities were among male drivers younger than 30 years, which was at 26.6%, versus females of the same age range, at a 5.6% (Bedard, Guyatt, Stones & Hirdes, 2002). Driving is a very complex task, which involves various cognitive, physical, sensory, and psychomotor skills working together. Distractions are defined as any secondary activity that competes for the driver‘s attention while driving. These distractions have the potential to worsen driving performance and have serious consequences for road safety. According to the NHTSA, it is estimated that 25% of police-reported fatal crashes are caused by driver inattention (NHTSA, 2010). According to one study, more experienced drivers are often capable of dividing their attention between driving tasks and non-driving tasks without any serious consequences (Young & Regan, 2007). Another study that

supports this found that 16% of all under-20 year old drivers involved in fatal crashes were reported to have been distracted while driving (Ascone & Lindsey, 2009).

Tewolde (2007) used a Poisson regression model to identify factors that mainly affect crash-related injuries (highly injured, slightly injured and death). According to the results of his study, driver age is found to be significantly associated with the occurrence (number) of injuries. Among the age categories, drivers in the age group 18-30 are responsible for the large number of injuries (51%) and the large number of accidents (46%). This indicates that the young were involved in the majority of the accidents. According to this study drivers with 2-5 years‘ experience take the major share i.e., above 29% drivers with 5-10 year experience and above 10 years of experience are having nearer figures when compared to the former one. When it comes to the number of injuries per accident the highest mean (.53) is attained by those with less than one year of experience and the next higher mean (.50) is attained by those with 1-2 years of experience. But the driver - vehicle relationship and sex of drivers were found to be not statistically significant.

Numerous cross-sectional studies have been conducted in varying scales and scopes in order to understand the relationships between factors and fatal traffic accidents by combining several years of data and performing statistical analysis and constructing statistical models. The multiple regression, logistic regression and Poisson regression are commonly used for modeling the mortality rates and the number of deaths in a specific population.

The fatality rate over the years has been used to compare road accident incidence in a large number of countries. The fatality rate is defined as the number of deaths which occurred through road accidents with respect to some measure of the use of the road

system. However, the fatality rate has been defined by several authors to suit the needs of their researches. Ghee et al (1997) stated that the fatality rate is defined as the number of injury accidents occurring per annum per million vehicle kilometer travelled. But since there is no much reliable accident data base in developing countries and much information required to compute this type of fatality rate, Ghee et al (1997) defined the fatality rate of road accidents in a given country to be measured in respect of the number of persons killed through road accidents per 10,000 licensed vehicles in a country. As the population increases the numbers of licensed vehicles in developing countries are rising rapidly.

However, Ghee et al (1997) suggested that this index cannot be used to compare accident fatality rates of different countries from the countries may vary in terms of population and total vehicles which ply their roads. He then proposed a model which assessed the relationship between fatalities, population and motorization of the country. This model supported the Smeed Formula for international comparisons of accident fatalities (Smeed, 1938 and 1968). Smeed in 1938 used accident data from different countries and proposed the formula $\frac{D}{N}$=0.0003$(\frac{N}{P})^{-0.67}$, where D is the annual number of fatalities from road accidents, N is number of vehicles in use and P is population size (Smeed, 1968).

Tsauo et al (1996) examined the effect of age, period of death and birth cohort in motor vehicle mortality in Taiwan from 1974 – 1992, using data from vital statistic. Log-linear regression was used for fitting the model to perform the effects of variables. However Pocock et –al (1981) pointed out that unweighted multiple regression is not appropriate for modeling mortality rates in different areas which vary in population size. In addition fully weighted regression is usually too extreme. Thus

they introduced an intermediate solution via maximum likelihood for modeling death rates.

Robert (2000) analyzed the relationship between road infrastructure and safety by using a cross-sectional time-series database collected from all 50 U.S. states over 14 years. Data on total fatalities and total injuries by state was collected. Data on road infrastructure included total lane miles (excluding local roads), average number of lanes by functional road category (interstates, arterials, and collectors), percent of centerline miles with a given lane width of road category, and the fractional percent of each road category in a given state (including local roads within the denominator). Interstates are controlled access highways built to the most rigorous and consistent design standards. Arterials are generally major multi-lane or intercity roads, perhaps with some controlled access, but generally not. These also tend to be major connector roads within cities and suburban areas. Collector roads are smaller scale roads that generally connect local distributor roads with arteries. A casual interpretation of the trends and those for total fatalities would suggest that as highway facilities are upgraded, there have reduced fatalities. In addition, estimates of seat belt usage, by the state, were used to control for the effects of increased seat belt use. The analyses also attempts to control for seatbelt effects by including dummy variables for those states with either primary or secondary seat belt laws. Data on total population, vehicle miles of travel (VMT), per capita income, alcohol consumption and population by age cohorts was also collected. These are used in the models primarily to control for other factors that are likely to affect fatalities and injuries. The occurrence of traffic crashes and the resulting injuries and fatalities are Poisson distributed. The use of a Poisson regression is usually affected by over-dispersion in the error term due to the inequality of the mean and variance within the data. This is

easily corrected by using a negative binomial regression. A number of different models were estimated using the data described above. The key variables of interest are the infrastructure variables. Other variables known to affect crashes are also included, specifically age cohorts, per capita income, state population, and VMT. VMT and population size cannot be included in the same model due to high collinearity between them. Separate models for each are therefore estimated.

Kardara and Kondakis (1997) identified trends of road traffic accident deaths and injury rates in Greece from 1981-1991 by using linear regression with logarithmic transformation. LaScala et al (2000) examined correlations between demographic and environmental versus pedestrian injury rates by using a spatial autocorrelation corrected regression model by applying the logarithmic transformation for the injury rates. Evans (2003) conducted statistical modeling for 11 estimating road traffics and railway accident fatality rates based on past accident data in Great Britain during 1967-2000.

Time series analysis was used by Mekky (1985) to study the effect of rapid increase in the motorization levels on the rate of fatalities in some developing countries. Many researchers have dived into the investigation of traffic crash patterns in different countries in order to understand its relationship with the fatality rate of road accident. Among such researchers are Dinesh (1985) who investigated crash patterns in Delhi, Emenalo et al (1987) developed the trend curves for road accidents, casualties and other vital quantities in Zambia, and Pramada and Sarkar (1993) who studied the variations in the pattern of road accidents in various States and Union Territories of India.

Abdel (2005) studied road accidents in Kuwait. He used an ARIMA model and compared it with ANN to predict fatalities in Kuwait. He concluded ANN was better in case of a long term series without seasonal fluctuations of accidents or autocorrelations components. Cejun and Chiou-Lin (2004) used two time series techniques ARMA and Holt-Winters (HW) algorithm, to predict annual motor vehicle crash fatalities. They concluded that the values predicted by ARMA models are a little bit higher than the ones obtained by HW algorithm.

Bisrat (2010) used binary and ordinal logistic regression models to identify factors influencing traffic fatalities and injuries in Addis Ababa, Ethiopia. In this study ordinal logistic regression analyses show that drivers aged 18-30 years caused the largest number of accidents. Low educational background of drivers, absence and poor lighting along roads, wet surface and asphalt surface, morning and evening hours, places like offices, residential and commercial neighborhoods, automobiles and small taxis/minibuses were found to be associated with fatalities and serious injuries.

Ulfarsson (2001) and Ulfarsson and Mannering (2004) focused on male and female differences in analysis of accident severity. They used multinomial logit models and accident data from Washington State. They found significant behavioral and physiological differences between genders, and also found that the probability of fatal and disabling injuries is higher for females as compared to males. Bedard et al (2002) applied a logistic regression to determine the independent contribution of driver, crash, and vehicle characteristics to drivers_ fatality risk. They found that increasing seatbelt use, reducing speed, and reducing the number and severity of driver-side impacts might prevent fatalities.

Crash frequency models are typically of the Poisson form. The number of crashes in a given space-time region is regarded as a random variable that takes values 0, 1, 2, ... with probabilities obeying the Poisson distribution. A characteristic feature of this distribution is that the variance is equal to its mean. More recently, negative binomial models, a variant of the Poisson, have been used in crash modeling. Such models generalize the Poisson form by permitting the variance to be over-dispersed, equal to the mean plus a quadratic term in the mean whose coefficient is called the over-dispersion parameter (Vogt and Bared, 1998).

Yang et al. (2005) used Poisson regression modeling to examine and compare age- and sex-specific mortality rates due to injuries in the Guangxi Province in South Western China in 2002 based on death certificates data. However this study focused only on small areas. Poisson regression analysis is a technique used to model dependent variables that describe count data (Cameron and Trivedi, 1998). It is often applied to study the occurrence of a small number of counts as a function of a set of predictor variables, in experimental and observational study in many disciplines, including Economy, Demography, Psychology, Biology and Medicine (Gardner, 1995). The Poisson regression model may be used as an alternative to the Cox model for survival analysis, when hazard rates are approximately constant during the observation period and the risk of the event under study is small (e.g., incidence of road accidents).

Poisson regression model usually replaces Cox model, which cannot be easily applied to aggregated data. Furthermore, using rates from an external population selected as a referent, Poisson regression model has often been applied to estimate standardized mortality and incidence ratios in cohort studies and in ecological investigations

(Breslow et al 1987). Finally, some variants of the Poisson regression model have been proposed to take into account the extra-variability (over dispersion) observed in actual data, mainly due to the presence of spatial clusters or other sources of autocorrelation (Cameron et al, 1998)

Miauo (1994) compared Poisson and Negative Binomial regressions since they all cater for the distributional properties of accident data. Many other researchers have also assessed the use of linear regression models for road accident models and confirmed the limitations in such models. Previous research applied Poisson or negative binomial regression models because of the distributive property of vehicle accidents (Milton and Mannering, 1998). Although the Poisson regression model has desirable statistical properties for describing vehicle accidents, it has an important constraint, which is that the mean and variance of the accident data are constrained to be equal. To overcome this constraint, the negative binomial regression model has been employed to analyze vehicle accidents (Miaou, 1994).

Tewolde (2007) also examined the relationship between types of vehicle and the occurrences of traffic accidents in Addis Ababa by using Poisson model. According to the results, automobiles and taxi are responsible for the largest number of injuries with values 26.8 % and 26% respectively. Kweon and Kockelman (2003) studied probabilities of accidents and accident severity outcomes for a given fixed driver exposure (which is defined as the total miles driven). They used Poisson and ordered probit models, and considered a nationwide accident data sample. After normalizing accident rates by driver exposure, the results of their study indicate that young drivers are far more crash prone than other drivers, and that sport utility vehicles and pickups are more likely to be involved in rollover accidents.

