Predictors of Human Death by Road Traffic Crashes in Bahir Dar City, North Western Ethiopia; A Count Data Analysis Regression Model

Metadel Azeze[1], Awoke Seyoum[2], Endalew Tesfa[1], Legesse Kassa Debusho[3]

[1]Department of Statistics, College of Science, Debre Markos University, Debre Markos, Ethiopia

[2]Department of Statistics, College of Science, Bahir Dar University, Bahir Dar, Ethiopia

[3]Department of Statistics, College of Science, University of South Africa, Pretoria, South Africa

Email address:

mekonmeti@gmail.Com (M. Azeze), tesfam32@gmail.com (E. Tesfa)

To cite this article:

Metadel Azeze, Awoke Seyoum, Endalew Tesfa, Legesse Kassa Debusho. Predictors of Human Death by Road Traffic Crashes in Bahir Dar City, North Western Ethiopia; A Count Data Analysis Regression Model. *International Journal of Theoretical and Applied Mathematics*. Vol. 6, No. 6, 2020, pp. 95-104. doi: 10.11648/j.ijtam.20200606.12

Abstract: Road traffic crashes are a major socio-economic and public health problem, affecting all people of the world and Ethiopia is a country with a very large number of traffic crashes and fatality rate. This study has major objective of assessing the predictors of road traffic accident in Bahir Dar city, Ethiopia and identifies factors that contribute to the occurrence of road traffic crashes that leads human death. Data regarding to the number of deaths per road traffic crash were obtained from Bahir Dar city administration traffic police office for a two year period from July 2015-June 2017. In this study we applied six count models namely Poisson, negative binomial, generalized Poisson, zero inflated Poisson, zero-inflated negative binomial and zero inflated generalized Poisson regression models. Based on different models comparison criteria, e.g. AIC, log likelihood and Vuong test ZIGP regression model provides more appropriate fit to the number of human death per road traffic crashes data considered in this study. Sex, age, driving under alcohol, fatigue, not give priority, days of weeks, road condition, overloading, over speeding, and type of accident were found to be statistically significant predictors of human death due to road traffic crash.

Keywords: Road Traffic Crash, GLM, Over Dispersion, AIC, BIC, Count Data, Ethiopia

1. Introduction

1.1. Background of the Study

Accident/crash is defined as anything which happens by chance, anything occurring unexpectedly and un-designed [1]. Road traffic crash is a collision or similar incident involving a moving vehicle, resulting in property damage, personal death or injury [2]. Road traffic crash is an unexpected phenomenon that occurs as a result of the use or operation of vehicles including bicycles and handcarts on the public highways and roads. A road traffic crash is defined as any vehicle accident occurring on a public highway. It includes collisions between vehicles with vehicles, vehicles with animals, vehicles with pedestrians, or vehicles with fixed obstacles. Single vehicle accidents, which involve a single vehicle, that means without other road user, are also included [3]. Accidents may be fatal, resulting in deaths of the road users (passengers, drivers or pedestrians), or minor when it is not severe enough as to cause substantial hardship [4]. Globally over 1.2 million people are killed and more than 20-50 million injured in crashes every year. The global economic losses due to road traffic crashes exceed US$ 500 billion [5]. In Africa over 80% of goods and people are transported by roads [6], and in Ethiopia road transport accounts for over 90% of all the inter-urban freight and passenger movements in the country [7].

Road traffic injuries pose a significant burden in Ethiopia, as is the case for other developing countries. Currently, developing countries contribute over 90% of the world's road traffic fatalities [5]. However, road traffic crash and poverty are linked because family bread winners are highly

represented among the road traffic crashes. Road traffic crashes are a major public health concern. Ethiopia is one of those developing countries with low level of income accompanied by the high rate of population growth. As part of the developing world, Ethiopia is predominantly an agrarian country with a low level of urbanization. Transport is an important sector in facilitating different economic activities in the national economy. Ethiopia experiences the highest rate of deaths such accidents in Sub-Saharan Africa. Road traffic injuries are growing as the vehicle use of developing countries rises [8-10].

By 2020, road traffic accidents are expected to be the third leading cause of death and disability worldwide, by some calculations matching the toll of AIDS. Residents of developing countries are at much higher risk of road traffic injuries than residents of high-income countries. They are also at greater risk of death injuries and property damage when a crash occurs. Developing countries also have inadequate trauma systems and are often unable to care for crash victims. It was indicated that unless action is taken to improve road safety systems, poor countries will continue to bear the heavy toll of road traffic fatalities [11]. According to the WHO data published in April 2011, road traffic accident deaths in Ethiopia reached 22,786 per year (2.77% of total deaths). Road accidents appear to occur regularly at some flash points such as where there are sharp bends, potholes and at bad sections of the highways. At such points over speeding drivers usually find it difficult to control their vehicles, which then results in fatal traffic accidents, especially at night [12]. Accident rates in developing countries are often 10-70 times higher than in developed countries. Whereas traffic crash situation is slowly improving in the industrialized societies (e.g. Australia, USA, UK etc.), most developing countries face a worsening situation. For developing measures aimed at reducing the rate of road traffic crashes and the consequent deaths, injuries, fatalities and property damage, there is the need for regular evaluation of the road traffic accidents.

1.2. Statement of the Problem

Road traffic crashes are major public safety and development obstacle. According to WHO [13] the current situation required high level of political dedication and took immediate action to reduce road traffic crashes. Road safety tends not to receive due consideration because not all road accidents and casualties are reported to the police and there is usually no other system of estimating road accidents and the corresponding casualties nationwide. Road accidents are too often accepted as inevitable negative side effects of motorization [14].

