Novelty Journals

ISSN 2394-9716

International Journal of Novel Research in Interdisciplinary Studies
Vol. 10, Issue 1, pp: (6-18), Month: January – February 2023, Available at: www.noveltyjournals.com

Incorporating Zero-Inflated Poisson (ZIP) Regression Model in Crash Frequency Analysis

Azad Abdulhafedh

DOI: https://doi.org/10.5281/zenodo.7632596

Published Date: 11-February-2023

Abstract: **This paper addresses the Zero-inflated Poisson (ZIP) regression model as an effective way to handle the excess zeros that usually exist in vehicle crash data and to allow for possible overdispersion in the data. The ZIP model is based on a zero-inflated probability distribution, that allows for frequent zero-valued observations. When the number of zeros is large that the data do not fit standard distributions (e.g., normal, Poisson, binomial, negative-binomial, and beta), the data is referred to as zero inflated. A dual state crash system is assumed in the ZIP model, in which one state is the zero crash state that can be regarded as virtually safe during the observation period, while the other state is the non-zero crash state. This paper starts by applying a multiple linear regression model, a Poisson regression model, a Negative Binomial regression model and then introduces the ZIP model to analyze the 2013-2015 crash data for the Interstate I-94 in the State of Minnesota in the US. Results show that the ZIP model overperformed the other models by fitting the crash data much better and was able to capture almost all the independent variables in the model and make them statistically significant in the analysis after being insignificant by the other models.**

Keywords: **Zero-Inflated Poisson Regression, ZIP model, Crash Frequency, Multiple Linear Regression, Poisson Regression, Negative Binomial Regression.**

1. INTRODUCTION

Vehicle crashes are a global concern, and socio-economic aspect, leading to tremendous life and property loss each year around the world. Despite the efforts to apply preventive measures, the annual number of traffic accidents has not yet significantly decreased. For instance, in the US in 2021, there were an estimated 42,915 people died in motor vehicle traffic crashes, a 10.5% increase from the 38,824 fatalities in 2020. On average, one person was killed every 14 minutes and an estimated 4 people were injured every minute in traffic crashes [1] [2]. Therefore, modeling crash data is emphasized in highway safety research. The average number of crashes per section of road is called the crash frequency, which has been widely used as an indicator of the crash occurrence at highways. A variety of independent variables can affect crash frequency that are related to the driver behaviors, road characteristics, vehicle, and environment. The influence of such variables on crash frequency could significantly vary on case by case basis, but in general, past research have shown that both driver's factors, and nonbehavioral factors related to the road geometry, vehicle, and environment can significantly affect crash frequencies [3] [4] [5] [6].

2. BACKGROUND LITERATURE

Crash frequency models were first based on the simple Multiple Linear Regression models assuming normally distributed errors. However, research showed that crash occurrence is more fitted with the Poisson distribution, and hence began to utilize the Poisson regression model that was developed by the Generalized Linear Models (GLM), instead of the conventional multiple linear regression technique. The Multivariate Poisson regression models have been used for several decades to explore the relationship between the risk factors and crash rates [7] [8] [9] [10]. However, it was found that the Poisson regression model has one important constraint that is the mean must be equal to the variance, and when this

assumption is violated, the standard errors estimated by the maximum likelihood method, will be biased and the test statistics derived from the model will be incorrect. In addition, since the crash data are usually overdispersed (i.e., the variance is greater than the mean), therefore, this will result in incorrect estimation of the likelihood of accident occurrence when using the Poisson regression model [11]. In overcoming the problem of over-dispersion, research began to employ the Negative Binomial (NB) distribution (or Poisson-Gamma) instead of the Poisson distribution, which relaxes the condition of mean equals to variance, and hence can consider the over-dispersion in the crash data [12]. However, the NB model was also found to have some limitations such as its inability to handle the case of under-dispersion of the data, where the mean of the crash data is higher than the variance, and this can exist when the sample size used is very small which can result in inadequate parameter estimates [13] [14]. Hence, to overcome the limitations of the NB models, the zero-inflated Poisson and zero-inflated negative binomial models have been introduced mainly to deal with the over-dispersion problem caused by the excessive zeroes (i.e., locations where no accidents can be observed) in traffic accident data. The zero-inflated procedure allows modeling the accident frequencies in two states, namely; the zero-accident state, and the non-zero accident state, and the probability of a section being in zero or non-zero states can be found by a binary logit model or a probit model. These zero-inflated models have shown great flexibility in both states and produced promising estimates [15] [16].

