Abyss. J. Sci. Technol. Vol. 5, No. 1, 2020, 26-33
ISSN 2616 – 4728 (Online); 2616 – 471X (Print)
© 2020Wollo University

Abyssinia Journal of
Science and Technology

The Application of Count Regression Models on Traffic Accidents in Case of Addis Ababa, Ethiopia

Getahun Worku*and Dejen Tesfaw
Department of Statistics, Wollo University, Dessie, Ethiopia
Department of Statistics, Addis Ababa University, Addis Ababa, Ethiopia

ABSTRACT

Road traffic accident is the major phenomena of the world as well as our country, Ethiopia which is from the low-income countries. Statistical modeling for count response variables is a primary interest in, insurance, and other areas. The main objective of this study is used to identify the most appropriate count regression model to fit the number of human deaths per road traffic accident (RTAs). The data for this study get from Addis Ababa Traffic Control, and Investigation Department (AATCID), daily basis recorded from July 30, 2013 to July 29, 2014. The difficultyassumption of Poisson regression model shifts to look for extended models like the negative binomial model, zero inflatedPoisson, and zero-Inflated Negative binomial regression models. Specifically, traffic accidents generate count response variables with an invalid assumption of Poisson distribution such thatthe variance and mean of human death per road traffic accident are (0.58) and (0.36), and the over-dispersion parameter under the negative binomial was detected to indicate the existence of over-dispersion implies ZIP (or ZINB) model is favored over the Poisson (or NB) model, respectively, by using Vuong test. By using the goodness of fit model criteria like LRT, AIC, and BIC zero-inflated Poisson (ZIP) is the most fitted model for road traffic accident dataset.Therefore, quarter of year, slope of road, age of driver, vehicle type and ownership, time of accident, and type of accident are found statistically significant factors at α = 0.05fornumber of human death per road accidents.

Keywords: Negative binomial regression, Traffic accidents, Zero inflated poisson regression, Zero inflated negative regression.

INTRODUCTION

Road traffic accident is among the top major problem currently in the world. About 1.35 million people lost each year due to road traffic accidents. Approximately about 90% of the fatalities death on the road's accident occurred in low- and middle-income countries, even though around 54% of the vehicles are exist in these regions(WHO, 2019, May). According to different scholars, over half of pedestrians killed and seriously injured in Great Britain in 2015 were involved in crashes at junctions. The pedestrian actions and behavior factors which contributed to pedestrian casualty crashes were found to be between 1.6 and 2.8 times the frequencies of driver actions and behavioral factors(Downey et al., 2019). The burden of road traffic accident in Africa is the most Agony issue for developing countries like Ethiopia (Fernando et al., 2017). Most of the vehicles that use for various activities are serving for long time. According to WHO reports, on average more than one person die per day due to road traffic accident. Addis Ababa, the capital city of Ethiopia, has a large-scale shortage of the transportation problem as well as crowded traffic.

Statistical modeling for count response variables is a primary interest in several fields in the real world. Those interested fields stated by several researchers like public health (Windmeijer & Cameron, 1996), insurance (Jones et al., 1991; Yau & Lee, 2001; Yip & Yau, 2005; Abegaz et al., 2014), epidemiology (Preisser et al., 2012), psychology (Famoye & Singh, 2006), and many other research areas since most of the variables response countable such as the number of infected person per day, the number of accidents occurred per day, the number of patients admitted per day, and the number of people who took guidance per hour, respectively are the practical example of applications of count regression modeling in real-world (Lambert, 1992; Kibria & Research, 2006).According to the studies conducted by Lambert (1992) and Wagh and Kamalja (2018) Poisson regression model is the most regularly used for handling such count data. But there commendations of the scholars are restricted by the validation of Poisson regression.