Ethiopia is one of the developing countries having a very low road network density and vehicle ownership level, currently Ethiopia has a relatively high accident record [15]. Road traffic accident problem in Ethiopia, especially in the metropolitan cities, is increasing at an alarming rate. Bahir Dar is one of the metropolitan cities of Ethiopia and has high road traffic accident record by different causalities [16]. In connection with the above facts, traffic volume is becoming huge and is increasing from time to time; as a result of different factors, road traffic accidents have increased over the years and are becoming a common day to day phenomenon resulting in loss of life, human suffering, destruction of properties and the environment. The number of victims treated in hospital, health center and clinics also show upward trend. Bahir Dar special zone health department reported that in the previous years (2000–2002) only 3,188 road casualties received medical treatment as in-patients and out-patients [17]. Some researchers have investigated the suitability of the binary logistic regression, Poisson regression and negative binomial model to predict accident frequencies at intersections or roadways [10, 18-22]. These researches have foreground the fact, because accident occurrences are necessarily discrete, often discontinuous and more likely random events, it is better to use Poisson regression for equal variance and for over dispersion negative binomial models than multiple linear regression models but there many count models that used to handle over dispersion. So the purpose of the study was to determine the significant factors by applying GLMs in road traffic crashes data.

1.3. Objectives of the Study

General objective
The general objective of this study was to identify the predictors of human death related to road traffic crashes by using count data regression model analysis.
Specific objective
a. To identify the factors significantly affect/causes of road traffic accidents (that leads human death by road traffic crashes).
b. To identify the models and select the robust model for count response data related to human death per traffic crash.

1.4. Significance of the Study

The findings of this study will be used for making awareness for the concerned bodies about problems related to the causes of road traffic accident to take appropriate measures. To show the severity of the road traffic accident for the readers so that they will save their lives and livelihoods from loss. To serve as information for those researchers interested in conducting further studies in the area. To help for the policy makers to design appropriate strategies to reduce road traffic crashes which results human death. Generally, the results obtained from this study and recommendations were made used for all members of the community of the city.

2. Methodology

2.1. Description of Study Area

Bahir Dar is special zone and capital city of Amhara National Regional State (ANRS). Bahir Dar is one of the leading tourist destinations in Ethiopia, with a variety of

attractions in the nearby Lake Tana and Blue Nile River.

2.2. Sources of Data and Study Design

Secondary data sources were used in cross–sectional survey of the road traffic crashes. The data used in this study was recorded from July 2015 to June 2017 by the traffic police in Bahir Dar city administration traffic police office on daily basis. The data provide information on road traffic crashes that occur within two years on consecutive days. The variables are used in this study are the number of human death per road traffic crashes as response variable and for the explanatory or predictor variables sex of driver, age of driver, driver-vehicle relationship, education level of driver, driving under alcohol, owner ship of vehicle, driving under fatigue, not give priority, day of weeks, accident time, type of road, road geometry, road condition, type of vehicle, overloading, over speeding and type of accident.

2.3. Methods of Data Analysis

2.3.1. Generalized Linear Models (GLMs)

GLMs represent a class of regression models that allow us to generalize the linear regression approach to accommodate many types of response variables including count, binary, proportions and positive valued continuous distributions [24, 25]. Because of its flexibility in addressing a variety of statistical problems and the availability of software to fit the models, it is considered a valuable statistical tool and is widely used. In fact, the generalized linear model has been referred to as the most significant advance in regression analysis in the past twenty years [25]. Generalized linear models GLMs extend ordinary regression models to encompass non normal response distributions and modeling functions of the mean. Three components specify a generalized linear model: A random component identifies the response variable Y and its probability distribution; a systematic component specifies explanatory variables used in a linear predictor function; and a link function specifies the function of E(Y) that the model equates to the systematic component. [24] introduced the class of GLMs, although many models in the class were well established by them.

A generalized linear model (GLM) consists of three components: A *random component*, specifying the conditional distribution of the response variable, Yi (for the ith of n independently sampled observations), given the values of the explanatory variables in the model. In the initial formulation of GLMs, the distribution of Y_i was a member of an exponential family. This family has probability density function or mass function of form $f(y_i; \theta_i) = a(\theta_i)b(y_i) exp[y_i Q(\theta_i)]$.

Several important distributions are special cases, including the Poisson and binomial. The value of the parameter θ_i may vary for is i=1,..., N, depending on values of explanatory variables. The term $Q(\theta_i)$ is called the natural parameter is sufficient for basic discrete data models. The *systematic component* of a GLM relates a vector $(\eta_1, ..., \eta_N)$ to the explanatory variables through a linear model. Let x_{ij} denote the value of predictor j (j=1,2,..., p) for subject i. Then $\eta_i = \sum_j \beta_j x_{ij}, i = 1, ..., N$. This linear combination of explanatory variables is called the linear predictor. Usually, one $x_{ij} = 1$ for all i, for the coefficient of an intercept (often denoted by) in the model. The third component of a GLM is a *link function* that connects the random and systematic components. Let $\mu_i = E(Y_i), i = 1, ..., N$. The model links μ_i to η_i by $\eta_i = g(\mu_i)$ where the link function g is a monotonic, differentiable function. Thus, g links $E(Y_i)$ to explanatory variables through the formula $g(\mu_i) = \sum_j \beta_j x_{ij}, i = 1, ..., N$.

The link function $g(\mu) = \mu$ called the identity link, has $\eta_i = \mu_i$. It ii specifies a linear model for the mean itself. This is the link function for ordinary regression with normally distributed Y. The link function that transforms the mean to the natural parameter is called the canonical link. In summary, a GLM is a linear model for a transformed mean of a response variable that has distribution in the natural exponential family. We now illustrate the three components by introducing the key GLMs for discrete response variables. The following subsections show example models for count data. Count data are non-negative integers; they represent the number of occurrence of an event within a fixed period. e.g., number of death per road traffic crashes.