Data

The crash data is obtained from the Highway Safety Information System (HSIS) database, which is maintained by the Federal Highway Administration (FHWA) of the United States Department of Transportation (USDOT). The data includes the accident records, the road data, and the vehicle records. The accident records contain information about the fatality of crashes, the environment, and the circumstances of the crash occurrence. The vehicle records describe various characteristics of the vehicle(s) involved in the crash. The road data provide information on the road characteristics and geometry where the accidents occurred. The crash data consists of 3 years crash records on the interstate 94 (I-94) in the State of Minnesota for the years 2013, 2014, 2015. These crash data were the latest available data from the state of Minnesota in the HSIS database. The I-94 is a multilane highway, which runs 259 miles (417 km) east–west through the central portion of the State of Minnesota. The observed crash frequency of I-94 at all road sections from 2013 to 2015 ranges from 0 to 5, the average crash frequency is 0.597, the number of sections with zero crash frequency is 484, the number of sections with only one crash frequency is 179, the number of sections with two crashes is 74, the number of sections with three crashes is 27 as shown in Figure 1. The normal distribution curve of crash frequency is clearly skewed as can be seen in Figure 1. The dependent variable is the crash frequency, and different independent variables are included in the research related to the road characteristics, the environment, the traffic volume, the driver's factors, and the vehicle types as shown in Table 1 along with their summary statistics. The crash data (770 observations) was randomly splitted into two subsets using the R software; training data (70%), and testing data (30%). The training data consisted of 545 observations, and testing data consisted of 232 observations.

