Addis Ababa University
Office of Graduate Program

Faculty of Science
Department of Statistics

EMPIRICAL ANALYSIS ON TRAFFIC ACCIDENTS INVOLVING HUMAN INJURIES (THE CASE OF ADDIS ABABA)

By
Tewolde Mekonnen

A THESIS SUBMITTED TO THE OFFICE OF GRADUATE PROGRAMMES OF ADDIS ABABA UNIVERSITY, IN PARTIAL FULFILLMENT FOR THE AWARD OF MASTER OF SCIENCE IN STATISTICS

August 2007
Addis Ababa

Addis Ababa University
Office of Graduate Program

Faculty of Science
Department of Statistics

EMPIRICAL ANALYSIS ON TRAFFIC ACCIDENTS INVOLVING HUMAN INJURIES (THE CASE OF ADDIS ABABA)

By
Tewolde Mekonnen

Approve by the Board of Examiners:

Butte Gotu (Dr.)	----------
Advisor	**Signature**
Fentaw Abegaz (Mr.)	----------
Internal Examiner	**Signature**
Emmanuel Gebreyohannes (Dr.)	----------
External Examiner	**Signature**

Tel. 251-1-239465 **P.O.Box 1176 Addis Ababa, Ethiopia**

Table of Contents

List of Abbreviations

AATCID	Addis Ababa Traffic Control and Investigation Department
AU	African Union
CAACG	Council of Addis Ababa City Government,
CSA	Central Statistics Agency
HAL	High Accident Locations
NRA	National Roads Authority
NRSCO	National Road Safety Coordination Office
RTA	Road Traffic Accidents
RTI	Road Traffic Injuries
TCID	Traffic Control and Investigation Department
UNECA	United Nations Economic Commission for Africa
VKT	Vehicle Kilometers of Travel
VMT	Vehicle Miles of Travel

Acknowledgements

First, and foremost, I thank God for giving me the opportunity to pursue my graduate study at Department of Statistics, Addis Ababa University.

I owe the deepest gratitude to Dr. Butte Gotu, my thesis advisor for his valuable and constructive comments and encouragements throughout my study. My special thanks also go to all staff of the Department of Statistics and to my classmates for all the encouragements and support that they provided me during my study years.

I am highly indebted to my wife W/ro Mahlate Mekuria, for her unreserved support and encouragement throughout my study.

Finally and yet importantly, my heart-felt thanks go to Ato Kassahun H/Mariam, Director General of the Federal Transport Authority, Ato Abebe Asrat and Ato Bamlaku Alemayehu at the NRSCO, Inspector Meskelu and all staff of the TCID whose unreserved cooperation made it possible for me to successfully complete my field work. I am also grateful to all my colleagues at World Vision Ethiopia, and especially to Ato Desta Abera, for all the encouragement and support that they gave me while I was writing this thesis.

Abstract

Ethiopia is a country with the largest number of traffic accidents and fatality rate, and the share of the city of Addis Ababa is quite big (more than 60%). Pedestrians and the disabled, children and the aged in particular, are the major victims of these accidents.

This study attempts to identify the variables that mainly determine the number of injuries and to describe the impact of Driver's Sex, Age, Driving Experience, Educational Background, Vehicle Years, Vehicle Category (plate code), Vehicle type (Automobile, taxi, Truck…etc), and Place of accident (Organization, Religious institutions, Market…etc) on the number of traffic injuries per accident.

Poisson regression analysis is used in this study. Findings of this study have shown that Drivers' Age, Educational background and Place of accident significantly affect the number of injuries per accident. Drivers who are in the age group of 18-30 are liable for most of the accidents including the sever ones. Drivers with elementary school level of education take the major responsibility for the increased number of injuries per accident. With regards to place of accident, residential areas are where the highest mean of injuries per accident is attained.

CHAPTER 1: Introduction

1.1 Background

Over one million people every year are killed in road crashes, and 20-50 million are injured. Road traffic injuries (RTIs) are growing as the vehicle use of developing-countries rise. By 2020, RTIs 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 RTIs 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. Unless action is taken to improve road safety systems, poor countries will continue to bear the heavy toll of road traffic injuries (Lawren *et al*., 2005).

The 2004 World Health Report shows that of the 1.2 million people killed in road crashes worldwide, 85% are in developing countries. Sub-Saharan Africa alone with only 4% of the global vehicle registered accounts for 10% of the total road fatalities, and the economic, social and health consequences are grave. Conversely, the high income nations, with 60% of the total global vehicle fleet contribute only 14% of the annual road deaths. Human error, road environment and vehicle factors are reported by the traffic police as the main causes of road crashes.

Two countries, South Africa and Nigeria, account for most of the reported deaths in Sub-Saharan Africa. The South African figure of over 9,000 has been consistent over time, while Nigeria with 6,185 deaths has declined from a high of over 9,200 in the early 1990s. Ethiopia, Kenya, Uganda, Tanzania and Ghana are the other countries that experience high numbers of road deaths.

Pedestrians account for the highest proportion of road fatalities in nearly all African countries, ranging between 31% in Zimbabwe and 51% in Ethiopia. Involvement of pedestrians is much greater in urban environment than in rural areas. Studies in Addis Ababa and Abidjan reported extremely high proportion of pedestrian casualties of 90% and 75%, respectively. Passengers rank second, accounting for 32% to 46%. Pedestrians and passengers altogether represent over 80% of all road deaths. Drivers account for a small share of fatalities, of less than 10 per cent. Among sub-Saharan countries, only South Africa has the largest share of driver fatalities (22%) (Odero, 2004).

Consistent to the above facts, another study shows that Ethiopia is a country with the largest number of traffic accidents and fatality rate, and the share of the city of Addis Ababa is quite big (more than 60%). 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. More than 12 million Birr is lost every year because of traffic accidents (Asfaw, 1999).

The prevailing situation calls for intervention in view of minimizing the number and magnitude of accidents in different aspects. To this effect there is a need for

identifying major factors that affect the number of traffic accidents. Accordingly this study concentrates on identifying factors that most explain the number of traffic injuries per accident.

1.2 Statement of The Problem

A road traffic accident is defined as any vehicle accident occurring on a public highway. It includes collisions between vehicles and animals, vehicles and pedestrians, or vehicles and fixed obstacles. Single vehicle accidents, which involve a single vehicle, that means without other road user, are also included (Safecarguide, 2004).

At all levels, whether at national or international level, road traffic accidents continue to be a growing problem. In connection with this, according to a World Health organization/World Bank Report, deaths from non-communicable diseases are expected to grow from 28.1 million a year in 1990 to 49.7 million by 2020, which is an increase in absolute number of 77%. Traffic accidents are the main cause of this rise. Road traffic injures are expected to take higher place in the rank order of disease burden in the near future.

The tragedy is more or less similar in Addis Ababa, Ethiopia. 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. In Ethiopia, above 1,800 people died while above 7,000 were crippled or injured in 2003. Moreover the death rate is 136 per 10,000 vehicles and Ethiopia is loosing over 400 million birr yearly as a result of road

accidents. The share of Addis Ababa city in the total number of accidents was 60 percent in 1989 with annual average traffic accident growth of 31.4 percent. Nowadays Addis Ababa is experiencing around 700 accidents per month resulting in various levels of injury (Tesema et al., 2005).

The factors in Vehicular Accidents can be categorized into the following major groups, namely:

- Physical environment (road type and condition, location,...)
- Person (driver, passenger, pedestrian...etc)
- Vehicle related (type, technical condition...etc)
- Other (Weather condition, Visibility...etc)

Records at the Addis Ababa Traffic Control and Investigation Department (AATCID) show that most of the accidents occur on roads with good condition (like asphalted ones) and during good weather conditions. Accordingly, for Addis Ababa, it could be assumed that the road and weather conditions have no significant impact on traffic accident. But location of accident may have significant role in causing accidents especially resulting in injuries.

Thus, among the categories stated above, vehicle and driver related factors along with place of accident are considered in the study.

The consequences of traffic accidents are categorized as: Deaths, Heavy Injuries, Light Injuries and Material Loss.

Most of the time heavy injuries may result in the death of the injured person, and also light injuries may result in disability. Thus, in this study the term “Injury” will be used to refer to any of these incidents on human life.

This study concentrates on the number of injured people per accident and will attempt to address the following questions:

- What are the variables that mainly determine the number of injuries?
- What is the impact of the following variables on the number of traffic injuries per accident?
    - Driver’s Sex
    - Driver’s Age
    - Driving Experience,
    - Drivers' Educational Background,
    - Vehicle-Driver Relation (owner, hired…etc)
    - Vehicle Age (Years of Service),
    - Vehicle type (Automobile, taxi, Truck…etc)
    - Place of accident (Organization, Religious institutions, Market…etc)

1.3 Objectives of The Study

The general objective of this study is to identify the major factors determining the number of traffic injuries.

The specific objectives are:

- To identify and describe the major variables (factors) that contribute to traffic injuries.
- To analyze the impact of driver and vehicle related variables on traffic injuries.

1.4 The Study Area

Addis Ababa, located about 2,408m above sea level at 9.02° N 38.44° E, became Ethiopia's capital when Menelik II was Emperor of Ethiopia. The town grew by leaps and bounds.

Addis Ababa has an estimated total population (for July 2006) of 2,973,000, consisting of 1,428,000 men and 1,545,000 women. It is estimated that presently there are no rural parts to the city, so 100% of the inhabitants are considered urban dwellers; 24% of all urban dwellers in Ethiopia are in Addis Ababa. With an estimated area of 530.14 square kilometers, the city has an estimated density of 5,607.96 people per square kilometer (CSA, 2005).

Addis Ababa is the headquarters of many international organizations including the UNECA and the AU. Public transportation is through public buses or blue and white share taxis, locally known as "blue donkeys". The taxis are usually minibuses that can sit at least twelve people. Two people are responsible for

each taxi, the driver and his assistant who collects fares and calls out the taxi's destination (Wikipedia, 2006).

Out of all the accidents registered in Ethiopia, Addis Ababa holds about 60% on average. This is partly because the city has great contact through its all gates with different regions every day. In addition to this, of the registered motor vehicles in Ethiopia, the city takes about 77% of it. All these facts reveal that Addis Ababa, having a great deal of concentration of vehicles and traffic, takes the lion's share in car accidents also. Statistical data from the office shows that Addis Ababa is experiencing around 700 accidents per month and the costs of such fatalities and injuries due to traffic accidents have a great impact on various aspects of the society (Tesema et al., 2005).

1.5 Limitations of the Study

This study is based on a secondary data obtained from AATCID. The data contains records of road traffic accidents and the number of injuries (which includes fatalities) in each accident.

The data related to Vehicle Service Years and Driving Experience are collected by interviewing drivers and the accuracy of the data is questionable to some extent. There were also incomplete and inconsistent records.

Regardless of the above limitations, the resultant error is assumed to be insignificant due to the large sample size used and because data are recorded in category (group intervals). In addition, only those records found complete and consistent are considered for the study.

CHAPTER 2: Literature Review

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).