*Corresponding author: getahun0514@gmail.com

The Poisson distribution assumed for validation that both the variance and mean are equal. Nevertheless, the validation of assumption may violate in many real applications subsequently the data is usually over dispersed. Various researchers have recommended the appropriateness of the Poisson model for predicting accident rates at intersections or roads (Dean & Lawless, 1989; Jones et al., 1991; Miaou & Lum, 1993; Berhanu, 2004; Malyshkina & Mannering, 2010). These studies outstrip the fact that incidents are necessarily discrete, often interrupted, and more likely to be accidental events. One important assumption in the Poisson model that the average value for outcome variable need to be equal to its variance (i.e., if the random response variable is denoted, Y, then,$\mathrm{E(y_i\,) = var(y_i)}$). If this assumption is no longer valid, then standard errors, usually calculated with the succor of the ML method are biased and the test statistic got from the model is wrong (Ismail & Zamani, 2013; Trivedi & Cameron, 2013; Wagh & Kamalja, 2018). According to (Trivedi & Cameron, 2013) the redundant zeros condition in the data set leads to an incorrect estimate which may the evidence of over-dispersion. To treat excess zeros, we should ZIP model (Jansakul & Hinde, 2002; Lambert, 1992; Lawal & Quantity, 2012) regression models that the probability of zero-defect state and the meanof an imperfect state (nonzero defect state) be subject to on the covariates (Chin & Quddus, 2003; Hasan & Sneddon, 2009; Trivedi & Cameron, 2013; Desjardins, 2016;). The main objective of this study is used to identify the most appropriate count regression model to fit the number of human deaths per road traffic accident (RTA). Specifically, it is used to check whether the road traffic accident data contains overdispersion and/or zero-inflation cases, and to select the appropriate count regression model to predict road traffic accidents among the candidate's models that are Poisson, Negative Binomial, Zero-Inflated Poisson and Zero-Inflated Negative Binomial regression models.

STATISTICAL METHODOLOGY

Data source:

The data used for this study was accessed from Addis Ababa Traffic Control, and Investigation Department (AATCID), Addis Ababa on daily basis recorded of by traffic on road from July 30, 2013 to July 29, 2014.

Analysis methodology:

Count regression models had been extensively utilized in statistics to model response variables that are assumed to be countable. The Poisson distribution has dramatically restricted assumption should be checked out the expected value of the response variable is equal to its variances. In practice, this restriction does not confirm, that variance is greater than the mean, which described as overdispersion. It implies that the Poisson regression isn't adequate. Due to the failure of the Poisson distribution, the analysis looks over other alternatives, Negative Binomial regression (NBR) which account for overdispersion problem and variables with excess zero and/or zero-inflated count, for the reason that extra zeros are the cause of smaller expected value than the variance. In other, the existence of excess zero may the cause of overdispersion (Poston Jr & McKibben, 2003; Trivedi & Cameron, 2013; Desjardins, 2016) recommended to Zero-Inflated candidate models such as zero-inflated Poisson (ZIP) or zero-inflated Negative binomial (ZINB). Therefore, the candidate models for this study were Poisson (PR), Negative binomial (NBR), zero-inflated Poisson (ZIPR), and ZINB Regression Models. The LRT and Wald used to test the parameter of the overdispersion and coefficients of the estimated regression models (Trivedi & Cameron, 2013; Desjardins, 2016; Wagh & Kamalja, 2018).

Poisson Regression (PR) Model:

For an independent sample of n pairs of observations $(y_i, x_i),\ i \in 1,\ 2,\ \ldots,\ n$, where y_i denotes "the number of events that occurred" and x_i is the value of explanatory variables for the i^{th} subject(Lambert, 1992). Assume $y_i \sim Poisson(\mu_i)$, $i = 1, 2, \ldots, n$ then the probability density function of Poisson accidental variables, Y_i, is given by

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

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

$$\mathrm{E(y_i) = Var(y_i)} = \mu_i \ (1.1)$$

Overdispersion

When the variance of the count response variable exceeds the mean, $Var[y_i] > E[y_i], i = 1, 2, 3, \ldots, n$ a feature of overdispersion will occur(Dean & Lawless, 1989; Perumean-Chaney *et al.*, 2013). As a result, the overdispersion trouble occurs, the Poisson maximum likelihood estimator received may be wrong (Trivedi & Cameron, 2013).