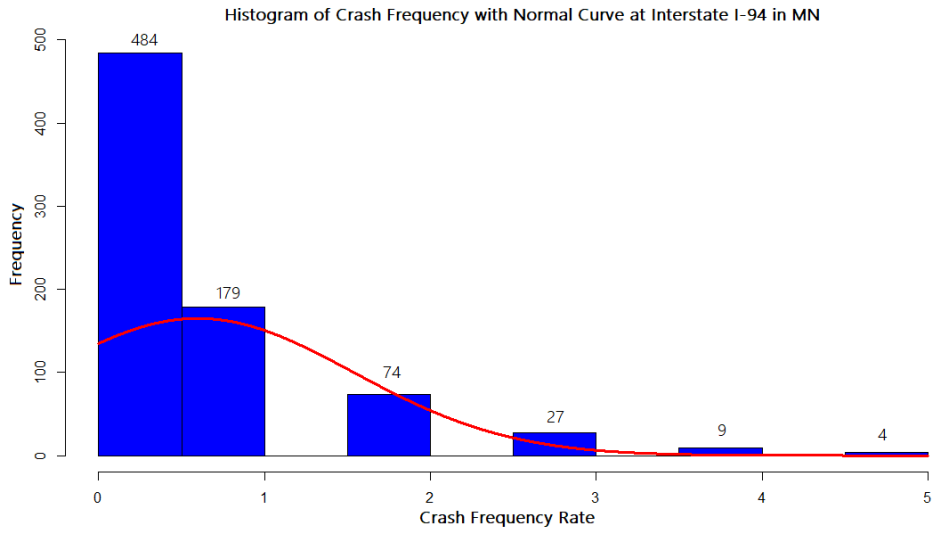

Figure 1: Histogram of Crash Frequency with Normal Curve at I-94 MN

Table 1: Variables included in the research with summary statistics

Variable	Description	Variable classes	Min.	Mean	Std. dev.	Max.
crash_frq	crash frequency rate	0 to 5	0	0.597	0.939	5
mi_post	the number of the mile marker at which the crash occurred	1 to 259	1	130	74.81	259
rd_char	The characteristics of the road section where the crash occurred	1-Straight 2-Upgrade 3-Downgrade 4-Horizontal curve	1	1.664	1.073	4
rd_surf	The condition of the road surface where the crash occurred	1-Dry 2-Wet 3-Snowy	1	2.382	0.816	3
aadt	The Annual Average Daily Traffic of the road section where the crash occurred	Numeric values in 1000s vehicles	5.7	13.04	5.515	27.22
weather	The weather conditions when the crash occurred	1-Clear 2-Rain 3-Snow 4-Fog	1	1.525	0.789	4
light	The type of light existed at the time of the crash	1-Daylight 2-Dark, but Lights On 3-Dark, but with No Lights	1	1.653	0.694	3
drv_age	The age of the driver of the vehicle involved in the crash	1-< 21 years 2-between 21 to 65 3-> 65 years	1	1.728	0.564	3
drv_sex	Sex of the driver of the vehicle involved in the crash	1-Male 2-Female	1	1.405	0.491	2
veh_type	Type of vehicle involved in the crash	1-Passenger Car 2-Van 3-Bus 4-Truck	1	1.179	0.585	4

Exploratory Data Analysis

To begin with the analysis, an exploratory data analysis (EDA) was conducted using R. The null values and outliers were checked, and found to be very few, so they were excluded. The matrix scatterplot of all variables is produced including correlation values between variables as shown in Figure 2. Also, the correlation matrix of all variables is created, which helps visualizing how the different variables are correlated as shown in Figure. 3. The matrix scatterplot shows the density plots and distribution of the subclasses of each variable. For example, the distribution of the crash frequency shows that road sections with zero crashes are higher than the other rates of 1, 2, 3 etc. The distribution of the road characteristics shows that the straight sections are much higher than upgrades and downgrades. Sections with horizontal curves are more frequent than upgrades and downgrades. The distribution of the road surface shows that the snowy sections are more frequent than the dry and wet sections when crashes occurred. The distribution of the weather conditions shows that the clear weather is higher than the rain, snow, and fog when crashes occurred. The driver's age distribution shows that the middle age group (21 – 65 years) contributed to the crash occurrence more than the young and elderly groups. The distribution of the vehicle type shows that the passenger car type was more involved in the crashes than vans, buses, and trucks. The correlation values between variables can be read directly from the scatterplot in Figure 2 and from the correlation matrix in Figure 3. All variables have very small correlation with each other as shown in Figure 2 and 3. The only moderate correlation value exists between road characteristic and crash frequency (56.5% or around 60%), which is still acceptable. Hence, all selected independent variable were kept in the analysis.

Novelty Journals

ISSN 2394-9716

International Journal of Novel Research in Interdisciplinary Studies
Vol. 10, Issue 1, pp: (6-18), Month: January – February 2023, Available at: www.noveltyjournals.com