Shankar et al (1997) studied the distinction between safe and unsafe road sections by estimating zero-inflated Poisson and zero-inflated negative binomial models for accident frequencies in Washington State (for these models the zero state corresponds to near zero accident likelihood on safe road sections). Shankar et al (1996) used a nested logit model for statistical analysis of accident severity outcomes on rural highways in Washington State. They found that environment conditions, highway design, accident type, driver and vehicle characteristics significantly influence accident severity. They found that overturn accidents, rear-end accidents on wet pavement, fixed-object accidents, and failures to use the restraint belt system lead to higher probabilities of injury and fatality accident outcomes, while icy pavement and single-vehicle collisions lead to higher probability of property damage only outcomes.

Lee and Mannering (2002) estimated zero-inflated count-data models and nested logit models for frequencies and severities of run-off-roadway accidents in Washington State. They found that run-off-roadway accident frequencies can be reduced by avoiding cut side slopes, decreasing (increasing) the distance from the outside shoulder edge to guardrail (light poles), and decreasing the number of isolated trees along roadway. The results of their research also show that run-off-roadway accident severity is increased by alcohol impaired driving, high speeds, and the presence of a guardrail.

CHAPTER 3; METHODOLOGY

This chapter deals with a detailed description of source of data, variable included in the study and the methods used for this research and explains the theory behind the various distributions and models for the analysis. The chapter also addresses the possible probability distributions of fatalities per accident (count data) and their likely regression models which may include the Poisson, Negative Binomial, ZIP and ZINB.

3.1 Sources of Data

In Addis Ababa road traffic accidents are recorded by AATCID on a daily basis. This study is based on a secondary data obtained from AATCID. The data provide information on accidents that occur within consecutive days from April 2012-March 2013. The total number of accidents covered in the study was about 3880. Among these 2668 accidents have no fatal accident while 1212 accidents involve at least one fatality.

3.2 Variables included in the study

Response variable the number of fatalities per accident.

Independent Categorical Variables

Human related variables are a gender of driver, age of drivers, driving experience, driver-vehicle relationship, ownership and driver educational background.

Environment related variables are atmospheric condition, light condition, day of the accident and time of the accident.

Roadway related variables are type of pavement, intersection type, road geometry, road type and road condition.

Vehicle related variable are driving direction at the time of accident, vehicle type, and vehicle service years.

3.3 Coding and Description of the variable

A description of human, vehicle, environmental, road factors, and their respective coding are provided in appendix A.

3.4 Methods of Data Analysis

Even though there are several statistical models, some models may not be appropriate to deal with some specific types of data. Their use is solely depending on the types and nature of the data. In this study, the variable of interest is a count data, which is most often characterized as non-normal distribution. When the response or dependent variable is a count data (which can take on non-negative integer values (0, 1, 2, ...), it is not appropriate to use linear models based on normal distribution to describe the relationship between the response variable and a set of predictor variables and we cannot use the binary logistic regression model because the response variable is not binary (0, 1). In this case the Poisson regression model is the popular tool to describe it (Cameron and Trivedi, 1998).

3.4.1 Poisson regression model

The standard Poisson distribution is a fundamental distribution to understand regression counts models. It was developed to model discrete count data, since it is easy to interpret in many aspects. According to Sturman (1999), the apparent simplicity of Poisson comes with two restrictive assumptions. First, the variance and mean of the count variable are assumed to be equal. In reality, however, the variance is usually much larger than the mean. Although Poisson regression models are widely

used to handle count data, it may not be well suitable to handle some types of count outcomes such as an over dispersed or under dispersed data. The other restrictive assumption of Poisson models is that occurrences of the event are assumed to be independent of each other.

The Poisson regression model assumed that the mean and the variance of the response variable are equal but in practice, the observed variance of the data may be larger than the corresponding mean. In these cases, the data is said to have involved over – dispersion, the variance is larger than the mean, for such situations, the Poisson model is not appropriate and the negative binomial regression model is appropriate (Paternoster and Brame, 1997 and Osgood, 2000). However, if the variance is larger than the mean, it induces deflated standard errors and inflated standardized normal (i.e. Z-normal) value and these make Poisson regression less adequate (Elhai et al, 2008). Some researchers suggest that, when there is an over-dispersion it is better to use other models, such as negative binomial which can take care of the over-dispersion problem (Cameron and Trivedi, 1998).

Assumptions of Poisson distribution are:

- Observations are independent.
- Probability of occurrence in a short interval is proportional to the length of the interval.
- The probability of another occurrence in such a short interval is zero.

The standard model for count data is the Poisson regression model, which is a nonlinear regression model. This regression model is derived from the Poisson distribution by allowing the intensity parameter μ to depend on covariates. If the

dependence is parametrically exact and involves exogenous covariates but no other source of stochastic variation, we obtain the standard Poisson regression. If the function relating μ and the covariates is stochastic, possibly because it involves unobserved random variables, then one obtains a mixed Poisson regression, the precise form of which depends on the assumptions about the random term. The scalar dependent variable, y_i, is the number of occurrences of the event of interest, and $\mathbf{x}_i$ is the vector of linearly independent explanatory that are thought to determine y_i. A regression model based on this distribution follows by conditioning the distribution of y_i on a k-dimensional vector of covariates, $\mathbf{x}_i = [x_{1i}, \ldots\ldots x_{ki}]$, and parameters β, through a continuous function $E[y_i|\mathbf{x}_i] = {}_i$ (Cameron and Trivedi ,1998).

The Poisson mass function is given by;

$$f(y_i|\mathbf{x}_i) = \frac{e^{-\ i}\ {}_i^{y_i}}{y_i!},\ y_i = 0, 1, 2, \ldots\ldots, \qquad (1)$$

In the log-linear version of the model the mean parameter is parameterized as

$${}_i = \exp(\mathbf{x}_i'\beta) \qquad (2)$$

By the property of the Poisson, $V[y_i|\ \mathbf{x}_i] = E[y_i|\ \mathbf{x}_i]$, implying that the conditional variance is not a constant, and hence the regression is intrinsically heteroskedastic. In the log-linear version of the model the mean parameter is parameterized as (2). The Poisson model is closely related to the models for analyzing counted data in the form of proportions or ratios of counts sometimes obtained by grouping data. In some situations, for example when the population ~~–at~~ risk" is changing over time in a known way, it is helpful to parameterize the model as follows. Let *y* be the observed number of events (e.g., number of fatalities per traffic accident), **x** a known set of *k* explanatory variables and μ the mean number of events. Given independent observations with the density function, the log-likelihood function can be obtained by:

$$l(\beta) \quad = \sum_{i=1}^{n}[\, Y_i \log(\ _i) - \ _i - \log(Y_i!)]$$

$$= \sum_{i=1}^{n}[\, Y_i x\ _i\beta - \exp(x\ _i\beta) - \log(Y_i!)], \qquad (3)$$

If there is over-dispersion causing the variance to be larger than the mean, then the estimation will be inefficient using a Poisson regression.

3.4.2 Negative binomial regression model

The negative binomial is a conjugate mixture distribution for count data. When the Poisson model assumption fails, negative binomial regression model may fit better, and address the over-dispersion problem. However, this is true only if it is not attributed to excess zeros. As we have discussed in section 3.4.1, a severe limitation of the standard Poisson models assumption is, that the variance of the data is equal to the mean of the data. Hence, at a fixed mean the variance cannot decrease as additional predictors enter the model.

Like Poisson regression, negative binomial regression model also examines predictive relationships with a count dependent variable. The standard Poisson regression accounts for observed differences among the observations; however negative binomial regression includes a random component that involves an unobserved variance among observations. The inclusion of this random component prevents the incorrect Poisson assumption, that is, all differences among subjects in the dependent variable are equally explained. In an over-dispersed data this random component results in more accurate standard errors and z-statistics for the regression coefficients than using the standard Poisson regression (Elhai et al, 2008).

Over-dispersion might happen when some relevant explanatory variables are not included in the model. Suppose λ_i has a gamma distribution with mean $\mathrm{E}(Y_i/\lambda_i)=\ _i$

and variance var(Y_i/λ_i)= $\frac{i}{\alpha^{-1}}$ to be a Poisson with conditional mean E(Y_i/λ_i)= λ_i. It can be shown that the marginal distribution of Y_i follows a negative binomial distribution with probability mass function:

$$f(Y_i|\mu,\alpha) = \frac{\Gamma(y_i+\alpha^{-1})}{\Gamma(\alpha^{-1})\gamma(y_i+1)}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right)^{\alpha^{-1}}\left(\frac{\mu_i}{\alpha^{-1}+\mu_i}\right)^{y_i},\ y_i=0,1,2,\ldots, \quad (4)$$

With mean $E(Y_i/x_i)=\mu_i$ and variance, $Var(Y_i/x_i)=\mu_i(1+\alpha\mu_i)$, where Γ(.) is the gamma function and the index α is called the dispersion parameter. As α approaches to zero**,** the variance and mean become identical**.** Hence the negative binomial will reduce to a Poisson. In such cases the data can be modeled easily by Poisson regression model. If $\alpha > 0$, the variance will exceed the mean, that is var(Y_i) > $E(Y_i)$, and the distribution allows for over dispersion (Agresti, 2007).

The negative binomial log-likelihood function is given by:

$$l(\alpha,\beta)=\sum_{i=1}^{n}\left[\log\left(\frac{\Gamma(Y_i+1/\alpha)}{\Gamma(Y_i+1)\Gamma(1/\alpha)}\right)-(Y_i-1/\alpha)\log(1+\alpha\mu_i)+Y_i\log(\alpha\mu_i)\right] \quad (5)$$

The negative binomial regression model is a useful model for count data in which unobserved heterogeneity is present. It is not necessarily an optimal model for dealing with data that contain an excess mass of zeros at the corner of its empirical distribution. Greene (1994) introduced the idea of the Zero Inflated Negative Binomial (ZINB) regression model to handle both excess zeros and over-dispersion as a result of unobserved heterogeneity.

3.4.3 Zero Inflated models

In some cases, excess zeros exist in count data and considered as a result of over-dispersion. In such a case, the NB model cannot be used to handle the over-dispersion which is due to the high amount of zeros. To do this, zero-inflation (ZI) models

including Zero Inflated Poisson (ZIP) and Zero Inflated Negative Binomial (ZINB) models can be alternatively used. The ZIP model, introduced by Lambert (1992), is served as a dual-state method for modeling data characterized by a significant amount of zeros or more zeros than the one would expect in a traditional Poisson or negative binomial model, while the ZINB model, introduced by Greene (1994), is a more flexible model that can be used to handle over-dispersion caused by both unobserved heterogeneity and excess zeroes. Both the ZIP and ZINB models assume that all zeros counts come from two different processes: (i) the process generating excess zero count (zero-number of fatalities per traffic accident) derived from a binary model, and (ii) the process generating non-negative counts for the number of fatalities per traffic accident including zero values, which is estimated from the Poisson/NB distribution (Welsh et al., 1996).

3.4.4 Zero-inflated Poisson regression model

A characteristic of the Poisson distribution as presented in Section 3.4.1 above is that the mean of the distribution is equal to the variance; however when there are excess zeros, probability of zero in the standard model will be less than the expected. Therefore, in such situation the standard Poison and negative binomial models are not suitable models. In such cases, a ZIP or ZINB models can be used to account for excess zeros. The zero values in the ZIP model can be viewed as comprising two parts. One portion of the zero counts arises from the inflated part of the distribution and the other portion comes from what would be expected given a Poisson distribution with parameter λ. When there are excess zeros and high variation in the non-zero outcomes, ZIP models have less adequate than ZINB models. Suppose Y_i is the number of fatal per accident then, the probability mass function of ZIP is given by:

$$P(Y_i=/p_i,\lambda_i)=\begin{cases} p_i + (1-p_i)\exp(-\lambda_i), & if\ y_i = 0 \\ 1-p_i\frac{\lambda_i{}^{y_i}\exp(-\lambda_i)}{y_{i!}}, & if\ y_i > 0 \end{cases} \quad i=1,2,\ldots n \qquad (6)$$

Where λ_i is the mean of the non-zero outcomes that can be expressed with the associated explanatory covariates using a natural logarithmic link function as:

$$\ln(\lambda_i) = X_i'\beta, \qquad (7)$$

Where X_i= (1, x_{i1} , x_{i2} ,....., x_{ip-1})' is a px1 vector of explanatory variable of the i^{th} observation and β is px1 vector of regression coefficient parameters. p_i is the probability of an excess zero which can be estimated by the logistic regression (Lambert 1992, Long 1997). That is

$$\text{Logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = Z_i\,\gamma \text{ or } p_i=\frac{\exp(Z_i\,\gamma)}{1+\exp(Z_i\,\gamma)}, \quad i=1,2,\ldots\ldots n_i, \qquad (8)$$

Where Z_i= (1, Z_{i1} , Z_{i2}....., Z_{ip-1}) is a qx1 vector of explanatory variable for the zero-inflation part model of the i^{th} observation and γ =(1,$\gamma_1, \gamma_{2,\ldots\ldots\ldots}\gamma_{q-1}$) is qx1 vector of regression coefficient parameters.

Unlike the Poisson distribution, which is determined by a single parameter, the ZIP distribution is determined by two parameters, λ_i and p_i. The ZIP model is a special case of a two-class finite mixture model with mean and variance $E(Y_i)$= $(1-p_i)\,\lambda_i$ and var(Y_i)=($1-p_i$)($\lambda_i + p_i\lambda_i{}^2$) respectively (Liu, et al, 2007).

3.4.5 Zero-inflated negative binomial regression model

The main motivation for zero-inflated count models is that real-life data frequently display over-dispersion and excess zeros (Cameron and Trivedi, 1998). Zero-inflated count models provide a way of modeling the excess zeros in addition to allowing for over-dispersion. If the dependent variable presents a high proportion of zeros which

could create problems for the negative binomial estimation, a modified count model is the zero inflated Negative Binomial (ZINB) models which take the existence of excess zeros into account. When dealing with count response variables the number of zeros is often excessive. This is because there are two processes that generate zero responses, with one always generating zero counts and the other both zero and non-zero counts. Both of these outcomes present an identical zero response but the process through which they are reached is very different.

Zero-Inflated Negative Binomial (ZINB) regression Model is an extension of the NB regression model. As the number of zeros in the count distribution is excessive, then the ZIP or ZINB model will be more accurately fit the data than the negative binomial or Poisson model. If over-dispersion is not accounted by the ZIP model, then there may be other aspects of the distribution that contribute to over-dispersion, in such case the ZINB model is more appropriate(Long, et al, 2006). The main difference between ZIP and ZINB model is that the Poisson distribution for the count data is replaced by the negative binomial distribution. The probability function of a ZINB is a simple modification of the ZIP. Suppose Y_i is the number of fatalities per accident. Then the probability density function of ZINB the random variable Y_i distributed as ZINB is given by:

$$P(Y_i=/p_i,\lambda_i)=\begin{cases} p_i+\frac{(1-p_i)}{(1+\alpha\lambda_i)^{\alpha^{-1}}}, & if\ y_{i=0} \\ 1-p_i\frac{\gamma\left(y_i+\frac{1}{\alpha}\right)}{\gamma\left(\frac{1}{\alpha}\right)\gamma(y_i+1)}\frac{(\alpha\lambda_i)^{y_i}}{(1+\alpha\lambda_i)^{y_i+\frac{1}{\alpha}}}, & if\ y_{i>0} \end{cases} \quad i=0,1,2,\ldots n \qquad (9)$$

Where λ_i is the mean of the non-zero response that can be expressed with the explanatory covariates using a natural logarithm link function as defined (7) and p_i is the probability of excess zeros, which can be estimated by the logistic regression

defined as (8). The ZINB model is a special case of a two-class finite mixture model like the ZIP model with mean and variance, $E(Y_i) = (1-p_i)\lambda_i$ and $\text{var}(Y_i) = (1-p_i)(\lambda_i + \frac{\lambda_i^2}{\alpha^{-1}})$ respectively (Zuur, et al, 2009).

Thus the ZINB model is also over-dispersed and allows extra variation relative to the traditional NB model. If $p_i = 0$, the ZINB model reduces to a classical NB regression model. For $\alpha = 0$ the ZINB regression model reduces to ZIP regression model and for $p_i = 0$ and $\alpha = 0$ it reduces to a classical Poisson regression model. For properties and statistical inference, including the maximum likelihood estimation of the parameters for ZIP or ZINB models, we refer Gupta et al. (1996), Lambert (1992), Long (1997) among others. The parameters of the models have been estimated by maximum likelihood estimation method using statistical software R3.0.0

3.5 Model specification test

3.5.1 Over-dispersion Test

Poisson model is a special case of negative binomial model. The negative binomial regression model reduces to the Poisson regression model when the over-dispersion parameter $\alpha \rightarrow 0$. To assess the adequacy of the negative binomial model over the Poisson regression model, we can test the hypothesis:

$$H_o: \alpha = 0 \; vs. \; H_A: \alpha > 0$$

This is to test for the significance of the over-dispersion parameter α. The presence of the over-dispersion parameter α in the NB regression model is justified when the null hypothesis $H_o: \alpha = 0$ is rejected. A likelihood-ratio (LR) tests for the over-dispersion parameter, α, in the negative binomial (NB) specification against the Poisson model specification (Cameron and Trivedi, 1998). In order to test the hypothesis the

likelihood ratio test (LRT) is used. For a general negative binomial regression model, the LRT for α is given by:

$$LRT_{\alpha} = -2[l(\hat{\mu}) - l(\hat{\mu}, \hat{\alpha})]$$

$Where: l(\hat{\mu})\ and\ l(\hat{\mu}, \hat{\alpha})$, are the maximized log-likelihood under the Poisson regression and NB regression models respectively. Standard asymptotic theory suggests that under H_{01} LRT $_{\alpha}$ has probability mass of one half at zero and one half – Chi-square distribution with 1 degree of freedom (Cameron and Trivedi, 1998). We can also use the asymptotic normal Wald type "t" statistics defined as the ratio of the estimate of α to its standard error.

3.5.2 Vuong non-nested test

The Vuong non-nested test is based on a comparison of the predicted probabilities of two models that do not nest. Examples include comparisons of zero-inflated count models with their non-zero-inflated analogs (e.g., zero-inflated Poisson versus ordinary Poisson, or zero-inflated negative-binomial versus ordinary negative-binomial). Under the null that the models are indistinguishable, the test statistic is asymptotically distributed standard normal. Given that $p_1(y_i|x_i)$ and $p_2(y_i|x_i)$ are the predicted probability of the Poisson/NB and ZIP/ZINB, respectively. We want to test the following hypotheses.

H_o= Two distribution functions are equivalent

H_a=$p_1(y_i|x_i)$ is better than $p_2(y_i|x_i)$

H_a=$p_1(y_i|x_i)$ is worse than $p_2(y_i|x_i)$

The Vuong test statistics can be expressed as (Vuong, 1989):

$$V=\frac{\bar{m}\sqrt{n}}{SD(m)}, \qquad m_i = ln\left[\frac{\widehat{p_1}(Y_i/x_i)}{\widehat{p_2}(Y_i/x_i)}\right], \qquad i = 1, 2, \dots, n \quad (10)$$

Where: $\bar{m}$ is the mean of m_i, $SD(m)$ is the standard deviation of m_i and n is the number of observation. $\widehat{p_1}(Y_i/x_i)$ and $\widehat{p_2}(Y_i/x_i)$ are predicted probabilities of the corresponding models $p_1(Y_i/x_i) and \quad p_2(Y_i/x_i)$ respectively.

For large sample size and under the null hypothesis the statistic *V* has the asymptotic standard normal distribution. Note that Shankar et al. (1997), Carson and Mannering (2001) and Lee and Mannering (2002) among others have defined that *V* statistic has a *t* distribution instead of approximate standard normal. This is not a correct statement as *t* statistic is developed based on the assumption that data are from normal distribution. In the context of count data, the parent population is discrete and for large sample size, *V* has asymptotic normal distribution. The critical values of *t* statistic depend on its degrees of freedoms (*df*). For small degrees of freedoms the *t* distribution is leptokurtic. However, as the number of degrees of freedom increases, the *t* distribution approaches the standard normal distribution. Thus to make any decision about the null hypothesis it is reasonable to compare the observed value of the test statistic with the critical value from standard normal distribution.