(i) Poisson Regression Model

The standard Poisson distribution is a fundamental distribution to understand regression count models. According to [26], the apparent simplicity of Poisson comes with two restrictive assumptions. First, the variance and mean of the count variable are assumed to be equal. The other restrictive assumption is that occurrences of the event are assumed to be independent of each other. A regression model based on this distribution follows by conditioning the distribution of y_i on a k-dimensional vector of covariates, $x_i=[x_1, ..., x_k]$, and parameters β, through a continuous function $E[y_i|x_i] = \lambda_i$ [27]. The Poisson mass function is given by; $f(y_i/x_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}, y_i = 0, 1, 2, ..., n$

In the log-linear version of the model the mean parameter is parameterized as $\lambda_i = exp(xi'\beta) and \log(\lambda_i) = x_i'\beta$. Given independent observations with the density function, the log-likelihood function can be obtained by: $l(\beta) = \sum_{i=1}^{n}[y_i log(\lambda_i) - \lambda_i - \log(y_i!)] => \sum_{i=1}^{n}[y_i x_i'\beta - exp(x_i'\beta) - \log(y_i!)]$.

Over-dispersion Models: Over dispersion is not expected that the residual deviance roughly equal to residual degree of freedom when the Poisson model fits the data reasonably. But, large residual deviance implies that the conditional variance exceeds the mean the Poisson model does not fit [27]. This usual incidence in the analysis of discrete outcome data is referred as over-dispersion ($Var(y_i) > E(y_i)$). If there is over-dispersion causing the variance to be larger than the mean, then the estimation will be inefficient using a Poisson regression.

(ii) Negative Binomial Regression Model

Negative binomial regression model is applicable for modeling over-dispersion, which is a conjugate mixture distribution for count data. When the Poisson model assumption fails, negative binomial regression model may fit

better, and address the over-dispersion problem. The probability mass function of negative binomial distribution is given by:

$$f\left(\frac{y_i}{\mu_i}, \delta\right) = \left(\frac{\Gamma(y_i + \delta^{-1})}{\Gamma(\delta^{-1})\Gamma(y_i + 1)}\right)\left(\frac{\delta^{-1}}{\delta^{-1} + y_i}\right)^{\delta^{-1}}\left(\frac{\mu_i}{\delta^{-1} + \mu_i}\right)^{y_i}, y_i$$

$$= 0, 1, 2, \dots, n$$

The regression model is also given by $\mu_i = exp\,(xi'\beta)$ or $\log\,(\mu_i) = x_i'\beta$. With mean E $(y_i/x_i) = \mu_i = \exp(x_i'\beta)$ and variance, $var(y_i/x_i) = \mu_i(1 + \delta\mu_i)$, where $\Gamma(.)$ is the gamma function and the index δ (read as delta) is called the dispersion parameter. As δ approaches to zero, the variance and mean become identical, then the negative binomial model reduces to the classical Poisson model. If $\delta > 0$, the variance will exceed the mean, that is $var(y_i) > E(y_i)$ and the distribution allows for over dispersion [28]. The negative binomial log-likelihood function is given by:

$$l(\delta, \beta) = \sum_{i=1}^{n}\left[\begin{array}{c} log\left(\frac{\Gamma(yi + \delta^{-1})}{\Gamma(\delta^{-1})\Gamma(yi + 1)}\right) \\ -(y_i - \delta^{-1})log(1 + \delta\mu_i) + y_i log(\delta\mu_i) \end{array}\right]$$

(iii) The Generalized Poisson Regression Model

The Generalized Poisson regression model is another alternative way for modeling over-dispersion and it's a good competitor with negative binomial model. The advantage of using the generalized Poisson regression model, is that it can be fitted for over-dispersion, $Var(y_i) > E(y_i)$ [29]. Suppose y_i is a count response variable that follows a generalized Poisson distribution, the probability density function of y_{i}, $=1,2,\dots,n$ is given as [29, 30];

$$f(Y_i/\mu_i, \delta) = \left[\frac{\mu_i}{1 + \delta\mu_i}\right]^{yi}\frac{(1 + \delta y_i)^{y_i - 1}}{y_i!}\exp\left[\frac{-\mu_i(1 + \delta y_i)}{1 + \delta\mu_i}\right], yi = 0, 1, , n$$

With mean $E(y_i) = \mu_i = exp(x_i'\beta)$ and variance $Var(y_i) = \mu_i(1 + \delta\mu_i)^2$, where δ is dispersion parameter. The generalized Poisson distribution is a natural extension of the Poisson distribution. If $\delta = 0$, reduces to the Poisson . If $\delta > 0$, it means $Var(y_i) > E(y_i)$, and the distribution represents count data with over dispersion. If it is assumed that the mean or the fitted values is multiplicative, (I. e.) E $(y_i/x_i) = \mu_i = \exp(x_i'\beta)$. Where x_i is a p X1 vector of explanatory variables, and β is a p X1 vector of regression parameters. The log-likelihood functions of the GPR model is given by [29]; $l(\beta, \delta) = \sum_{i=1}^{n}\left\{y_i log[\frac{\mu_i}{1 + \delta\mu_i}] + (y_i - 1)log(1 + \delta y_i) - \left[\frac{\mu_i(1 + \delta y_i)}{1 + \delta\mu_i}\right] - log(y_i)\right\}$

2.3.2. Zero Inflated Regression Models

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

(i) Zero Inflated Poisson Regression Model

Zero inflated (ZI) models can be used to account for

excess zeros. ZIP models have less adequate than ZINB and ZIGP models when the presence of over dispersion due to excess zeros and unobserved heterogeneity. The probability mass function of ZIP is given by

$$f(Y_i = /\psi_i, \lambda_i) = \begin{cases} \psi_i + (1 - \psi_i)e^{-\lambda_i}, if\ y_i = 0 \\ (1 - \psi_i)\frac{\lambda_i^{y_i}\ e^{-\lambda_i}}{y_i!}, if\ y_i > 0, i = 1, 2, \dots, n' \end{cases}$$

where λ_i is the mean of the non-zero outcomes that can be expressed with the associated explanatory covariates using a natural logarithmic link is given by: $\log(\lambda_i) = x_i'\beta$, where $X_i = (1, x_{i1}, x_{i2},\dots, x_{ip-1})'$ is a px1 vector of explanatory variable of the ith observation and β is px1 vector of regression coefficient parameters and ψ_i is the probability of an excess zero which can be estimated by the logistic regression [31, 32]. That is Logit(ψ_i) $= ln\left(\frac{\psi_i}{1 - \psi_i}\right) = z_i'\gamma$ where $\psi_i = \frac{e^{z_i'\gamma}}{1 + e^{z_i'\gamma}}, i = 1, 2, \dots, n$, where $Z_i = (1, Z_{i1}, Z_{i2}, Z_{iq-1})$ is a qx1 vector of explanatory variable for the zero-inflation part model of the ith observation and $\gamma = (1, \gamma_1, \gamma_2, \dots, \gamma_{q-1})$ is qx1 vector of regression coefficient parameters. Unlike the Poisson distribution, which is determined by a single parameter, the ZIP distribution is determined by two parameters, λ_i and1 ψ_i. The ZIP model has mean and variance $E(Y_i) = (1 - \psi_i)\lambda_i$ and $var(Y_i) = (1 - \psi_i)(\lambda_i + \psi_i\lambda_i^2)$ respectively [33].

(ii) Zero-Inflated Negative Binomial Regression Model

If the dependent variable presents a high proportion of zeros which could create problems for the negative binomial estimation, a modified count model is the zero inflated Negative Binomial (ZINB) models which take the existence of excess zeros in to account. Zero-Inflated Negative Binomial (ZINB) regression model is an extension of the NB regression model. Then the probability density function of ZINB the random variable Y_i distributed as ZINB is given by:

$$f(Y_i = /\psi_i, \lambda_i) =$$

$$\begin{cases} \psi_i + \frac{(1 - \psi_i)}{(1 + \delta\lambda_i)^{\delta^{-1}}}, if\ y_i = 0 \\ (1 - \psi_i)\frac{\Gamma(y_i + \delta^{-1})}{\Gamma(\delta^{-1})\Gamma(y_i + 1)}\frac{(\delta\lambda_i)^{y_i}}{(1 + \delta\lambda_i)^{y_i + \delta^{-1}}}, if\ y_i > 0, i = 1, 2, 9, n \end{cases}$$

The ZINB model with mean and variance, $E(y_i) = (1 - \psi_i)\lambda_i$ and $var(y_i) = (1 - \psi_i)(\lambda_i + \frac{\lambda i^2}{\delta^{-1}})$ respectively [34]. In order to obtain the parameter estimates of ZINB regression models, $\hat{\beta}, \hat{\gamma}$ and $\hat{\delta}$ the Newton-Raphson method can be used [27].

(iii) Zero Inflated Generalized Poisson Regression Model

Besides ZINB, zero-inflated generalized Poisson (ZIGP) regression has been proposed as an alternative to handle zero-inflation and additional over dispersion in count data. Zero inflated generalized Poisson (ZIGP) distribution is another alternative for modeling over dispersed count data with excess zeroes.[35] have used ZIGP distribution to model domestic violence data. A zero-inflated generalized Poisson (ZIGP) regression model is defined as.

$$f(Y_i = /\lambda_i, \psi_i) = \begin{cases} \psi_i + (1 - \psi_i) exp\left(\dfrac{-\mu_i}{1 + \delta\mu_i}\right), \text{if } y_i = 0 \\[3mm] (1 - \psi_i)\left[\dfrac{\mu_i}{1 + \delta\mu_i}\right]^{y_i} \dfrac{(1 + \delta y_i)^{y_i-1}}{y_i!} \ exp\left[\dfrac{-\mu_i(1 + \delta y_i)}{1 + \delta\mu_i}\right] \ if \ y_i > 0, i = 1,2, \dots, n \end{cases}$$

The ZIGP model is a special case of a two-class finite mixture model with mean and variance

$$E(Y_i) = (1 - \psi_i)\mu i \text{ and } var(Y_i) = (1 - \psi_i)[\mu_i{}^2 + \mu_i(1 + \delta\mu_i)^2] - (1 - \psi_i{}^2)\mu_i{}^2) \text{ respectively.}$$

2.4. Goodness of Fit Tests

Over-dispersion Test: Poisson model is a special case of negative binomial and generalized Poisson model. To assess the adequacy of the negative binomial and generalized Poisson model over the Poisson regression model, we can test the hypothesis: $H_0: \delta = 0$ vs $H_A: \delta > 0$. This is to test for the significance of the over-dispersion parameter δ. The presence of the over-dispersion parameter δ in the NB and GP regression model is justified when the null hypothesis $H_0: \delta = 0$, is rejected. A likelihood-ratio (LR) tests for the over-dispersion parameter δ, in the negative binomial (NB) and generalized Poisson (GP) specification against the Poisson model specification [27]. In order to test the hypothesis the likelihood ratio test (LRT) is given by: $LRT_\delta = -2[l(\hat{\mu}) - l(\widehat{\mu}, \delta)]$ where $l(\hat{\mu})$ and $l(\widehat{\mu}, \delta)$ is the maximized log-likelihood under the given models respectively.

Likelihood Ratio Test (LRT): The LRT is a test of the overall model and a test of a null hypothesis H_0 against an alternative H_A based on the ratio of two log-likelihood functions. The overall test statistic for LRT is given as LRT=G^2=-2 $(L_R - L_F) \sim x^2_{p-1}$, where: L_R is the log-likelihood of the null model (reduced model) and L_F is the log-likelihood of the model comprising k predictors, p is number of parameters and x^2_{p-1} is a chi-square distribution. If the test statistics exceeds the critical value, the null hypothesis is rejected. That means the overall model is significant. The statistic of LRT for δ is given by the following equation: LRT=-2 $(L_1 - L_2)$. This statistic has a Chi-squared distribution and L is log-likelihood. If the statistic is greater than the critical value then, the model 2 is better than the model 1.