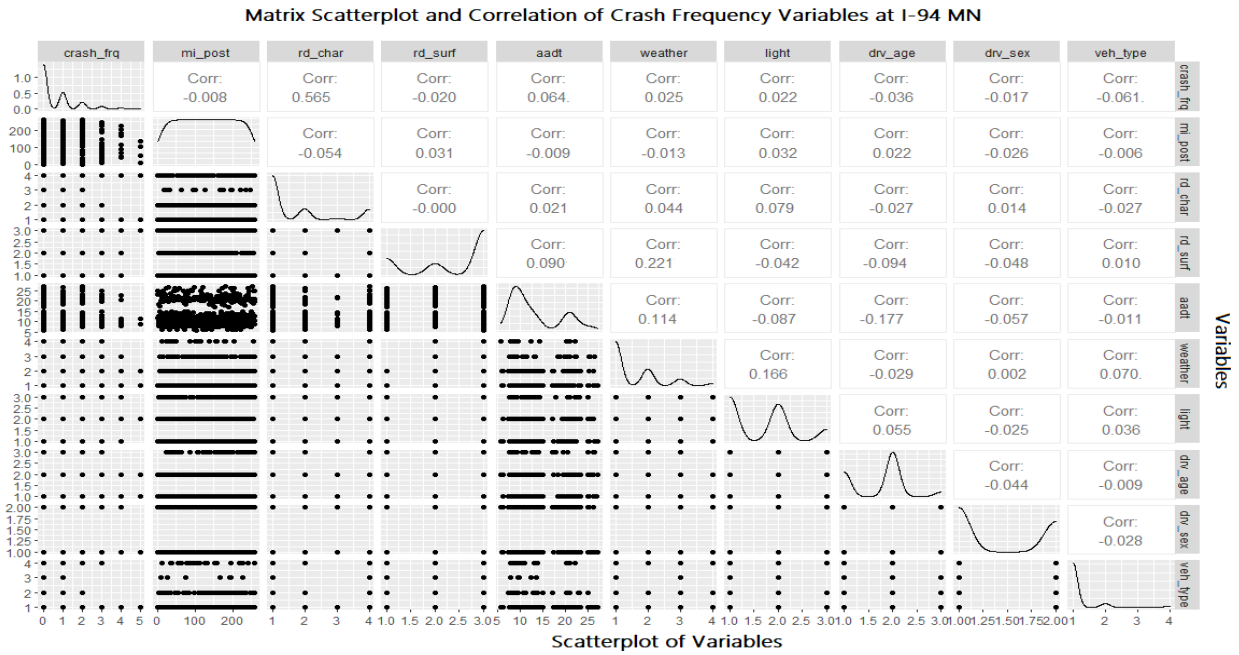

Figure 2: Matrix Scatterplot of all variables at I-94 MN

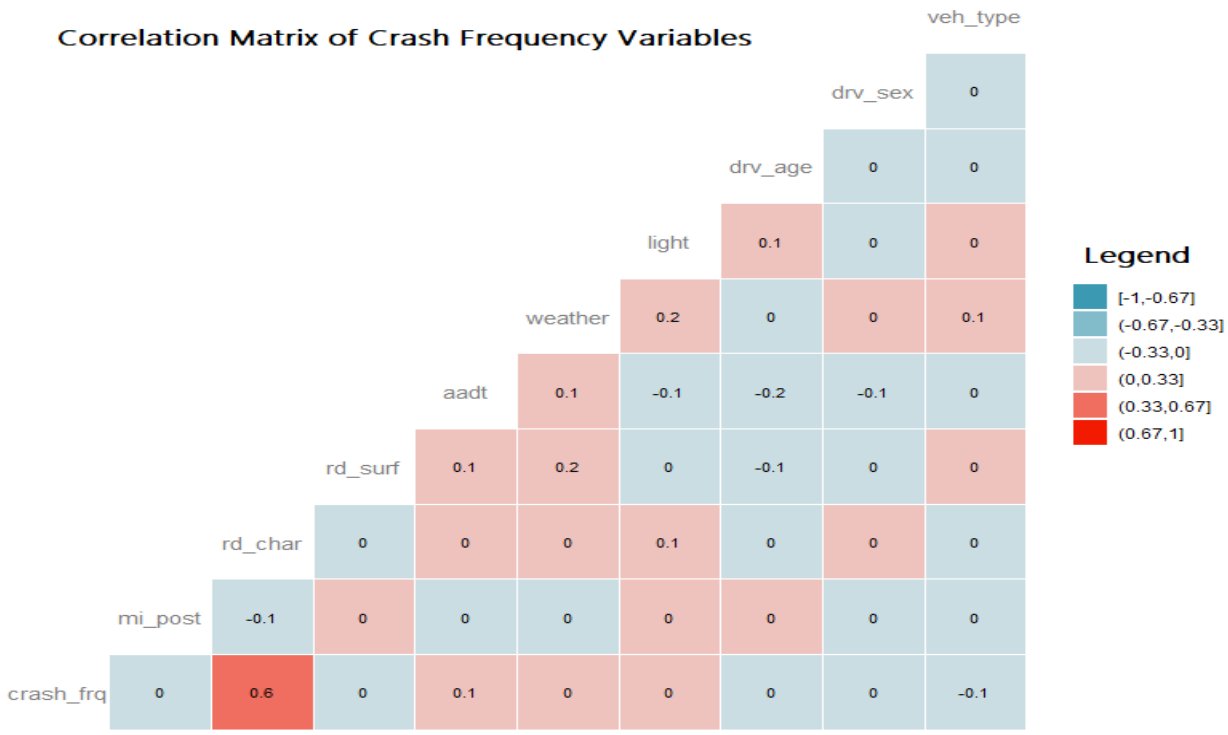

Figure 3: Correlation Matrix of all variables at I-94 MN

Histograms of all variables are produced in R as shown in Figure 4. Histograms can help understanding how the values of different variables subclasses are distributed. For example, the road characteristics subclasses can be easily found from Figure 4. For instance, the straight sections were crashes occurred are 548, the upgrades are 129, the downgrades are 26, and horizontal curves are 122. The dry road surface conditions where crashes occurred are 168, and the wet surface conditions are 122. The clear weather conditions when crashes occurred are 479, and the rain weather conditions are 186. The passenger car types involved in crashes are 681, and van types are 39.

Novelty Journals

ISSN 2394-9716

International Journal of Novel Research in Interdisciplinary Studies
Vol. 10, Issue 1, pp: (6-18), Month: January – February 2023, Available at: www.noveltyjournals.com