$H_o: \delta = 0$ (there is no overdispersion implies equi-dispersed), versus

$H_o: \delta > 0$ (overdispersion exist in the dataset)

Negative Binomial Regression (NBR) Model

Random variable $Y_i, i = 1, 2, 3, \ldots, n$ is negatively binomial distribution count with parameter μ and δ the probability density function is expressed as follows (Lambert, 1992):

$$f(y_i; \mathrm{\mu}_i, \delta) = \frac{\Gamma\left(y_i + {}^1/_\delta\right)}{\Gamma\left({}^1/_\delta\right) y_i!} (1 + \delta\mu_i)^{-{}^1/_\delta} \left(1 + \frac{1}{\delta\mu_i}\right)^{-y_i},\ y_i = 0, 1, 2, \ldots \quad (2)$$

with mean and variance, respectively, given by:

$E(Y_i) = \mu_i = exp(x_i^T\beta)$, and $Var(Y_i) = \mu_i(1 + \delta\mu_i)$ … … … … … (2.1)

The term δ (read as delta) is called the dispersion parameter. If the dispersion parameter closes to null ($\delta \to 0$), then the NBR model reduces to the classical Poisson regression model. The likelihood function of the NBR model based on a sample of n independent observation is given by:

$$\ell(\mu_i, \delta; y_i) = \prod_{i=1}^{n}\left(\frac{\Gamma\left(y_i+\frac{1}{\delta}\right)}{\Gamma\left(\frac{1}{\delta}\right)y_i!}\left(1 + \delta\mu_i\right)^{-1/\delta}\left(1 + \frac{1}{\delta\mu_i}\right)^{-y_i}\right) \text{… … … … …} \quad (2.2)$$

and the log-likelihood function is

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

$$= \sum_{i=1}^{n}\left\{-\log(y_i!) + \log\left(\frac{\Gamma\left(y_i+\frac{1}{\delta}\right)}{\Gamma\left(\frac{1}{\delta}\right)}\right) - \frac{1}{\delta}\log(1 + \delta\mu_i) - y_i\log\left(1 + \frac{1}{\delta\mu_i}\right)\right\} \text{… … … … …} \quad (2.3)$$

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

$$\frac{\Gamma\left(y_i+{}^{1}/_{\delta}\right)}{y_i!\Gamma\left({}^{1}/_{\delta}\right)} = \prod_{k=1}^{y_i}\left(y_i + \frac{1}{\delta} - k\right) = \delta^{-y_i}\prod_{k=1}^{y_i}(\delta y_i - \delta k + 1) \text{… … … … …} \quad (2.4)$$

Then

$$= \sum_{i=1}^{n}\left\{-\log(y_i!) + \sum_{k=1}^{y_i}\left(\log(\delta y_i - \delta k + 1)\right) - \left(y_i + {}^{1}/_{\delta}\right)\log(1 + \delta\mu_i) + y_i\log(\mu_i)\right\} \text{… … … … …} \quad (2.5)$$

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

$$\mathcal{L} = \sum_{i=1}^{n}\left\{-\log(y_i!) + \sum_{k=1}^{y_i}\left(\log(\delta y_i - \delta k + 1)\right) - \left(y_i + {}^{1}/_{\delta}\right)\log(1 + \delta\exp(x_i^T\beta)) + y_i x_i^T\beta\right\} \text{… … … … …} \quad (2.6)$$

According to (Borgan, 1984; Lloyd-Smith, 2007)the likelihood equations to estimate $\beta$and δ are obtained by taking the partial derivatives of the log-likelihood function and set them equivalent to zero. Thus, we obtain the first derivatives of the log-likelihood function, $\mathcal{L}$, respecting the underlying parameters are obtained as follows:

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

$$\frac{\partial\mathcal{L}}{\partial\delta} = \sum_{i=1}^{n}\left\{-\delta^{-2}\sum_{k=0}^{n-1}\frac{1}{(k+\frac{1}{\delta})} + \delta^{-2}\log(1 + \delta\mu_i) + \frac{y_i-\mu_i}{\delta(1+\delta\mu_i)}\right\} \text{… … … … …} \quad (2.7)$$

Zero-Inflated Poisson Regression (ZIPR) Model:

The zero-inflated Poisson distribution assumed for two distinct underlying states. The first state $\omega_i$produces for only zeros, while $1-\omega_i$to a standard Poisson count with mean μ_i and hereafter a chance of extra zeros. In general, the first state is called structural zeros and the other state from the Poisson model is called sampling zeros (Lambert, 1992; Jansakul & Hinde, 2002; Lawal & Quantity, 2012). This two-state process gives the following probability mass function (pmf):

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

where $0 \le \omega_i \le 1$, and μ_i. The parameter μ_i and ω_i depends on the covariates x_i and z_i, respectively.

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

$E(y_i) = \mu_i(1-\omega_i)$, 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 assumed below:

$$\mathcal{L} = \sum_{i=1}^{n}\{I_{(y_i=0)}\log[\omega_i + (1-\omega_i)exp(-\mu_i)] + I_{(y_i>0)}[\log(1-\omega_i) - \mu_i + y_i\log(\mu_i) - \log(y_i!)]\} \text{… … … … …} (3.1)$$

The first derivative the log-likelihood function regarding the underlying parameter is:

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

$$= \sum_{i=1}^{n}I_{(y_i=0)}\left[\frac{exp(z_i^T\gamma)}{exp(z_i^T\gamma)+exp(-exp(x_i^T\beta))}\right]z_{ir} - \sum_{i=1}^{n}\left[\frac{exp(z_i^T\gamma)}{1+exp\ (z_i^T\gamma}\right]z_{ir}, \text{… … … … …} \quad (3.2)$$

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

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

$$= \sum_{i=1}^{n}I_{(y_i=0)}\left[\frac{-exp(x_i^T\beta)exp(-exp(x_i^T\beta))}{exp(z_i^T\gamma)+exp(-exp(x_i^T\beta))}\right]x_{ij} + \sum_{i=1}^{n}I_{(y_i>0)}[y_i - exp(x_i^T\beta)]\,x_{ij}, \quad (3.3)$$

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

Zero-Inflated Negative Binomial Regression (ZINBR) Model:

The ZINBR model is proposed to demonstrate variable with excess zeros and overdispersion. The researches (Lambert, 1992; Jansakul & Hinde, 2002; Li et al., 2019) demonstrated that the ZINBR model gives an appropriate fit for the over dispersed response variable as compared and the ZIP model.

$$p(Y_i = y_i) = \begin{cases} \omega_i + (1-\omega_i)(1 + \delta\mu_i)^{-\frac{1}{\delta}}, & y_i = 0 \\ (1-\omega_i)\frac{\Gamma\left(y_i + {}^1/_\delta\right)}{y_i!\Gamma\left({}^1/_\delta\right)}(1 + \delta\mu_i)^{-1/\delta}\left(1 + \frac{1}{\delta\mu_i}\right)^{-y_i}, & y_i > 0 \end{cases} \quad \ldots (4)$$

where $\delta > 0$, is a dispersion parameter.

The variance and the mean of the ZINBR model are:

$$E(y_i) = \mu_i(1-\omega_i), Var(y_i) = \mu_i(1-\omega_i)(1 + \omega_i\mu_i + \delta\mu_i) \ldots\ldots\ldots\ldots\ldots (4.1)$$

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

$$\log(\mu_i) = x_i^T\beta, \quad \log\left(\frac{\omega_i}{1-\omega_i}\right) = z_i^T\gamma \ldots\ldots\ldots \quad (4.2)$$

The log-likelihood function (Borgan, 1984; Lloyd-Smith, 2007),$\mathcal{L} = \ell(\delta, \mu_i, \omega_i; y_i)$ for ZINBR model is

$$\mathcal{L} = \sum_{i=1}^{n}\left\{I_{(y_i=0)}\log\left(\omega_i + (1-\omega_i)(1 + \delta\mu_i)^{-1/\delta}\right) + I_{(y_i>0)}\log\left((1-\omega_i)\frac{\Gamma\left(y_i + {}^1/_\delta\right)}{y_i!\Gamma\left({}^1/_\delta\right)}(1 + \delta\mu_i)^{-1/\delta}\left(1 + \frac{1}{\delta\mu_i}\right)^{-y_i}\right)\right\} \ldots\ldots\ldots\ldots\ldots (4.3)$$

since

$$\frac{\Gamma\left(y_i + {}^1/_\delta\right)}{y_i!\Gamma\left({}^1/_\delta\right)} = \prod_{k=1}^{y_i}\left(y_i + \frac{1}{\delta} - k\right) = \delta^{-y_i}\prod_{k=1}^{y_i}(\delta y_i - \delta k + 1) \ldots\ldots\ldots\ldots\ldots (4.4)$$

Furthermore, the log-likelihood function can be written as

$$\mathcal{L} = \sum_{i=1}^{n} I_{(y_i=0)}\left[\log\left(\omega_i + (1-\omega_i)(1 + \delta\mu_i)^{-1/\delta}\right)\right] + I_{(y_i>0)}\sum_{i=1}^{n}\left[\log(1-\omega_i) - \log(y_i!) + \sum_{k=1}^{y_i}\log(\delta y_i - \delta k + 1) - \left(y + \frac{1}{\delta}\right)\log(1 + \delta\mu_i) + y_i\log(\mu_i)\right] \quad (4.5)$$

Tests for the Comparison of the Models

a) Tests for Comparison of Nested Models:

Likelihood-Ratio Test (LRT):

The LRT is used to evaluate the adequacy of two or more than two nested modelings. It compares the maximized log -likeliness value of the full model and reduced model (Anisimova & Gascuel, 2006). For illustration, the null hypothesis can be stated as the overdispersion argument is equal to zero (i.e. the Poisson model can be fit well the information) versus the option hypothesis can be stated as the overdispersion parameter is different from zero (i.e. the data would be better fitted by the NB regression). The likelihood-ratio test is given by:

$$LRT = 2(\mathcal{L} - \mathcal{L}_o) \ldots\ldots\ldots (5)$$

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

b) Test for Comparison of Non-nested Models:

Vuong's Test:

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

$$m_i = log\left[\frac{P_1(y_i|x_i)}{P_2(y_i|x_i)}\right] \ldots\ldots\ldots (6)$$

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

$$V = \frac{\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} m_i\right)}{\sqrt{(\frac{1}{n}\sum_{i=1}^{n}(m_i - \overline{m})^2}} \ldots\ldots\ldots (6.1)$$

For a large sample size and under the null hypothesis the test statistic V has the standard normal distribution at α level of significance. The ZIP (or ZINB) model is favored over the Poisson (or NB) model, respectively. On the basis of Vuong (1989), if the calculated value of the Vuong test is positive and high ($V > 1.96$) vice versa, and the two models are equivalent when |V|<1.96.

Model Fitting Test (Goodness of Fit of the Model)

In this study, to select the appropriate fitted model, which fits the data well was done using the likelihood-ratio test (LRT), Akaike information criteria (AIC) and Bayesian information criteria (BIC) according to the recommendation of scholars Archer and Lemeshow (2006). The formula is given as:

$$AIC = -2\mathcal{L} + 2k, \ldots\ldots\ldots (7)$$

where $\mathcal{L}$ is the log-likelihood of a model that will compare with the other models and k is the numeral of parameters including the intercept. On the other side, BIC is given by,

$$BIC = -2\mathcal{L} + klog(n) \ldots\ldots\ldots (8)$$

where $\mathcal{L}$ is the log-likelihood of a model that will compare with the other models, n is dimensions of observation, and k is the numeral of parameters including the intercept.

The model which has the minimum value of AIC and BIC is the most appropriate fitted model to the dataset.

RESULT OF STATISTICAL ANALYSIS

As shown in Table 2, the variance of human death per road traffic accident (0.58) was greater than its mean (0.36). In this descriptive value of the response variable namely number of human deaths per accident dataset indicated the plausibility of existing overdispersion and hence the Poisson model was not appropriate to fit the road traffic accident data. Additional inspection of the data also indicated the existence of an excess number of zeros (70%) and the shape of the curve is right-skewed Fig. 1. The actual test of overdispersion is in negative binomial stage of model constructing teat is delta is significantly different from zero (Appendix 1).

Even though, Appendix 1 shows all most all the variables that we used for these models for Poisson, Negative binomial, zero inflated Poisson, and negative binomial models, are statistically significant factors for the number of human death per accident at three levels of significance (1%, 5% and 10%). But then again, the standard error of the parameter of the variables is slightly different which leads to select the appropriate counting data models for such a traffic accident dataset to predict the finding model for it. In the case of like statistical models tried to incorporate the minimum value of standard error of the estimated parameter.

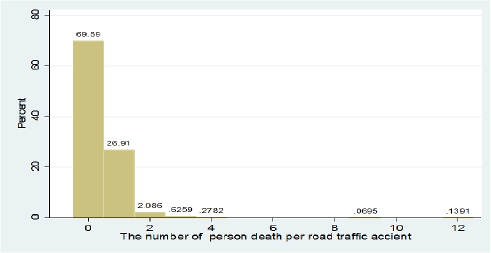

Fig.1: Histogram for the number of human death due to RTAs

The zero-inflated model was proper to fit the number of human death per accident dataset because of the presence of excess zero in the data and the violation of the Poisson model assumption that is the mean is smaller than the variance of the number of human deaths per RTAs.

DISCUSSION

Even though the Poisson regression model considered as baseline for analysis count data (Lambert 1992; Wagh & Kamalja 2018), the assumption of Poisson regression model for number of human death per road traffic accident dataset is invalid for this study, which is consistent with the studies conducted by (Dean & Lawless, 1989; Jones et al., 1991; Miaou & Lum, 1993).The existence of overdispersion and excess zeros in the dataset modify to other suitable candidate models(Dean & Lawless, 1989; Perumean-Chaney et al., 2013), negative binomial (NB), and zero inflation extended models (zero-inflated Poison and Zero-inflated Negative Binomial models) integrate to the accident dataset.

Table 1: Summary statistics for the number of human death due to road traffic accidents

Mean	Variance	Skewness	Kurtosis
0.36	0.58	6.88	88.36

After all, different model selection criteria were considered like the likelihood-ratio test (LRT), Akaike information's criterion (AIC) and Bayesian information's criterion (BIC) (Archer & Lemeshow, 2006)to detect the furthermost fitted model. For non-nested models such as ZIP versus Poisson and ZINB versus NB regression models were identified using the Vuong test statistic (Vuong, 1989).

Table 2: Vuong test statistics for Poison Vs ZIP, and Negative Binomial Vs ZINB

Tested nested models	Test statistics	P. Value
ZIP	3.22	(0.000)
ZINB	9.08	(0.000)

Table 2 showed the criteria 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 ZIPR versus PR model as the researcher (Vuong, 1989). This value revealed that the ZIPR model was preferred to the PR model to estimate the number of human death due to road traffic accidents. In the second case, the comparison of ZINB versus NB models, the calculated value of the Vuong test is 9.08, revealed that the ZINB model was preferred to NBR model as similar techniques to Vuong (1989) and Perumean-Chaney et al. (2013).

Finally, to compare the ZIPR and ZINBR models, AIC and BIC were used as shown in Table 3. Therefore, the ZIPR model is better-fitted human death per road accident data than did the ZINBR model. AIC and BIC values of ZIPR was found to be small as compared to other count models.

As shown in Fig. 4, the ZIPR model was a better choice than the other count models, since the predicted probability for the ZIPR model was closed to the observed probability. From Fig. 2 and the value of AIC, BIC, and Vuong criteria in Table 3, it can be observed that there was a difference between PR, NBR, ZIPR, and ZINBR models for the dataset(Lambert, 1992). Therefore, it is possible to conclude that the ZIP model was more appropriate than the ZINB model to fit the Addis Ababa road traffic accident dataset.