Robert, (2000), analyzed the relationship between road infrastructure and safety by using a cross-sectional time-series data base collected for 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 center-line miles with a given lane width by 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 arterials. A casual interpretation of the trends and those for total fatalities would suggest that as highway facilities are upgraded, there are

reduced fatalities. In addition, estimates of seatbelt usage, by state, were used to control for the effects of increased seatbelt use. The analyses also attempts to control for seatbelt effects by including dummy variables for those states with either primary or secondary seatbelt 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 can not be included in the same model due to high collinearity between them. Separate models for each are therefore estimated. The dependent variables were the total deaths (DEATHS) and total injuries (INJURED) from traffic-related crashes. Models containing all the relevant infrastructure variables, and models without the lane width variables were developed. According to the results of his study, total lane miles are found to be highly significant across some of models for both total fatalities and injuries. No significant effect is found for increases in the average amount of interstate lanes on fatalities. The percent of lane miles by each road category shows that those

states with more lane miles of interstate (relative to other categories) have a statistically significant reduction in injuries. However, there was no statistically significant reduction in fatalities when a state has proportionally more interstate lane miles. Furthermore, the results indicated that states with a larger share of arterial lane miles in their networks have more fatalities and injuries, and those with more collectors have more injuries.

The National Roads Authority of Ireland has conducted a study with the main objectives being identifying High Accident Locations (HALs) on the inter-urban national road network, and investigating whether at each of these individual sites there were any risk factors, such as skidding etc, that may have influenced the number of accidents occurring at that section. The report of the study introduced Road Traffic Accidents (RTAs) as rare, random, multi-factor events in which a road user/s fails to cope with the surrounding traffic environment. The occurrence of RTA, is by its very nature unpredictable. However, according to the report, while each individual accident is fundamentally unpredictable by its nature, the number of accidents in given locations over given time periods may display notable patterns and can be influenced by a range of factors. The report indicates that Multiple Linear Regression Models, Poisson Models, Negative Binomial Models and Accident Rate Models can be used in analyzing RTA. It is further discussed in the report that researchers have gradually moved away from the use of Multiple Linear Regression Models as a result of well-documented statistical difficulties. The Poisson distribution which is a discrete distribution with the variable in question taking on whole number values greater than or equal to

0, and which is often used to model the number of events occurring within a given period was considered for modeling purpose. The core aspect of the study was to develop a statistical model which, for inter-urban national roads, generates the expected number of accidents in a section given its traffic volume and section length. In circumstances where the number of recorded accidents significantly exceeds the expected number, then that section is considered as a HAL. The basic or null hypothesis studied was that sections of road of equal length with equal vehicle miles traveled should have similar levels of accidents over a given time period; and that the number of accidents in a given period should be similar to that estimated using the statistical techniques outlined above. The null hypothesis was rejected if the difference between the actual and estimated number was so large as to have arisen in less than one in twenty times. In other words, if it can be said to a ninety-five per cent confidence level that there is a difference between the actual and expected number of accidents then the null hypothesis is rejected in favor of the alternate hypothesis – there is a difference (NRA, 2007**).**

Geedipally (2005) shows the effect of new pavement on traffic safety in Sweden. The study investigated whether higher pavement road standards lead to a higher or lower accident rates. The Poisson Regression model was developed taking traffic accidents as dependent variables and the factors which cause the accidents as independent variables. The result of his study shows that Taffic accidents increased by 12 % after one year of resurfacing on all types of roads. The rural roads showed much worse effect with an increase of 17% after

resurfacing. The urban roads showed a positive effect with a decrease in traffic accidents by 50% after resurfacing. Further he stated that a chi-square test showed that there is no firm evidence on real change of accidents due to the new pavement. Finally he stated that, the Poisson regression model fits reasonably well for the expected number of accidents and it does not depend on the different periods considered for analysis and also on the urban road accidents.

Al-Masaed *et, al* (2004) investigated the relationship between city planning and street network variables and traffic accidents at the zone level. Damascus, the capital of Syria, which consists of fourteen urban zones, was selected as a case study. For each zone, data on traffic accidents, population density, land-use developments, level of travel, road network, intersection density, and the distribution of public buildings were measured through field surveys or obtained from documents of relevant authorities. In the study, a cross-sectional approach was adopted to focus on the differences in safety between urban zones at a specific point in time. Mathematical modeling was used to develop relationships between zone traffic accidents and the variables mentioned above. In the analysis, a correlation matrix was established to identify variables that had an influence on traffic accidents and to check possible multicollinearity between pairs of independent variables. In the modeling process, the first step was to identify the best transformation for variables that had an impact on traffic accidents. In the second step, multivariate regression analysis was carried out to develop traffic accident predictive models. Based on their study, they concluded that levels of travel and population density have a strong influence on urban

accidents. Reduction of the need for travel and locating major streets on the edge of an urban zone as well as limiting population density could enhance traffic safety. Another conclusion of their study was traffic accidents in an urban zone are exponentially proportional to intersection density and total street length. In street network planning, the reduction of both intersection density and total street length could provide safer networks. Finally they stated that commercial frontages and the location of public buildings have adversely influenced urban accidents.

The concept of traffic control system in Addis Ababa was put in place around 1900 with the introduction of motor vehicles. Regarding the regulations and laws governing road traffic, the first cited traffic control rule with 18 articles was introduced in 1918. This rule was aimed at facilitating the traffic system, which involves movement of pedestrians, animals and motor vehicles. It was in 1935 that formal and legal driving license had been put in place and the present driving license issuance rule has been under implementation since 1960. With respect to comprehensiveness, it is the Transport Act number 361 /1961, which is more comprehensive in having most of the major regulations commonly used. The road traffic safety regulation numbers 5/1998 and 4/2004 are also the recent rules and regulation in use at the City of Addis Ababa (Tesema *et al*, 2005).

Tessma et al (2005), studied injury severity levels resulting from an accident using real data obtained from the Addis Ababa traffic office. Their research was focused on developing adaptive regression trees to build a decision support system to handle road traffic accident analysis for Addis Ababa city traffic office.

As stated in their study, the Classification and Adaptive Regression Trees (CART) methodology is technically known as binary recursive partitioning. The process is binary because parent nodes are always split into exactly two child nodes, and recursive because the process can be repeated by treating each child node as a parent. The key elements of a CART analysis are a set of rules for splitting each node in a tree: a) deciding when a tree is complete and b) assigning each terminal node to a class outcome (or predicted value for regression). They further stated that the specific attributes by which a given accident can be described are *date and time, accident id , driver's name, vehicle type, driver's age, driver's gender, driver's educational level, driver's license status, relation of the driver and vehicle, driver's experience, possession of the vehicle, vehicle defect, vehicle age, accident area, accident road name, road segment separation, road direction, road surface type, roadway surface condition, light condition, weather condition, vehicle maneuver, accident type, total vehicles involved, total number of victims, accident victims category, victims profession, victims health condition, pedestrian maneuver, vehicle plate number, cost estimate of the damage and cause for accident*. In addition to the input variables mentioned above the output variable for the research that is injury severity was also another attribute of a given accident. The target attribute, *injury severity*, has four classes: *fatal injury, property damage (no injury), serious injury* and *slight injury*. In Defining the Data Mining Function, they stated that each individual accident record in the data set is an input/output pair with each record having an associated output. The output variable, the injury severity, is

categorical and as described above, has four classes. After successive experiments in building the best decision tree model, the next step of their study was to generate rules by tracing through the branches up to leafs. A rule is a correlation found between the main variable (dependent) and the others (independent). From the general rule, they found out that it was easy to see that, there is a 7.8 % chance that *Injury_Severity* will be a fatal injury, a 67.63% chance that Injury_Severity will be partial damage, a 8.15% chance that Injury_Severity will be serious injury and a 16.39% chance that Injury_Severity will be slight injury and this reveals that about 33% of the accidents, results in injury of different level half of which is either fatal or serious injury. Further, their study indicated that accident types involving pedestrians and single vehicle turn over mostly results in either fatal or serious injuries. Denying pedestrian priority and over speeding were also the top most determinant factors in injury severity. It is also apparent, as per their study, that if accident cause is driving with alcohol and accident type is vehicle_peds there is high probability of an accident resulting in fatalities or injuries. Over speeding is also another important factor especially if associated with vehicle_peds type of accidents. In general, the rules presented above indicate the possible conditions in which an accident will result in either of the injury severity classes. Moreover, the rules generated have indicated that attributes such as *'accident cause', 'accident type', 'driver age',' road surface type', 'road condition', 'vehicle type'*, and '*light condition'* are found to be important variables for classification of accident severity. The study also indicated that, these variables are playing a significant role in all experiments

being placed at the higher level of the tree which indicates their statistical significance than other variables like *'sex', 'weather condition'* and *'accident_id'*. Decision of selecting the best decision tree was based on the soundness of the rules generated as well as the number of misclassified records of different levels of injury severity.

The National Road Safety Coordination Office (NRSCO), (2006) states that road traffic accident in Ethiopia is a serious problem. According to the office, the death rate is estimated to be 130 per 10,000 vehicles. Of those killed over half are pedestrians, of which 30% are children. In Ethiopia one out of five people injured die as a result of road accident, which again is a very high figure mainly due to poor situation of the emergency medical services. The economic loss due to road accident is also significant. Unless the present trend is arrested, the social and economic problem of road accident will be more serious as the number of cars increases.

Based on a five-year average records the office further states that, of the personal injury accidents, 81 % are caused due to drivers' error, 5% due to vehicle defect, 4% due to pedestrian error, 1 % due to road defects and 9% due to other problems. These figures show that the majority of personal injury accidents are caused as a result of drivers' error. The study further shows that professional drivers are involved in 88% of the fatal accidents. Special purpose vehicles and motor bicycles cause 8% of such accidents. On the other hand automobile drivers have very good safety records with only 4% of the fatal

accidents, which is equivalent to a rate of 12 fatal accidents per 10,000 vehicles. Automobiles share is around 43% of the total vehicle number of the country.

Fatal accident types were 68% pedestrian strike, 13% overturn, 6% fall from vehicles, 3% animal or cart strike and 10% for all the other remaining crash types. The underling reasons for these accidents are noted to be: -

1) Improper behavior or low skill of drivers (Over speeding, not respecting pedestrian priority…etc.)
2) Poor vehicle technical conditions
3) Animals and carts using the highways
4) Pedestrians not taking proper precautions
5) Poor traffic law enforcement
6) Poor emergency medical services and
7) Safety consideration not sufficiently given in roads developments

Most of the studies conducted on traffic accidents indicate that factors that contribute to the occurrence of traffic accident can be categorized as per their relation to Vehicle, Driver or Road condition. Others like Visibility, Weather condition, population density...etc are also additional factors that may be taken into consideration.

Even though much has been said about the role of drivers in causing traffic accidents, little has been done on the factors related to drivers' Age, educational background, driving experience. In addition, studies in the area of vehicle related factors like vehicle years of service and vehicle type are rare. This study will attempt to fill this gap.

CHAPTER 3 : Methods

3.1 Data Source

Accidents are recorded by the traffic police on daily basis. This study is based on a secondary data obtained from Addis Ababa Traffic Control and Investigation Department.

The observations are information about three thousand one hundred accidents that occur with in 310 consecutive days (25/06/98-29/04/99E.C).

3.2 Variables Considered

The following variables are considered in this study.

Independent Categorical Variables

Demographic Variables: The demographic variables related to driver involved in the accident are:

- Age: it is categorized as:
    - 18-30,
    - 31-50 and
    - 51 and above
- Sex: (Male or Female)

- Educational Background: The maximum education level attained by the driver is recorded under one of the following categories:

    1. Basic Education
    2. Elementary School
    3. Junior Secondary School
    4. Secondary School
    5. Above Secondary School

- Driving Experience: This is the number of years since the driver received a driving license. This information is sometimes recorded by asking the driver since the year the driver received the first license could not be found on the current license if the driver is having higher level driving license. The information obtained from the driver is recorded under one of the following six categories:

    1. Less than or equal to 1 year
    2. Greater than 1 year and less than or equal to 2 years
    3. Greater than 2 years and less than or equal to 5 years
    4. Greater than 5 years and less than or equal to 10 years
    5. More than 10 years

- Driver-Vehicle Relationship: the categories are:

    1. Owner
    2. Hired
    3. Other (which may be relative, friend, one who rented the vehicle …etc)

Other Variables

Vehicle related variables: Vehicle is defined as carriage, chariot, bicycle, motor vehicle, semi-trailer and trailer operated on a road (CAACG, 1998). The variables related to the vehicle responsible for the traffic accident are described below.

- Vehicle Years of Service. This is the number of years since the date the vehicle is manufactured. The categories are:

    1. Less than or equal to 1 year
    2. Greater than 1 year and less than or equal to 2 years
    3. Greater than 2 years and less than or equal to 5 years
    4. Greater than 5 years and less than or equal to 10 years
    5. More than 10 years

- Vehicle Type: it has the following categories
    - 1-Buses: (13-45 Seats, Above 46 Seats)
    - 2-Minibuses and Taxies
    - 3-Cargo (Upto 10 Quintals, 11-40Quintals, 41-100Quintals)
    - 4-Trailer (above 100 quintals)
    - 5-Station Wagon
    - 6-Automobile
    - 7-Liquid Cargo

Location related variable: This variable indicates the area where a traffic accidents happened. The categories are given by:

1. Organization
2. Residence
3. Market
4. Religious Place
5. Entertainment
6. School
7. Hospital

Dependent Discrete Variable

The dependent variable is the Number of Injuries per accident. This includes the number of people dead, lightly injured or heavily injured due to a traffic accident.

3.4 The Model

One of the main assumptions of linear models such as linear regression and analysis of variance is that the residual errors follow a normal distribution. To meet this assumption when a continuous response variable is skewed, a transformation of the response variable is used. Often, however, the response variable of interest is categorical or discrete, not continuous. In such cases, a simple transformation cannot produce normally distributed errors. A common example is when the response variable is the number of occurrences of an event. The distribution of counts is discrete, not continuous, and is limited to non-negative values. There are two problems with applying an ordinary linear regression model to such data. First, many distributions of count data may be positively skewed with many observations in the data set having a value of 0. The existence of many 0's in the data set prevents the transformation of a skewed distribution into a normal one. Second, it is quite likely that the regression model will produce negative predicted values, which are theoretically impossible. As an alternative a Poisson regression model or one of its variants is widely used. These models have a number of advantages over an ordinary linear regression model, including a skew, discrete distribution and the restriction of predicted values to non-negative numbers. The Poisson model assumes that the mean and variance are equal. But, usually in practice the variance is larger (or

sometimes smaller) than the mean. When the variance is larger than the mean, an alternative is a negative binomial model. The negative binomial distribution is a form of Poisson distribution in which the distribution's parameter itself is considered a random variable. The variation of this parameter can account for a variance of the data that is higher than the mean (Grace-Martin, 2000).

The Poisson distribution arises when you count a number of events across time or over an area. one should think about the Poisson distribution for any situation that involves counting events. Sometimes, count is represented as a rate.

The Poisson distribution is based on four assumptions. We will use the term "interval" to refer to either a time interval or an area, depending on the context of the problem.

1. The probability of observing a single event over a small interval is approximately proportional to the size of that interval.
2. The probability of two events occurring in the same narrow interval is negligible.
3. The probability of an event within a certain interval does not change over different intervals.
4. The probability of an event in one interval is independent of the probability of an event in any other non-overlapping interval.

If either of these last two assumptions are violated, they can lead to extra variation, sometimes referred to as over dispersion. (Simon, 2007)

Since this study is based on the number of injuries of traffic accidents, the data consists of counts. Thus the data is assumed to follow Poisson distribution.

In Poisson regression it is assumed that the dependent variable Y, number of occurrences of an event, has a Poisson distribution. Thus, given the independent variables X_1, X_2,... X_m,

$$P(Y = k \mid X_1, X_2, ..., X_m) = \frac{e^{-\lambda}\lambda^k}{k!} \qquad k = 0,1,2.... \qquad (1)$$

where the log of the mean λ is assumed to be a linear function of the independent variables. That is,

$$\log(\lambda) = \beta_o + \sum_{i=1}^{m} \beta_i X_i \quad \text{.....(2)}$$

which implies that λ is the exponential function of independent variables,

$$\lambda = e^{(a+\beta_1 X_1 + \beta_2 X_2 + ... + \beta_m X_m)} \qquad (3)$$

The Poisson distribution depends on a single parameter λ. Although there is no theoretical upper bound for the Poisson distribution, in practice these probabilities could be considered as negligible when k is very large. How large k needs to be before the probabilities become negligible depends entirely on the value of λ. (Bauer *et al*, 2007)

In a Poisson distribution, the model coefficients are estimated by the maximum likelihood method. The likelihood function L is the product of the terms in equation (1) over all *n* measured values y_i. This function is viewed as a function of the parameter λ and the parameters β_i. The parameters are estimated by

maximizing the likelihood, or more usually, by maximizing the logarithm of the likelihood (denoted by log likelihood). The log likelihood is given by the equation:

$$\log(L) = \sum_{i=1}^{n} [y_i \log(\lambda) - \lambda - \log(y_i!)] \qquad (4)$$

The maximum value possible for the likelihood for a given data set occurs if the model fits the data exactly. This occurs if λ is replaced by y_i in equation (2). The difference between the log-likelihood functions for two models is a measure of how much one model improves the fit over the other. A special case of this was defined as the deviance. The deviance is defined as minus twice the log of the ratio of the likelihood for a model to the maximum likelihood. For the Poisson distribution, the deviance takes the form given in the following equation:

$$D = 2[\sum_{i=1}^{n} y_i \log(y_i / \lambda) - \sum_{i=1}^{n} (y_i - \lambda)] \qquad (5)$$

where the second term is identically zero in the usual case that the model includes a constant or intercept term. The deviance so defined is measured from that of the saturated model and so terms involving constants, the data alone, or a scale factor alone are omitted. For a sample of n independent observations, the deviance for a model with p degrees of freedom (that is, p parameters estimated, including the mean or constant) has residual (n-p) degrees of freedom. When the residual degrees of freedom of the current model are approximately equal to the deviance, it is unlikely that further fitting of systematic components is worthwhile. (Bauer *et al*, 2007)

Since the deviance is effectively -2 times the log of the likelihood ratio, it has an asymptotic distribution that is chi-squared with degrees of freedom equal to n- p. This result can be used to construct a goodness-of-fit test for the model. In addition, by forming the ratio of the deviance to its residual degrees of freedom, an estimate of the scale constant can be found. For the Poisson distribution, this should theoretically be equal to one. Values substantially in excess of one reflect over dispersion of the data.

Negative Binomial Regression Model

As mentioned above, a limitation of the Poisson distribution is that the mean equals the variance of the distribution. Previous work in the field of accident research has shown that this is not always the case. Suppose a Poisson model is used for modeling accidents and the variance (or dispersion) of the data exceeds the estimated mean of the accident data distribution. The data are then said to be over dispersed, and the underlying assumption of the variance being equal to the mean for the Poisson distribution is violated. The negative binomial, which is a discrete distribution, provides an alternative model to deal with over dispersion in count data such as accident frequencies (Bauer *et al*, 2007).

Unlike the Poisson distribution, the negative binomial distribution adds a quadratic term to the variance representing over dispersion. The negative binomial model takes the form:

$$P(y_i) = \frac{\Gamma(y_i + \frac{1}{K})}{y_i!\Gamma(\frac{1}{K})}(\frac{K\lambda}{1+K\lambda})^{y_i}(\frac{1}{1+K\lambda})^{\frac{1}{K}} \quad (6)$$

Where K is the over dispersion parameter and the variance is

$\lambda + K\lambda^2$.

If K= 0, the negative binomial reduces to Poisson distribution. The larger the value of K, the more variability there is in the data over and above that associated with the mean λ

The parameter k is not known a priori, but can be estimated so that the mean deviance becomes unity or the Pearson chi- square statistic equals its expectation (i.e., equals its degrees of freedom).

As for the Poisson model, the model regression coefficients, β_0, β_1, β_2, ... , are estimated by the method of maximum likelihood. The estimation of the model parameters can be done by minimizing the negative of the log likelihood. For the negative binomial distribution, the log likelihood is given by the equation (Bauer *et al*, 2007) :

$$\log(L) = \sum_i [\sum_{j=0}^{y_i} \log(1+Kj)) - \log(1+Ky_i) + y_i \log\lambda - (y_i + \frac{1}{K})\log(1-K\lambda) - \log(y_i!)] \quad (7)$$

There are some empirical ways of checking for a Poisson distribution.

- The simplest way is to see if the variance is roughly equal to the mean of the data.

- A histogram of the Poisson data should be skewed to the right, though the skewness decreases as the mean increases.

In order to check whether the mean and variance are equal, descriptive statistics could be used.

The Excel result of descriptive statistics (Annex 1), shows that the empirical mean of the dependent variable is 0.43 and the variance is 0.46. The variance is slightly greater than the mean.

The histogram (Annex1) is skewed right which can be an indication of Poisson distribution. The value of Skewness which is 2.88 also confirms that the distribution is skewed right.

Even though the skewness of the distribution indicates that the data follows Poisson distribution, we need to test for over dispersion because sample variance and mean are not equal.

In the presence of over dispersion, an alternative approach to model the over-dispersion is the Negative Binomial regression. The Negative Binomial model has mean $E(Y) = \lambda$ and variance $Var(Y) = \lambda + k\lambda^2$ where $k \geq 0$ is the dispersion parameter. When k=0, the negative binomial distribution reduces to Poisson. Therefore, in testing over dispersion, the null hypothesis is:

H_0 : k=0

and the alternative hypothesis is:

H_a : k>0.

Model Evaluation - Overdispersion

A decision about whether the Poisson form is appropriate can be based on one of several statistics. The deviance of a model m is:

$$D^m = 2\ (L^f - L^m)$$

where L^f is the log-likelihood (4) that would be achieved if the model gave a perfect fit ($\lambda = y_i$ for each i, and K = 0) and L^m is the log-likelihood (7) of the model under consideration ($\lambda = y_i$). If the latter model is correct, D^m is approximately a chi-squared random variable with degrees of freedom equal to the number n of observations minus the number p of parameters.

A value of the deviance greatly in excess of n - p suggests that the model is over dispersed due to missing variables and/or non-Poisson form. Thus when deviance divided by degrees of freedom

$$\frac{D^m}{n-p}$$

is significantly larger than 1, over dispersion is indicated.

Likewise, the Pearson chi-square statistic, defined by

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{\hat{y}_i} ,$$

is an approximately chi-squared random variable with mean n - p for a valid Poisson model.

If

$$\frac{\chi^2}{n - p}$$

is significantly larger than 1, over dispersion is also indicated. (Bauer *et al*, 2007)

By analyzing the data for both the Poisson and Negative Binomial models using SAS, the result (Annex 2) gives D= 2 (-2419.7482+ -2420.1941)= 0.8918 which is not significant with p-value 0.34499.

In addition, the Poisson Deviance and Pearson Chi square divided by their respective degrees of freedom are nearly equal to one (Annex 2.1). This can lead to the assumption that the dispersion parameter K is zero.

Thus, we can conclude that, there is no statistically significant over dispersion in the data that can lead to the rejection of the assumption of Poisson distribution.

Consequently, the analysis is done by applying Poisson Regression Techniques.

Model Evaluation- Residuals

For any GLM, goodness-of-fit statistics only broadly summarize data. We obtain further insight by comparing observed and fitted counts individually. For observation i, the residual difference $Yi - \lambda_i$ between an observed and fitted count has limited usefulness. For Poisson sampling, for instance the standard deviation of a count is $\sqrt{\lambda_i}$, so larger differences tend to occur when λ_i is larger.

The Pearson *residual* is a standardization of this difference, defined by

$$Pearson\,\mathrm{Re}\,sidual = \frac{(observed - fitted)}{\sqrt{\hat{V}ar(observed)}}$$

For poisson GLMs, this simplifies to

$$e_i = \frac{y_i - \hat{\lambda}_i}{\sqrt{\hat{\lambda}_i}}$$

which standardizes by dividing the difference by the estimated Poisson standard deviation. These residuals relate to the Pearson goodness-of-fit statistic by

$$\sum e_i^2 = X^2$$

Pearson residual values fluctuate around zero, following approximately a normal distribution when μ_i is large. When the model holds, these residuals are less variable than standard normal, however, because the numerator must use the fitted value $\hat{\lambda}_i$ rather than the true mean λ_i.

Since the sample data determine the fitted value, $y_i - \hat{\lambda}_i$ tends to be smaller than $y_i - \hat{\lambda}_i$. The Pearson residual divided by its estimated standard error is called an *adjusted residual.*

Adjusted residuals larger than about 2 in absolute value are worthy of attention, though one expects some values of this size by chance alone when the number of categories is large. Adjusted residuals are preferable to Pearson residuals. [Agresti, 1996]

3.5 Chi-Squared Tests of Independence

According to Agresti (1996), categorical data consist of frequency counts of observations occurring in the response categories. Let X and Y denote two categorical variables, X having I levels and Y having J levels. We display the IJ

possible combinations of outcomes in a two-way table. Let π_{ij} = P(X=I, Y=j) denote the probability that (X,Y) falls in the cell in row i and column j. The probabilities $\{\pi_{ij}\}$ form the joint distribution of X and Y. The marginal distributions are the row and column totals of the joint probabilities. They are denoted by $\{\pi_{i+}\}$ and $\{\pi_{+j}\}$ for the row and column variable respectively.

Agresti further states that when two variables are independent, the probability of any particular column response j is the same in each row. In two-way contingency tables the null hypothesis of statistical independence of two responses has the form

H_o: $\pi_{ij} = \pi_{i+}\pi_{+j}$, for all i and j.

To test the H_o, we identify $\mu_{ij} = n\pi_{ij} = n\pi_{i+}\pi_{+j}$ as the expected frequency. Here μ_{ij} is the expected value of n_{ij}(sample cell counts) assuming independence. And n is the sample size. Since $\{\pi_{i+}\}$ and $\{\pi_{+j}\}$ are unknown, they are estimated by substituting sample proportions for the unknown probabilities, giving

$$\hat{\mu}_{ij} = np_{i+}p_{+j} = n\frac{n_{i+}}{n}\frac{n_{+j}}{n} = \frac{n_{i+}n_{+j}}{n}$$

The $\{\hat{\mu}_{ij}\}$ are called expected frequencies.

For testing independence in I x J contingency tables, the Pearson and likelihood ratio statistics equal :

$$X^2 = \sum\sum\frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}}, \quad G^2 = 2\sum\sum n_{ij}\log\left(\frac{n_{ij}}{\hat{\mu}_{ij}}\right).$$

Their large sample chi-squared distributions have df=(I-1)(J-1). This means, Under H_o, $\{\pi_{+j}\}$ and $\{\pi_{i+}\}$ determine the cell probabilities.

There are a number of other methods like, Backward Elimination (Top down approach), Forward Selection (Bottom up approach) and Stepwise Regression (Combines Forward/Backward) which could be used for selecting variables. But in this study, using any of these methods is believed unnecessary since the relationship of each variable to the existence of injury in an accident can be easily identified by the chi-squared test of independence.

Thus, chi-squared test is used to identify and select explanatory variables which have statistically significant association with the dependent variable.

CHAPTER 4: Results and Discussion

4.1 Variable Selection

The analysis was started by testing the significance of the association each explanatory variable could have with the dependent variable. For this purpose the chi-square test was applied.

Even though the dependent variable is count of injuries per accident, during the application of chi-square test, it was converted to a categorical variable. This was done for making the chi-square test more reliable in its results. Accordingly, the categories for the dependant variable are “No injury existed” and “1 or more injuries existed”. Detail results of the analysis are shown on Annex 4. but summary of the results is given in the following table (Table 1).

Table 1: Summary of Chi-Square Test

No	Variable	Likelihood Ratio Test Result		
		Value	DF	P.Value
1	Driver Sex	3.047	1	.081
2	Driver Age	11.943	2	.003
3	Driver Educational Background	19.818	4	.001
4	Driving Experience	11.524	4	.021
5	Vehicle-Driver Relation	.822	2	.663
6	Vehicle Type	21.088	6	.002
7	Vehicle Years of Service	5.018	4	.285
8	Place of Accident	77.948	6	000

The chi-square test results indicate that the variables:

- Driver Age
- Driver Educational Background
- Driving Experience
- Vehicle Type
- Place of Accident

are found significantly associated with the dependent variable. i.e. they have statistical significance in determining the outcome of the accident in relation to the existence (number) of injuries. While the following variable, namely:

- Driver Sex
- Vehicle Years of Service
- Vehicle-Driver Relation

are found statistically insignificant.

This result shows that except Vehicle Type, all other vehicle related variables have no significant effect on the number of injuries. With regards to driver related variables, all variables except Vehicle-Driver Relation and Driver Sex are found significant. The variable related to Location of accident (Place of Accident) is found significantly associated to existence (number) of injuries.

Therefore, further analyses are made only on those variables which are found significantly associated with the dependent variable.

4.2 Descriptive Statistics

When we look at the accidents data, among the 3100 records, 37% involved injuries. The rest are considered as accidents with material damage only.

Table2: Existence of Injuries

	Frequency	Percent
No Injury	1959	63.2
1 or more injuries	1141	36.8
Total	3100	100.0

The number injuries per accident ranges from 0 to 9. In order to compare the impact of each variable on the number of injuries per accident, the mean number of injuries with respect to the levels of each variable was calculated. SPSS results of such values are shown on annex 5. The results are summarized based on the three indicators which can show the impact of the different levels of each variable. The three indicators are:

- Number of Accidents (by each level of a variable)
- Number of Injuries (by each level of a variable)
- Number of Injuries Per Accident (by each level of a variable)

A. Driver related variables:

Driver's Age

The driver's age has three categories. Among these categories, drivers with in the age group 18-30 are responsible for the larger number of injuries (51%) and for the large number of accidents (46%). With regards to the number of injuries

per accident, the highest mean (.48) is also attained by this group. Drivers with age 51 and above have the smallest share in all the three measurements.

When we look at the number of driving licenses by age group of drivers, here are the figures of 1997E.C from the Addis Ababa Traffic Control and Investigation Office.

Table 3: Number of Licenses by Age Group

No	Age Group	Number of Licenses	Percent
1	18-30	95,178	23%
2	31-50	213,180	52%
3	51 and Above	101,824	25%
Total		**410,182**	**100%**

The results of table 3 show that the number of drivers in the age group 18-30 is lower than those in the other group. This may indicate that young drivers take the main responsibility for the number and magnitude (severity) of accidents.

Driver's Educational Background

Among the five categories of Educational Background, those drivers with secondary school level of education are responsible for the largest share of injuries (57.4%) and also for the largest number of accidents (57.7%). With regards to the number of injuries per accident, drivers with Elementary School level of education take the liability for the largest mean (.53). The other categories are almost similar to each other.

These results show that, the severity of accidents is higher for elementary school level drivers. But the number of accidents is much higher for drivers having secondary school and above level of education. Even though the number of

drivers in each category of educational background is not available, it can be said that accidents are not necessarily occurred due to lack of knowledge or education.

Driving Experience

When Driving Experience is considered, with regards to the number of injuries and the number of accidents, drivers with 2-5 years of experience take the major share i.e. above 29%. Drivers with 5-10 and above 10 years of experience are having nearer figures when compared to the former ones.

When it comes to the number of injuries per accident the highest mean (.53) is attained by those with less than one year experience and the next highest mean (.50) is attained by those with 1-2 years of experience.

Based on the figures of new licenses issued every year, a rough estimate shows that more than 30% of the drivers fall with in the 5-10 years category. Those with 1-2 years experience are 24%. Therefore it may not be surprising that the maximum number of accidents is attained by these two groups.

B. <u>Vehicle Related Variable</u>

Type of Vehicle

This variable has seven categories. Among the seven categories, Automobiles and Taxies are responsible for the largest number of injuries with values 26.8% and 26% respectively. The maximum number of accidents is due to automobiles which is 29.6% followed by that of taxies and cargo up to 100Q, about 23% each.

For better comparison of the rate of accidents by automobiles (private cars) and taxies, we need to consider the total number of vehicles and the average distance (in Km) traveled by the vehicles per day. According to the National Road Safefty Coordination Office, taxies travel 200Km while automobiles travel 50Km per day. Concerning the number of vehicles in each category, figures of 1997E.C are considered.

Accordingly the results are shown in Table 3 below.

Table 4: Comparison of the Rate of Accidents between Taxi and Automobiles (private cars)

No	Vehicle type	Number of Vehicles (A)	KM per Day (B)	Total Distance Per Day (C)	Number of accidents (D)	Number of Injuries (E)	Accidents per vehicle (D/C)	Injuries Per Vehicle (E/C)
1	Automobile	66950	50	3347500	721	349	0.022%	0.010%
2	Taxi	14138	200	2827600	918	359	0.032%	0.013%

From Table 2, it can be observed that the rate of accident for taxies is higher than that of automobiles though the number of automobiles is more than 4 times the number of taxies or the kilometers traveled by automobiles are greater than those by taxies. A test of proportions for the two types of vehicles was also carried out using Minitab Software. The results are shown below.

Test and CI for Two Proportions (for Number of Accidents)

```
Sample      X      N  Sample p
1         721  3E+06  0.000215
2         918  3E+06  0.000325

Estimate for p(1) - p(2):  -0.000109272
95% CI for p(1) - p(2):  (-0.000135503, -0.0000830419)
Test for p(1) - p(2) = 0 (vs not = 0):  Z = -8.16  P-Value = 0.000
```

Test and CI for Two Proportions (number of Injuries)

```
Sample      X      N  Sample p
1         349  3E+06  0.000104
2         359  3E+06  0.000127

Estimate for p(1) - p(2):  -0.0000227059
95% upper bound for p(1) - p(2):  -0.00000836289
Test for p(1) - p(2) = 0 (vs < 0):  Z = -2.60  P-Value = 0.005
```

Where:

- X= is the number of accidents or injuries
- N = the total kilometers traveled per day by each category (type) of vehicle
- Sample 1 refers to Automobiles
- Sample 2 refers to Taxis

The test result confirms that the rate of accident for taxies is significantly higher than that of automobiles.

When we look at number of injuries per accident, Cargo Above 100Q is responsible for the largest percentage (65%), though it has very little share in both the number of accident and number of injuries. Generally the number of Cargo Trucks Carrying Above 100Quintals comprises not more than 4% of all vehicles. This indicates that the severity of accidents caused by these trucks is higher than those by other vehicles.

C. Location Related Variable

Place of Accident

Among the seven categories of places, both the number of injuries and number of accidents are the highest (54.9%, and 61.2% respectively) for organization. But when it comes to the number of injuries per accident, residential areas are

the places with highest mean (.67) followed by Religious Institutions and Schools with mean 0.58 and 0.57 respectively. The lowest mean of injuries per accident is for Organizations and Hospitals with mean o.39 each.

4.3 Poisson Regression Analysis

As mentioned in section 3.4, Poisson Regression is the selected method for analyzing the accident data.

Only the explanatory variables found significantly associated with the dependant variable as per the chi-squared test result are considered in the Poisson regression analysis. Accordingly five variables, namely, Driver Age (DA), Driver's Educational Background (DE), Driving Experience (EX), and Vehicle Type (VT), Place of Accident (PA) are included in the model

Thus the model is given by:

$$\log(\bar{Y}) = a + \sum_{i=1}^{2} b_{1i} DA_i + \sum_{j=1}^{4} b_{2j} DE_j + \sum_{k=1}^{4} b_{1k} EX_k + \sum_{l=1}^{6} b_{1l} VT_l + \sum_{m=1}^{6} b_{1m} PA_m$$

where

$\bar{Y}$= estimated mean number of injuries per accident

a = Constant

DA_i= Driver Age of level i

DE_j = Driver's Educational Background of level j

EX_k= Driving Experience of level k

VT_l= Vehicle Type of level l

PA_m =Place of Accident of level m

DA_i= 1 for Driver Age of level = i and DA_i= 0 otherwise, for i=1,2

DE_j =1 for Education Level =j and ED_k =0 otherwise, for j=1,2,...,4

EX_k=1 for Experience Level =k and EX_k =0 otherwise, for k=1,2,...,4

VT_l=1 for Vehicle Type Level =l and VT_l=0 otherwise, for l=1,2,...,6

PA_m=1 for Place of Accident Level =m and PA_m =0 otherwise, m=1,2,...,6

The GENMOD procedure of SAS is used to generate coefficient estimates. The complete output is shown on Annex2.

Based on the results of SAS, the regression equation consisting of the above variables is given by

$$\log(\bar{Y}) = -1.1252 + 0.2277\mathbf{DA_1} + \ldots + 0.0750\mathbf{DA_2} + -0.0337\mathbf{DE_1} \ldots + 0.1252\mathbf{DE_4} + 0.1403\mathbf{EX_1} \ldots + -0.0614\mathbf{EX_4} + 0.1162\mathbf{VT_1} + \ldots + -0.0834\mathbf{VT_6} + -0.0475\mathbf{PA_1} + \ldots + 0.3217\mathbf{PA_6} \quad \text{.....(EQ1)}$$

When we look at the Type 1 and Type 3 analysis results of SAS (Annex 7 for description), we observe that Driving Experience and Vehicle type are less significant. Therefore another model was fitted by excluding these variables. The new model is given by:

$$\log(\bar{Y}) = -1.2246 + 0.2406\mathbf{DA_1} + \ldots + 0.0748\mathbf{DA_2} + 0.1298\mathbf{DE_1} \ldots + 0.1656\mathbf{DE_4} + -0.0445\mathbf{PA_1} + \ldots + 0.3354\mathbf{PA_6} \quad \text{.....(EQ2)}$$

Goodness of Fit

It is useful to be able to judge whether a model is a good fit to the data. A useful quantity in judging goodness of fit is the deviance. The saturated model has a total of n parameters. The regression model with fewer variables has k independent variables plus an intercept = k+1 terms. However, if the 'smaller' model, such as EQ2, is good, it should fit almost as well as the saturated one.

In case of the two fitted models, the full model EQ1 has 27 dummy variables plus 1 intercept term which results n=28. The 'smaller' model EQ2 has 15 dummy variables and 1 intercept term which results in K= 15 or K+1=16.

To test the goodness of fit of EQ2, we can use the deviance to compare its goodness of fit with EQ1.

Deviance = 2 x [log Likelihood (EQ2) - log Likelihood (EQ1)] with degrees of freedom = n – (k+1)= 28-16=12.

From the SAS output (annexed), we get

Deviance= 2(-2428.9329 + 2420.1941) = 17.48.

This is not significant for $\chi^2_{(12)}$ with p-value = 0.13242

Thus it can be concluded that there is no statistically significant advantage that can be gained by considering the more complex equation (EQ1). Therefore, EQ2 is selected as the better regression equation.

Further significance of the variables in EQ2 was checked by fitting the Intercept-Only model. The deviance was 2(-2428.93 + 2465.58) =73.29. And this is

significant for $\chi^2_{(15)}$ with p-value much less than .0001. Therefore we can conclude that the variables in EQ2 are significant.

In addition, the interaction effect among pairs of variables was also considered. But the results of Type1 and Type3 analysis (Annex A2.4) show that the interaction effect is not as such significant.

Contrast Estimate

The Contrast Estimate Results table (Annex 2) gives summary of the predicted values of the mean number of injuries per accident.

The highest mean number of injuries per accident (0.9) is attained by drivers in age group 18-30, with Elementary school level of Educational background and location of residential areas. And the lowest mean of injuries (0.28) per accident is attained by drivers in age of 51 and above, have education level above secondary school and the place being around organizational institutions.

This is consistent with the results of the descriptive statistics.

Interpretation of Regression Coefficients

We can compare the difference in impact or risk among the levels of each variable by looking at the regression coefficients. For instance we can consider age group, other variables being controlled.

Age group 1vs 2 = $\exp(-1.2246 + 0.2406) / \exp(-1.2246 + 0.0748) = 1.18$

Age group 1vs 3= $\exp(-1.2246 + 0.2406) / \exp(-1.2246) = 1.27$

Age group 2vs 3 = $\exp(-1.2246 + 0.0748) / \exp(-1.2246) = 1.08$

These figures show that the impact of drivers in age group1 (18-30) is 18% more than those in age group 2 (31-50) while 27% higher than those in age group3 (51 and above). The impact of drivers in age group 2 is only 8% higher than those in age group3.

Other coefficients can be interpreted in the same way.

Outliers and Influential Observations

When we look at the plot of residuals (annex 2.5), relatively few predicted values were found to be outside the “box” [-2, 2].

But they may not be considered as real outliers because often points identified as outliers are simply a reflection of a skewed distribution. In addition (according to Agresti, 1996) when the number of categories is large, it is may be expected for such values to appear by chance.

CHAPTER 5: Conclusion and Recommendation

The results in this study show that the number of injuries per accident is mainly determined by the variables related to drivers. Drivers' Age, Educational background and Place of accident significantly affect the number of injuries per accident.

Drivers who are in the age group of 18-30 are liable for the most of accidents including the sever ones.

Drivers with elementary school level of education take the major responsibility for the increased number of injuries per accident.

With regards to places of accidents, residential areas are where the highest mean of injuries per accident is attained.

The highest mean number of injuries per accident (0.9) is obtained in residential areas by drivers in the age group of 18-30 who have elementary school level of education.

Though the number of taxies is less that one fourth of that of automobiles, the number of accidents is almost equal for the two types of vehicles. The impact of taxies is even much greater than that of automobiles. For instance the empirical results show that the mean of injuries per accident for taxies is .48 while for automobiles it is .39.

It can be learnt from this study that in addition to the efforts being made to reduce the frequency of traffic accidents in general, special attention should be given to reduce the severity of accidents by taking the following into consideration:

- Strict control and management of vehicle movement is necessary especially in residential areas. The same would be required for areas around schools and religious institutions.
- Further studies can be made on the area of traffic accidents by considering detail and accurate information on various variables. For example if the causes and consequences of an accident are recorded in detail instead of broad categories results could be more accurate and efficient.

References

Al-Masaeid *et al*. (2004): *Relationships Between Urban Planning Variables And Traffic Crashes In Damascus*. Road & Transport Research. Retrieved December 8, 2006 from the World Wide Web http://findarticles.com/p/articles/mi_qa3927 /is_200412/ai_n9521917.

Agresti A . (1996): *An introduction to Categorical Data Analysis*. Wiley Series in Probability and Statistics.

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..

Bauer, K. et al (2007): Statistical Models Of Accidents On Interchange Ramps And Speed-Change Lanes. U.S. DOT FHWA. Turner-Fairbanks Highway Research Center. Retrieved December 8, 2006 from the World Wide Web http://www.tfhrc.gov/safety/pubs /97106/index.htm

Council of Addis Ababa City Government (1998): *Road Traffic Safety Regulations of Addis Ababa*, Regulations Number 5/1998.

Federal Democratic Republic of Ethiopia Central Statistics Agency, (2005): *Statistical Abstract*

Geedipally, S.R, (2005): *Analysis Of Traffic Accidents Before And After Resurfacing- A S.tatistical Approach*.Department of Science and Technology. Linkopings Universitet, Seweden

Grace-Martin, K, (2000): *Regression Models for Count Data.* StatNews #43.Office of Statistical Consulting.

Iqbal, J. *et al* (2007): *Market for statisticians in developing economies -The case study of Pakistan's corporate sector.* Department of Statistics University of Karachi and Monash University.

Kopits, E. *et al* (2003): *Traffic Fatalities and Economic Growth*. Policy Research Working Paper. The World Bank Development Research Group Infrastructure and Environment

Lauren P. et al (2005): *Road Traffic Injuries- Can We Stop A Global Epidemic?* Retrieved December 8, 2006 from the World Wide Web http://www.thedoctorwillseeyounow.com/ articles/other/road_33/.

National Road Safety Coordination Office, (2006): *Overview of the Road safety activities in Ethiopia.* Unpublished paper.

Odero, W. (2004): *Road Traffic Injury Research In Africa - Context And Priorities.* School of Public Health. Moi University. KENYA.

Robert B. (2000): *Traffic Fatalities And Injuries - Are Reductions The Result Of 'Improvements' In Highway Design Standards?* Centre for Transport Studies Dept. of Civil and Environmental Engineering. Imperial College of Science, Technology and Medicine. London.

Safecarguide (2004): Retrieved January 22, 2007 from the World Wide Web http://www. safecarguide.com/exp/intro/idx.htm.

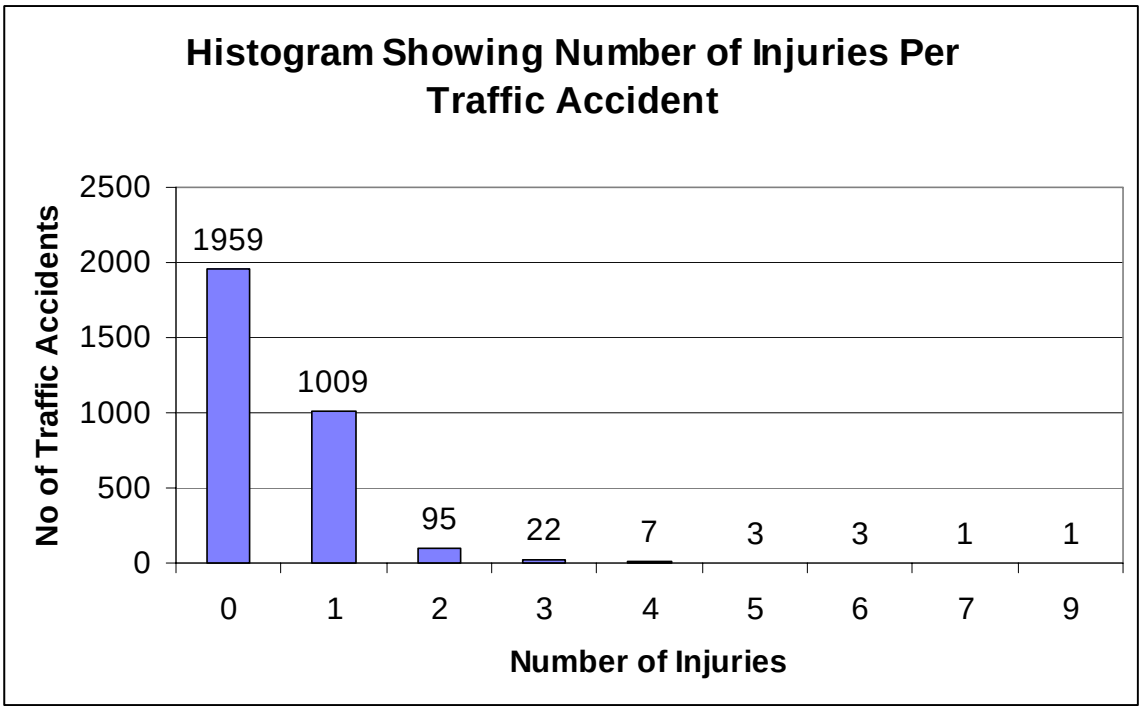
Simon S. (2007) retrieved January 22, 2007 from the World Wide Wed http://www.childrens-mercy.org/stats/definitions/poisson.htm,

The National Roads Authority: HIGH ACCIDENT LOCATIONS. Retrieved January 22, 2007 from the World Wide Web www.nra.ie/PublicationsResources/DownloadableDocumentation/ RoadSafety/file.

Tesema,T. *et al* (2005): *Rule Mining And Classification Of Road Traffic Accidents Using Adaptive Regression Trees.* International Journal of Simulation Systems. Science & Technology Special Issue on Soft Computing for Modeling and Simulation. Volume 6, Number 10-11.

Wikipedia: The Free Encyclopedia. Retrieved December 10, 2006 from the World Wide Web www.wikipedia.org.

Annex 1: Histogram and Descriptive Statistics



Descriptive Statistics on Number of Injuries	
Mean	0.432903
Standard Error	0.012205
Sample Variance	0.461776
Skewness	2.877051
Range	9
Minimum	0
Maximum	9

Annex 2: SAS Output

A2.1 Poisson Regression Full Model

The GENMOD Procedure

Model Information

Data Set	WORK.INJURIESANALYSIS
Distribution	Poisson
Link Function	Log
Dependent Variable	Injuries
Observations Used	3100

Class Level Information

Class	Levels	Values
DA	3	1 2 3
DE	5	1 2 3 4 5
EX	5	1 2 3 4 5
VT	7	1 2 3 4 5 6 7
PA	7	1 2 3 4 5 6 7

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	3077	2822.0127	0.9171
Scaled Deviance	3077	2822.0127	0.9171
Pearson Chi-Square	3077	3273.0950	1.0637
Scaled Pearson X2	3077	3273.0950	1.0637
Log Likelihood		-2420.1941	

Algorithm converged.

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	-1.1252	0.5833	-2.2685	0.0181	3.72	0.0537
DA	1	1	0.2277	0.1077	0.0167	0.4388	4.47	0.0344
DA	2	1	0.0750	0.1023	-0.1254	0.2754	0.54	0.4631
DA	3	0	0.0000	0.0000	0.0000	0.0000	.	.
DE	1	1	-0.0337	0.3924	-0.8029	0.7355	0.01	0.9316
DE	2	1	0.2926	0.1069	0.0831	0.5021	7.49	0.0062
DE	3	1	0.1371	0.1107	-0.0799	0.3541	1.53	0.2155
DE	4	1	0.1252	0.0858	-0.0430	0.2933	2.13	0.1446
DE	5	0	0.0000	0.0000	0.0000	0.0000	.	.
EX	1	1	0.1403	0.1348	-0.1239	0.4044	1.08	0.2980
EX	2	1	0.0828	0.0977	-0.1087	0.2743	0.72	0.3967
EX	3	1	-0.0620	0.0838	-0.2262	0.1022	0.55	0.4592
EX	4	1	-0.0614	0.0819	-0.2219	0.0990	0.56	0.4530
EX	5	0	0.0000	0.0000	0.0000	0.0000	.	.
VT	1	1	0.1162	0.5126	-0.8884	1.1208	0.05	0.8206
VT	2	1	0.0219	0.5094	-0.9766	1.0204	0.00	0.9657
VT	3	1	-0.0432	0.5094	-1.0417	0.9552	0.01	0.9324
VT	4	1	0.4159	0.5466	-0.6554	1.4873	0.58	0.4467
VT	5	1	-0.1852	0.5137	-1.1921	0.8217	0.13	0.7185
VT	6	1	-0.0834	0.5100	-1.0830	0.9162	0.03	0.8701
VT	7	0	0.0000	0.0000	0.0000	0.0000	.	.
PA	1	1	-0.0475	0.2614	-0.5599	0.4648	0.03	0.8557
PA	2	1	0.5075	0.2711	-0.0239	1.0389	3.50	0.0612
PA	3	1	0.0543	0.2708	-0.4765	0.5851	0.04	0.8412
PA	4	1	0.3453	0.2798	-0.2030	0.8936	1.52	0.2171
PA	5	1	0.0338	0.2757	-0.5066	0.5742	0.02	0.9024
PA	6	1	0.3217	0.2853	-0.2375	0.8810	1.27	0.2595
PA	7	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

LR Statistics For Type 1 Analysis

Source	Deviance	DF	Chi-Square	Pr > ChiSq
Intercept	2912.7795			
DA	2901.4157	2	11.36	0.0034
DE	2886.2603	4	15.16	0.0044
EX	2880.4650	4	5.80	0.2150
VT	2868.4862	6	11.98	0.0624
PA	2822.0127	6	46.47	<.0001

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
DA	2	7.49	0.0237
DE	4	7.90	0.0955
EX	4	5.26	0.2620
VT	6	11.97	0.0626
PA	6	46.47	<.0001

A2.2 Reduced Poisson Regression Model and Contrast Estimate Results

The GENMOD Procedure

Model Information

Data Set	WORK.INJURIESANALYSIS
Distribution	Poisson
Link Function	Log
Dependent Variable	Injuries
Observations Used	3100

Class Level Information

Class	Levels	Values
DA	3	1 2 3
DE	5	1 2 3 4 5
PA	7	1 2 3 4 5 6 7

Parameter Information

Parameter	Effect	DA	DE	PA
Prm1	Intercept			
Prm2	DA	1		
Prm3	DA	2		
Prm4	DA	3		
Prm5	DE		1	
Prm6	DE		2	
Prm7	DE		3	
Prm8	DE		4	
Prm9	DE		5	
Prm10	PA			1
Prm11	PA			2
Prm12	PA			3
Prm13	PA			4
Prm14	PA			5
Prm15	PA			6
Prm16	PA			7

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	3087	2839.4902	0.9198
Scaled Deviance	3087	2839.4902	0.9198
Pearson Chi-Square	3087	3307.7052	1.0715
Scaled Pearson X2	3087	3307.7052	1.0715
Log Likelihood		-2428.9329	

Algorithm converged.

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	-1.2246	0.2787	-1.7708	-0.6783	19.30	<.0001
DA	1	1	0.2406	0.0989	0.0468	0.4345	5.92	0.0150
DA	2	1	0.0748	0.1004	-0.1219	0.2715	0.56	0.4560
DA	3	0	0.0000	0.0000	0.0000	0.0000	.	.
DE	1	1	0.1298	0.3871	-0.6289	0.8884	0.11	0.7374
DE	2	1	0.3713	0.1020	0.1713	0.5713	13.24	0.0003
DE	3	1	0.2045	0.1066	-0.0044	0.4135	3.68	0.0550
DE	4	1	0.1656	0.0829	0.0032	0.3281	4.00	0.0456
DE	5	0	0.0000	0.0000	0.0000	0.0000	.	.
PA	1	1	-0.0445	0.2610	-0.5561	0.4671	0.03	0.8646
PA	2	1	0.5058	0.2707	-0.0248	1.0363	3.49	0.0617
PA	3	1	0.0532	0.2700	-0.4761	0.5825	0.04	0.8437
PA	4	1	0.3462	0.2793	-0.2013	0.8937	1.54	0.2153
PA	5	1	0.0137	0.2753	-0.5258	0.5533	0.00	0.9602
PA	6	1	0.3354	0.2842	-0.2217	0.8924	1.39	0.2380
PA	7	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

LR Statistics For Type 1 Analysis

Source	Deviance	DF	Chi-Square	Pr > ChiSq
Intercept	2912.7795			
DA	2901.4157	2	11.36	0.0034
DE	2886.2603	4	15.16	0.0044
PA	2839.4902	6	46.77	<.0001

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
DA	2	11.14	0.0038
DE	4	13.50	0.0091
PA	6	46.77	<.0001

Contrast Estimate Results (SAS)

Driver's Age	Educational Back.	Place of Accident						
		Org.	**Res.**	**Mar.**	**Relg.**	**Ent.**	**Sch.**	**Hos.**
18-30	Basic	0.41	0.71	0.45	0.60	0.43	0.60	0.43
	Elementary	0.52	0.90	0.57	0.77	0.55	0.76	0.54
	Junior	0.44	0.76	0.48	0.65	0.47	0.64	0.46
	Secondary	0.42	0.73	0.47	0.62	0.45	0.62	0.44
	Above Sec	0.36	0.62	0.39	0.53	0.38	0.52	0.37
	Max	**0.52**	**0.90**	**0.57**	**0.77**	**0.55**	**0.76**	**0.54**
	Min	**0.36**	**0.62**	**0.39**	**0.53**	**0.38**	**0.52**	**0.37**
31-50	Basic	0.34	0.60	0.38	0.51	0.51	0.50	0.36
	Elementary	0.44	0.76	0.48	0.65	0.47	0.64	0.46
	Junior	0.37	0.64	0.41	0.55	0.39	0.54	0.39
	Secondary	0.36	0.62	0.39	0.53	0.38	0.52	0.37
	Above Sec	0.30	0.53	0.33	0.45	0.32	0.44	0.32
	Max	**0.44**	**0.76**	**0.48**	**0.65**	**0.51**	**0.64**	**0.46**
	Min	**0.30**	**0.53**	**0.33**	**0.45**	**0.32**	**0.44**	**0.32**
51 and Above	Basic	0.32	0.55	0.35	0.47	0.30	0.47	0.33
	Elementary	0.41	0.71	0.45	0.60	0.43	0.60	0.43
	Junior	0.34	0.60	0.38	0.51	0.37	0.50	0.36
	Secondary	0.33	0.58	0.37	0.49	0.35	0.49	0.35
	Above Sec	0.28	0.49	0.31	0.42	0.30	0.41	0.29
	Max	**0.41**	**0.71**	**0.45**	**0.60**	**0.43**	**0.60**	**0.43**
	Min	**0.28**	**0.49**	**0.31**	**0.42**	**0.30**	**0.41**	**0.29**

A2.3 Values for Goodness of Fit for the Saturated Negative Binomial Regression Model

The GENMOD Procedure

Model Information

Data Set	WORK.INJURIESANALYSIS
Distribution	Negative Binomial
Link Function	Log
Dependent Variable	Injuries
Observations Used	3100

Class Level Information

Class	Levels	Values
DA	3	1 2 3
DE	5	1 2 3 4 5
EX	5	1 2 3 4 5
VT	7	1 2 3 4 5 6 7
PA	7	1 2 3 4 5 6 7

Criteria For Assessing Goodness Of Fit

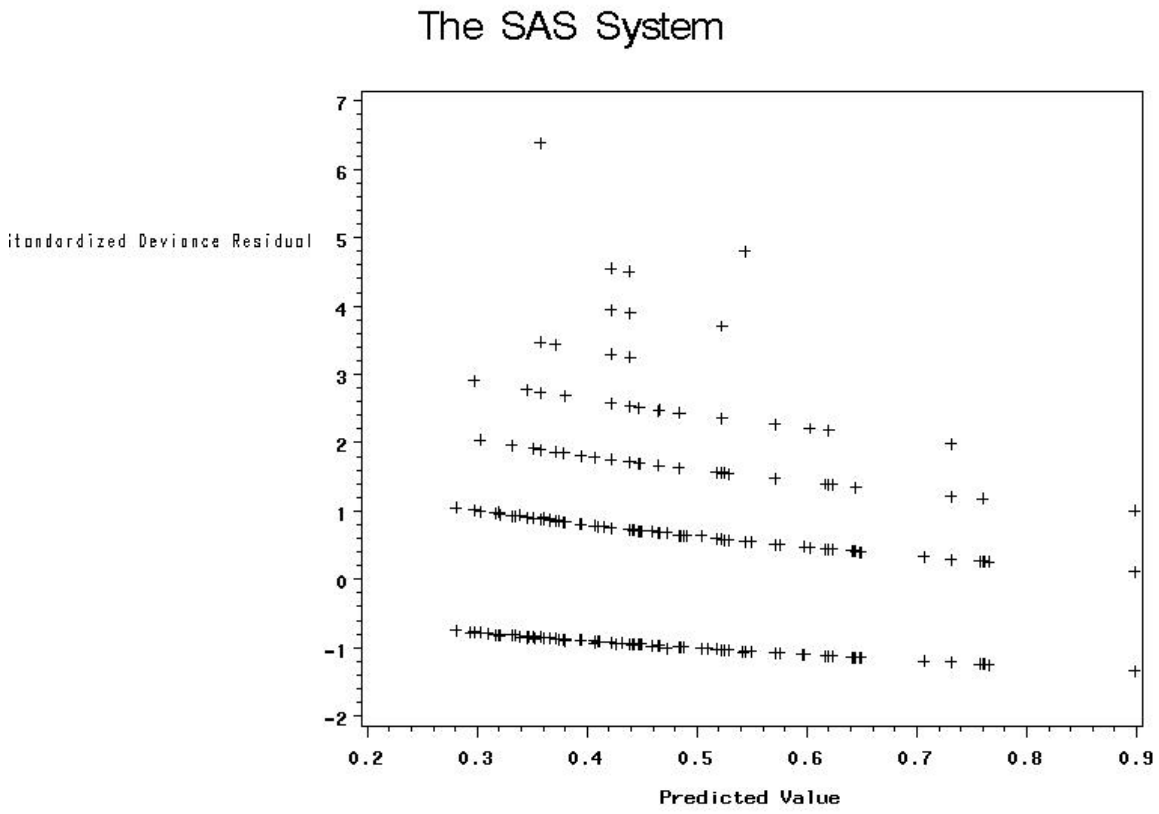
Criterion	DF	Value	Value/DF
Deviance	3077	2771.1614	0.9006
Scaled Deviance	3077	2771.1614	0.9006
Pearson Chi-Square	3077	3221.2397	1.0469
Scaled Pearson X2	3077	3221.2397	1.0469
Log Likelihood		-2419.7482	

A2.4 Likelihood Ratio Test for Reduced Poisson Regression Model Including Interaction Effect

LR Statistics For Type 1 Analysis

Source	Deviance	DF	Chi-Square	Pr > ChiSq
Intercept	2912.7795			
DrvAge	2901.4157	2	11.36	0.0034
DrvEdu	2886.2603	4	15.16	0.0044
AccPlace	2839.4902	6	46.77	<.0001
DrvAge*DrvEdu	2833.6230	8	5.87	0.6621
DrvEdu*AccPlace	2808.3425	23	25.28	0.3360
DrvAge*AccPlace	2794.0336	12	14.31	0.2814

A2.5 Plot of Residuals



Annex 3: SAS Codes

```
DATA InjuriesAnalysis;
INFILE 'd:\thesis\DataForFinalThesis.txt';
INPUT DrvSex DrvAge     DrvEdu DrvVehRel DrvExp VehType     VehYears
AccPlace Injuries;
RUN;
proc genmod data=InjuriesAnalysis;
  class DrvAge DrvEdu DrvExp VehType AccPlace;
  model Injuries = DrvAge DrvEdu DrvExp VehType AccPlace/dist = Poisson
                              Type1
                              Type3;
RUN;
```

```
DATA InjuriesAnalysis;
INFILE 'd:\thesis\DataForFinalThesis.txt';
INPUT DrvSex DrvAge     DrvEdu DrvVehRel DrvExp VehType     VehYears
AccPlace Injuries;
RUN;
proc genmod data=InjuriesAnalysis;
      class DrvAge DrvEdu DrvExp VehType AccPlace;
      model Injuries = DrvAge DrvEdu DrvExp VehType AccPlace/ dist = nb
                              Type1
                              Type3;
RUN;
```

```
DATA InjuriesAnalysis;
INFILE 'd:\thesis\DataForFinalThesis.txt';
INPUT DrvSex DrvAge DrvEdu DrvVehRel DrvExp VehType VehYears AccPlace
Injuries;
RUN;
proc genmod data=InjuriesAnalysis;
      class DrvAge DrvEdu AccPlace;
      model Injuries = DrvAge DrvEdu AccPlace/ dist = Poisson
                              Type1
                              Type3;
RUN;
```

```
DATA InjuriesAnalysis;
INFILE 'd:\thesis\DataForFinalThesis.txt';
INPUT DrvSex DrvAge     DrvEdu DrvVehRel DrvExp VehType     VehYears
AccPlace Injuries;
RUN;
proc genmod data=InjuriesAnalysis;
  class DrvAge DrvEdu AccPlace;
  model Injuries = DrvAge DrvEdu AccPlace DrvAge*DrvEdu/ dist = Poisson
                              Type1
                              Type3;
RUN;
```

Annex 4: Chi-Square Test Results for Selection of Variables

Driver Sex * Existence of Injuries

Crosstab

			Existence of Injuries		
			No Injury	1 or more injuries	Total
Driver Sex	Male	Count	1887	1112	2999
		Expected Count	1895.2	1103.8	2999.0
	Female	Count	72	29	101
		Expected Count	63.8	37.2	101.0
Total		Count	1959	1141	3100
		Expected Count	1959.0	1141.0	3100.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	2.940(b)	1	.086		
Continuity Correction(a)	2.592	1	.107		
Likelihood Ratio	3.047	1	.081		
Fisher's Exact Test				.094	.052
Linear-by-Linear Association	2.939	1	.086		
N of Valid Cases	3100				

a Computed only for a 2x2 table

b 0 cells (.0%) have expected count less than 5. The minimum expected count is 37.17.

Driver Age * Existence of Injuries

Crosstab

			Existence of Injuries		
			No Injury	1 or more injuries	Total
Driver Age	18-30	Count	865	577	1442
		Expected Count	911.3	530.7	1442.0
	31-50	Count	873	452	1325
		Expected Count	837.3	487.7	1325.0
	51 and Above	Count	221	112	333
		Expected Count	210.4	122.6	333.0
Total		Count	1959	1141	3100
		Expected Count	1959.0	1141.0	3100.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	11.951[a]	2	.003
Likelihood Ratio	11.943	2	.003
Linear-by-Linear Association	10.067	1	.002
N of Valid Cases	3100		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 122.57.

Driver Education * Existence of Injuries

Crosstab

			Existence of Injuries		
			No Injury	1 or more injuries	Total
Driver Education	Basic Education	Count	12	7	19
		Expected Count	12.0	7.0	19.0
	Elementary	Count	221	173	394
		Expected Count	249.0	145.0	394.0
	Junior Secondary	Count	239	144	383
		Expected Count	242.0	141.0	383.0
	Secondary	Count	1126	664	1790
		Expected Count	1131.2	658.8	1790.0
	Above Secondary	Count	361	153	514
		Expected Count	324.8	189.2	514.0
Total		Count	1959	1141	3100
		Expected Count	1959.0	1141.0	3100.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	19.664[a]	4	.001
Likelihood Ratio	19.818	4	.001
Linear-by-Linear Association	15.710	1	.000
N of Valid Cases	3100		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 6.99.

Vehicle Driver Relation * Existence of Injuries

Crosstab

			Existence of Injuries		
			No Injury	1 or more injuries	Total
Vehicle Driver Relation	Owner	Count	287	176	463
		Expected Count	292.6	170.4	463.0
	Hired	Count	1412	805	2217
		Expected Count	1401.0	816.0	2217.0
	Other	Count	260	160	420
		Expected Count	265.4	154.6	420.0
Total		Count	1959	1141	3100
		Expected Count	1959.0	1141.0	3100.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	.824[a]	2	.662
Likelihood Ratio	.822	2	.663
Linear-by-Linear Association	.000	1	.990
N of Valid Cases	3100		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 154.59.

Driving Experience * Existence of Injuries

Crosstab

			Existence of Injuries		
			No Injury	1 or more injuries	Total
Driving Experience	Less than 1 year	Count	72	65	137
		Expected Count	86.6	50.4	137.0
	1-2 Years	Count	254	172	426
		Expected Count	269.2	156.8	426.0
	2-5 Years	Count	582	342	924
		Expected Count	583.9	340.1	924.0
	5-10 Years	Count	578	306	884
		Expected Count	558.6	325.4	884.0
	Above 10 Years	Count	473	256	729
		Expected Count	460.7	268.3	729.0
Total		Count	1959	1141	3100
		Expected Count	1959.0	1141.0	3100.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	11.736[a]	4	.019
Likelihood Ratio	11.524	4	.021
Linear-by-Linear Association	8.595	1	.003
N of Valid Cases	3100		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 50.42.

Vehicle Type * Existence of Injuries

Crosstab

			Existence of Injuries		
			No Injury	1 or more injuries	Total
Vehicle Type	Bus	Count	180	109	289
		Expected Count	182.6	106.4	289.0
	Taxi	Count	422	299	721
		Expected Count	455.6	265.4	721.0
	Cargo Upto 100 Q	Count	443	276	719
		Expected Count	454.4	264.6	719.0
	Cargo Above 100 Q	Count	19	18	37
		Expected Count	23.4	13.6	37.0
	Station Wagon	Count	284	124	408
		Expected Count	257.8	150.2	408.0
	Automobile	Count	607	311	918
		Expected Count	580.1	337.9	918.0
	Liquid Cargo	Count	4	4	8
		Expected Count	5.1	2.9	8.0
Total		Count	1959	1141	3100
		Expected Count	1959.0	1141.0	3100.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	21.048[a]	6	.002
Likelihood Ratio	21.088	6	.002
Linear-by-Linear Association	11.266	1	.001
N of Valid Cases	3100		

a. 1 cells (7.1%) have expected count less than 5. The minimum expected count is 2.94.

Vehicle Service Years * Existence of Injuries

Crosstab

			Existence of Injuries		
			No Injury	1 or more injuries	Total
Vehicle Service Years	Less than 1 year	Count	122	54	176
		Expected Count	111.2	64.8	176.0
	1-2 Years	Count	55	33	88
		Expected Count	55.6	32.4	88.0
	2-5 Years	Count	282	173	455
		Expected Count	287.5	167.5	455.0
	5-10 Years	Count	898	503	1401
		Expected Count	885.3	515.7	1401.0
	Above 10 Years	Count	602	378	980
		Expected Count	619.3	360.7	980.0
Total		Count	1959	1141	3100
		Expected Count	1959.0	1141.0	3100.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	4.950[a]	4	.292
Likelihood Ratio	5.018	4	.285
Linear-by-Linear Association	2.352	1	.125
N of Valid Cases	3100		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 32.39.

Place of Accident * Existence of Injuries

Crosstab

			Existence of Injuries		
			No Injury	1 or more injuries	Total
Place of Accident	Organization	Count	1279	618	1897
		Expected Count	1198.8	698.2	1897.0
	Residence	Count	96	134	230
		Expected Count	145.3	84.7	230.0
	Market	Count	239	142	381
		Expected Count	240.8	140.2	381.0
	Religious Place	Count	73	80	153
		Expected Count	96.7	56.3	153.0
	Entertainment	Count	180	95	275
		Expected Count	173.8	101.2	275.0
	School	Count	69	57	126
		Expected Count	79.6	46.4	126.0
	Hospital	Count	23	15	38
		Expected Count	24.0	14.0	38.0
Total		Count	1959	1141	3100
		Expected Count	1959.0	1141.0	3100.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	80.473[a]	6	.000
Likelihood Ratio	77.948	6	.000
Linear-by-Linear Association	12.971	1	.000
N of Valid Cases	3100		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 13.99.

Annex 5: SPSS Output Descriptive Statistics

Number of Injuries, Number of Accidents and Mean of Injuries per Accident

Variable	Levels	Number of Injuries	% of Total Injuries	Total Number or Accidents	% of Total Accidents	Mean of Injuries per Accident
DrvAge	18-30	685	51.0%	1,442	46.5%	0.48
	31-50	531	39.6%	1,325	42.7%	0.40
	51 and Above	126	9.4%	333	10.7%	0.38
	Total	1,342	100.0%	3,100	100.0%	0.43
DrvEdu	Basic Education	7	0.5%	19	0.6%	0.37
	Elementary	207	15.4%	394	12.7%	0.53
	Junior Secondary	175	13.0%	383	12.4%	0.46
	Secondary	770	57.4%	1,790	57.7%	0.43
	Above Secondary	183	13.6%	514	16.6%	0.36
	Total	1,342	100.0%	3,100	100.0%	0.43
DrvExp	Less than 1 year	73	5.4%	137	4.4%	0.53
	1-2 Years	211	15.7%	426	13.7%	0.50
	2-5 Years	391	29.1%	924	29.8%	0.42
	5-10 Years	364	27.1%	884	28.5%	0.41
	Above 10 Years	303	22.6%	729	23.5%	0.42
	Total	1,342	100.0%	3,100	100.0%	0.43
VehType	Bus	145	10.8%	289	9.3%	0.50
	Taxi	349	26.0%	721	23.3%	0.48
	Cargo Upto 100 Q	316	23.5%	719	23.2%	0.44
	Cargo Above 100 Q	24	1.8%	37	1.2%	0.65
	Station Wagon	145	10.8%	408	13.2%	0.36
	Automobile	359	26.8%	918	29.6%	0.39
	Liquid Cargo	4	0.3%	8	0.3%	0.50
	Total	1,342	100.0%	3,100	100.0%	0.43
AccPlace	Organization	737	54.9%	1,897	61.2%	0.39
	Residence	155	11.5%	230	7.4%	0.67
	Market	163	12.1%	381	12.3%	0.43
	Religious Place	89	6.6%	153	4.9%	0.58
	Entertainment	111	8.3%	275	8.9%	0.40
	School	72	5.4%	126	4.1%	0.57
	Hospital	15	1.1%	38	1.2%	0.39
	Total	1,342	100.0%	3,100	100.0%	0.43

Annex 6: Number of Registered Vehicles by Fiscal Year and Type of License

No	Type of License	Fiscal Year (E.C.) *										
		1988	**1989**	**1990**	**1991**	**1992**	**1993**	**1994**	**1995**	**1996**	**1997****	**1998****
1	Private Cars	36167	39001	40608	40611	41985	43770	47362	53540	58696	66950	76486
2	Taxi	6595	6524	9847	9598	9858	10325	11571	12506	12395	14138	16152
3	Commercial	28990	30315	34033	34615	39122	42724	44647	43176	50211	57272	65430
4	Government	13510	14239	12983	15356	15573	15750	16165	17070	17424	19874	22705
5	Others	6195	6423	5409	5670	5684	6163	6154	6646	7081	8077	9227
		91457	**96502**	**102880**	**105850**	**112222**	**118732**	**125899**	**132938**	**145807**	**166310**	**190000**
											14.062%	**14.244%**

Source: Annual Statistical Bulletin, 1996 F.Y, Ministry of Infrastructure

*This figures consists of the number of vehicles all over the country. The share of Addis Ababa is 60-70 percent of these figures except for Taxis which may be more these percentages.

** Only the total number of vehicles is available for the years 1997 and 1998 and was collected from the Planning and Research Department of Transport Authority. The number of vehicles by type of license is calculated by the respective rate of growth of each year i.e. 14.062% and 14.244% respectively.

Annex 7: About Type 1 and Type 3 Analysis of SAS

From: http://v8doc.sas.com/sashtml/stat/chap29/sect30.htm, Accessed on May 14, 2007

The GENMOD Procedure

Type 1 Analysis

A Type 1 analysis consists of fitting a sequence of models, beginning with a simple model with only an intercept term, and continuing through a model of specified complexity, fitting one additional effect on each step. Likelihood ratio statistics, that is, twice the difference of the log likelihoods, are computed between successive models. This type of analysis is sometimes called an analysis of deviance since, if the dispersion parameter is held fixed for all models, it is equivalent to computing differences of scaled deviances. The asymptotic distribution of the likelihood ratio statistics, under the hypothesis that the additional parameters included in the model are equal to 0, is a chi-square with degrees of freedom equal to the difference in the number of parameters estimated in the successive models. Thus, these statistics can be used in a test of hypothesis of the significance of each additional term fit.

This type of analysis is not available for GEE models, since the deviance is not computed for this type of model.

If the dispersion parameter ϕ is known, it can be included in the models; if it is unknown, there are two strategies allowed by PROC GENMOD. The dispersion parameter can be estimated from a maximal model by the deviance or Pearson's chi-square divided by degrees of freedom, as discussed in the "Goodness of Fit" section, and this value can be used in all models. An alternative is to consider the dispersion to be an additional unknown parameter for each model and estimate it by maximum likelihood on each step. By default, PROC GENMOD estimates scale by maximum likelihood at each step.

A table of likelihood ratio statistics is produced, along with associated *p*-values based on the asymptotic chi-square distributions.

If you specify either the SCALE=DEVIANCE or the SCALE=PEARSON option in the MODEL statement, the dispersion parameter is estimated using the deviance or Pearson's chi-square statistic, and *F* statistics are computed in addition to the chi-square statistics for assessing the significance of each additional term in the Type 1 analysis. See the section "F Statistics" for a definition of *F* statistics.

This Type 1 analysis has the general property that the results depend on the order in which the terms of the model are fitted. The terms are fitted in the order in which they are specified in the MODEL statement.

Type 3 Analysis

A Type 3 analysis is similar to the Type III sums of squares used in PROC GLM, except that likelihood ratios are used instead of sums of squares. First, a Type III estimable function is defined for an effect of interest in exactly the same way as in PROC GLM. Then, maximum likelihood estimation is performed under the constraint that the Type III function of the parameters is equal to 0, using constrained optimization. Let the resulting constrained parameter estimates be $\tilde{\beta}$ and the log likelihood be $l(\tilde{\beta})$ .Then the likelihood ratio statistic

$$S = 2(l(\hat{\beta}) - l(\tilde{\beta}))$$

where $\hat{\beta}$ is the unconstrained estimate, has an asymptotic chi-square distribution under the hypothesis that the Type III contrast is equal to 0, with degrees of freedom equal to the number of parameters associated with the effect.

When a Type 3 analysis is requested, PROC GENMOD produces a table that contains the likelihood ratio statistics, degrees of freedom, and *p*-values based on the limiting chi-square distributions for each effect in the model. If you specify either the DSCALE or PSCALE option in the MODEL statement, *F* statistics are also computed for each effect.

Options for handling the dispersion parameter are the same as for a Type 1 analysis. The dispersion parameter can be specified to be a known value, estimated from the deviance or Pearson's chi-square divided by degrees of freedom, or estimated by maximum likelihood individually for the unconstrained and constrained models. By default, PROC GENMOD estimates scale by maximum likelihood for each model fit.

The results of this type of analysis do not depend on the order in which the terms are specified in the MODEL statement.

A Type 3 analysis can consume considerable computation time since a constrained model is fitted for each effect. Wald statistics for Type 3 contrasts are computed if you specify the WALD option. Wald statistics for contrasts use less computation time than likelihood ratio statistics but may be less accurate indicators of the significance of the effect of

interest. The Wald statistic for testing $\mathbf{L}'\beta = \mathbf{0}$,where **L** is the contrast matrix, is defined by

$$S = (\mathbf{L}'\hat{\beta})'(\mathbf{L}'\hat{\mathbf{\Sigma}}\mathbf{L})^{-}(\mathbf{L}'\hat{\beta})$$

where $\hat{\beta}$ is the maximum likelihood estimate and $\hat{\mathbf{\Sigma}}$ is its estimated covariance matrix. The asymptotic distribution of *S* is chi-square with *r* degrees of freedom, where *r* is the rank of **L**.

DECLARATION

I, the undersigned, declare that the thesis is my original work, has not been presented for degrees in any other university and all sources of material used for the thesis have been duly acknowledged.

Name: Tewolde Mekonnen
Signature: ……
Place: Faculty of Science, Addis Ababa University
Date: August 2007

This thesis has been submitted for examination with my approval as a University Advisor.

……..
Dr. Butte Gotu