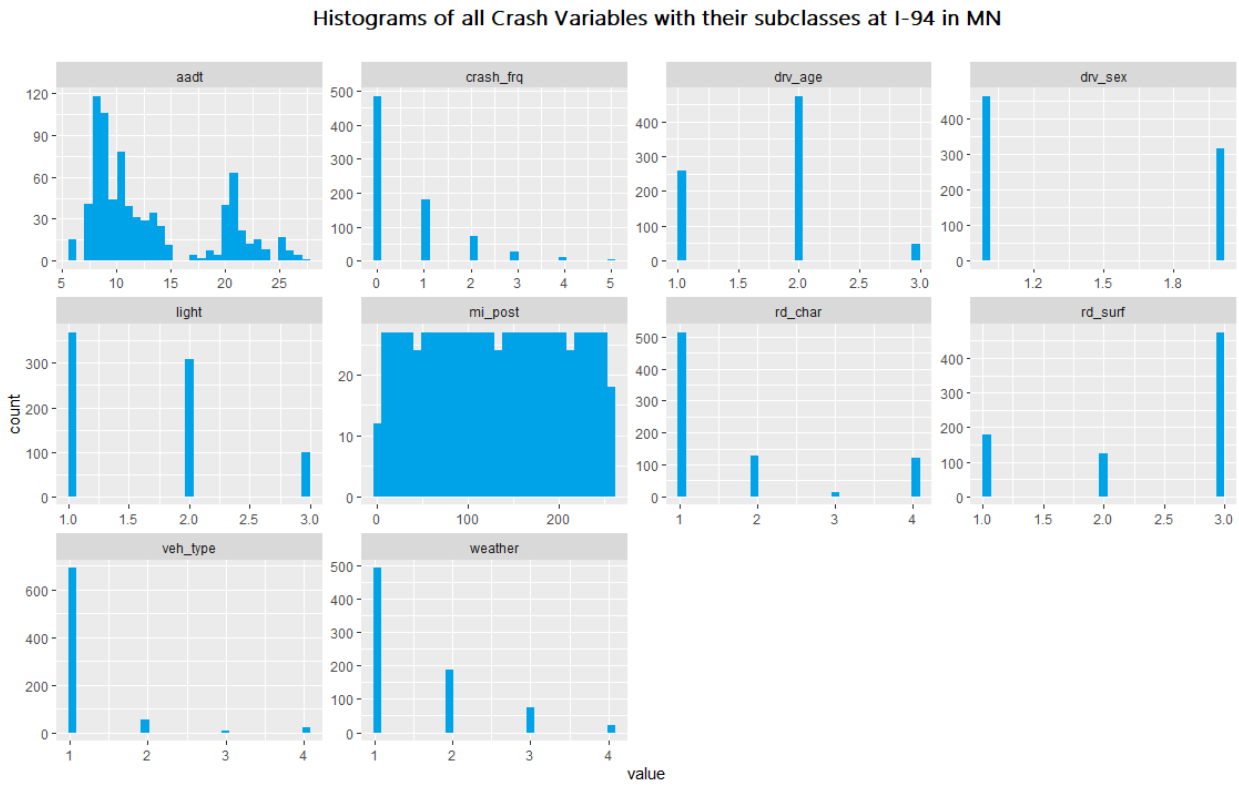

Figure 4: Histograms of all Variables with subclasses at I-94 MN

An Overview of the Regression Models used

The basic goal of regression analysis is to fit a model that best describes the relationship between the predictor variables and a response variable [17]. The following regression models are used in this paper to model the crash frequency at the intestate I-94 in the State of Minnesota for the period (2013-2015); the multiple linear regression model, the Poisson regression model, the Negative Binomial regression model, and the zero inflated Poisson regression model.

The Multiple Linear Regression Model

Multiple linear regression is the basic model that can be used to estimate the relationship between two or more independent variables (also called the explanatory variables) and one dependent variable (also called the response variable).

let Y be the response variable, and X_i be the vector of the explanatory variables. Then, the formula can be expressed as [6] [18]:

$Y = \beta_0 + \beta_1 X_1 + ...+ \beta nXn + \epsilon$ where $\epsilon \sim N(0,\sigma^2)$

Where:

Y = the predicted value of the response variable,

β_0 = the y-intercept,

$\beta_1 X_1$ = the regression coefficient of the first independent variable (X_1),

βnXn = the regression coefficient of the last independent variable Xn,

ε = model error (i.e., how much variation there is in the estimate of Y).

The assumptions of the multiple linear regression include: normal distribution of data, linear relationships between the explanatory and response variables, independence of observations, and homogeneity of variance (also called the homoscedasticity).

The Poisson Regression Model

The Poisson regression can easily handle the count response variable of the crash frequency, since crash data are often described as random events, discrete, and non-negative integers, and often their distributions are found to be skewed as shown in Figure 1, which is close to the Poisson distribution rather than other distributions such as the normal distribution (Hafsa, 2019). The Poisson distribution formula is [6] [18]:

$P(x) = (e \wedge -\lambda * \lambda^{x}) / x!$

Where:

e: is the Euler's number(e = 2.71828),

x: is a Poisson random variable that gives the number of occurrences (x = 0, 1 , 2,.....),

λ: is an average rate of the event in the desired time interval and,

! = factorial of functions.

The crash frequency can be estimated by the expression:

$\lambda i = EXP(\beta Xi)$

where:

λi: the dependent variable (the expected number of crashes per road section),

Xi : a vector of the independent (explanatory) variables,

β: a vector of the estimates (coefficients) of the independent variables Xi.

The assumptions of the Poisson regression include; the response variable should be a count variable, independence of observations, the mean = variance, the log of the mean rate must be a linear function of x.

The Negative Binomial Regression Model

The Negative Binomial regression model is used as an alternative to the Poisson regression, because it relaxes the condition of mean equals to variance, and hence can consider the overdispersion that may exist in crash data. In order to obtain the negative binomial model for the crash frequency, the Poisson regression can be rewritten by adding an error term to its expected number of crashes, and becomes [6] [8] [18]:

$$\lambda_i = EXP (\beta Xi + \varepsilon_i)$$

where:

$EXP (\varepsilon_i)$ is a gamma-distributed error with mean equals one and variance equals α.

This error term allows the variance to differ from the mean. When α is zero, the model becomes Poisson regression, and if α is found to be significantly different from zero, then the negative binomial regression can be used instead of the Poisson regression model [6] [18].

The Zero Inflated Poisson Regression (ZIP) