ADDIS ABABA UNIVERSITY

GRADUATE STUDY PROGRAMM

DEPARTMENT OF STATISTICS

COUNT REGRESSION MODELS OF HUMAN DEATHBY ROAD TRAFFIC ACCIDENTSIN ADDIS ABABA, ETHIOPIA

Getahun Worku

Advisor: Dejen Tesfaw(PhD)

June, 2015

Addis Ababa, Ethiopia

ADDIS ABABA UNIVERSITY

GRADUATE STUDIES PROGRAMME

DEPARTMENT OF STATISTICS

COUNT REGRESSION MODELS OF HUMAN DEATH BY ROAD TRAFFIC ACCIDENTS IN ADDIS ABABA, ETHIOPIA

Getahun Worku

A THESIS SUBMITTED TO

THE DEPARTMENT OF STATISTICS

PRESENTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN STATISTICS (APPLIED STATISTICS)

Advisor: Dejen Tesfaw (PhD)

Addis Ababa, Ethiopia

June, 2015

Addis Ababa University

School of Graduate Studies

This is to certify that the thesis prepared by *Getahun Worku*, entitled: ***Count Regression Models of Human Death by Road Traffic Accidents in Addis Ababa, Ethiopia*** and submitted in partial fulfillment of the requirements for the Degree of Master of Science in Statistics (Applied Statistics) complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the Examining Committee:

Advisor: Dejen Tesfaw (PhD) Signature__________ Date _____________

Examiner: Shibru Temesgen (PhD) Signature__________ Date _____________

Examiner: M.K. Sharma (professor) Signature___________ Date ____________

____

Chair of Department or Graduate Program Coordinator Declaration

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 materials used for the thesis have been duly acknowledged.

Name: Getahun Worku

Signature: ____________________

Place of submission: Department of Statistics, College of Natural Sciences, Addis Ababa University

Date of submission: June, 2015

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

Dejen Tesfaw (PhD) ________________________

Advisor's Name **Signature**

Acknowledgment

First, and foremost, I would like to sincerely thank, my thesis advisor Dejen Tesfaw (PhD) who from the very inception of the idea to its realization has given me his substantive advice, comments and professional assistance.

My sincere special thanks also go to member of Department of Statistics in Addis Ababa University and my course instructors, for their constructive support and ready to help at every time of my stay in the university.

I would like to thank the Addis Ababa Traffic Control and Investigation Department (AATCID), Addis Ababa Police Commission Office and all Sub-city Police Command Office for providing me the traffic accident data.

Finally, my deep appreciation and heart-felt thanks also goes to my friends, families and relatives for all their encouragement and support throughout my study.

Table of Contents

List of Figures

List of Tables

ABSTRACT

Count Regression Models of Human Death by Road Traffic Accidents in Addis Ababa, Ethiopia

Getahun Worku

Addis Ababa University, 2015

Poisson regression model is commonly used to study the association between count outcome variable and covariates. This assumption does not hold true in real dataset due to the presence of overdispersion and excess zeros. While negative binomial regression model can deal with overdispersion, on the other hand, zero-inflated Poisson and zero-inflated negative binomial regression model can be used to handleexcesszeros in the observations. The objective of this study is to identify the most important traffic accident variables using count regression models and identify the best count data models in order to analyze road traffic accident datawhich is obtained from the Addis Ababa Traffic Control and Investigation Department (AATCID) daily basis recorded within 365 consecutive days from July 30, 2013 - July 29, 2014. Based on different model comparison, (using AIC and Vuong test) ZIP regression model provides more appropriate fit to the road traffic accident (the number of human death per RTA) data considered in this study. Month in quarter, road inclination, age of the drivers, vehicle type, ownershipof the vehicle, accident time, accident type and injured vehicle *count were found to be statistically significant predictors of human deathat* $\alpha = 0.05$.

List of Abbreviations

AADT	Annual Average Daily Traffic
AATCID	Addis Ababa Traffic Control and Investigation Department
AD	Access Density
AIC	Akaike information criterion
BIC	Bayesian information criterion
CSA	Central Statistics Agency
DALY	Disability Adjusted Life Years
ECA	Economic Commission for Africa
IRR	Incidence Rate Ratio
LRT	Likelihood-ratio test
LW	Lane Width
RTA	Road Traffic Accidents
RTABS	road traffic accident black spot
RTI	Road Traffic Injury
SS	Spot Speed
SW	Shoulder Width
UN	African Union the
WHO	World Health Organization

Chapter 1
Introduction

1.1 Background of the Study

Since the 1960s, the increasing growth of traffic density has been a major challenge for road traffic managements, not only for dealing with the lasting congestion problem but also for addressing road traffic safety related issues. Although, in the past few decades, significant progress has been made to reduce the number of fatal casualties in road traffic by identifying and rectifying flaws inroad design along with improvements in car safety and traffic management.

The Global status on road safety 2013presented information from 182 countries, accounting for almost 99% of the world's population. The report indicated that the total number of road traffic deaths in the world was 1.24 million per year. Only 28 countries covered 7% of the world's population because they have comprehensive road safety laws on five key risk factors: drinking and driving, speeding, motorcycle helmets, seat belts, and child restraints (WHO, 2013).

Road traffic injury (RTI) ishigh contributor to regional and global disease burden. Although there is an accumulated body of knowledge concerning how this burden may be reduced through appropriate road safety measures, little is known about the relative costeffectiveness of such measures. Particularly, in the context of low and middleincome countries, the road use and injury patterns are different from those highincome countries(Chisholm, 2008).Residents of developing countries are at higher risk of RTI than highincome countries. They are also at high risk of death when a crash occurs and unable to care for crash victims. Unless action is taken to improve road traffic safety, poor countries will continue to bear the heavy toll of road traffic injuries (Lauren and Hill, 2005).

According to Tewolde (2007) study in Addis Ababa shown that among the 3,100 records which are collected within 310 consecutive days (04/03/06 – 07/01/07) from Addis Ababa Traffic Control and Investigation Department (AATCID), 1,141 (37%) involved human injuries. The rest are considered as accidents with material damage only. The highest mean number of injuries per accident (0.9) is obtained in residential areas bydrivers in the age group of 18-30 and who have elementary school level of education.Drivers with elementary school level of education take the major responsibility for the increased number of injuries per accident.

Roadtraffic accident is the result of multiplicity of factorsthat leads to the occurrence of accident. Among the factors human related factors,environmental related factors, vehicle related factors and road related factors are considered. Road traffic accidents and their resulting fatalities may be regarded as social and economic problem, especially in developing countries like Ethiopia, where the resources are limited (Tewolde, 2007).

1.2 Statement of Problem

Road traffic accidents are major public safety and development obstacle. According to (WHO, 2011), the current situation required high level of political dedication and took immediate actionto reduce RTA.

Different researchers have investigated the suitability of the Poisson regression model to predict accident frequencies at intersections or roadways (Jones*et al.*,1991; Miaou and Lum, 1993; Tewolde, 2007;Malyshkina*et al.*,2009). These researches have foregrounded the fact that because accident occurrences are necessarily discrete, often discontinuous and more likely random events, it is better to use Poisson regression models than multiple linear regression models. One important assumption in the Poisson regression model is that the mean of the distribution must be equal to the variance (i.e. if the response random variable is denoted, Y, then, $E(y_i) = var(y_i)$). If this assumption is not valid, then the standard errors, usually estimated by the maximum likelihood (ML) method, will be biased and the test statistics derived from the model will be incorrect

(Yaacob *et. al.*, 2010). Therefore, the traffic accident data is found to be significantly overdispersed, i.e., the variance was much greater than the mean, suggesting that the likelihood of accident occurrence might be incorrectly estimated.

To overcome the overdispersion problem, it is important to use negative binomial (NB)model instead of Poisson model.NB model is more suitable in describing discrete and non-negative events. If there exist excesszero counts in annual traffic accident data, it had better to reflect the situation,a dual – statesystem may be assumed.

1. Zero accident state, where the intersections can be regarded as almost safe;
2. Non-zero accident state (unsafe conditions), where the accident frequencies are assumed to follow some known distributions such as the Poisson and negative binomial.

Poisson and negative binomial regression models are used only for non–zero traffic accident state(unsafe conditions). Due to the existence of zero-traffic accident state (excess zeros in accident data), the estimated model will be biased, which leads wrong estimation regarded as evidence of overdispersion. To handle count data with excess zeroes, Lambert (1992) deals about zero-inflated Poisson (ZIP) regression models that the probability of a perfect state (i.e., zero-defect state) and the mean of the imperfect state (i.e., non-zero defect state) depend on the covariates.

The purpose of this study is to evaluate the model and to determine the suitability of applying different count models (Poisson and negative binomial and their Zero-inflated regression models) in road traffic accident data (the number of person death due to road traffic accident) of Addis Ababa, Ethiopia.

1.3 Objectives of the Study

The general objective of this study is to identify the most important human death per road traffic accident variables using countregression models. Thestudy is also used to identifythe best count regression modelin order to analyze human death perroad traffic accident data in Addis Ababa, Ethiopia.

The specific objectives of this study are:

1. To check whether the road traffic accident data contains overdispersion and/or zero-inflation cases.
2. To considerappropriate count regression modelin order to analyze road traffic accident data in Addis Ababa.

1.5 Significance of the Study

This study is mainly concerned with human death per road traffic accidents in Addis Ababain order to identify factors related to drivers, road environments, condition of vehicles and other factors. Therefore, the significant of this study is:

1. To provide clear information to researcher how to use overdispersed and zero-inflated regression models.
2. The research outcome may help policy makers in order to design appropriate strategies to reduce traffic accidents.
3. This study can be a baseline for the other interested finders (researchers) for further investigation on road traffic accident.

1.6 Limitation of the Study

The data obtained from Addis Ababa Police Commission Office, Addis Ababa Traffic Control and Investigation Department (AATCID) related to vehicle service years and driving experience were collected by interviewing the drivers themselves. The daily recorded data was available by hardcopy (paper writing by hand) in all sub-cities of Addis Ababa, even from commissioner office, as the result of this condition, there are some challenges such as scarcity of time, financial and human labor faced at the time data collection.

There are also other challenges faced during data collection.

1. Some information like driver's age, driving experience, driver's sex obtained from the drivers was missed due to the drivers who were cause to accident escaped.
2. The injured person recorded under serous injured, but after some time of stayed in hospital s/he might be died and did not report the condition to police.
3. Property damage injury typewas not included in this study due to the following reasons:
   - Therewas no possibility to human death event occurred. Theaccident occurred only on property (vehicle) like a little crash of vehicle to vehicle recorded as property damage or type.
   - The shortage of time and human labor was a big challenge to collect all the accident data (property damage injury type).

1.7 Definition of Terms

Traffic Accident:-pertain to any vehicle accidents occurred on public high way (i.e. originating on, terminating on, or involving a vehicle partially).

Road Traffic Accident:- is defined as rare, random and multiple factor events always precede by a situation in which one or more road users have failed to cope up with the road and its environment.

Accident Frequency:- is a measure of the relative magnitude of a specific type of accident in relation to all accidents for a specified time period such as one or three years.

Accident Rate:- is a measure of the rate of occurrence of accidents in relation to any one of a number of exposure factors, such as per million vehicles entering or per million km in a study area.

Property Damage:- means physical damages of vehicles or properties but not to persons.

Light Injury:-is a traffic casualty, which is minor injury not requiring hospital admission as an in-patient.

Heavy Injury:- isthe injury that require admission to hospital as an in-patient.

Fatality Accident:**-**an occurrence of death by accident (one or more persons die within 30 days of the accident (WHO) although some countries use shorter periods).

Human factors:-are factors related to drivers and other road users that may contribute to a collision. Like driver behavior, visual and auditory acuity, decisionmaking ability, and reaction speed.

Junction:- a place at which two or more roads meet, whatever the angle of the axes of the roads (including roundabouts), or within 20 meters of such a place.

Chapter 2

Literature Review

Road traffic accident is a major public health challenge that requires concerted efforts for effective and sustainable prevention. Of all the systems with which people have to deal with every day, road traffic systems are the most complex and the most dangerous problem. Worldwide, an estimated1.24 million people are killed in road crashes each year and Tens of millions of people are injured or disabled every year. Every day around the world more than 3400 people die from road traffic injury. Low-income and middle-income countries account for about 90% of the deaths even though these countries have only about half the world's vehicles. Without take appropriate action, annual road traffic deaths are predicted to increase to around 1.9 million by 2030 and to become the seventh leading cause of death. Road traffic injuries are the leading cause of death globally among people aged 15–29 years and for comparisons between the continents Africa has the highest contribution to road traffic accident which is 24.1 fatality deaths per 100,000 populations(WHO, 2013).

Various studies have addressed the different aspects of RTAs, with most focusing on predicting or establishing the critical factors influencing injury severity (Chong, 2005). Ossenbruggen and Pendharkar(2001) used logistic regression model to identify the prediction factors of crashes and crash-related injuries. The study explained that village sites were less hazardous than residential or shopping sites. Malyshkina *et. al.*(2009)also studied the assessment of the impact of highway design exceptions on the frequency and severity of vehicle accidents.They used the negative binomial model for accidentfrequency and multinomial logit model for severity of vehicle accidents. The result of the study revealedlocality of the road, average annual daily traffic and road pavement were statistically significantly for negative binomial model. In addition, design exception and road pavement were statistically significantly for multinomial logit model.

Ali and Ötüken (2014) performed study to determine risk factors affecting the fatal versus non-fatal accidents in a rural region of Turkey, logistic regression analysis was used to analyze the data and critical factors that contributed significantly to fatal versus non-fatal road traffic accidents results revealed that the driver's age, clear weather condition, winter season, straight/or slight road, curve/or bend road, the driver's education and the purpose of the vehicle were the significant factors affecting road traffic accidents over the sample period. Similarly, (Ma *et. al.*, 2009) identified the possible contributory factors on urban road environment to accident severity using the logistic regression model to traffic accident data. As the result, roaddivision, accident location, road alignment, road type and lighting condition were significantly associated with accident severity.

The ordered probit model is used to examine the contribution of several factors to the injury severity faced by motor-vehicle occupants involved in road accidents. The results showed vehicle type, rollover, run-off road, fixed-object collision, and other type of collisions had positive effect on the injury severity whereas gender, seating position, road environment, lighting conditions, and multivehicle were negatively affect (Garrido *et. al.*, 2014).

Roshi *et al.* (2008) performed multiple binary logistic regression models and showed that age, speed and alcohol consumption were significant predictors of fatal accidents. On the other study, Sharma and Landge (2013) tried to correlate the road traffic crash rate with road geometry and traffic characteristics involving heavy vehicles on national highway number 6 (NH-6), one of the busy rural roads in central India. Zero-inflated negative binomial (ZINB) regression method has been used to model the occurrence of road traffic crashes. The independent variables selected in this study were shoulder width (SW), lane width (LW), access density (AD), spot speed (SS) and annual average daily traffic (AADT). Based on the selected model density, lane width and shoulder width are significant predictors that affect the traffic accident.

A logistic regression model analysis in Calgary shows that, the road class was a significant effect on pedestrian safety. Although two-way divided roads have a better geometric design, such roads appeared to increase the serious injury risk to pedestrians.Ifvehicle crash happens, a pedestrian is more likely to suffer from more severe injury due to the higher impact speed(Anowar *et. al.*, 2010).

Geedipally (2005) showed 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 by taking traffic accidents as dependent variables and the factors which cause the accidents as independent variables. The result of the study showed that road traffic accidents were 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 also, showed a negative effect with a decrease in traffic accidents by 50% after resurfacing.

Pan and Prakash (2013) studied motorway safety by developing accident prediction models that link accident frequencies to their non-behavioral contributing factors, including traffic conditions, geometric and operational characteristics of road, and weather conditions. A number of accident prediction models were developed and assessed for their predictive ability using negative binomial regression models under three categories: first for the whole of the motorway, second for rural and urban motorway segments separately and third for motorway segments without ramp, with on-ramp and with off-ramp separately. The results uncovered the safety impacts of different non-behavioral contributing factors, in which segment length, average annual daily traffic (AADT) per lane and the number of lanes always have the most profound effects on accident frequency.

In Lithuanian regional gravel roads traffic accidents occur from April to November with a peak in July. This can be done using trend analysis, and explainedby an increased traffic volume in spring and summer. A smaller number of traffic accidents between December and March can be explained by a decrease in traffic volume and speed as well as by a safer behavior of road users in winter conditions (Stanislav and Henrikas, 2013).

Abdul-Rahaman (2014) conducted study to identify the variables that mainly contribute to accident severity in the Northern Region of Ghana, and to describe the impact of these variables. The study applied binary logistic regression to accident-related data collected from motor transport and traffic unit (MTTU) Northern Region traffic-police records. A total of 398 accident data from 2007-2009 was used. Among the variables obtained from police-accident reports, two independent variables were found most significantly associated with accident severity; namely, Over loading and Obstruction.

Most of the RTAs in Nigeria occur in the months of June, July, September, October, November and December. June, July and September are the peak rainy months and more road accidents occur in rainy months because wet road conditions affect many drivers' ability (Augustus, 2012).

Tewolde (2007) used a Poisson regression model to identify factors that mainly determine the crash-related injuries. Age of driver, educational background and place of accident were statistically significantwith the number of injuries per accident. Among the age categories, drivers in the age group 18-30 are accountable for larger number of injuries per accidents. This indicates that young drivers are accountable for the magnitude (severity) of accidents.Drivers with elementary school level of education take the major responsibility for the increased number of injuries per accident.

Haile and Demeke(2014)identified the major factors that affect the occurrence of trafficaccidentsusingchi-squaretest of association andbinarylogistic regression modeleither thedrivers faced road traffic accidents or not in Bahir Dar city, Ethiopia. The result of the study indicated that majority of accidents were occurred by drivers whose age is less than 30 years and the minimum accidents were occurred by drivers whose age is greater than 50 years old. Most accidents were occurredby drivers whose educational level is secondary.Furthermore, pedestrian's while crossing the road, driver's usage of seat belt had statistically significant for the occurrence of traffic accidents in the city.

Guyu (2013) identified the road traffic accident black spots (RTABSs) in Addis Ababa, using desctripitive statistics (frequency) which is obtained from the Addis Ababa traffic police office. The results of the study revealed that Kirkos, Bole, and Arada sub-cities were the leading in terms of the total number of road accidents. Particularly, the highest accident concentrations were located in Kirkos sub-city, and the least was in Gullele sub-city.

Most of the studies (Tewolde, 2007; Stanislav and Henrikas, 2013) conducted on traffic accidentsindicate that factors that contribute to the occurrence of traffic accident can be categorized as vehicle, driver, road and weather (environment) related factors.

.

Chapter 3
Statistical Data and Methodology

3.1 Description of the Study Area

Thisstudy is conductedin Addis Ababa, capital city of the Federal Democratic Republic of Ethiopia,which is located in the center of the country. It is also home to the African Union (UN), the Economic Commission for Africa (ECA) and other international organizations.It is the largest city in Ethiopia, with a population of 3,384,569 and annual growth rate of 3.8% (CSA, 2007).

For administration purposes, Addis Ababa is divided into ten sub-cities, which are, AkakiKaliti, Nefas Silk lafto, KolfeKeraniyo, Gulele, Ledeta, Kerkos, Arada, Addis Ketema, Bole and Yeka. Over the past years, the city of Addis Ababa has witnessed with an amazing expansion in size. The rapid increase in urban population with an annual growth rate of 3.8 percent per year has not been provided with an equal growth in provisions of road infrastructures, urban transportation, and other infrastructures(CSA, 2007).

3.2 Source of Data

The data in this study is based on 2014 road traffic accident which is obtained from the Addis Ababa Traffic Control and Investigation Department (AATCID), Addis Ababa on daily basis. The data provide information on road traffic accidents that occur within 365 consecutive days fromJuly 30, 2013 to July 29, 2014.

3.3 Variables of the Study

In regression models there are two types of variables called outcome and explanatory variables. Detaildescription of the outcome and explanatory variables that are consideredfor this study are mentioned as follows.

a) Outcome Variable

The outcomevariable of the study is the number of human death per accident due to road traffic accident in Addis Ababa.

b) Explanatory Variables

The most important explanatoryvariables thatexplainthe number of human death per accident is stated as follows:

Human factors included in this studyare:

- Sex of driver,
- Age of driver,
- Driving experience,
- Ownership of vehicle,
- Driver-vehicle relationship, and
- Education level of driver.

Vehicle factors included in this study are:

- Vehicle movement,
- Vehicle type, and
- Vehicle age of service.

Environmental factors used in this study are:

- Air situation,
- Light condition,
- Month of accident,
- Accident time, and
- Road condition.

Road related factors included in this study are:

- Road junction,
- Road division, and
- Road inclination.

Other factors such as accident type are also included in the study.

Description and Coding of the Study Variable

Description factors related to human, vehicle, environmental, and roadand other factors including their coding of categories are presented as follows.

Table 3.1: Description and Coding of the Study Variables

Predictors	Categories, codes and descriptions
Human Factors	
Ageof driver	(0) 18-30 (1) 31-50 (2) > 50
Sex of the driver	(0) Male (1) Female
Education level of driver	(0) Elementaryand below (1) High School (2) Above high school
Driver-vehicle relationship	(0) Hired (1) Owner (2) Other (relative's, friend's)
Experience (in year)	(0) < 6 (1) 6-10 (2) 10-15 (3) 16-20 (4) >20
Ownership of vehicle	(0) Private (1) Government (2) Organization
Vehicle Factors	
Vehicle type	(0) Minibus and automobile (1) Bus (public bus, higer bus, ISUZU, and others contain>12 persons) (2) Cargo (trailer, fluid cargo, lobags) (3) Others (Bajaj, motor cycle, excavator, dozer)
Vehicle service (in year)	(0) < 6 (1) 6-10 (2) 10-15 (3) 16-20

	(4) >20
Vehicle movement	(0) Direct/ straight (1) Turn to left/ right (2) Reverse (3) Other movement(stop and road off)
Road related Factors	
Road division	(0) Island (deset) (1) One-Way road (2) Two-Way road (3) Others (roundabout, cross, compound)
Road inclination	(0) Direct (1) sloped (scarp and uphill)
Road junction	(0) Yes (1) No
Environmental Factors	
Road condition	(0) Dry (1) Wet
Light condition	(0) Dark (no road light) (1) Day (sun shine) (2) Night (with road light)
Air situation	(0) Normal air condition (1) Not normal (cold/ rainy) air condition
Accident time	(0) Afternoon (1) Morning (2) Evening (3) Night
Month in quarter	(0) Q1 = September, October, November (1) Q2 = December, January , February (2) Q3 = March, April ,May (3) Q4 = January, July , August
Other Factors	
Accident type	(0) Vehicle-vehicle (vehicle crash to vehicle) (1) Vehicle-pedestrian (vehicle crash to pedestrian) (2) Others (vehicle down, vehicle crash to inert, pedestrian jump from the vehicle, piston fail)

3.4 Methodology

Count data regression models have been widely used in statistics to model response variables that are assumed to be observed without error. Poisson regression model is commonly used to study the association between count outcome variable and covariates. There are two strong assumptions for Poisson model to be checked: one is the events that occur independently over time or exposure period, the other is the conditional mean and variances are equal. In practice, counts have greater variance than the mean are described as overdispersion. This indicates Poisson regression is not adequate. There are two common causes that can lead to overdispersion: additional variation to the mean or heterogeneity, a negative binomial model is often used and other cause counts with excess zeros or zero-inflated counts, since the excess zeros will give smaller mean than the true value, this can be modeled by using zero-inflated Poisson (ZIP) or zero-inflated negative binomial (ZINB)(Cameron and Trivedi, 1998, 2005).

There are various zero-inflated models found in the literature, which have been advocated to model data, which are overdispersed, and with extra zeros. These data are found in various disciplines from public health, economics, epidemiology, psychology, sociology, political sciences, agriculture, species abundance and road safety(Phang and Loh, 2013).

In this study, the analysis of data with overdispersed and many zeros of count data were carried out by using themodels of Poisson, negative binomial, zero-inflated Poissonand zero-inflated negative binomial regression models.The likelihood-ratio test and Wald test can be used to test the overdispersion and zero-inflation problems, and for testing the coefficient of regression parameters, respectively.Further, model selection is also considered to give a simple and effective way of selecting a model for the overdispersion and zero-inflation process.

3.4.1 Poisson Regression Model

Poisson regression model provide a standard framework for the analysis of count data. Suppose we have an independent sample of n pairs of observations $(y_i, x_i)_{i \in 1,\ 2,\ \dots,\ n}$, where y_i denotes the number of times an event has occurred and x_i is the value of the explanatory variables for the i^{th} subject. Assume

$y_i \sim Poisson(\mu_i),\ i = 1, 2, \ldots, n$ then the probability density function of Poisson random variables, Y_i, is given by

$$P(y_i|\mu_i) = \frac{e^{-\mu_i}{\mu_i}^{y_i}}{y_i!},\quad y_i = 0, 1, 2, \ldots \qquad (3.1)$$

where $\mu > 0$, represents the expected number of occurrences in a fixed period of time. The mean and variance of the Poisson regression model is given as follows:

$$\mathrm{E}(\mathrm{y_i}) = \mathrm{Var}(\mathrm{y_i}) = \mu_i$$

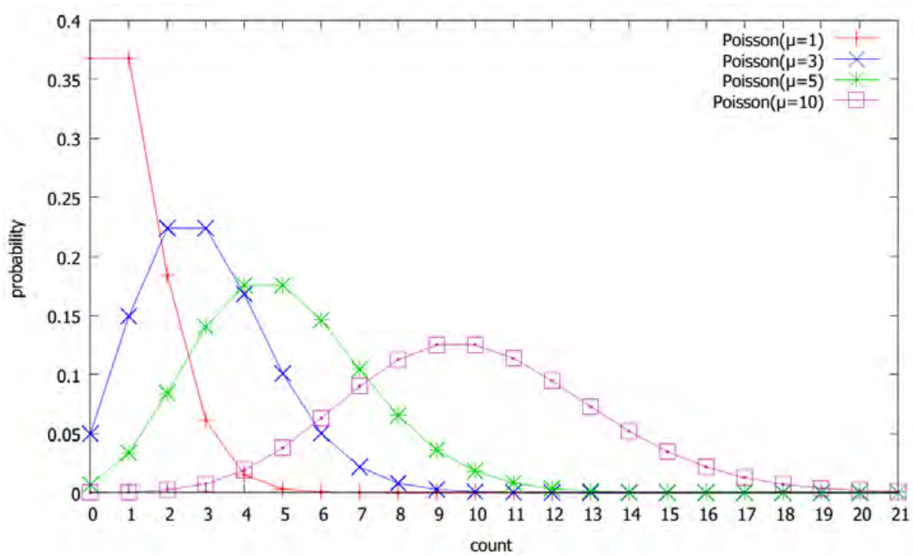

Figure 3.1: Poisson distribution plot with different means

Figures 3.1 illustrated four characteristics of the Poisson distribution that are important to understand the regression models for count dataset. Then we have the following empirical observations.

1. As μ increases, the curve of the distribution shifts to the right,where $\mu$is the mean of the distribution.
2. $Var(y) = \mu$, which is known as equi-dispersion. In real data, many count variables have a variance greater than the mean, which is called overdispersion.

3. As μ increases, the probability of a zero count decreases. For many count dataset,there are more observed zeros than predicted by the Poisson distribution.
4. As μ increases, the Poisson distribution approximates a normal distribution. This is shown by the distribution for μ = 10.

Let x be $n \times p$ matrix of explanatory variables. The relationship between y_i and i^{th} row vector of x, x_i,linked by$g(\mu_i)$is given by:

$$\log(\mu_i) = \eta_i = x_i^T \beta, i = 1, 2, \ldots, p$$

This model is known as the Poisson regression model or log-linear model. To find the maximum likelihood function of$\mu_i = E(y_i) = exp(x_i^T \beta)$, we define the likelihood function as follows:

$$\mathcal{L} = \ell(\mu_i, \beta) = \prod_{i=1}^{n} \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!} = \prod_{i=1}^{n} \frac{e^{-e^{x_i^T \beta}} \left(e^{x_i^T \beta}\right)^{y_i}}{y_i!}$$

Taking log on both sides and we get:

$$\mathcal{L} = \log\big(\ell(\mu_i, \beta)\big) = \sum_{i=1}^{n} \left[y_i x_i^T \beta - e^{x_i^T \beta} - \log(y_i!)\right] \quad (3.2)$$

The first derivative of the log likelihood function is

$$\frac{\partial \mathcal{L}}{\partial \beta_j} = \sum_{i=1}^{n} \{y_i - exp(x_i^T \beta)\} x_{ij}$$

The second derivative of log likelihood function is given as follows:

$$\frac{\partial^2 \mathcal{L}}{\partial \beta_j \partial \beta_k} = -\sum_{i=1}^{n} exp(x_i^T \beta)\, x_{ij} x_{ik}$$

Hence, the log-likelihood function of the Poisson regression model is nonlinear in β so that they can be obtained via using an iterative algorithm. The most commonly used iterative algorithms are either Newton-Raphson or Fisher scoring. In practice $\hat{\beta}$ is the solution of the estimating equations obtained by differentiating the log likelihood, (3.2) in terms of β and equating them to zero. Therefore, β will be obtained by maximizing using numerical iterative method (McCullagh and Nelder, 1989).

3.4.2 Overdispersion

The most viewing property of the Poisson model is equi-dispersion, $Var[y_i] = E[y_i], i = 1, 2, 3, \dots, n$which is often reflective. In practice, it is often found that the data exhibit overdispersion. Generally, two sources of overdispersion are determined: heterogeneity of the population and excess of zeroes. The heterogeneity is observed when the population can be divided into many homogeneous subpopulations. The excess of zeroes is detected when the numbers of observed zeroes exceed largely the number of zeroes produced by the fitted Poisson distribution. The existence of overdispersion is an incentive to seek alternative models that suit better the data.

However, the application of Poisson regression model works well for count data, it is constrained with the assumption of equality of variance and mean. When the variance of count data exceeds the mean, $Var[y_i] > E[y_i], i = 1, 2, 3, \dots, n$ a feature of overdispersionwill occur. When overdispersion occurs, the Poisson maximum likelihood estimator obtained will be incorrect (Cameron and Trivedi, 1998).

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

$H_o: \delta = 0$ (The Poisson model can be fitted well the data),versus

$H_o: \delta > 0$ (The data would be better fitted by the negative binomial regression)

The likelihood-ratio test is given by:

$$LRT = 2(\mathcal{L} - \mathcal{L}_o),$$

where $\mathcal{L}$ and $\mathcal{L}_0$ are the log-likelihood of models under the null and alternative hypothesis, respectively.Underthe null hypothesis LRT follows a chi-square distribution with degree of freedom one (Cameron and Trivedi, 1998).

3.4.3 Negative Binomial Regression Model

The limitation of the Poisson regression model is that the mean and variance of the outcome variable is not identical. Another alternative solution to overcome this problem is negative binomial regression model. It is obtained from the mixture of Poisson and Gamma distribution called Poisson-Gamma distribution. Negative binomial model allows extra variation, δ, relative to the standard Poisson model. In this case the variance of negative binomial model is significantly greater than the mean.

A random variable $Y_i, i = 1, 2, 3, \dots, n$ is called a negative binomial distributed count with parameter $\mu$and $\delta$the probability density function is expressed as follows:

$$f(y_i; \mu_i, \delta) = \frac{\Gamma\left(y_i + {}^{1}/_{\delta}\right)}{\Gamma\left({}^{1}/_{\delta}\right) y_i!} (1 + \delta\mu_i)^{-1/\delta} \left(1 + \frac{1}{\delta\mu_i}\right)^{-y_i}, \; y_i = 0, 1, 2, \dots \qquad (3.3)$$

with mean and variance, respectively, given by:

$E(Y_i) = \mu_i = exp(x_i^T \beta)$, and $Var(Y_i) = \mu_i(1 + \delta\mu_i)$

The term δ (read as delta) is called the dispersion parameter and assumed not to depend on a covariate. If the dispersion parameter$\delta \to 0$, then the negative binomial model reduces to the classical Poisson model. The likelihood function of the negative binomial model (3.3) based on a sample of n independent observation is given by:

$$\ell(\mu_i, \delta; y_i) = \prod_{i=1}^{n} \left(\frac{\Gamma\left(y_i + \frac{1}{\delta}\right)}{\Gamma\left(\frac{1}{\delta}\right) y_i!} (1 + \delta\mu_i)^{-1/\delta} \left(1 + \frac{1}{\delta\mu_i}\right)^{-y_i} \right)$$

and the log-likelihood function is

$$\mathcal{L} = log \;\; \ell(\mu_i, \delta; y_i)$$

$$= \sum_{i=1}^{n} \left\{ -\log(y_i!) + \log\left(\frac{\Gamma\left(y_i + \frac{1}{\delta}\right)}{\Gamma\left(\frac{1}{\delta}\right)} \right) - \frac{1}{\delta}\log(1 + \delta\mu_i) - y_i \log\left(1 + \frac{1}{\delta\mu_i}\right) \right\}$$

where the following expression can be used to simplify the equation:

$$\frac{\Gamma\left(y_i + {}^{1}/_{\delta}\right)}{y_i!\,\Gamma\left({}^{1}/_{\delta}\right)} = \prod_{k=1}^{y_i} \left(y_i + \frac{1}{\delta} - k \right) = \delta^{-y_i} \prod_{k=1}^{y_i} (\delta y_i - \delta k + 1)$$

Then

$$\mathcal{L} = \sum_{i=1}^{n} \left\{ -\log(y_i!) + \sum_{k=1}^{y_i} \big(\log(\delta y_i - \delta k + 1)\big) - \left(y_i + {}^{1}/_{\delta}\right)\log(1 + \delta\mu_i) + y_i \log(\mu_i) \right\} \tag{3.4}$$

where $\mu_i = exp(x_i^T \beta)$

$$\mathcal{L} = \sum_{i=1}^{n} \left\{ -\log(y_i!) + \sum_{k=1}^{y_i} \big(\log(\delta y_i - \delta k + 1)\big) - \left(y_i + {}^{1}/_{\delta}\right)\log(1 + \delta exp(x_i^T \beta)) + y_i x_i^T \beta \right\} \tag{3.5}$$

The likelihood equations to estimate $\beta$and δ are obtained by taking the partial derivatives of the log-likelihood function,(3.5), and setting them equal to zero. Thus, we obtain the first derivatives of the log-likelihood function, $\mathcal{L}$, with respect to the underlying parameters are obtained as follows:

$$\frac{\partial \mathcal{L}}{\partial \beta} = \frac{\partial \mathcal{L}}{\partial \mu}\frac{\partial \mu}{\partial \beta} = \sum_{i=1}^{n}\left[\frac{y_i - \mu_i}{1 + \delta\mu_i}\right] x_i,$$

$$\frac{\partial \mathcal{L}}{\partial \delta} = \sum_{i=1}^{n}\left\{-\delta^{-2}\sum_{k=0}^{n-1}\frac{1}{(k + \frac{1}{\delta})} + \delta^{-2}\log(1 + \delta\mu_i) + \frac{y_i - \mu_i}{\delta(1 + \delta\mu_i)}\right\}.$$

The likelihood equation are non-linear in parameters β and δ. The above equations were solved simultaneously by using iterative algorithm. The iterative algorithm commonly used is either Newton–Raphson or Fisher Scoring (McCullagh and Nelder, 1989).

3.4.4 Zero-Inflated Poisson Regression Model

Zero-inflated Poisson model is used to model count data with excessive number of zero observations. Further theory suggested that a separate process from the count values generate the excess zeros and the excess zeros can be modeled and analyzed. Hence, the ZIP model has two parts, i.e. the usual Poisson count model and the logit model which is used to predict excess zeros in the count dataset.

If the outcome variables $y_i, i = 1, 2, 3, \dots, n$ have independent observations having a zero-inflated Poisson distribution, the zeros are assumed to arise in two ways corresponding to distinct underlying states. The first state occurs with probability ω_i and produces only zeros, while the other state occurs with probability $1 - \omega_i$and leads to a standard Poisson count with mean μ_i and hence a chance of further zeros. In general, the first state is called structural zeros and the other state from Poisson model are called sampling zeros (Jansakul and Hinde, 2002). This two-state process gives the following probability mass function (pmf):

$$P(Y_i = y_i) = \begin{cases} \omega_i + (1 - \omega_i)e^{\mu_i}, & y_i = 0 \\ (1 - \omega_i)\dfrac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}, & y_i = 1, 2, 3, \dots \end{cases} \quad (3.6)$$

where $0 \leq \omega_i \leq 1$, and μ_i. The parameter μ_i and ω_i depends on the covariates x_i and z_i, respectively. More specifically, we can write the model as follows:

$$\log(\mu_i) = x_i^T\beta, \quad \log\left(\frac{\omega_i}{1-\omega_i}\right) = z_i^T\gamma$$

The mean and the variance of ZIP regression model, respectively, are:

$$E(y_i) = \mu_i(1-\omega_i), \text{ and } Var(y_i) = \mu_i(1-\omega_i)(1+\omega_i\mu_i).$$

The log-likelihood function, $\mathcal{L} = \ell(\mu_i, \omega_i; y_i)$ for ZIP model is given below:

$$\mathcal{L} = \sum_{i=1}^{n}\left\{I_{(y_i=0)}\log[\omega_i + (1-\omega_i)exp(-\mu_i)] + I_{(y_i>0)}[\log(1-\omega_i) - \mu_i + y_i\log(\mu_i) - \log(y_i!)]\right\} \quad (3.7)$$

with respect to the covariate $x_i$and z_i the log-likelihood functioncan be written as follows:

$$\mathcal{L} = \ell\log(y_i|\, x_i, z_i) = \sum_{i=1}^{n}\left\{I_{(y_i=0)}\log\left[exp(z_i^T\gamma) + exp(-exp(x_i^T\beta))\right] + I_{(y_i>0)}[[y_i x_i^T\beta - exp(x_i^T\beta)] - \log(1 + exp(z_i^T\gamma))\right\} \quad (3.8)$$

The first derivative the log-likelihood function with respect to the underlying parameter is:

$$\frac{\partial\mathcal{L}}{\partial\gamma_r} = \frac{\partial\mathcal{L}}{\partial\omega_i}\frac{\partial\omega_i}{\partial\gamma_r}$$

$$= \sum_{i=1}^{n} I_{(y_i=0)}\left[\frac{exp(z_i^T\gamma)}{exp(z_i^T\gamma) + exp(-exp(x_i^T\beta))}\right]z_{ir} - \sum_{i=1}^{n}\left[\frac{exp(z_i^T\gamma)}{1 + exp(z_i^T\gamma}\right]z_{ir,}$$

$$r = 1, 2, \dots, q$$

$$\frac{\partial\mathcal{L}}{\partial\beta_j} = \frac{\partial\mathcal{L}}{\partial\mu_i}\frac{\partial\mu_i}{\partial\beta_j}$$

$$= \sum_{i=1}^{n} I_{(y_i=0)}\left[\frac{-exp(x_i^T\beta)\,exp(-exp(x_i^T\beta))}{exp(z_i^T\gamma) + exp(-exp(x_i^T\beta))}\right]x_{ij} + \sum_{i=1}^{n} I_{(y_i>0)}[y_i - exp(x_i^T\beta)]\,x_{ij},$$

$$j = 1, 2, 3, \dots, p$$

3.4.5 Zero-Inflated Negative Binomial Regression Model

Zero-inflated negative binomial regression model is designed to model data with population heterogeneity which is caused by the occurrence of excess zeros and overdispersion due to unobserved heterogeneity. Many studies showed that ZINB model provides a better fit to the overdispersed count data when compared with ZIP model. We consider a zero-inflated negative binomial regression model in which the response variable $y_i, i = 1, 2, 3, \dots n$ has the distribution.

$$p(Y_i = y_i) = \begin{cases} \omega_i + (1 - \omega_i)(1 + \delta\mu_i)^{-\frac{1}{\delta}}, \quad y_i = 0 \\ (1 - \omega_i)\dfrac{\Gamma\left(y_i + {}^1\!/_\delta\right)}{y_i!\,\Gamma\left({}^1\!/_\delta\right)}(1 + \delta\mu_i)^{-1/\delta}\left(1 + \dfrac{1}{\delta\mu_i}\right)^{-y_i}, \qquad y_i > 0 \end{cases} \tag{3.9}$$

where $\delta > 0$ is a dispersion parameter and is assumed not to depend on covariates.

The mean and the variance of the ZINB regression model are:

$$E(y_i) = \mu_i(1 - \omega_i), Var(y_i) = \mu_i(1 - \omega_i)(1 + \omega_i\mu_i + \delta\mu_i)$$

The parameters μ_i and$\omega_i$depend of covariates x_i and z_i, respectively. We can write the model as

$$\log(\mu_i) = x_i^T\beta, \quad \log\left(\frac{\omega_i}{1 - \omega_i}\right) = z_i^T\gamma$$

The log-likelihood function,$\mathcal{L} = \ell(\delta, \mu_i, \omega_i; y_i)$ for the ZINB model is given by

$$\mathcal{L} = \sum_{i=1}^{n} \left\{ I_{(y_i=0)} \log\left(\omega_i + (1 - \omega_i)(1 + \delta\mu_i)^{-1/\delta}\right) \right.$$

$$\left. + I_{(y_i>0)} \log\left((1 - \omega_i)\frac{\Gamma\left(y_i + {}^1\!/_\delta\right)}{y_i!\,\Gamma\left({}^1\!/_\delta\right)}(1 + \delta\mu_i)^{-1/\delta}\left(1 + \frac{1}{\delta\mu_i}\right)^{-y_i} \right) \right\}$$

since

$$\frac{\Gamma\left(y_i + {}^{1}/_{\delta}\right)}{y_i!\,\Gamma\left({}^{1}/_{\delta}\right)} = \prod_{k=1}^{y_i}\left(y_i + \frac{1}{\delta} - k\right) = \delta^{-y_i}\prod_{k=1}^{y_i}(\delta y_i - \delta k + 1).$$

Further more, log-liklihood function can be written as

$$\mathcal{L} = \sum_{i=1}^{n} I_{(y_i=0)}\left[\log\left(\omega_i + (1-\omega_i)(1 + \delta\mu_i)^{-1/\delta}\right)\right]$$
$$+ I_{(y_i>0)}\sum_{i=1}^{n}\left[\log(1-\omega_i) - \log(y_i!)\right.$$
$$\left. + \sum_{k=1}^{y_i}\log(\delta y_i - \delta k + 1) - \left(y + \frac{1}{\delta}\right)\log(1 + \delta\mu_i) + y_i\log(\mu_i)\right] \quad (3.10)$$

with respect to the covariates x_i and z_i, the log-likelihood function is given as follows:

$$\mathcal{L} = \ell(y_i|x_i, z_i)$$
$$= \sum_{i}^{n}\left\{I_{(y_i=0)}\left[\log(exp(z_i^T\gamma) + (1 + \delta exp(x_i^T\beta))^{-\frac{1}{\delta}}) - \log(1 + exp(z_i^T\gamma))\right]\right.$$
$$+ I_{(y_i>0)}\sum_{i=1}^{n}\left[\sum_{k=1}^{y_i}\log(\delta y_i - \delta k + 1) - \log(1 + \exp(exp(z_i^T\gamma))) - \log(y_i!)\right.$$
$$- \left(y_i + {}^{1}/_{\delta}\right)\log(1 + \delta\exp(x_i^T\beta))$$
$$\left.\left. + y_i x_i^T\beta\right]\right\} \quad (3.11)$$

The first derivatives of the log-likelihood function with respect to δ, β and γ are given below:

$$\frac{\partial\mathcal{L}}{\partial\delta} = \frac{\partial\ell(y_i|x_i, z_i)}{\partial\delta}$$

$$= \sum_{i=1}^{n} \left\{ I_{(y_i=0)} \frac{(1-\omega_i)(1+\delta\mu_i)^{\frac{-1}{\delta}}}{\omega_i + (1-\omega_i)(1+\delta\mu_i)^{\frac{-1}{\alpha}}} \left[\frac{\log(1+\delta\mu_i)}{\delta^2} - \frac{\mu_i}{\delta(1+\delta\mu_i)} \right] \right.$$

$$\left. + I_{(y_i>0)} \sum_{k=1}^{y_i} \frac{y_i - k}{\delta y_i - \delta k + 1} + \frac{\log(1+\delta\mu_i)}{\delta^2} - \frac{\left(y_i + {}^{1}/_{\delta}\right)\mu_i}{1+\delta\mu_i} \right\},$$

$$= \sum_{i=1}^{n} \left\{ I_{(y_i=0)} \frac{(1+exp(z_i^T\gamma))^{-1}(1+\delta\, exp(x_i^T\beta))^{\frac{-1}{\delta}}(1+exp(z_i^T\gamma))}{exp(z_i^T\gamma) + (1+\delta\, exp(x_i^T\beta))^{\frac{-1}{\alpha}}} \left[\frac{\log(1+\delta\, exp(x_i^T\beta))}{\delta^2} \right. \right.$$

$$\left. - \frac{exp(x_i^T\beta)}{\delta(1+\delta\, exp(x_i^T\beta))} \right] + I_{(y_i>0)} \sum_{k=1}^{y_i} \frac{y_i - k}{\delta y_i - \delta k + 1} + \frac{\log(1+\delta\, exp(x_i^T\beta))}{\delta^2}$$

$$\left. - \frac{\left(y_i + {}^{1}/_{\delta}\right) exp(x_i^T\beta)}{1+\delta\, exp(x_i^T\beta)} \right\},$$

$$\frac{\partial \mathcal{L}}{\partial \beta_j} = \frac{\partial \ell(y_i|x_i, z_i)}{\partial \beta_j} = \frac{\partial \mathcal{L}}{\partial \mu_j}\frac{\partial \mu_j}{\partial \beta_j}$$

$$= \sum_{i=1}^{n} \left\{ I_{(y_i=0)} \left[\frac{-(1-\omega_i)\mu_i(1+\delta\mu_i)^{\frac{-(1+\delta)}{\delta}}}{\omega_i + (1-\omega_i)(1+\delta\mu_i)^{\frac{-1}{\delta}}} \right] + I_{(y_i>0)} \left[\frac{y_i - \mu_i}{1+\delta\mu_i} \right] \right\} x_{ij}$$

$$= \sum_{i=1}^{n} \left\{ I_{(y_i=0)} \left[\frac{-(1+exp(z_i^T\gamma))^{-1}\, exp(x_i^T\beta)\, (1+\delta\, exp(}{exp(z_i^T\gamma) + (1+\delta\, exp(} \right. \right.$$

$$\left. + I_{(y_i>0)} \left[\frac{y_i - exp(x_i^T\beta)}{1+\delta\, exp(x_i^T\beta)} \right] \right\} x_{ij}$$

$$\frac{\partial\mathcal{L}}{\partial\gamma_r}=\frac{\partial\ell(y_i|x_i,z_i)}{\partial\gamma_r}=\frac{\partial\mathcal{L}}{\partial\omega_i}\frac{\partial\omega_i}{\partial\gamma_r}$$

$$=\sum_{i=1}^{n}\left\{I_{(y_i=0)}\frac{\omega_i(1-\omega_i)\left(1-(1+\delta\mu_i)^{\frac{-1}{\delta}}\right)}{\omega_i+(1-\omega_i)(1+\delta\mu_i)^{\frac{-1}{\delta}}}-I_{(y_i>0)}\omega_i\right\}z_{ir},$$

$$=\sum_{i=1}^{n}\left\{I_{(y_i=0)}\frac{\frac{exp(z_i^T\gamma)}{(1+exp(z_i^T\gamma))}\frac{1}{(1+exp(z_i^T\gamma))}\left(1-(1+\delta exp(x_i^T\beta))^{\frac{-1}{\delta}}\right)}{\frac{exp(z_i^T\gamma)}{(1+exp(z_i^T\gamma))}+\frac{1}{(1+exp(z_i^T\gamma))}(1+\delta exp(x_i^T\beta)^{\frac{-1}{\delta}}}\right.$$

$$\left.-I_{(y_i>0)}\frac{exp(z_i^T\gamma)}{(1+exp(z_i^T\gamma))}\right\}z_{ir},$$

$$=\sum_{i}^{n}I_{(y_i=0)}\left[\frac{exp(z_i^T\gamma)\left(1-(1+\delta exp(x_i^T\beta))^{\frac{-1}{\delta}}\right)}{(1+exp(z_i^T\gamma))\,(exp(z_i^T\gamma)\,(1+\delta exp(x_i^T\beta))^{-\frac{1}{\delta}})}\right]z_{ir}$$

$$-\sum_{i}^{n}I_{(y_i>0)}\left[\frac{exp(z_i^T\gamma)}{1+exp(z_i^T\gamma)}\right]z_{ir}$$

The ZINB regression model does not belong to a family of exponential distribution of a standard generalized linear model (GLM). In order to obtain the parameter estimates of ZINB regression models, $\hat{\delta}, \hat{\beta}$ and $\hat{\gamma}$, the Newton-Raphson method or the method of Fisher scoring can be used (Cameron and Trivadi, 1998).

3.4.6 Model Fitting Test (Goodness of Fit of the Model)

There arefour models to be compared in order to select the appropriate fitted model, which fits the data well. This was done using likelihood-ratio test (LRT), Akakie information criteria (AIC) and Bayesian information criteria (BIC).

AIC is the most common means of identifying the model which fits well by comparing two or more than two models. The formula is given as:

$$AIC = -2\mathcal{L} + 2k,$$

where $\mathcal{L}$is the log-likelihood of a model that will compare with the other models and k is the number of parameter in the model including the intercept.

Bayesian information criteria (BIC) takes into account the size of the data under consideration. BIC is given by,

$$BIC = -2\mathcal{L} + klog(n),$$

where $\mathcal{L}$is the log-likelihood of a model that will compare with the other models, n is the sample size of the data and k is the number of parameters in the model including the intercept.

The comparison will start from the model without any independent variable with the model with adding the independent variable one by one through the full model.

The model which has the minimum value ofAIC and BIC is the most appropriate fittedmodel to the dataset.

3.4.6.1 Test for Estimation of the Parameters

Wald test

Wald test is used to test the null hypothesis that a parameter is equal to assumed value. It is also important to test the dispersed parameter delta, δ,. The hypothesis is given as follows:

$$H_0: \beta_i = 0 \text{ Versus} H_a: \beta_i \neq 0$$

The Wald statistic is given by:

$$z = \frac{\hat{\beta}_i}{S.E(\hat{\beta}_i)}, i = 1, 2, 3, \dots, p$$

where $\beta_i, i = 1, 2, 3, \dots p$ is the coefficient of the i^{th}variable. For large sample size, z^2followschi-square distribution with one degree of freedom.

3.4.6.2 Tests for the Comparison of the Models

a) Tests for comparison of nested models

Likelihood-ratio test (LRT)

The likelihood-ratio test is used to assess the adequacy of two or more than two nested models.It compares the maximized log-likelihood value of the full model and reduced model. For instance, the null hypothesis can be stated as the overdispersion parameter is equal to zero (i.e. the Poisson model can be fitted well the data) versus the alternative hypothesis can be statedas the overdispersion parameter is different from zero(i.e. the data would be better fitted by the negative binomial regression). The likelihood-ratio test is given by:

$$LRT = 2(\mathcal{L} - \mathcal{L}_o),$$

where $\mathcal{L}$ and $\mathcal{L}_0$ are the log-likelihood of models under the null and alternative hypothesis, respectively.

Standard asymptotic theory suggests that under the null hypothesis LRT follows a chi-square distribution with degree of freedom given by the difference between the full model and reduced model. This test sometimes criticized because it requires separate estimation of two models, the restricted and unrestricted model (Cameron and Trivedi, 1998).

b) Test for comparison of Non-nested models

Vuong's test

Vuong (1989) has introduced a test which is an appropriate method to compare non-nested modelsfor count dataset. It is used to assess goodness of fit for comparisons of ZIP model versus Poisson and ZINB model versus the negative binomial model.

Assume that $P_1$is the predicted probability of ZIP model (or ZINB model) and $P_2$is the predicted probability of Poisson model (or NB model) of an observed count for case i, the relationship between the likelihood-ratio test can be defined as follows:

$$m_i = log\left[\frac{P_1(y_i|x_i)}{P_2(y_i|x_i)}\right]$$

Hence, the Vuong's test under the null hypothesis is given by:

$$V = \frac{\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} m_i\right)}{\sqrt{(\frac{1}{n}\sum_{i=1}^{n}(m_i - \overline{m})^2}}$$

For large sample size and under the null hypothesis the statistic V has the standard normal distribution at α level of significance.TheZIP (or ZINB) model is preferred over the Poisson (or NB) model, respectively, if the calculated value of the Vuong test is positive andhigh $(V > 1.96)$ viceversa,and the two models are equivalent when|V|<1.96.

Chapter 4

Statistical Data Analysis

Inthis study the analysis pertains to search the significant factors. The Poisson, negative binomial and zero-inflated Poisson, and zero-inflated negative binomial regression modelswere used to analyze the data. The statistical analyses were performed using statistical software package called STATA version 13 and SPSS version 20.

4.1 Variable Selection

In this study we have used Stepwise variable selection which is a combination of backward elimination and forward selection to identify the predictors in the model. This was done on Poisson regression model as it is the bench mark for other count regression models. Stepwise selection method addresses where variables were added or removed withrespect to the p-value in the process.

Based on stepwise variable selection procedure;

- Month in quarter,
- road inclination,
- age of the drivers,
- vehicle type and
- ownership of the vehicle,
- accident time,
- accident type and
- injured vehicle count

are included in the models at 5% level of significant.Furtheranalyses are made onlyon those variables which are foundsignificant factors for the number of human death per road traffic accident in Addis Ababa, Ethiopia.

4.2 Descriptive Statistic

The major human, roadway, vehicle and environmental factors of the human death are presented in Table 4.1. A total of 1438 road traffic accidents were recorded without property damage and light accident type in the given time of study by Addis Ababa traffic control and investigation department, and Addis Ababa police Commission Office. The mean number of human death in Addis Ababa due to road traffic accidentswere presented in Table 4.1shown below.

Table 4.1: Descriptive Results for Human Death per Road Traffic Accidents in Addis Ababa

Variables	RTA	Death per accident	
	Count	Mean	Variance
Human Factors			
Age of driver			
18-30	557	0.43	0.642
31-50	619	0.35	0.667
> 50	262	0.26	0.218
Sex of driver			
Male	1408	0.37	0.587
Female	30	0.20	0.166
Education level of driver			
Elementary and below	356	0.41	0.935
High School	400	0.40	0.391
Above high school	682	0.32	0.500
Driver-vehicle relationship			
Owner	293	0.25	0.220
Hired	865	0.41	0.794
Other	280	0.35	0.265
Experience			
< 6	642	0.41	0.585
6-10	353	0.39	0.392
11-15	152	0.29	1.134
16-20	99	0.20	0.163
> 20	192	0.30	0.650
Ownership of vehicle			
Private	483	0.21	0.584
Government	743	0.50	0.636
Organization	212	0.24	0.231
Vehicle Factors			
Vehicletype			
Minibus and automobile	910	0.27	0.396
Bus	134	0.55	1.318
Cargo	320	0.53	0.764
Others	74	0.49	0.391

Vehicle service			
< 6	448	0.37	0.645
6-10	676	0.41	0.573
11-15	166	0.10	0.104
16-20	94	0.46	0.402
> 20	54	0.39	1.638
Vehicle moment			
Straightforward	1194	0.37	0.526
Turnto left/right	105	0.24	0.298
Reverse	77	0.31	0.244
Others	62	0.50	2.484
Road related Factors			
Road division			
Island (deset)	969	0.34	0.472
One Way road	79	0.20	0.189
Two Way road	288	0.44	0.645
other type of road	102	0.53	1.656
Road inclination			
Direct	1358	0.33	0.466
Sloped road	80	0.86	2.247
Roadjunction			
Yes	639	0.27	0.436
No	799	0.44	0.681
Environmental Factors			
Roadcondition			
Dry	1376	0.35	0.578
Wet	62	0.55	0.546
Light condition			
Dark (no light)	90	0.56	0.295
Day (sun shine)	1044	0.34	0.674
Night (Road light)	304	0.38	0.323
Air situation			
Normal air condition	1284	0.34	0.594
Cold/ rainy condition	154	0.56	0.404
Accident time			
Afternoon	522	0.31	0.596
Morning	516	0.35	0.748
Evening	315	0.43	0.283
Night	85	0.52	0.491
Monthinquarter			
Q1 (Sep, Oct, Nov)	374	0.29	0.271
Q2 (Dec, Jan, Feb)	384	0.36	0.632
Q3 (Mar, Apr, May)	328	0.34	0.392
Q4 (Jun, Jul, Aug)	352	0.46	1.008
Other Factors			
Accident type			
Vehicleto vehicle	426	.12	.220
Vehicleto pedestrian	853	.46	.568
Others	159	.47	1.377
Total	**1438**	**0.36**	**0.578**

Table 4.1 revealed the mean and variance of number of human death for each predictor. The driver's age has three categories. Among these categories, drivers within the age group 18-30 yearshad the highest mean number human death (0.43) compared to the oldest age group (>50 years) who had the smallest mean number of human death. Considering the experience of drivers, the highest mean number of human deathoccurred bythe driver who had less than 6 years driving experience (0.41), whilst the smallest mean number of human death was(0.2)reported by the driver who had 16-20 years driving experience. The smallest mean number of human death (0.32) was reported due to the drivers who had completed or more thanhigh school level; whereas the highest mean number of human death was (0.41) occurred by elementary and below education level of the driver. Government vehicles had high contribution to the mean number of human death (0.50); on other hand, less contribution was made by private vehicles (0.21). The mean number of human death was higher for male drivers (0.37) than female drivers (0.20).The highest mean number of human death per road traffic accident occurred at night time (0.52), whilst the smallest mean number of human death was reported at afternoon time. Regarding to month in quarter, the highest mean number of human death wasoccurred in June, July and August seasons while the smallest mean number of death were recorded in September, October and November.Wet road condition is accountable for higher mean number of human death (0.55) than the dry road condition (0.35). According to road junction, the highest mean number of human death (0.44) happened on non-junction road as compared to the road with junction (0.27). The predictor light condition revealed that dark without road light tookthe higher mean number of human deathperroadtraffic accident (0.56)when compared withsun rise light condition.The highest mean number of human death occurred by bus (.55) with variance (1.318) whilst the smallest mean numberofhuman death were recordedbyminibus and automobile (0.27) with variance(0.396).

Among the five categories of vehicle service age (in year), the highest mean number of human death were occurred in the vehicles service age group 16-20years (0.46) while the smallest mean number of human death were occurred in the vehicles service age group 11-15 years (0.1).

Table 4. 2: Summary Statistics for the Number of Human Death Due to Road Traffic Accident

Mean	Variance	Skewness	Kurtosis
0.36	0.58	6.88	88.36

As shown in Table 4.2, the variance of human death perroad traffic accident (0.58) wasgreater than itsmean (0.38). This indicated the possibility of overdispersionandhence the standard Poisson regression model was not appropriate to fit the road traffic accident data. Additional inspection of the data also indicated the existence of excess number of zeros (70%) and the shape of the curve is skewed to the right (Figure 4.1).

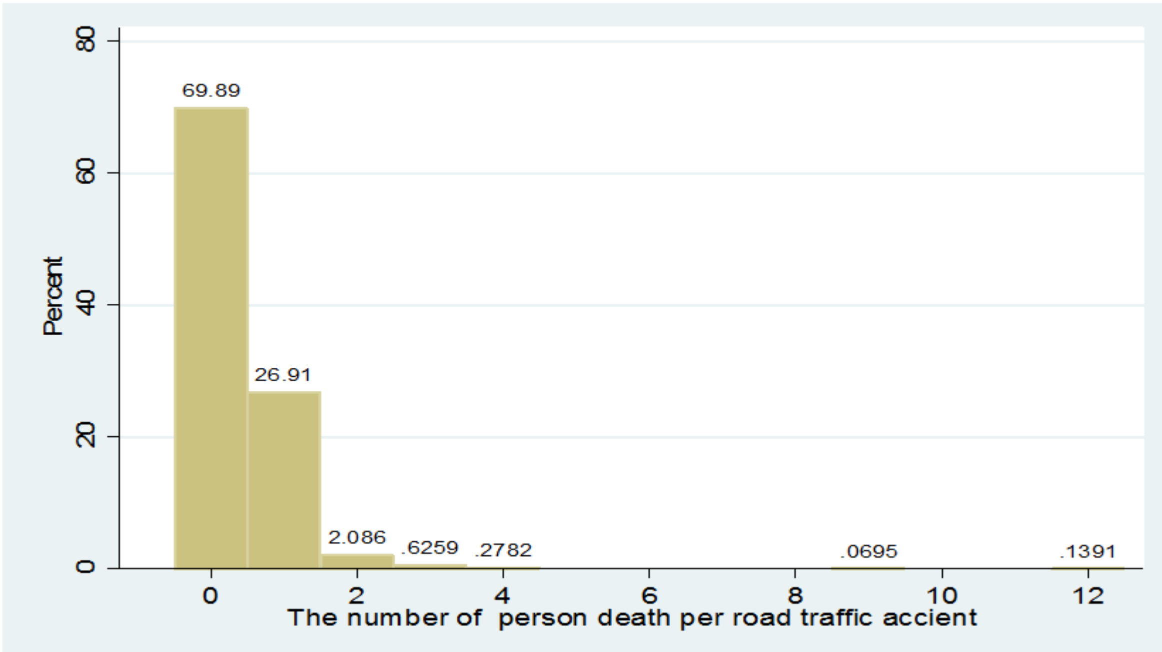

Figure 4.1: Histogram for human death due to RTA

Figure 4.1 showed that there are large numbers of zero values (highly picked at zero) and a positively (or right) skewed distributionof human death per road traffic accidents. In this case zero-inflated count data models like ZIP and ZINB models better fit the data than Poisson model.

4.2 PARAMETER ESTIMATION

Table 4.3 presented parameter estimates with their corresponding standard error of the Poisson, negative binomial (NB), zero-inflated Poisson (ZIP), and zero-inflated negative binomial (ZINB) regression models. It shouldbe noted that, ZIP and ZINB models are a mixture of two regression models, namely; standard count models part such as Poisson and/or negative binomial model and logistic (inflated) part. The overdispersion parameter delta (δ), is significantly different from zero in both NB and ZINB regression models. Hence, there was an overdispersion problem in the data. As a result of this the standard error of the standard Poisson regression model is smaller than that of the other models. Particularly, when we compare the negative binomial with the Poisson regression model, the standard error of the Poisson regression model is smaller than negative binomial regression model and this leads to wrong interpretation (Cox, 1983; Aklilu, 2011). When the assumption of the equality of variance and mean in the Poisson regression model is violated, overdispersion occurs and the standard error estimates will be biased which leads to incorrect value of the test statistic. Consequently, the covariates may wrongly interpreted (Yaacob et. al., 2010).

Table 4.3: Parameter Estimation of Poisson, NB, ZIP and ZINB Regression Models

Parameters	Poisson		NB		ZIP		ZINB	
	Coef.	Std. Err.	Coef.	Std. Err.	Coef.	Std. Err.	Coef.	Std. Err.
Poisson model for non-zero part								
Month in quarter								
Q1 (Sep, Oct, Nov)	(Ref.)							
Q2 (Dec, Jan, Feb)	0.2305	0.1296	0.2381	0.1354	0.2307	0.1437	0.2372	0.1468
Q3 (Mar, Apr, May)	0.2675	0.1356	0.2738	0.1417	0.2676*	0.1241	0.2748	0.1266
Q4 (Jun, Jul, Aug)	0.4111*	0.1258	0.4109*	0.1319	0.4122*	0.1417	0.4119*	0.1362
Age of driver								
18-30	(Ref.)							
31-50	-0.1252	0.0973	-0.1279	0.1026	-0.1253	0.1173	-0.1276	0.1149
> 50	-0.3350*	0.1418	-0.3361*	0.1478	-0.3354*	0.1342	-0.3364*	0.1336
Vehicletype								
Minibus and automobile	(Ref.)							
Bus	0.7101*	0.1371	0.7256*	0.1453	0.7121*	0.1811	0.7358*	0.1904
Cargo	0.7627*	0.1020	0.7621*	0.1076	0.7629*	0.1131	0.7631*	0.1109
Others	0.4121*	0.1889	0.4332*	0.1988	0.4125*	0.1805	0.4334*	0.1783
Ownership of vehicle								
Private	(Ref.)							
Government	0.6997*	0.1144	0.7053*	0.1184	0.6907*	0.1737	0.7060*	0.1662
Organization	0.1092	0.1731	0.1092	0.1786	0.1096	0.2046	0.1094	0.2019
Accident time								
Afternoon	(Ref.)							
Morning	0.2360*	0.1113	0.2352*	0.1165	0.2366	0.1370	0.2345	0.1362
Evening	0.4245*	0.1181	0.4250*	0.1242	0.4252*	0.1301	0.4255*	0.1264
Night	0.5490*	0.1737	0.5584*	0.1843	0.5491*	0.1591	0.5583*	0.1631
Accidenttype								
Vehicleto vehicle	(Ref.)							
Vehicle to pedestrian	1.1731*	0.1498	1.1758*	0.1528	1.1741*	0.1983	1.1764*	0.1972
Others	0.9989*	0.1877	0.9797*	0.1942	0.9992*	0.2586	0.9753*	0.2494
Road inclination								
Direct	(Ref.)							
Sloped road	0.4475*	0.1434	0.4702*	0.1548	0.4485*	0.1531	0.4712*	0.1547
Education level of driver								
Elementary and below	(Ref.)							
High School	-0.1324	0.1152	-0.1229	0.1219	-0.1334	0.1621	-0.1218	0.1557

Above high school	-0.2698*	0.1092	-0.2694*	0.1152	-0.2702	0.1488	-0.2690	0.1470
Intercept	-2.9552*	0.2187	-2.9684*	0.2261	-2.9561*	0.2568	-2.9691*	0.2535
Inflate/logistic part for zero count								
Injured vehiclecount					0.2526*	0.1150	0.2255*	0.1177
Intercept					-1.71486*	0.6556	-1.72445*	0.2369
ln(δ)			-1.6973	0.3623			-1.6970*	0.7697
δ			0.1832*	0.0664			0.1832	0.1410

*** p-value< 0.05**

4.3 Statistical Model Results

4.3.1 Comparison of Models

Poisson and negative binomial regression models are designed to analyze count data. However, Poisson and negative binomial regression models differ in terms of their assumptions.Poisson model assume, that the mean and variance of the outcome variable are equal while negative binomial regression model does not assume an equal number of mean and variance.

Inorder to select thebest model which fits the data well, four different models were considered, namely; Poisson, negative binomial, zero-inflated Poisson and zero-inflated negative binomial models. In this study, different model selection criteria were considered like the likelihood-ratio test (LRT), Akaike information criterion (AIC) and Bayesian information criterion (BIC) in order to identify the most fitted model. For non-nested models: ZIP versus Poisson and ZINB versus NB regression modelswere identified using the Vuong test statistic.

Table 4.4: Model selection criteria for PR, NB, ZIP and ZINB models for the number of human death dataset

Selection Criteria	Models			
	Poisson	NB	ZIP	ZINB
Log-likelihood	-884.115	-882.013	**-809.027**	-809.027
AIC	1794.230	1792.026	**1664.054**	1666.054
BIC	1862.753	1865.820	**1785.287**	1792.558
Vuong (p-value)			**3.22 (0.000)**	9.08 (0.000)

Table 4.4 showedthe criteria in order to select the best model among the candidates. First, the calculated value of the Vuong test (3.22) was greater than the hypothetical value (1.96) for ZIP versus Poisson model. This value revealed that ZIP modelwas preferred to Poisson model in order to estimate the number of humandeath due toroad traffic accident. In the second case,comparison of ZINB versus NB models, the calculated value of theVuongtest is 9.08, revealedthat the ZINB model waspreferred to NB regression model. Finally, to compare the ZIP and ZINB models, AIC and BIC were used as shown in Table 4.4. Therefore, the ZIP model is better fitted human death per road traffic accident data than did ZINB model. AIC and BIC values of ZIP was found to be small as compared to other count models.

As shown in Figure 4.2, the ZIP model was a better choice than the other count models, since the predicted probability for ZIP model was closed to the observed probability. From Figure 4.2 and the value of AIC, BIC and Vuong criteria in Table 4.4, it can be clearly observed that there was a difference among Poisson, NB, ZIP and ZINB models for the dataset. Therefore, it is possible to conclude that the ZIP model was more appropriate than the ZINB model to fit Addis Ababa road traffic accidents dataset.

Table 4.5: Observed and predicted probability from Poisson, NB, ZIP and ZINB model for human death per RTA

Count	Frequency	Observed Probability	Predicted probability			
			Poisson	NB	ZIP	ZINB
0	1,005	0.6989	0.6951	0.7171	0.6989	0.7171
1	387	0.2691	0.2528	0.2186	0.2462	0.2186
2	30	0.0209	0.0460	0.0510	0.0480	0.0510
3	9	0.0063	0.0056	0.0107	0.0063	0.0107
4	4	0.0028	0.0005	0.0021	0.0006	0.0021
5	0	0.0000	0.0000	0.0004	0.0000	0.0004
6	0	0.0000	0.0000	0.0001	0.0000	0.0001

Table 4.5showed the observed and predicted probability of number of human death per road traffic accident due to road traffic accident.

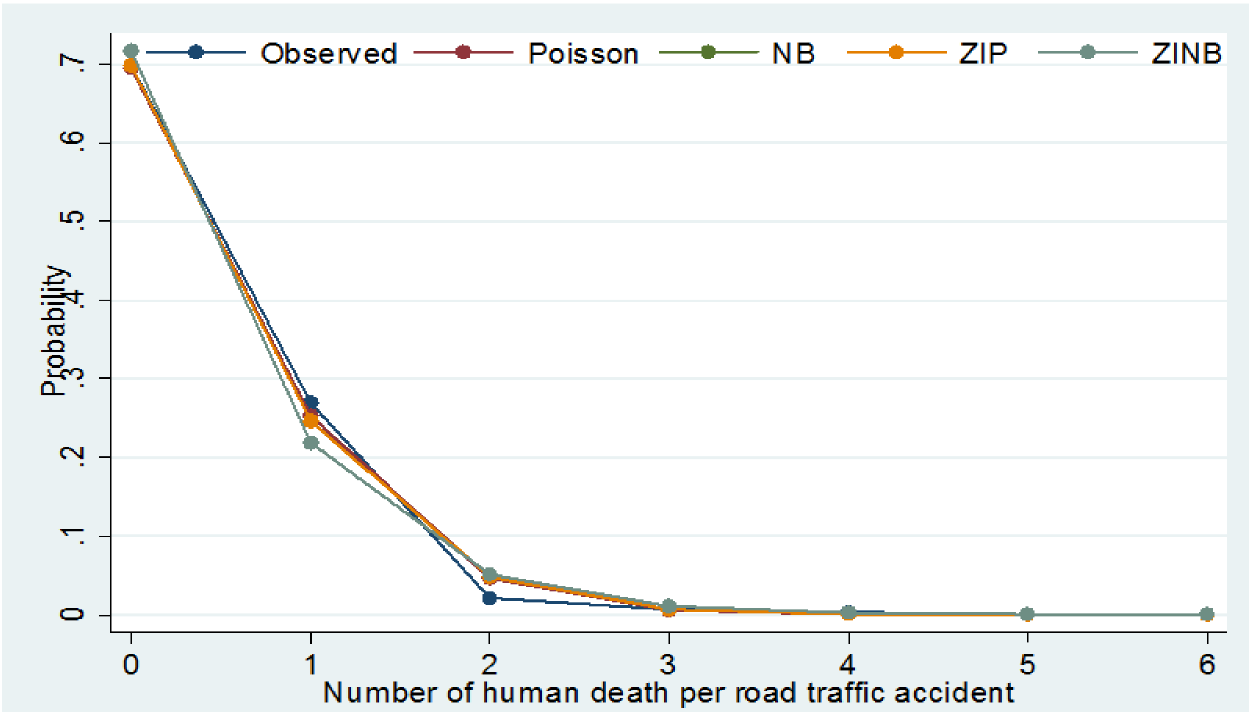

Figure4.2: The Observed and predicted probability of Poisson, NB, ZIP and ZINB models

4.3.2 Results and Discussion

In Table 4.6, count regression techniques model the log of incident counts, the coefficients can be interpreted as follows: for a one unit change in the predictor variable, the log of the response variable is expected to change by the value of the regression coefficient (coef.). In ZIP model, for every one unit increase in a unit's of the significant predictors, the log number of human death per road traffic accident is expected to increase or decrease by approximately the corresponding coefficient in the column of coefficient (coef.). In this model the variables whose p-value<0.05, wereconsidered statistically significant. To interpret the categorical data we use the incidence rate ratios ($\text{IRR} = \exp(\text{coef.})$) which is important to explain the change in percentage ($\text{IRR} - 1$) of significant predictors.

The number of human death per road traffic accident was increased by 51% and 31%for the fourth quarter of month (June, July and August) and third quarter of month (March, April, May), respectively, as compared to the first quarter month (September, October and November).Thenumber of human death per road traffic accident was increased by 56.6% for sloped roadas compared to non-sloped (straight) road inclination. With regard to accident time, the number of human death per road traffic accident wasincreased by 53% and 73.2%, for evening and night time, respectively, as compared with the afternoon time. The number of human death per road traffic accident was increased by 103.8%, 114.4% and 51.1%, for Bus, cargo and other vehicle type, respectively, when compared with minibus and automobile vehicle types. Regarding the age of the drivers, the number of human death per road traffic accident wasdecreased by 28.5% for those whose age group was 50 years and above as compared to that of age group of 18-30 years.Thenumber of human death per road traffic accident was increased by 99.5% for government's vehicle as compared to the private one. Moreover, the number of human death per road traffic accident was increased by 223.5% and 171.6%for vehicle crashed to pedestrian and other (vehicle down, vehicle crash toinert, pedestrian jump from the vehicle, piston fail), respectively, as compared to the vehicle crashed to vehicle.

In inflation (logistic) part, each additional injured vehicle count increases one's odds of no human death by 25.9%.

Table 4.6: The zero-inflated Poisson model for human death of selected independent variables

Parameters	Coef.	z	p>|z|	IRR	95% CIs for IRR Lower	upper
Poisson model for non-zero part						
Month in quarter						
Q1 (Sep, Oct, Nov)	(Ref.)					
Q2 (Dec, Jan, Feb)	0.2307	1.6054	0.109	1.2595	0.9502	1.6689
Q3 (Mar, Apr, May)	0.2676	2.1563	0.031	1.3068	1.0246	1.6663
Q4 (Jun, Jul, Aug)	0.4122	2.9090	0.004	1.5101	1.1427	1.9913
Age of driver						
18-30	(Ref.)					
31-50	-0.1253	-1.0682	0.286	0.8822	0.7011	1.1103
> 50	-0.3354	-2.4993	0.013	0.7151	0.5499	0.9306

Vehicle type						
Minibus and automobile	(Ref.)					
Bus	0.7121	3.9321	0.000	2.0383	1.4263	2.9012
Cargo	0.7629	6.7454	0.000	2.1445	1.7176	2.6763
Others	0.4125	2.2853	0.022	1.5106	1.0600	2.1510
Ownership of vehicle						
Private	(Ref.)					
Government	0.6907	3.9764	0.000	1.9951	1.4322	2.8299
Organization	0.1096	0.5357	0.593	1.1158	0.7470	1.6655
Accident time						
Afternoon	(Ref.)					
Morning	0.2366	1.7270	0.085	1.2669	0.9680	1.6563
Evening	0.4252	3.2683	0.001	1.5299	1.1846	1.9728
Night	0.5491	3.4513	0.001	1.7317	1.2678	2.3650
Accident type						
Vehicle to vehicle	(Ref.)					
Vehicle to pedestrian	1.1741	5.9208	0.000	3.2352	2.1909	4.7674
Others	0.9992	3.8639	0.000	2.7161	1.6359	4.5073
Road inclination						
Direct	(Ref.)					
Sloped road	0.4485	2.9295	0.003	1.5660	1.1587	2.1119
Education level of driver						
Elementary and below	(Ref.)					
High School	-0.1334	-0.8229	0.414	0.8751	0.6375	1.2035
Above high school	-0.2702	-1.8159	0.070	0.7632	0.5704	1.0221
Intercept	-2.9561	-11.5113	0.000	0.0520	0.0315	0.0861
Inflate/logistic part						
Injured vehicle count	0.2526	1.6054	0.028	1.2595	1.0276	1.6129
Intercept	-1.7146	-2.6157	0.001		0.1583	0.2047

Discussion

Month in quarter was a significant factor for number of human death per road traffic accident. Table 4.6 showed that the highest number of human death per road traffic accident occurred in the fourth (January, July, August) quarter and the third (March, April, May) quarter. This finding is consistent with other studies (Augustus, 2012;Ali and Ötüken, 2014) and somewhat different from the study done byStanislav and Henrikas, 2013.

The driver age group greater than 50 years was a significant factor to the number of human death per road traffic accident as compared to the age group 18-30 years. Table 4.6 revealed that the young driver (18-30)had higher contribution to the number of human death per road traffic accident as compared to elderly driver (51 and above years). This result was similar to the study (Tewolde, 2007; Roshi *et al.*, 2008; Ali and Ötüken, 2014; Haile and Demeke, 2014).

Among vehicle factors included in this study, vehicle type was found to be an important factor which had a significant contribute to the number of human death per road traffic accident. Our results showed that bus (public bus, higer bus, ISUZU modified for public transport, and others contain greater than12 persons), cargo (trailer, fluid cargo, lobags) and others (Bajaj, motor cycle, excavator, dozer) were a significantly increase the number of human death per accident compared to Minibus and automobile. Note that the possibility behind this result might be, cargo vehicles have larger mass, greater momentum and longer stopping distance. For any given speed, the greater the mass of the vehicle, the greater would be its force of impact at collision with the pedestrians leading to higher injury severities. Furthermore, it is possible that drivers of small vehicles but high speedy are more likely to weave around in traffic, change lanes, dart ahead of others or even take corners and curves faster. This finding seemed to be in accordance with other studies (Tewolde, 2007; Garrido *et. al.*, 2014).

According to the results, accident time was found to be another significant factor to determine the number of human death per road traffic accident. As can be seen in Table 4.6, more number of human death per road traffic accident happened in evening and night time as compared to afternoon time. This indicates that in the night (illegal use of road, light, speed) due to no traffic police present at a time and evening (the end of the working day) the roads are typically busy with traffic volume, so this situation might be increased in the number of human death per accidents.

The result of this study also showed that thevehicle crashes to pedestrian, and others (vehicle down, vehicle crash to inert, pedestrian jump from the vehicle, piston fail) accident type were increased the number of human death per road traffic accident as compared to vehicle crashes to Vehicle. As expected, the number of human death per road traffic accident significantly increased for pedestrians who tended to cross roadway without proper right of way. Since drivers do not expect pedestrians at locations not designated for pedestrians, they might fail to detect the pedestrian in time and take evasive actions accordingly and negligent crossing behavior has been reported to be a major cause of pedestrian collision (Anowar *et. al.*, 2010; Haile and Demeke, 2014; Garrido *et. al.*, 2014).

Chapter 5

Conclusions and Recommendations

5.1 *Conclusions*

This study was designed to identify the most important road traffic accident variables through count regression models. Thestudy also identifies the best count fit model in order to analyze the road traffic accident (human death due to road traffic accident) data. In this study, 1438 road traffic accidents were recorded without property damage accident type, out of which 522 people were died due to road traffic accident from July 2013 to July 2014 in Addis Ababa. In addition, the data contain about 70% of the road traffic accident without death event happened indicate the occurrence of excess zeros. This leads using zero-inflated model is appropriate to fit this dataset.

From the exploratory results we could identify that there is an excess zeros and high variability in the non-zero values. The variance of the numberof human death was larger than its mean, indicating that there waspossibility of overdispersion. In adition, the overdispersion parameter delta (δ), is significantly different from zero in both NB and ZINB regression models.This implies that standard Poisson regression model is not an appropriate model to fit the road traffic accident dataset.

The most fitted model was selected from different candidate models: Poisson, negative binomial (NB), zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB)using different comparison techniques. The comparison was conducted by using likelihood-ratiotest(LRT), Akakie information criteria (AIC), Bayesian information criteria (BIC) and Vuong test. LRT which was used to compare any two nested models such as Poisson versus negative binomial (NB) and zero-inflated Poisson (ZIP) versus zero-inflated negative binomial (ZINB),while Poisson versus ZIP and NB versus ZINB are non-nested models;

which wascompared using Vuong test. The model with the smallest AIC and BIC value is the most fit model.

The road traffic accident data contain an excess zeros leads the standard Poisson and negative binomial regressionmodels did not fit the data. Consequently,the standard error of the parameter estimates of the standard Poisson modelis smaller than negative binomial (NB) model and other zero-inflated models.Zero-inflated model was appropriate to fit the number of human death per road traffic accident data due to the presence of excess zero in the data and the violation of Poisson model assumption that is the mean is smaller than the variance of number of human death per road traffic accident.

This study also captured predictor variables that had significant effects on number of human death per road traffic accident. For selected ZIP model, the Poisson part, the predictor variables like month in quarter, age of the drivers, vehicle type, ownership ofvehicle, accident time, accident type and road inclination were statistically significant factors while for logistic part, injured vehicle countwasstatistically significant factor on the number of human death per road traffic accident in Addis Ababa.

5.2 Recommendation

- The Addis Ababa Traffic Control and Investigation Department (AATCID) should prepare appropriate policiesand accomplish on those selected statistically significant variables in order to reduce the number of human death per road traffic accident.
- Further studies can be made on the area of road traffic accident byconsidering detail and accurate information on the variables. For example,if the causes and consequencesof an accident are recorded in detail instead of broad categories results could be more accurate and efficient.
- A further extension of the ideas in this study is to consider other count regression model like hurdle model, and zero-inflated generalized Poisson regression model.

References

Aklilu Z. W. (2011). Use ofZero-Inflated Models forAnalysis ofImmunological Data.InteruniversityInstituteforBiostatistics and Statistical Bioinformatics.

Abdul-Rahaman H. (2014). Identification of Factors that Cause Severity of Road Accidents in Ghana: a Case Study of the Northern Region, International Journal of Applied Science and Technology, 4:3.

Ali K. C. and Ötüken S. (2014). Risk Factors Affecting Fatal Versus Non-Fatal Road Traffic Accidents: The Case of Kars Province, Turkey. International Journal for Traffic and Transport Engineering, 4:3:339 – 351.

AnowarS., YasminS. and TayR. (2010). A Logistic Model forExamining Pedestrian Serious Injury Risk inTraffic Crashes Department Of Civil Engineering,University Of Calgary.

Augustus O. A. (2012). A Monthly Analysis of Road Traffic Accident in Selected Local Government Areas of Lagos State, Nigeria. Mediterranean Journal of Social Sciences, 3:11.

Cameron A.C. and Trivedi P.K.(1990). Regression Based Test for Overdispersion in the Poisson Model. Journal of Applied Econometrics, 46:347- 364.

Cameron A.C. and Trivedi P.K. (1998). Regression Analysis of Count Data. Cambridge University Press.

Chisholm D. (2008). Road Traffic Injury Prevention: An Assessment of Risk Exposure and Intervention Cost-Effectiveness in Different World Regions.

Chong M., Ajith A., and Marcin P. (2005). Traffic Accident Analysis Using Machine Learning Paradigms. Informatica (Slovenia), 29:1:89-98.

Cox R. (1983). Some Remarks on Overdispersion. Biometrical, 70: 269-274.

Garrido R., Bastos A., Almeida A. and Elvas J. P. (2014).Prediction of Road Accident Severity using the Ordered Probit Model. Transportation Research Procedia, 3:214-223.

Geedipally S.R. (2005). Analysis of Traffic Accidents before and after Resurfacing Statistical Approach. Department of Science and Technology. Linkoping's University, Seweden.

Guyu F. (2013). Identifying Major Urban Road Traffic Accident Black-Spots: A Sub-city Based Analysis of Evidences from the City of Addis Ababa, Ethiopia.Addis Ababa University,Ethiopia,15:2.

Haile M. and Demeke L. (2014). Analysis of Factors that Affect Road Traffic Accidents in Bahir DarCity, North Western Ethiopia.Science Journal ofApplied Mathematics andStatistics, 2:5: 91-96.

Jansakul N. and Hinde J. P. (2002). Score Tests for Zero-Inflated Poisson Models. Computational Statistics and Data Analysis, 40:75-96.

Jones B., Janseen L. and Mannering F. (1991). Analysis of the Frequency and Durationof the Freeway Accidents in Seattle. Accident Analysis and Prevention 23:4: 239-255.

Lambert D. (1992). Zero-Inflated Poisson Regression with an Application to Defects in Manufacturing. Techno metrics, 34:1- 14.

Lauren, P. and Hill, S. (2005). Road traffic Injuries-Can we a global epidemic? Retrievedfrom http://www.thedoctorwillseeeyounow.com/articles/other/road-33/, accessed 21St May 2015.

Ma Z., Shao C., Yue H. and Ma S. (2009). Analysis of the Logistic Model for Accident Severity on Urban Road Environment. In Intelligent Vehicles Symposium, 983-987.

Malyshkina N. V. (2009). Empirical Assessment of the Impact of Highway Design Exceptions on the Frequency and Severity of Vehicle Accidents.

Mccullagh P. and Nelder J.A. (1989). Generalized Linear Models. Chapman and Hall, London.

Miaou S. P. and Lum H. (1993). Modeling Vehicle Accidents and Highway Geometric Design Relationships. Accident Analysis and Prevention 25:6: 689- 709.

OssenbruggenP. J., J. Pendharkar, (2001). Roadway Safety in Rural and Small Urbanized Areas Accidents Analysis andPrevention, 33:4: 485-498.

Pan C. and Prakash R. (2013). Modeling Motorway Accidents Using Negative Binomial Regression.Proceedings of theEastern Asia Society forTransportation Studies, Vol.9.

Phang Y. N. and Loh E. F (2013). Zero Inflated Models for Overdispersed Count Data. International Journal of Mathematical, Computational, Physical and Quantum Engineering, 7:8.

Roshi, E., Qirjako, G., Hysa, B. and Burazeri, G. (2008). Factors Associated with Fatal Traffic Accidents in Tirana, Albania: Cross-sectional Study 49: 734-740.

SAS Institute Inc. (2008). SAS/ETS ®9.2User's Guide. Cary, NC: SAS Institute Inc.

Sharma A.K. and Landge V.S. (2013). Zero Inflated Negative Binomials for Modeling Heavy Vehicle Crash Rate on Indian Rural Highway. International Journal of Advances in Engineering and Technology, 5:2:292-301.

Stanislav M. and Henrikas S. (2013). The analysis of traffic accidents on Lithuanian regional gravel roads, Transport, 28:1:108-115

Tewolde M. (2007). Empirical Analysis on Traffic Accidents Involving Human Injuries. Msc. Thesis Presented Addis Ababa University.

Vuong Q.H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica, 57:2:307-333

WHO Road Traffic Injuries Home Page;WWW.Who.Int/Violence-Injury-Prevention, accessed 10th March 2015.

Yaacob W. F. W., Lazim M. A. and Wah Y. B. (2010). A Practical Approach in Modeling Count Data. Proceedings of the Regional Conference on Statistical Sciences, 176-183.