3.5.3 Likelihood ratio test

The maximum likelihood estimation method is used to assess the adequacy of any two or more than two nested models by using the likelihood ratio test. It compares the maximum likelihood under the alternative hypothesis with the null hypothesis. For instance, the null hypothesis can be the over-dispersion parameter is equal to zero (i.e. the Poisson distribution can be fit the data well) and the alternative hypothesis is that the data would be better fitted by the Negative binomial regression (i.e. the over-dispersion parameter is different from zero). The likelihood ratio test is defined as:

$$\chi^2 = -2(L - L_o) \qquad (11)$$

Where L and L_o are the log likelihood of models under the alternative and null hypotheses. This has a chi-square distribution with degrees of freedom equal to the difference between the degree of freedom of the model under null hypothesis and the alternative hypothesis, respectively. This method is not appropriate for models which are not nested one on the other. In such situation; we will use another method such as the Akakie information criteria (AIC) and Bayesian information criteria (BIC). (Jemain, et al, 2007). In this study a likelihood ratio was used to compare the Poisson with the negative binomial

3.5.4 Information criteria

For comparison of non-nested models based on maximum likelihood, several authors beginning with Akaike (1973) have proposed model selection criteria based on the fitted log-likelihood function. Because we expect the log-likelihood to increase as parameters are added to a model, these criteria penalize models with larger *k*, the number of parameters in the model. This penalty function may also be a function of *n*, the number of observations.

Akaike information criterion: One of the most commonly used information criteria is AIC. The idea of AIC (Akaike, 1973) is to select the model that minimizes the negative likelihood penalized by the number of parameters. The AIC is defined as:

$$AIC=-2\ln\ell+2k \quad (12)$$

With the model with lowest AIC preferred. The term *information criterion* is used because the log-likelihood is closely related to the Kullback-Liebler information criterion. Modifications to AIC include the *Bayesian information criterion* (Cameron and Trivedi, 1998)

$$BIC = -2\ln\ell + (\ln n)k \quad (13)$$

Where ℓ is the logarithm of maximum likelihood estimation for each model, k is number of the model parameters, and n is the number of observation. A model with lower AIC and BIC values is preferred.

3.6 Parameter Estimation

Estimation involves estimating the regression parameters specifically using the maximum likelihood estimation

Maximum likelihood estimation

Software finds model parameter estimates using a numerical algorithm. The algorithm starts at an initial guess for the parameter values that maximize the likelihood function. Successive approximations produced by the algorithm tend to fall closer to the ML estimates. The *Fisher scoring* algorithm for doing this was first proposed by R. A. Fisher for ML fitting of Probit models. For binomial logistic regression and Poisson log linear models, Fisher scoring simplifies to a general-purpose method called the *Newton–Raphson* algorithm. The Newton–Raphson algorithm approximates the log-likelihood function in a neighborhood of the initial guess by a polynomial function that has the shape of a concave (mound-shaped) parabola. It has the same slope and curvature at the initial guess as does the log-likelihood function. It is simple to determine the location of the maximum of this approximating polynomial. That location comprises the second guess for the ML estimates. The algorithm then approximates the log-likelihood function in a neighborhood of the second guess by another concave parabolic function, and the third guess is the location of its maximum. The process is called *iterative*, because the algorithm repeatedly uses the same type of step over and over until there is no further change (in practical terms) in

the location of the maximum. The successive approximations converge rapidly to the ML estimates, often within a few cycles. Each cycle in the Newton–Raphson method represents a type of weighted least squares fitting. This is a generalization of ordinary least squares that accounts for nonconstant variance of Y in GLMs. Observations that occur where the variability is smaller receive greater weight in determining the parameter estimates. The weights change somewhat from cycle to cycle, with revised approximations for the ML estimates and thus for variance estimates. ML estimation for GLMs is sometimes called *iteratively reweighted least squares (Agresti, 2007)*.

The Newton–Raphson method utilizes a matrix, called the *information matrix* that provides standard error values for the parameter estimates. That matrix is based on the curvature of the log likelihood function at the ML estimate. The standard errors are the square roots of the diagonal elements for the inverse of the information matrix. The greater the curvature of the log likelihood, the smaller the standard errors. This is reasonable, since large curvature implies that the log-likelihood drops quickly as β moves away from $\hat{\beta}$; hence, the data would have been much more likely to occur if β took value $\hat{\beta}$ than if it took some value not close to $\hat{\beta}$. Software for GLMs routinely calculates the information matrix and the associated standard errors (Agresti, 2007).

Given independent observations, the log-likelihood of Poisson is

$$l(\beta) = \sum_{i=1}^{n}[Y_i \log(\ _i) - \ _i - \log(Y_i!)]$$

$$= \sum_{i=1}^{n}[Y_i x\ _i\beta - \exp(x\ _i\beta) - \log(Y_i!)],$$

The Poisson MLE is the solution to the first-order conditions

$$\sum_{i=1}^{n}(y_i - \exp(\boldsymbol{x}\ _i\beta))\,\boldsymbol{x_i} = \mathbf{0}$$

The standard method for computation of $\hat{\beta}$ is the Newton-Raphson iterative method. Convergence is guaranteed, because the log-likelihood function is globally concave. In practice often fewer than ten iterations are needed. The maximum likelihood method is used to estimate the parameters of the ZI models. In general speaking, the probability of observing zero counts in ZI models is the sum of the probability of observing an excess zero in the first process (estimated from logistic regression) and the probability of observing a zero in the second process (estimated from a count model) (Rose et al., 2006).

Estimation of dispersion parameter

The dispersion parameter α may be estimated by maximum likelihood or method of moments. Maximum likelihood estimation involves iterative solution of the equation $\partial\ /\partial\alpha = 0$, which is different for each response distribution (Jong and Heller 2008).

3.7 Goodness of fit

In the preceding section the focus was on parameter estimation of the model. Now we consider the overall performance of the model. Common goodness-of-fit measures for GLMs are the Pearson and deviance statistics, which are weighted sums of residuals. These can be used to form pseudo R-squared measures, with those based on deviance statistics preferred. A final measure is comparison of average predicted probabilities of counts with empirical relative frequencies, using a chi-square goodness-of-fit test that controls for estimation error in the regression coefficients.

Pearson Statistic

A standard measure of goodness of fit for any model of y_i with mean μ_i and variance φ_i is the *Pearson statistic*

$$p = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\varphi}_i} \quad (14)$$

where $\hat{\mu}_i$and $\hat{\varphi}_i$ are estimates of μ_i and φ_i. If the mean and variance are correctly specified then E[$\sum_{i=1}^{n}(y_i - \ _i)^2/\varphi_i$] = n, because E[$(y_i - \ _i)^2/\varphi_i$] = 1. In practice *P* is compared with (*n* -*k*), reflecting a degrees of freedom correction due to estimation of $\ _i$. The simplest count application is to the Poisson regression model. This sets $\ _i$ = φ_i, so that

$$P_p = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}, \quad (15)$$

In the GLM literature it is standard to interpret $P_p > n - k$ as evidence of over-dispersion, that is, the true variance exceeds the mean, which implies E[$(y_i - \ _i)^2/\ _i$] >1; $P_p < n - k$ indicates under-dispersion. Note that this interpretation presumes correct specification of μ_i. In fact $P_p \neq$ n-k may instead indicate misspecification of the conditional mean (Cameron and Trivedi ,1998).

Some references to the Pearson statistic suggest that it is asymptotically chi-square distributed, but this is only true in the special case of grouped data with multiple observations for each μ_i. McCullagh (1986) gives the distribution in the more common case of ungrouped data, in which case one needs to account for the dependence of $\hat{\mu}_i$ and $\hat{\beta}$. The Pearson statistics distribution can be obtained by T_p=P'$\hat{V}_p^{-1}$P ~$X^2(1)$, where the formula for the variance V_p is quite cumbersome

The goodness-of -fit or calibration of a model measures how well the model describes the response variable. Assessing goodness of fit involves investigating how close values predicted by the model are to the observed values.

The quality of the fit between the observed values (y) and predicted values ($\hat{\mu}$) can be measured by various test statistics; however, the one of the useful statistic is called deviance and defined as:

$$D(y:\hat{\mu}) = -2[l(\hat{\mu}:y) - l(y:y)] \quad (16)$$

Deviance is useful to see whether additional explanatory variables improve the fit significantly or not. To do this we should for each model have the resulting deviance. The deviance difference from one fitted model to an extended model is approximately χ^2 -distributed with degree of freedom which equals the number of additional free regression parameters. For a better model, one would expect smaller value of the $deviance, D(y:\hat{\mu})$, (McCullah and Nelder, 1987 and Agresti, 2007).

Test of the overall goodness of fit: is used to assess the overall goodness fit of the model. The likelihood ratio test looks at the model chi-square (chi-square difference) by subtracting deviance (-2ℓ) for the final (full) model from deviance for the intercept-only model. The degrees of freedom in this test equal the number of terms in the model minus one (for the constant). This is the same as the difference in the number of terms between the two models, since the null model has only one term. Model chi-square measures the improvement in fit that the explanatory variables make compared to the null model. The likelihood ratio test is thus a test of the overall model. The overall test statistic for likelihood ratio test is given as:

$$Likelihood\ ratio\ test = G^2 = -2(\ell_{\text{null}} - \ell_{\text{k}}) \quad (17)$$

Where: ℓ_{null} is the log-likelihood of the null model and ℓ_{k} is the log-likelihood of the model comprising k predictors.

Under the global null hypothesis,

H_o: $all\ parametrs\ in\ the\ model\ are\ equal\ to\ zero$ the likelihood ratio test statistic, G^2 , follows a chi-square distribution with p degrees of freedom.

Test for individual predictors

Let β denote an arbitrary parameter. Consider a significance test of H_o:β=β_o(under H_o β_o =0) The simplest test statistic uses the large-sample normality of the ML estimator 0 Let *SE* denote the standard error of $\hat{\beta}$, evaluated by substituting the ML estimate for the unknown parameter in the expression for the true standard error. When H_o is true the test statistics

$$\boldsymbol{Z} = \frac{\hat{\beta} - \beta_o}{\widehat{SE}}$$

has approximately a standard normal distribution. Equivalently, Z^2 has approximately a chi-squared distribution with *df* = 1. This type of statistic, which uses the standard error evaluated at the ML estimate, is called a *Wald statistic*. The z or chi-squared test using this test statistic is called a *Wald test*. The significance test for each coefficient in the model is done using Wald chi-square, the Wald statistic (w) is:

$$\mathrm{W}=Z^2=\left(\frac{\hat{\beta}}{\widehat{SE}}\right)^2 \qquad (18)$$

The Wald statistic under the null hypothesis is approximately chi-square distributed with 1 degree of freedom. Wald statistics are easy to calculate but their reliability is questionable, particularly for small samples. For data that produce large estimates of the coefficient, the standard error is often inflated, resulting in a lower value of the Wald statistic, and therefore the explanatory variable may be incorrectly assumed to

be unimportant in the model. Likelihood ratio tests are generally considered to be superior (Agresti, 2007).

3.8 Software

The data are analyzed using the Statistical Packages for Social Sciences (SPSS) version 20 and R (version 3.0.0). In addition all hypotheses were tested at 0.05 level of significance.

Chapter 4; ANALYSIS AND RESULTS

4.1 Descriptive Statistics

It is always a good idea to start with descriptive statistics and plots. Thus we start with the description of the variables presented in a cross tabulation table.

4.1.1 Number of fatal per traffic accident

Table 4.1 shows frequency and percentage distribution of number fatalities per traffic accident in Addis Ababa based on information from 3880 traffic accidents recorded. It can be seen that 68.8% of the traffic accidents have not involved any fatalities, whereas 15.5%, 10.3% and 2.4% of traffic accidents involve 1, 2 and 3 fatalities per accident, respectively. Moreover 3% of the traffic accidents involve at least four fatalities per accident. The results also indicated that the maximum frequency of number of fatalities per accident recorded was 9. From the summary in Table 4.1, it can be observed that the sample mean of the response variable was .57 while the sample variance was 1.072. If the mean is smaller than the variance, it suggests a case of over-dispersion. The data has excess zeros and thus one might expect that both ZIP and ZINB would possibly be better models to predict the number of fatalities per traffic accident. If we observe the overall pattern of the number fatalities per traffic accident, it is highly skewed to the right with excess zeroes (see figure 4.1).

Table 4.1 Frequency and percentage distribution of number of fatal per accident

No. of fatal per traffic accident	Frequency	Percent
0	2668	68.8
1	603	15.5
2	398	10.3
3	94	2.4
≥4	117	3
Total	3880	100.0
Minimum	0	
Maximum	9	
Mean	.57	
Variance	1.072	
Skewness	2.274	

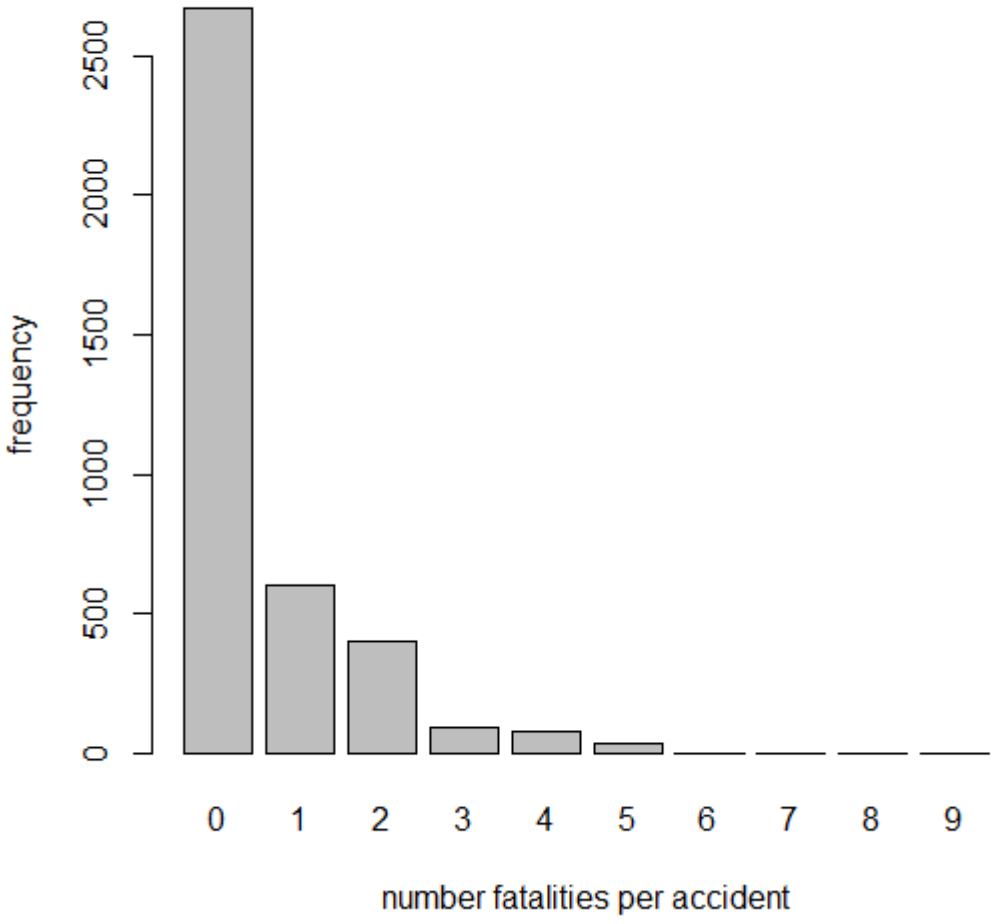

Figure4.1. Bar-graph of number of fatalities per traffic accident

In order to compare the impact of each variable on the number of fatalities per accident, the frequency and percentage distribution number of fatalities with respect to the levels of each variable was calculated. SPSS results of such values are shown in Appendix B.

The driver‘s age had three categories. Among these categories, drivers within the age group 18-30 were responsible for the large number of fatalities (30.83%) and for the large number of accidents (52.91%). Among the three categories of driver‘s experience, those drivers with less than 5 years‘ experience were responsible for the largest share of fatalities (39.26%) and also for the large number of accidents (59.41%). Accident involvement also differs by gender. With regards to the number of fatalities and the number of accidents, male drivers take the major share i.e. above 35.9% while the share of female drivers only 13.1%.

Lighting conditions were important. As mentioned in the Appendix B, most accidents occurred in day time. Table A2 showed that 34.49% of fatal accidents occurred in daylight. About 96.24% of traffic accidents occurred on asphalt roads and only 3.76% on gravel. Midblock road sections had a considerable share in fatal and non-fatal accidents, probably because much pedestrian crossing takes place in these sections. Overall, 31.28% of fatalities and 68.72% of non-fatal accident occurred on midblock road sections as shown in Table A3. Accident fatalities are also analyzed in terms of vehicle type. Among the seven categories of vehicle type, Automobiles and Taxies are responsible for the largest number of fatalities with values 32.88% and 30.7%, respectively.

4.2 Model development and Selection

The fitted over-dispersion parameter (α) is tested to check whether it is significant and if found significant, the negative binomial model is the immediate solution to accommodate the observed over-dispersion. The Negative Binomial distribution was used to correct the error of over-dispersion in the data in situations where the result of the Poisson regression model shows over dispersion. The various models obtained when the number of fatalities per accident was regressed on different factors such as human factor, environmental factor, road factor and vehicular factor were involved in the fatalities per traffic accident model. Sometimes when analyzing a count response variable, the number of zeros may be excessive. When analyzing a dataset with an excessive number of zero outcomes, a zero-inflated model should be considered.

In this study we apply four count approaches to model the number of fatalities per accident. These models are the standard Poisson, negative binomial (NB), zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) models. The modeling approach is based on a stepwise backward procedure by starting with all variables and removing non-significant variables with *p*-value more than 0.05 at each step. For comparing the Poisson versus NB and ZIP versus ZINB, since these models are nested together, we test if there is over-dispersion due to heterogeneity by testing the significance of dispersion parameter. A significant value for dispersion parameter indicates that the over-dispersion in the fatal per traffic accident data is due to unobserved heterogeneity, which implies the NB and ZINB models are plausible models of the Poisson and ZIP models, respectively. However, an extra test is needed to check whether excess zero counts may be the cause of over-dispersion or not. To do this, since the Poisson and NB are not nested within the ZIP and ZINB models, respectively, we use the Vuong test to compare the non-nested models.

4.2.1 Goodness of fit and test over-dispersion

Table 4.6 shows that the results of the Poisson, NB, ZIP and ZINB model fit statistics. As shown in the summary table, the likelihood-ratio chi-square values for all models were found to be significant. Thus, all regression models are significant. Choosing between Poisson and negative binomial models depends on the nature of the distribution of the dependent variable. Measuring the distribution of count data was a fairly straightforward process. Particularly, deviance goodness-of-fit tests can be incorporated along with exploratory negative binomial regression models to measure the distribution of the dependent variable. This simple test identifies the distribution of the data and ensures the selection of the correct statistical model.

As mentioned before, NB model is a superior model for dealing with over-dispersion due to unobserved heterogeneity; the ZIP model is a mixture Poisson model for handling over-dispersion arising from excess zero; and ZINB is a model for handling over-dispersion resulting from both excess zero and unobserved heterogeneity in the fatal traffic accident data. We summarized the results of considered models selection criteria in Table 4.2. The results show that, in the NB and ZINB model, the dispersion parameter α, was statistically significant. This implies there was an over-dispersion in the accident data. To confirm this, we compared the Poisson versus ZIP and the NB versus ZINB model by Vuong test. The Vuong statistic for the Poisson versus ZIP (-18.61738, *p*-value = 1.16162e-7) favors the ZIP model, while the Vuong statistic for the NB versus ZINB (-15.61731, *p*-value = 2.77e-15) favors the ZINB model. Thus we might select the ZINB model. The AIC and BIC were also supported for the ZINB model to fit the fatal accident. The comparison for the Poisson model revealed that the Poisson model was not an appropriate model for predicting the number fatalities per traffic accident with over-dispersed and excess zero counts. On the other hand, NB

model was also not an appropriate model for excess zero counts. Next we compare all developed models using AIC and BIC. These information-based criteria favor the ZINB model over all other models. As a result, we select the ZINB model as the best among the four models and tabulate its results in Table 4.7 and Table 4.8. Results for the Poisson, NB, and ZIP were not presented, since the ZINB model was found to be preferred over the Poisson, NB, and ZIP models.

Table 4.2 Test for goodness of fit and Model Selection Summary Results of Poisson, NB, ZIP and ZINB

Criteria	Poisson	ZIP	NB	ZINB
Log-likelihood (ℓ)	-4353.014	-3331.235	-3620.135	-4391.033
-2 ℓ	8706.028	11379.98	7240.27	8782.066
Likelihood Ratio (LR) Chi-Square test	165.21(0.000)	306.54(0.000)	597.29(0.000)	584.4(0.000)
AIC	8760.028	7770.471	7296.271	6772.471
BIC	8936.669	8910.345	7479.454	7258.102
Over-dispersion(α)	-	-	2.35	3.055
αp-value	-	-	2.2e-16*	0.000292 *
Vuong statistic	-18.61738(1.16162e-07)		-15.61731(2.77e-15)*	

Note: NB = negative binomial, ZIP = zero-inflated Poisson model, ZINB = zero-inflated negative binomial model, ―p-value" = corresponding p-value for the dispersion parameter for the NB and ZINB models, The P-values of the Vuong test of ZINB vs. standard Negative binomial is presented within parenthesis

4.3 Interpretations of Negative Binomial Part of the ZINB Model

Table 4.3 showed the impact of human, environmental, road and vehicle related factors on the number of fatalities per accident. As shown in Table 4.4 driver's age had significant impact on the number of fatalities per accident. The expected number of fatalities per traffic accident for drivers in the age group 31-50 years had decreased by 22.6% as compared to the expected number of fatalities per accident for drivers in the age group 18-30 while holding all other variables in the model constant. Also the expected number of fatalities per accident for driver's age group above 51 years was decreased by19.1% as compared to the expected number of fatalities per accident for

drivers in the age group 18-30 while holding all other variables in the model constant. The model also revealed that driver experience had statistically significant impact on the number of fatal traffic accidents. The expected number fatalities per accident for driver experience between 5-10 years and above 10 years were 44.5% and 41.5% less as compared to those with less than five years of driving experience, respectively, controlling for the other variables in the model.

The finding of this study also revealed that the driver vehicle relation had a significant impact on the number of fatalities per accident. The expected number of fatalities per traffic accident for employed driver was increased by 32% as compared to those drivers who own the vehicle while holding all other variables in the model constant. The expected number of fatalities per accident on Monday, Tuesday and Thursday were increased by 59.0%, 31.0% and 30.3% respectively, as compared to the expected number of fatalities per accident on Sunday.

Furthermore, the expected number of fatalities per accident for automobile was increased by 23.6% as compared to the expected number of fatalities per accident of taxi-minibuses. Also the expected number of fatalities per accident for Cargo(10-40qtl) was decreased by 24.6% as compared to the expected number of fatalities per accident of taxi-minibuses, while other vehicle types were not significant. The results showed that the expected number fatalities per accident for vehicles between 5-10 years‘ service was decreased by 23.1% as compared to those vehicles with less than five years‘ service. Table 4.3 also showed that the expected number of fatalities per accident for vehicle service above 10 years was increased by 1.58% as compared to the expected number of fatalities per accident for vehicles with less than 5 years‘ service while holding all other variables in the model constant. Finally road class has insignificant effect on the number of fatalities per accidents. The expected number of

fatalities per accident in a dual carriage way was significantly different from one way road, while the other levels were not significant

Table 4.3 Parameter estimates of NB part of ZINB regression model

Count model coefficients(negbin with log link)					
	Coef	Exp(Coef)	Std. Er	zvalue	Pr(>|z|)
(Intercept)	0.366	1.442	0.261	1.403	0.160473
18-30 years(ref)					
31-50 years	-0.26	0.774	0.088	-2.908	0.003637 **
51 and above	-0.21	0.809	0.063	-3.354	0.000797 ***
0-5 years(ref)					
5-10 years	-0.59	0.555	0.233	-2.525	0.011586 *
10 and above	-0.54	0.585	0.101	-5.293	1.2e-07 ***
Owner (ref)					
Employee	0.284	1.329	0.124	2.298	0.021570 *
Other	-0.31	0.736	0.13	-2.356	0.018467 *
Unknown	-0.52	0.592	0.202	-2.596	0.009420 **
sunday(ref)					
Monday	0.464	1.59	0.142	3.264	0.001097 **
Tuesday	0.27	1.31	0.096	2.810	0.004948 **
Wednesday	0.131	1.14	0.099	1.323	0.185787
Thursday	0.265	1.303	0.093	2.835	0.004587 **
Friday	0.126	1.134	0.095	1.324	0.185456
Saturday	0.127	1.136	0.093	1.364	0.172621
0-5 years(ref)					
5-10 years	-0.26	0.769	0.08	-3.272	0.001068 **
10 and above	0.016	1.016	0.068	0.230	0.818470
Taxi-minibuses(ref)					
Automobile	0.212	1.236	0.063	3.354	0.000797 ***
Cargo(10qtl)	-0.28	0.754	0.122	-2.306	0.021089 *
Cargo(1-40qtl)	-0.11	0.897	0.095	-1.140	0.254362
Cargo(41-100qtl)	-0.22	0.799	0.154	-1.462	0.143868
Buses(13-27seats)	0.027	1.027	0.118	0.226	0.821163
Buses (46 seats)	0.029	1.029	0.104	0.280	0.779705
One-way (ref)					
Two-way	0.122	1.13	0.216	0.565	0.571828
Double courage Way	0.921	2.513	0.271	3.397	0.000681 ***
Ring Road	0.276	1.317	0.269	1.024	0.305991
Two way(divided by solid line)	0.016	1.016	0.247	0.065	0.948215
Two way(divided by broken line)	0.075	1.078	0.212	0.353	0.723967

Alpha	3.055	21.21	0.843	3.623	0.000292 ***

Note: -Signif. codes:‘***’0.001 ‘**’0.01 ‘*’0.05 ‘.’0.1 ‘ ’1

4.4 Interpretations of Zero-inflation part of The Model

Zero inflated models are interpreted as a mix of structural and sampling zeros from two processes; the process that generates excess zeros from a binary distribution which are the structural zeros, and the process that generates both non-negative and zero counts from NB distributions which are the sampling zeros.

The results of Table 4.4 below indicated the parameter estimates of the Zero-Inflated (logit model) part of the ZINB regression model for examining the impact of explanatory variable on the odds of being in the always zero group. Table 4.4 showed that drivers experience has a significant impact on the odds of being in the always zero group. The odds of being in the always zero group decreased by a factor 0.809 for drivers having 5-10 years of experience as compared to the those drivers with less than five years of experience holding all other variables in the model constant. In other words, the more the number of years of experience, the less likely the driver is involved in the always zero group.

As shown in Table 4.4, the age of drivers had a significant negative impact on the odds of being in the always zero group. The odds of being in the always zero group decreased by 14.2 % for drivers with age group 31-50 as compared to those drivers in the age group 18-30. Additionally, the odds of being in the always zero group decreased by 19% for the drivers in the ≥51 age group as compared with those drivers in the age group 18-30 controlling for other variables in the model. In other words, as the age of drivers decrease, the more likely they involved in excess zero groups. The findings of this study show that the driver vehicle relationship has a positive

significant impact on the odds of being in the always zero group. The odds of being in the always zero group increased by 59.5% for employed drivers as compared with owner driver controlling for other variables in the model constant.

Table 4.4 also showed that the day of the weeks had significant impact on the odds of being in the always zero groups. For instance, as compared to the sunday, the odds of being in the always zero group increased by 36.4%, 49.5%, 35.4% and 44.3% for Monday, Tuesday, Thursday and Saturday respectively. The analysis also revealed that vehicle service year was one of the main determinants of the frequency of number of fatalities per traffic accident. The odds of being in the always zero group decrease by 39.9% and 57.3% for vehicle service year 5-10 years and above 10 years respectively, as compared to those drivers with less than five years of experience holding all other variables in the model constant. The odds of being in the always zero group increased by 31% for automobile as compared to taxi-minibus.

Table 4.4 Parameter estimates of Zero-Inflation part of ZINB regression model

Zeroinflation model coefficients (binomial with logit link)					
	Coef	Exp(coef)	Std. Er	z value	Pr(>|z|)
(Intercept)	0.989	2.689	0.369	2.677	0.00742 **
18-30 years(ref)					
31-50 years	-0.15	0.858	0.06	-2.551	0.010738 *
51 and above	-0.3	0.739	0.068	-4.420	9.85e-06 ***
0-5 years(ref)					
5-10 years	-0.21	0.809	0.063	-3.354	0.000797 ***
10 and above	0.256	1.292	0.088	2.908	0.003637 **
Owner (ref)					
Employee	0.467	1.595	0.186	2.510	0.01206 *
Other	0.583	1.792	0.194	3.013	0.00259 **
Unknown	0.276	1.317	0.342	0.807	0.41976
sunday(ref)					
Monday	0.31	1.364	0.119	2.598	0.00937 **
Tuesday	0.402	1.495	0.124	3.244	0.00118 **
Wednesday	0.17	1.185	0.127	1.340	0.18017
Thursday	0.303	1.354	0.12	2.525	0.01156 *
Friday	0.149	1.161	0.12	1.242	0.21423
Saturday	0.367	1.443	0.114	3.215	0.00130 **
0-5 years(ref)					
5-10 years	-0.51	0.601	0.123	-4.148	3.36e-05 ***
10 and above	-0.85	0.427	0.149	-5.721	1.06e-08 ***
Taximinibuses(ref)					
Automobile	0.271	1.311	0.124	2.179	0.02931 *
Cargo(10qtl)	0.458	1.58	0.187	2.452	0.01422 *
Cargo(1-40qtl)	-0.01	0.989	0.151	-0.071	0.94305
Cargo(41-100qtl)	0.477	1.611	0.145	3.289	0.00101 **
Buses(13-27seats)	0.083	1.086	0.165	0.499	0.61784
Buses (46 seats)	0.291	1.338	0.132	2.213	0.02687 *
One-way (ref)					
Two-way	0.572	1.772	0.305	1.873	0.06112 .
Double courage Way	-0.21	0.809	0.063	-3.354	0.000797***
Ring Road	0.365	1.44	0.412	0.885	0.37600
Two way(divided by solid line)	0.575	1.777	0.348	1.650	0.09900 .
Two way(divided by broken line)	0.522	1.685	0.3	1.740	0.08191 .

Note: -Signif. codes:'***'0.001 '**'0.01 '*'0.05 '.'0.1 ' '1
-see Appendix A for coding of variable

CHAPTER 5; DISCUSSION CONCLUSIONS AND RECOMMENDATION

5.1 Discussion

In this study, the mean number of fatalities per accident was lower than the variance. This might occur due to an excess of zeros and high variation in the non-zero outcomes. The best model was selected from different possible models namely: Poisson, NB, ZIP and ZINB. The goodness of fit was examined by using the likelihood ratio test (LR) and the comparison was conducted by Akakie information criteria (AIC) and Bayesian information criterion (BIC) and Vuong test. Non-nested models were compared by Vuong test. Of these the ZINB regression model was selected as the best model. Since the data had excess zeros, the standard Poisson and negative binomial regression models were not appropriate. This was due to the fact that the number of zeros in the data was beyond the model could predict.

While the actions of people might be influenced by subconscious motives and subliminal cues, they are also the most adaptive elements in the traffic system. They can create risk situations as well as respond to ever changing new demands of the traffic environment. The findings of the study showed that accident fatalities of young drivers were more than the rest of the categories. Drivers who have less than 5 years of experience cause more fatalities as compared to those who have more driving experience. In case of driver vehicle relationship, employed drivers were responsiple for large number of fatalities per accident. This result is consistent with Towelde(2007), Bisrat (2010) and Kweon and Kockelman (2003).

Vehicle type was found to be an important factor which affects the number fatalities per accidents. Among vehicle type automobile pose significantly greater fatalities per

accident as compared to other vehicle type. This result is consistent with the study conducted by Toweled (2007)

Zero inflated part of ZINB indicated that drivers experience had a significant impact on the odds of being in the always zero groups. The odds of being in the always zero group for driver with less than 5 years of experience were higher as compared to other categories. Also the age of drivers has a significant negative impact on the odds of being in the always zero groups. The findings of this study also show that among the category of driver vehicle relationship, employed drivers were high in the odds of being in the always zero groups.

The analysis also reveals that the vehicle service year had a negative significant influence on the odds of being in the always zero group. The odds of being in the always zero groups for vehicle having 5-10 and above 10 years of service were higher than the vehicle which have less than 5 years of service. Finally regarding the zero-inflation model, the result shows that, the vehicle type had a significant impact on the odds of being in the always zero groups. The odds of being in the always zero groups for automobile would be greater when compared to the other categories.

5.2 Conclusions

In this study, traffic accident data in Addis Ababa taken from AATCID was analyzed using Poisson, NB, ZIP and ZINB Models to determine the main factors which were highly related to fatalities per accident. More than 60% of the data involve zero number of fatalities per accident. From the study ZINB model is found to be better for zero inflated and over dispersed data.

The driver in the age group 18-30 years, the drivers having less than 5 years of experience and those who are employed were highly associated with the number of

fatalities per accident. When the day of the week is considered Monday, Tuesday and Thursday are found to be the days where higher number fatalities per accident were expected. Among vehicle related factors automobile, taxi-minibuses and vehicles of less than five years of service were highly associated with the number of fatalities.

5.3 Recommendations

Based on the result of the study we recommend the following;

- Education on road accidents should be intensified especially among the young driver
- Since the type of vehicle involved in the accident affects the number of people fatalities, drivers of vehicles such as automobile and taxi-minibus should be given special training to be able to avoid preventable accidents.

Reference

- Abdel. H .A (2005). Stats Methods of Predicting Fatalities of Road Traffic Accidents in Kuwait. –Egyptian Population and Family Review. pp. 2-19.
- AfDB, OECD, UNDP, & UNECA. (2012). *Economic Outlook: Ethiopia*. Retrieved from http://www.afdb.org/ fileadmin/uploads/afdb/Documents/ Publications/Ethiopia%20Full%20PDF%20Country%20Note.pdf
- Agresti, A. (2002). *Categorical Data Analysis.* New York: John Wiley & Sons.
- Agresti, A. (2007). *Categorical Data Analysis.* New York: John Wiley & Sons.
- Akloweg, Y., Hayshi, Y., & Kato, H. (2011). The Effect of Used Cars on African Road Traffic Accidents: a Case Study of Addis Ababa, Ethiopia. *International Journal of Urban Sciences, **15*(1)**, 61-69.
- Ascone, D., & Lindsey, T. (2009). *An Examination of Driver Distraction as Recorded in NHTSA Databases.* Washington DC: NHTSA.
- Asfaw, M. (1999): *Urban Mobility- Challenges and Prospects, The case of Addis Ababa, Ethiopia.* Retrieved December 8, 2006 from the World Wide Web www.Bremen-Initiative.De/Lib/Papers /Addis.Pdf.
- Atubi, A.O. (2009) –Urban Transportation: An Appraisal of Features and Problems in the Nigerian Society". *International Journal of Geography and Regional Planning*. Vol. 1, No. 1, Pp. 58-62.
- Bedard, M., Guyatt, G. H., Stones, M. J., & Hirdes, J. P. (2002). The independent contribution of driver, crash, and vehicle characteristics to driver fatalities. *Accident Analysis & Prevention*, 717-727.
- Bisrat, M.(2010). Determinants of Traffic Fatalities and Injuries in Addis Ababa. Unpublished Msc thesis, Department of Statistics, Addis Ababa University.

- Bonate, L. P. , Sung, C. and Richards, S.(2009).Conditional modelling of antibody titters using a zero-inflated Poisson random effects model: application to Fabrazyme® , . *J. Pharmacokinet Pharmacodyn,* **36**:443–459.
- Breslow NE, Day NE. (1987) Statistical Methods in Cancer Research, Vol. 2, The Design and Analysis of Cohort Studies (IARC Scientific Publications No. 82), Lyon (Fr), IARC.
- Brüde, U. and Larsson, J. (1993). ―Models for predicting accidents at junctions where pedestrians and cyclists are involved", *Accident Analysis and Prevention*, Vol. 25, Issue 5, pp. 499-519.
- Cameron A.C. and Trivedi P.K. (1998). *Regression Analysis of Count Data*, New York: Cambridge University Press.
- Cameron A.C. and Trivedi P.K. (2005). *Regression Analysis of Count Data*, New York: Cambridge University Press.
- Cejun, Liu and Chou – Lin Chen (2004). Time Series Analysis and Forecast of Annual Crash Fatalities in china. ―National Centre for Statistics and Analysis".
- Dinesh M. (1985) An analysis of road traffic fatalities in Delhi, India. *Accident Analysis and Prevention*. **17**:33-45.
- Elhai , J. D., Calhoun, P.S., and Ford, J.D.(2008).Statistical procedures for analyzing mental health services data. *J. Psychiatry Research*, Elsevier, **160**: 129-136.
- Elizabeth Kopits and Maureen Cropper, "Traffic Fatalities and Economic Growth," The World Bank, *Policy Research Working Paper* No. 3035 (Washington, DC: World Bank, 2003).

- Emenalo M., Pusteli A., Ciampi, and Joshi H.P. (1987). Analysis of road accidents data in Zambia. *Traffic Engineering and Control*. **28**:635-640. Engineering. Imperial College of Science, Technology and Medicine. London.
- Ericson, M., & Kim, P. (2011). How Road Traffic Injuries Affect Household Welfare in Cambodia Using the Millennium Development Goals Benchmarks. *Asian Studies Review, 35*(2), 209-234. Retrieved from http://dx.doi.org/10.1080/10357823.2011.575209.
- Evans, L. (2003). Estimating Accident Fatalities in Britain. Crowthonne, U.
- Gardner H. (1995). Reflections on multiple intelligences. Phi Delta Kappan, 77 200-Generalized Linear Models (with J.A. Nelder),(1983), Chapman and Hall, London.
- Ghee C., Silcock D., Astrop A., and Jacobs G. (1997). Socio-economic aspects of road accidents in developing countries, *Transport Research Laboratory Report* **247**:29. *injuries in developing countries*, **324**:113-120.
- Greene, WH. "*Accounting for excess zeroes and sample selection in the Poisson and negative binomial regression models*", New York University, School of Business, Department of Economics, 1994.
- Jha, N., D.K. Srinivasa, G. Roy and S. Jagdish (2004). Epidemiological Study of road traffic accident cases: A study from South India. India J. Community Med.**29**: 20-24.
- Kardara, C., and Kondakis (1997). Road Safety Situation in Greece. "Traffic Engineering and Control. 6:509-12.
- Kweon, Y.-J., and Kockelman, K. M., 2003, "Overall injury risk to different drivers: combining exposure, frequency, and severity models", Accident Analysis and Prevention, **35**, 441

- Lambert, D.L. (1992). –Zero-inflated Poisson regression, with an application to defects in manufacturing",*Technometrics*, Vol. 34, Issue 1, pp. 1-14.
- Lascala, M., Jadaan Kis and Homel R. (1996). Epidermology of Transportation Related Injuries in a Subsahara African.
- Lauren, P. and Hill, S. (2005). Road traffic Injuries-Can we a avoid global epidemic? Retrieve from the World Wide Web http://www.thedoctorwillseeeyounow.com/articles/other/road-33/.
- Lee, C. and Abdel-Aty, M.A. (2005). –Comprehensive analysis of vehicle-pedestrian crashes at intersections in Florida", *Accident Analysis and Prevention*, Vol. 37, Issue 4, pp. 775-786.
- Lee, J., and Mannering, F., 2002, –Impact of roadside features on the frequency and severity of run-off-roadway accidents: an empirical analysis", Accident Analysis and Prevention, **34**, 149
- Long, J. S., and Freese, J. (2006). *Regression models for categorical dependent*
- Lord, D. (2006) Modeling motor vehicle crashes using Poisson-gamma models: examining the effects of low sample mean values and small sample size on the Estimation of the fixed dispersion parameter. *Accident Analysis and Prevention* 38 (4): 751-766.
- Lum, H., & Reagan, J. A. (1995). Interactive Highway Design Model: Accident Predictive Module. *Public Roads Magazine*, 14-17.
- Lyon, C., Persaud, B., 2003. Pedestrian collision prediction models for urban intersections. Transport. Res. Rec. 1818, 102–107.
- McCullagh P., and Nelder J.A. (1989). *Generalized Linear Models* (Second edn). New York: Chapman and Hall.

- Mekky A. (1985). Effects of rapid motorization levels on road fatalities in some rich developing countries. *Accident Analysis and Prevention.* **17**:101-109.
- Miaou S.P. (1994). The relation between track accident and geometric design of road section: Poisson vesrus negative binomial regression. *Accident Analysis and Prevention*. **26(4)** :471-482.
- Murad,M. (2011) Costing road traffic accidents in Ethiopia. Department of Civil Engineer, Addis Ababa University
- Murray, C. J. L., Lozano, T. V. R., Naghavi, M., (2012). Disability-adjusted Life Years (DALYs) for 291 Diseases and Injuries in 21 Regions, 1990–2010: A Systematic Analysis for the Global Burden of Disease Study 2010. *The Lancet, 380*(December 15/22/29,2012), 2197-2224. Retrieved from http://www.thelancet.com/.
- Nantulya, V. M. and M. R. Reich (2002). The neglected epidemic: *Road traffic*
- NHTSA. (2010). *Fatality Analysis Reporting System*. Retrieved from National Highway Traffic Safety Administration: http://www.nhtsa.gov/
- P. de Jong and G. Z. Heller.(2008). *Generalized linear model for insurance data*, Cambridge University, New York
- Pocock, R.J., Pustelli, A and Joshi, H. P. (1981). Some Statistical aspects of road safety. "*Journal of Royal Society*" **12**: pp.1-24.
- Pramada V.P., and Sarkar P.K. (1993). Variation in the pattern of road accidents in different states and union territories in India. Proceedings of the third national conference on transportation systems studies: *Analysis and Policy*. 1X-5 to 1X-9.
- Robert B. (2000): *Traffic Fatalities And Injuries - Are Reductions The Result Of 'Improvements' InHighway Design Standards?* Centre for Transport Studies Dept. of Civil and Environmental

- Rose CE., Martin SW., Wannemuehler KA. *and* Plikaytis BD. *(*2006*).* "On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data", *Journal of Biopharmaceutical Statistics*, Vol. 16, pp. 463-81.
- Shankar, V., Mannering, F., and Barfield, W., 1996, "Statistical analysis of accident severity on rural freeways", Accident Analysis and Prevention, **28**, 391
- Shankar, V., Milton, J., and Mannering, F., 1997, "Modeling accident frequencies as zero-altered probability processes: an empirical inquiry", Accident Analysis and Prevention, **29**, 829
- Smeed R.J. (1968). Variation in the pattern of accident rates in different countries and their causes. Traffic Engineering and Control. 10:364-371.
- Sturman, M. C. (1999). Multiple approaches to analyzing count data in studies of individual differences: The propensity for type I errors, illustrated with the case of absenteeism prediction. *Educational and Psychological Measurement*, **59**:414–430.
- Tesema, T. and Tibebe, B. (2005). Rule Mining and Classification of Road Accidents Using Adaptive Regression Trees. *International Journal of Simulation Systems Science &Technology Special Issue on Soft Computing for Modeling and Simulation* **6**: 10-11.
- Tewolde, M. (2007). Analysis on traffic accidents involving human injuries in the case of Addis Ababa. (Unpublished M.Sc. Thesis), Department of Statistics, Addis Ababa University.
- Tsauo, P. Elvik, R., Lopez, A. (1996). Research on road accidents in Taiwan. "*Traffic engineering and Control*".**18**: 166-170.

- Ulfarsson, G. F., 2001, "Injury severity analysis for car, pickup, sport utility vehicle and minivan drivers: male and female differences", Ph.D. Dissertation, University of Washington, UMI Dissertation Publishing
- Ulfarsson, G. F., and Mannering, F. L., 2004, "Differences in male and female injury severities in sport-utility vehicle, minivan, pickup and passenger car accidents", Accident Analysis and Prevention, **36**, 135
- Vogt, Andrew and Bared, Joe G. *Accident Models for Two-lane Rural Roads: Segments and Intersections*. FHWA-RD-98-133, October 1998.
- Vuong, QH. (1989). "Likelihood ratio tests for model selection and non-nested hypotheses", *Econometrica*, Vol. 57, Issue 2, pp. 307–333.
- Welsh, AH., Cunningham, RB., Donnelly, CF. and Lindenmayer, DB. (1996). "Modelling the abundance of rare species: statistical models for counts with extra zeros", *Ecological Modelling*, Vol. 88, pp. 297-308.
- WHO (2004). World Report on Road Traffic Injury Prevention: Summary.
- WHO (2008). World Report on Road Traffic Injury Prevention: Summary.
- WHO, World Report on Road Traffic Injury Prevention, 2002 WHO statistic
- WHO. (2009). *Global status report on road safety: Time for action. Geneva: World Health Organisation*. Retrieved Feb 24, 2011, from www.who.int/violence_injury_prevention/road_safety_status/2009
- Young, K., & Regan, M. (2007). Driver Distraction: A Review of the Literature. *Australasian College of Road Safety*, 379-405.

APPENDICES

Appendix A: Coding and Description of the variable

Table A1 coding and description of human related variable

Human related variable			
No	Variable	Levels	Description of variable
1	GD	(1)Male (2)Female	**Gender of driver**
2	AD	(1)18-30 (2) 31-51 (3)51 and above	Age of driver
3	DE	(1) 0-5 years (2) 5-10 years (3) above 10 years	Driving Experience
4	DVR	(1)Owner (2)employee (3)Other (4) unknown	Driver-vehicle relationship

Table A2 coding and description of Environment related variable

Environment related variable			
No	Variable	Levels	Descriptions of variable
5	LC	(1) Dark night (2) Night without SL (3) Night with SL (4) Day light	Light condition
6	DTW	(1)Sunday (2)Monday (3)Tuesday (4)Wednesday (5)Thursday (6)Friday (7)Saturday	Day of the weeks

Table A3 coding and description of road related variable

Road way related variable			
No	Variable	Levels	Discriptions of variable
7	TP	(1)gravel road (2)Asphalt	Type of pavement
8	JT	(1) midblock (2) Y-intersection (3)T-intersection (4) Round about (5)Cross(four leg jun) (6)five leg junction	Junction Type
9	RG	(1)Curved Road (2)Gradient (3)Straight (4)Tangent	Road Geometry
10	RCL	(1) One-way (2)Two-way (3)Double courage Way (4) Ring Road (5) Two way(divided by solid line) (6) Two way(divided by broken line)	Road type
11	RC	(1)Wet (2)Dry	Road condition

Table A4 coding and description of vehicle related variable

Vehicle related variable			
No	Variable	Levels	Description of variable
12	VT	(1)Taxi-minibuses(up to 12 Seats) (2)Automobile (3)Cargo (upto 10 quintals) (4)Cargo(11-40 quintals) (5)(41-100 quintals) (6)Buses (13-45 seats) (7)Buses(more than 46 seats)	Vehicle type
13	VSY	(1) 0-5 years (2) 5-10 years (3) above 10 years	Vehicle service year

Appendix B; Frequency and percentage distribution of different factor related to the existence of fatalities

Table B1: frequency and percentage distribution of human factor related to existence of fatalities

Explanatory variables	Categories		Existence of fatalities		Total accident	
			No fatal	≥1	Freq.	%
Driver Age	18-30 years	Freq.	1420	633	2053	52.91
		%	69.18	30.82		
	31—50 years	Freq	1010	473	1483	38.22
		%	68.1	31.9		
	More than 51 years	Freq	238	106	344	8.87
		%	69.2	30.8		
Driver's Experience	0-5 years	Freq.	1400	905	2305	59.41
		%	60.74	39.26		
	5-10 years	Freq.	900	300	1200	30.93
		%	75	25		
	Above 10 years	Freq.	300	75	375	9.66
		%	80	20		
Driver Gender	Male	Freq.	2085	1168	3253	83.84
		%	64.09	35.9		
	Female	Freq.	545	82	627	16.14
		%	86,9	13.1		
Driver vehicle r/nship	Employee	Freq.	1552	703	2255	58.12
		%	68.82	31.18		
	Owner	Freq.	894	396	1290	33.25
		%	69.3	30.68		
	Other	Freq.	180	90	270	6.96
		%	66.67	33.33		
	Unknown	Freq.	42	23	65	1.68
		%	64.62	35.38		

Table B2: frequency and percentage distribution of Environmental factor related to existence of fatalities

Explanatory variables	Categories		Existence of fatalities		Total accident	
			No fatal	≥1	Freq.	%
Light condition	Dark night	Freq.	256	6	262	6.75
		%	97.7	2.29		
	Night with SL	Freq.	756	345	1101	28.38
		%	68.66	31.34		
	Night without SL	Freq.	214	108	322	8.3
		%	66.46	33.45		
	Day light	Freq.	1438	757	2195	56.57
		%	65.51	34.49		
Day of the week	Sunday	Freq.	353	164	517	13.32
		%	68.278	31.72		
	Monday	Freq.	401	185	586	15.10
		%	68.4	31.6		
	Tuesday	Freq.	333	164	497	12.8
		%	67	33		
	Wednesday	Freq.	320	139	459	11.83
		%	69.72	30.28		
	Thursday	Freq.	375	168	543	13.99
		%	69.1	30.9		
	Friday	Freq.	413	178	591	15.23
		%	69.88	30.11		
	Saturday	Freq.	473	214	687	17.70
		%	68.85	31.15		

Table B3: frequency and percentage distribution of Road factor related to existence of fatalities

Explanatory variables	Categories		Existence of fatalities		Total	
			No fatal	≥1	Freq.	%
Types of pavement	Gravel roads	Freq.	98	48	146	3.76
		%	67.12	32.88		
	Asphalt roads	Freq	2570	1164	3734	96.24
		%	68.83	31.17		
Road junction type	Midblock	Freq.	2052	934	263	6.78
		%	68.72	31.28		
	Y-junction	Freq.	184	79	2986	76.96
		%	69.96	30.04		
	T-junction	Freq.	319	143	462	11.91
		%	69.05	30.95		
	Roundabout	Freq.	30	15	45	1.16
		%	66.67	33.33		
	Cross (four leg jun)	Freq.	83	41	124	3.196
		%	66.94	33.06		
Road geometry	Curved road	Freq.	29	15	44	1.13
		%	65.91	34.1		
	Gradient road	Freq.	136	63	199	5.13
		%	68.34	31.66		
	Straight road	Freq.	2405	1079	3484	89.79
		%	69.03	30.97		
	Tangent road	Freq.	94	59	153	3.94
		%	63.95	36.05		
Road class	One way	Freq.	53	24	77	1.98
		%	68.83	31.17		
	Two way	Freq.	901	414	1315	33.89
		%	68.52	31.48		
	Double courage way	Freq.	1450	667	2117	54.56
		%	68.49	31.51		
	Ring road	Freq.	47	24	71	1.83
		%	63.51	36.49		
	Two way(divided by solid line)	Freq.	167	74	241	6.21
		%	69.29	30.71		
	Two way(divided by broken line)	Freq.	50	9	59	1.52
		%	84.75	15.25		

Table B4: frequency and percentage distribution of vehicle factor related to existence of fatalities

Explanatory variables	Categories		Existence of fatalities		Total	
			No fatal	≥1	Freq.	%
Vehicle service years	0-5 years	Freq	1116	507	1623	41.83
		%	68.76	31.24		
	5-10 years	Freq	737	131	868	22.37
		%	84.91	15.1		
	Above 10 years	Freq	815	574	1389	35.80
		%	58.68	41.32		
Vehicle type	Taxis-minibuses(upto 12seat)	Freq	517	229	746	19.23
		%	69.3	30.7		
	Automobile	Freq	594	291	885	22.81
		%	67.12	32.88		
	Cargo (upto 10 qtl)	Freq	294	135	429	11.057
		%	68.53	31.47		
	Cargo (upto 11-40 qtl)	Freq	560	248	808	20.82
		%	69.31	30.69		
	Cargo (upto41-100 qtl)	Freq	114	51	165	4.25
		%	69.1	30.9		
	Buses (13-45 seat)	Freq	299	120	419	10.80
		%	71.36	28.64		
	Buses (above 46)	Fre	290	138	428	11.03
		%	67.76	32.24		