Information Criteria: Akakie information criteria (AIC) and Bayesian's information criteria (BIC) are goodness of criteria used for model selection. AIC and BIC are the most common means of identifying the model which fits well by comparing two or more than two nested models and models with the largest log-likelihood value can be chosen as the best model for describing the data under consideration. The formula is given as: $\left.\begin{array}{l} AIC = -2L + 2k \\ BIC = -2L + k \ln(n) \end{array}\right\}$ where 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 of AIC and BIC and largest log-likelihood value is the most appropriate fitted model to the dataset.

Vuong's test: The Vuong's test is a non-nested test that is based on a comparison of the predicted probabilities of two models that do not nest [38]. That means Vuong test statistics are needed to provide the appropriateness of zero-inflated models against the standard count models. Under the null that the models are indistinguishable, the test statistic is asymptotically distributed standard normal. Given that $p_1(Y_i/X_i)$ and $p_2(Y_i/X_i)$ are the predicted probability of the zero inflated models versus ordinary models respectively. That means set as model1 zero inflated models and model2 ordinary models. We want to test the following hypotheses of Vuong test are: H_o: The two models are equivalent versus H_A: The two models are not equivalent. Vuong showed that asymptotically, V has a standard normal distribution. As Vuong notes, the test is directional [38]. If $V > Z_{\alpha/2}$, the first model is preferred, if $V < -Z_{\alpha/2}$, the second model is preferred and if $|V| < Z_{\alpha/2}$, none of the models are preferred (the two models are equivalent). The Vuong test statistics can be expressed as [38]: $V = \dfrac{\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} m_i\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(m_i - \bar{m})^2}} = \dfrac{\sqrt{n}}{sd(m)}\bar{m}, \ m_i = log(\frac{p_1(y_i/x)}{p_2(y/x_i)})$, where, $\bar{m}$ is the mean of m_i, $sd(m)$ is the standard deviation of m_i and n is the sample size. In general, $P_N(Y_i/X_i)$ is the predicted probability of observed count for cases i from model N, then the Vuong test statistic is simply the average log-likelihood ratio suitably normalized.

Test for individual predictors: Let β denote an arbitrary parameter. Consider a significance test of $H_0: \beta_0 = 0$. The simplest test statistic uses the large-sample normality of the ML estimator $\hat{\beta}$, let $SE(\hat{\beta})$ denote the standard error of $\hat{\beta}$, evaluated by substituting the ML estimate for the unknown parameter in the expression for the true standard error. The hypothesis is given as follows: $H_0: \beta_i = 0 \ Vs \ H_A: \beta_i \neq 0$. When H_0 is true, the test statistics is $Z = \frac{\hat{\beta} - \beta_0}{SE(\hat{\beta})}$ The significance test for each coefficient in the model will be done using Wald chi-square the Wald statistic (Z^2) is: $(Z^2 = \frac{\hat{\beta} - \beta_0}{SE(\beta)})^2$. Under H_0 true, Z^2 is a chi-square distribution with 1 degree of freedom. Likelihood-ratio tests are generally considered to be superior [39].
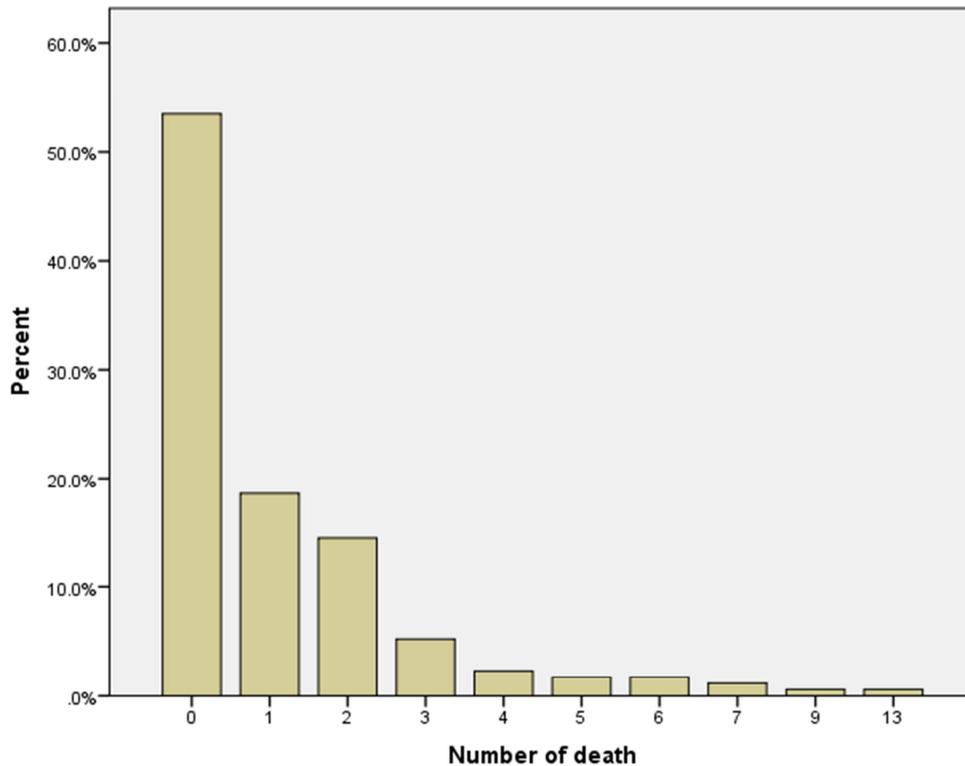
Figure 1. *Number of human death due road traffic crashes.*

In this study the analysis pertains to search the important factors. It is always a good idea to start with descriptive statistics.

3. Results and Discussion

3.1. Descriptive Statistic

Figure 1 showed that there are large numbers of zero values (54%) highly picked at zero and a positively (or right) skewed distribution of human death per road traffic crashes. In this case zero-inflated count data models s better fit the data. The results also indicated that the maximum frequency of number of deaths per accident recorded was 13.

Table 1. *Summary Statistics for the number of human death due to road traffic accident.*

Mean	Variance	Skewness	Kurtosis
1.13	3.446	2.892	11.74

As shown in Table 1, the variance of human death per road traffic accident (3.45) was greater than its mean (1.13). This indicated the possibility of over dispersion and hence the standard Poisson regression model was not appropriate to fit the road traffic accident data. Thus one might expect that ZIP, ZINB and ZIGP would possibly be better models to predict the traffic crashes dataset.

3.2. Statistical Model Results

Goodness of fit and Comparison of Models: In order to select the best model which fits the data well, from the above six models, different model selection criteria were considered

Table 2. *Tests for goodness of fit and model selection summary results of Poisson, NB, GP, ZIP, ZINB and ZIGP models.*

Criteria	Poisson	ZIP	NB	ZINB	GP	ZIGP
Log-likelihood	-178.998	188.87	197.69	-192.195	-186.67	-167.72
AIC	432.683	399.12	428.39	401.382	401.34	397.996
BIC	549.14	460.946	497.636	467.47	449.92	419.051
Over dispersion (δ)	-	-	0.074	0.172	0.072	0.135
δ p-value	-	-	0.0001*	0.0025*	0.000*	0.00015*
Vuong statistic	2.6656 (<2e-16)		6.1677 (<2e-16)		4.1526 (<2e-16)	

"*"shows the significance codes.

Then the Table 2 showed that the model selection criteria among the candidates. First, the calculated value of the Vuong test (2.67) was greater than the hypothetical value (1.96) for ZIP versus Poisson model. This value revealed that ZIP model was preferred to Poisson model. In the second case, comparison of ZINB versus NB models, the calculated value of the Vuong test is 6.168, revealed that the ZINB model was preferred to NB regression model and in the third case, the comparison ZIGP versus GP models, the statistic value of Vuong test is 4.15, and revealed that the ZIGP model was preferred to GP regression model. Finally, to compare the ZIP, ZINB and ZIGP models, log likelihood, AIC and BIC were used. Therefore, ZIGP model is better fitted model for human death per road traffic crashes data than ZIP and ZINB models. Log likelihood value is large and AIC and BIC values were found to be small for ZIGP model as compared to other count models. Therefore, it is possible to conclude that ZIGP model was more appropriate than the ZIP and ZINB models to fit road traffic crashes dataset.

Table 3. *Parameter estimates of ZIGP regression mode.*

Count model coefficients						
Parameters		**Coef.**	**Exp (Coef.)**	**Std. Err.**	**z value**	**Pr(>|z|)**
Sex of driver	Male	(Ref.)				
	Female	-0.868	0.42	0.432	-2.0062	0.004**
Age of driver	18-30	(Ref.)				
	31-50	-0.489	0.613	0.321	-1.5248	0.127
	>50	-1.236	0.291	0.603	-2.0510	0.034*
Driving under alcohol	Yes	(Ref.)				
	No	-2.641	0.071	1.008	-2.6192	0.009**
Driving under fatigue	Yes	(Ref.)				
	No	-1.923	0.146	0.602	-3.1968	0.0014**
Not give priority	To vehicle	(Ref.)				
	To pedestrian	0.67	1.954	0.433	4.2066	2.59e-1**
	Others	0.451	1.57	0.304	3.1003	0.0019**
Days of weeks	Sunday	(Ref.)				
	Monday	0.498	1.645	0.388	1.2813	0.0012*
	Tuesday	-0.069	0.933	0.419	-0.165	0.868
	Wednesday	0.582	1.789	0.438	1.3270	0.184
	Thursday	0.555	1.742	0.394	1.4079	0.159
	Friday	0.619	1.857	0.381	1.6236	0.104
	Saturday	0.691	1.996	0.355	2.4104	0.015*
Road condition	Wet	(Ref.)				
	Dry	0.624	1.866	0.1920	3.2501	0.0012**
Overloading	Yes	(Ref.)				
	No	-1.187	0.305	0.594	-1.9971	0.036*
over speeding	Yes	(Ref.)				
	No	-1.933	0.145	0.753	-2.5678	0.0102*
Type of accident	Vehicle–vehicle	(Ref.)				
	Vehicle-pedestrian	0.397	1.487	0.438	3.1883	0.0014**
	Vehicle-others	0.521	1.684	0.439	3.4629	0.0005***
	Reverse	0.59	1.804	0.344	6.1412	8.19e-1***
Intercept		-0.624	0.536	0.408	-1.5299	0.126
Zero inflation model coefficients						
Driving under alcohol	Yes	(Ref.)				
	No	-0.488	0.614	0.322	-1.5148	0.0024**
over speeding	Yes	(Ref.)				
	No	-1.393	0.248	0.573	-3.1863	0.012*
Intercept		-1.352	0.258	0.825	-2.7231	0.0003***
δ		0.135	1.1445	0.278	1.167	0.00015***

Note: -Signif. codes:'***'0.001 '**'0.01 '*'0.05 '.'011 ' '1.

Interpretations of Count model coefficients of ZIGP regression model

As shown in Table 3 sex of driver had significant impact on the number of human deaths per accident. The expected number of deaths per road traffic accident had decreased by 58.0% for the female drivers as compared to male drivers while holding all other variables in the model constant. This result was consistent with the study [40].