In conclusion, this study concludes that the standard Poisson regression is not a proper model to fit the road traffic accident dataset. By using the goodness of fit model criteria like Likelihood-Ratio Test (LRT), Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC), and Vuong test, Zero-Inflated Poisson (ZIP) is the most fitted model for road traffic accident dataset. Therefore, by using ZIP model quarter of year(specifically, fourth quarter (June, July and August) highly increased the numbers of human death, age of driver (specifically deriver whose age greater than 50 years are contribute to decrease the numbers of human death), vehicle type (specifically bus and cargo) highly contribute to increase the number of human death and ownership (Governmental vehicle contribute highly), time of accident (specifically night time the large number of human death) as compared to their reference categories (See Table 5) are found statistically significant factors at α = 0.05 for number of human death per road accidents

Table 3: Model selection criteria for PR, NBR, ZIPR and ZINBR models for the number of human deaths per RTA dataset

Selection	Models			
Criteria	Poisson	NB	ZIP	ZINB
LL	-884.12	-882.01	**-809.03**	-809.03
AIC	1794.23	1792.03	**1664.05**	1666.05
BIC	1862.75	1865.82	**1785.29**	1792.56

LL – Log-likelihood

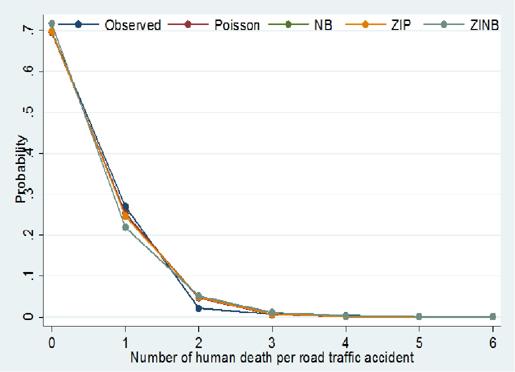

Fig. 2: The Observed and predicted probability of PR, NBR, ZIPR, and ZINBR models

REFERENCES

Abegaz, T., Berhane, Y., Worku, A., Assrat, A., & Assefa, A. (2014). Effects of excessive speeding and falling asleep while driving on crash injury severity in Ethiopia: A generalized ordered logit model analysis. *Accident Analysis Prevention, 71*, 15-21.

Anisimova, M., & Gascuel, O. (2006). Approximate Likelihood-Ratio Test for Branches: A Fast, Accurate, and Powerful Alternative. *Systematic Biology, 55*(4), 539-552. doi:10.1080/10635150 600755453

Archer, K. J., & Lemeshow, S. (2006). Goodness-of-fit Test for a Logistic Regression Model Fitted using Survey Sample Data. *The Stata Journal, 6(1)*, 97-105. doi:10.1177/1536867X0600600106

Berhanu, G. (2004). Models relating traffic safety with road environment and traffic flows on arterial roads in Addis Ababa. *Accident Analysis and Prevention, 36(5)*, 697-704.

Borgan, Ø. (1984). Maximum Likelihood Estimation in Parametric Counting Process Models, with Applications to Censored Failure Time Data. *Scandinavian Journal of Statistics, 11(1)*, 1-16.

Chin, H. C., & Quddus, M. A. (2003). Modeling Count Data with Excess Zeroes: An Empirical Application to Traffic Accidents. *Sociological Methods & Research, 32(1)*, 90-116. doi:10. 1177/0049124103253459

Dean, C., & Lawless, J. F. (1989). Tests for Detecting Overdispersion in Poisson Regression Models. *Journal of the American Statistical Association, 84(406)*, 467-472. doi:10.1080/ 01621459.1989.10478792

Table 5: Observed and predicted probability from PR, NBR, ZIPR and ZINBR model for the number of human deaths per RTAs

		Observed	Predicted probability			
Count	Frequency	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

Desjardins, C. D. (2016). Modeling Zero-Inflated and Overdispersed Count Data: An Empirical Study of School Suspensions. *The Journal of Experimental Education, 84(3)*, 449-472. doi:10.1080/ 00220973.2015.1054334

Downey, L. T., Saleh, W., Muley, D., & Kharbeche, M. (2019). Pedestrian crashes at priority-controlled junctions, roundabouts, and signalized junctions: The UK case study. *Traffic Inj Prev, 20(3)*, 308-313. doi:10.1080/15389588.2019.1574972

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

Fernando, D. M., Tennakoon, S. U., Samaranayake, A. N., & Wickramasinghe, M. (2017). Characteristics of road traffic accident casualties admitted to a tertiary care hospital in Sri Lanka. *Forensic Sci Med Pathol, 13(1)*, 44-51. doi:10.1007/s12024-016-9828-3

Hasan, T. M., & Sneddon, G. (2009). Zero-Inflated Poisson Regression for Longitudinal Data. Communications in Statistics. *Simulation and Computation, 38(3),* 638-653. doi:10.1080/ 03610910802601332

Ismail, N., & Zamani, H. (2013). Estimation of claim count data using negative binomial, generalized Poisson, zero-inflated negative binomial and zero-inflated generalized Poisson regression models. *Casualty Actuarial Society E-Forum, 41(20)*, 1-28.

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

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

Kibria, B. G., & Research, O. (2006). Applications of some discrete regression models for count data. *Pakistan Journal of Statistics, 2(1)*, 1-16.

Lambert, D. (1992). Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. *Technometrics, 34(1)*, 1-14. doi:10. 2307/1269547

Lawal, B. H., & Quantity. (2012). Zero-inflated count regression models with applications to some examples. *Quality, 46(1)*, 19-38.

Li, C.-S., Lee, S.-M., & Yeh, M.-S. (2019). A test for lack-of-fit of zero-inflated negative binomial models. *Journal of Statistical Computation and Simulation, 89(7),* 1301-1321. doi:10.1080/ 00949655.2019.1577856

Lloyd-Smith, J. O. (2007). Maximum likelihood estimation of the negative binomial dispersion parameter for highly overdispersed data, with applications to infectious diseases. *PloS one, 2(2),* e180-e180. doi:10.1371/journal.pone.0000180

Malyshkina, N. V., & Mannering, F. L. (2010). Empirical assessment of the impact of highway design exceptions on the frequency and severity of vehicle accidents. *Accid Anal Prev, 42(1)*, 131-139. doi:10.1016/j.aap.2009.07.013

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

Perumean-Chaney, S. E., Morgan, C., McDowall, D., & Aban, I. (2013). Zero-inflated and overdispersed: what's one to do? *Journal of Statistical Computation and Simulation, 83(9),* 1671-1683. doi:10.1080/00949655.2012.668550

Poston Jr, D. L., & McKibben, S. L. (2003). Using zero-inflated count regression models to estimate the fertility of US women. *Journal of Modern Applied Statistical Methods, 2(2),* 10.

Preisser, J. S., Stamm, J. W., Long, D. L., & Kincade, M. E. (2012). Review and recommendations for zero-inflated count regression modeling of dental caries indices in epidemiological studies. *Caries research, 46(4)*, 413-423.

Trivedi, P. K., & Cameron, C. A. (2013). *Regression Analysis of Count Data*. Cambridge, UK: Cambridge University Press.

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

Wagh, Y. S., & Kamalja, K. K. (2018). Zero-inflated models and estimation in zero-inflated Poisson distribution. *Communications in Statistics - Simulation and Computation, 47(8),* 2248-2265. doi:10.1080/03610918.2017.1341526

WHO. (2019, May). Road traffic injuries. *World Health Organization.* Retrieved from http://www.who.int/mediacentre/factsheets/fs358/en/

Windmeijer, F. A. G., & Cameron, A. C. (1996). R-squared measures for count data regression models with applications to health-care utilization. *Journal of Business and Economic Statistics, 14(2)*, 209-220.

Yau, K. K., & Lee, A. H. (2001). Zero-inflated Poisson regression with random effects to evaluate an occupational injury prevention programme. *Statistics in medicine, 20(19),* 2907-2920.

Yip, K. C., & Yau, K. K. (2005). On modeling claim frequency data in general insurance with extra zeros. *Insurance: Mathematics and Economics, 36(2)*, 153-163.

Appendix 1: Parameter Estimation of Poisson, NB, ZIP, and ZINB Regression Models

	Poisson		NB		ZIP		ZINB	
Parameters	**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
Vehicle type								
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
Accident type								
Vehicle to 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< 0.001, ** p-value< 0.01, * p-value< 0.05**