Driver's age had significant impact on the number of deaths per accident. The expected number of deaths per traffic accident had decreased by 70.9% for drivers in the age group above 50 years as compared to the drivers in the age

group 18-30 years while holding all other variables in the model constant. This result was similar to the study [8, 19, 21, 41-45]. The model also shows that the driver's driving under alcohol had significant impact on the number of deaths per accident. The expected number of deaths per accident had decreased by 92.9% for driver driving without drinking alcohol as compared to the drivers driving with drinking alcohol while holding all other variables in the model constant. This finding seemed to be in accordance with other studies [22, 43, 45]. From the result drivers driving under fatigue had significant impact on the number of human deaths per accident. The expected number of deaths per accident had decreased by 85.4% for drivers driving without fatigue as compared to drivers driving with fatigue while holding all other variables constant in the model. Fatigue consider only in this study. The model also revealed that driver not give priority had statistically significant impact on the number of death per traffic accidents. The expected number deaths per accident were increased by 95.4% and 57.0% for the driver not give priority to pedestrian and to others respectively, compared to the drivers not give priority to vehicles controlling for the other variables in the model constant. This result was consistent with [45].

The finding of this study also revealed that the day of weeks had a significant impact on the number of deaths per accident. The expected number of deaths per traffic accident had increased by 64.5% and 99.6% for Monday and Saturday respectively, compared to the day of weeks Sunday and the rest categories have the same effect with Sunday while holding all other variables in the model are constant.

The finding shows the road condition had statistically significant impact on the number of death per traffic accidents. The expected number deaths per road traffic accident had increased by 86.6% for dry condition of road as compared to for a wet condition of road controlling for the other variables in the model constant. This result is contradicted with [21, 45].

In the model the variable overloading had significant impact on the number of deaths per accident. The expected number of deaths per accident had decreased by 69.5% for drivers driving without overloading as compared to the drivers driving with overloading while holding all other variables in the model constant.

The model revealed that the predictor over speeding had significant impact on the number of deaths per traffic accident. The expected number of deaths per accident had decreased by 85.5% for drivers driving without over speeding as compared to the drivers driving with over speeding while holding all other variables in the model constant. This result is consistent with [22, 43]. Finally the type of accident had the significant effect on the number of deaths per accidents. The expected number of deaths per road traffic accident had increased by 48.7%, 68.4% and 80.4% for the accident type of vehicle to pedestrian, vehicle to others and reverse of vehicle respectively, compared to the accident type of vehicle to vehicle while holding all other variables in the model constant. This result was similar to the study [21, 44].

Interpretations of zero-inflation part of the model

Zero inflated models are interpreted as a mixture of structural and sampling zeros from two processes; the process that generates excess zeros from a binary distribution which are the structural zeros, and the process that generates both non-negative and zero counts from GP distributions which are the sampling zeros. The results of Table 3 above indicated the parameter estimates of the Zero-Inflated (logit model) part of the ZIGP regression model for examining the impact of explanatory variable on the odds of being in the always zero group.

Table 3 showed that drivers driving under alcohol has a significant impact on the odds of being in the always zero group. The odds of no occurrence of human death (being always zero group) decreased by a factor 0.614 (38.6%) for drivers driving without drinking alcohol as compared to driving with drinking alcohol holding all other variables in the model constant. And also the variable over speeding has a significant impact on the odds of being in the always zero group. The odds of no occurrence of deaths (being always zero group) decreased by a factor 0.248 (75.2%) for drivers driving without over speeding as compared to driving with over speeding holding all other variables in the model constant.

4. Conclusions

Road traffic crashes are increasing at an alarming rate, causing the loss of life and resources. This study revealed that the predictor variables that had significant impact on number of human death per road traffic crashes and also identifies the best count fit model in order to analyze the road traffic crash (human death due to road traffic crashes) data. For the selected ZIGP model, the generalized Poisson part, the predictor variables like sex of driver, age of the drivers, driving under alcohol, driving under fatigue, not give priority, day of weeks, road condition, over loading, over speeding and type of accident were statistically significant factors on the number of human death per road traffic crashes in this study. Then by giving more attention for these factors we can reduce the number of human death due to road traffic crashes. Finally, in our belief the road traffic crashes can be reduced if the significant factors are properly taken care of.

5. Recommendations

a. Bahir Dar city administration traffic police office and the police commission should prepare appropriate policies and strategies & accomplish on those selected statistically significant variables in order to reduce the number of human death due to road traffic injuries.

b. Further studies can be made on the area of road traffic crashes by considering detail and accurate information on the determinant variables that are recorded in detail instead of broad categories results could be more accurate and efficient in the study.

References

[1] Odugbemi, O., *Road Transportation and Tourism in Nigeria*, 2010, Joja Press, Lagos.

[2] Astrom, J., M. Kent, and R. Jovin, *Signatures of Four Generations of Road Safety Planning in Nairobi City, Kenya In.* Journal of Eastern African Research and Development, 2006. 20: p. 186-201.

[3] Safecarguide, *Retrieved January 22, 2007 from the World Wide Web http://www. safecarguide.com/exp/intro/idx.htm.* 2004.

[4] Sarin, S., *ROAD TRAFFIC SAFETY IN INDIA: ISSUES AND CHALLENGES AHEAD.* Indian Highways, 1998.

[5] WHO, *Global status report on road safety: Time for action. Geneva: World Health Organisation..* 2009.

[6] ECA, *African Road Safety congress, compendium of Papers. Addis Ababa, Ethiopia.* 2006.

[7] Atnafseged, K., *Road Safety Management Crisis in Ethiopia. Unpublished Report.* 2000.

[8] Zewude, B. T. and K. M. Ashine, *Statistical Modeling on Determinants of Traffic Fatalities and Injuries in Wolaita Zone, Ethiopia.* Global Journal of Human-Social Science Research, 2016.

[9] Abdella, A., *Statistical Analysis of Correlates of Number of Fatalities per Traffic Accident in Addis Ababa Using Count Data Models*, 2013, Addis Abeba university.

[10] Lauren, P. and S. Hill, *Road traffic Injuries-Can we avoid global epidemic.* Retrieved from the World Wide Web http://www. thedoctorwillseeeyounow. com/articles/other/road-33, 2005.

[11] Atubi, A., *Urban Transportation: An Appraisal of Features and Problems in the Nigerian Society.* International Journal of Geography and Regional Planning, 2009. 1 (1): p. 58-62.

[12] WHO, *Road Traffic Injuries Home Page; WWW.Who.Int/Violence-Injury-Prevention.* 2015.

[13] Mohammed, M., *Costing Road Traffic Accidents in Ethiopia.* 2011.

[14] Endris, M., *The Causes of Road Traffic Accidents in Bahir Dar City, Ethiopia.* International Journal of African and Asian Studies, 2015. Vol. 11.

[15] KELEMU, T., *A STUDY ON THE SOCIO-ECONOMICAL IMPACT OF ROAD TRAFFIC ACCIDENTS IN BAHIR DAR TOWN*, 2012, St. Mary's University.

[16] YAYEH, A., *THE EXTENT, VARIATIONS AND CAUSES OF ROAD TRAFFIC ACCIEDNTS IN BAHIR DAR.* 2003.

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

[18] Tewolde, M., *Empirical analysis on traffic accidents involving human injuries.* The case of Addis Ababa. University of Addis Ababa, Ethiopia, 2007.

[19] Malyshkina, N. V., *Empirical Assessment of the Impact of Highway Design Exceptions on the Frequency and Severity of Vehicle Accidents.* 2009.

[20] Abdissa, M., *Analysis of Human Deaths by Road Traffic Accident in Oromia Region, Ethiopia*, 2018, Addis Ababa University.

[21] Zewde, T., *Determinants that lead drivers into traffic accidents: a case of Arba Minch city, south Ethiopia.* Sci J Appl Math Stat, 2017. 5 (6): p. 210-5.

[22] Agency, E. M., *National Atlas of Ethiopia, Addis Ababa: EMA.* 2008.

[23] Nelder, J. A. and R. W. Wedderburn, *Generalized linear models.* Journal of the Royal Statistical Society: Series A (General), 1972. 135 (3): p. 370-384.

[24] Hoffmann, J. P., *Generalized linear models: An applied approach.* 2004: Pearson College Division.

[25] Sturman, M. C., *Multiple approaches to analyzing count data in studies of individual differences: The propensity for type I errors, illustrated with the case of absenteeism prediction.* Educational and Psychological Measurement, 1999. 59 (3): p. 414-430.

[26] Cameron, A. C. and P. K. Trivedi, *Regression analysis of count data.* Vol. 53. 2013: Cambridge university press.

[27] Agresti, A., *An introduction to categorical data analysis.* 2007: John Wiley.

[28] Wang, W. and F. Famoye, *Modeling household fertility decisions with generalized Poisson regression.* Journal of Population Economics, 1997. 10 (3): p. 273-283.

[29] Famoye, F., *Restricted generalized Poisson regression model.* Communications in Statistics-Theory and Methods, 1993. 22 (5): p. 1335-1354.

[30] Lambert, D., *Zero-Inflated Poisson Regression with an Application to Defects in Manufacturing. Techno metrics, 34: 1-14.* 1992.

[31] Long, J. S., *Advanced quantitative techniques in the social sciences: Volume 7. Regression models for categorical and limited dependent variables*, 1997, SAGE.

[32] Liu, C., et al., *Modeling lumber value recovery in relation to selected tree characteristics in black spruce using the Optitek sawing simulator.* Forest products journal, 2007. 57 (4): p. 57.

[33] Zuur, A. F., et al., *Zero-truncated and zero-inflated models for count data*, in *Mixed effects models and extensions in ecology with R.* 2009, Springer. p. 261-293.

[34] Famoye, F. and K. P. Singh, *Zero-inflated generalized Poisson regression model with an application to domestic violence data.* Journal of Data Science, 2006. 4 (1): p. 117-130.

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

[36] Vuong, Q. H., *Likelihood ratio tests for model selection and non-nested hypotheses.* Econometrica: Journal of the Econometric Society, 1989: p. 307-333.

[37] Agresti, A., *Introduction to categorical analysis. Wiley Series in Probability and Statistics.* 2007.

[38] Garrido, R., et al., *Prediction of Road Accident Severity using the Ordered Probit Model. Transportation Research Procedia, 3: 214-223.* 2014.

[39] Fenta, H. M. and D. L. Workie, *Analysis of Factors that affect road traffic accidents in Bahir Dar city, North Western Ethiopia.*

[40] Bisrat, M. *Determinants of Traffic Fatalities and Injuries in Addis Ababa. Unpublished Msc thesis,.* 2010.

[41] Qirjako, G., et al., *Factors associated with fatal traffic accidents in Tirana, Albania: crosssectional study.* Croatian medical journal, 2008. 49 (6): p. 734-740.

[42] Worku, G., *COUNT REGRESSION MODELS OF HUMAN DEATHBY ROAD TRAFFIC ACCIDENTSIN ADDIS ABABA, ETHIOPIA.* 2015.

[43] Belachew, M. and D. Zeleke, *Statistical analysis of road traffic car accident in Dire Dawa Administrative City, Eastern Ethiopia.* Science Journal of Applied Mathematics and Statistics, 2015. 3 (6): p. 250-256.

[44] Lee, J. and F. L. Mannering, *Impact of roadside features on the frequency and severity of run-off-roadside accidents: An empirical analysis. Accident Analysis and Prevention, 34 (2), 149–161.* 2002.

[45] Haadi, A.-R., *Identification of Factors that Cause Severity of Road Accidents in Ghana: A Case Study of the Northern Region*, 2012.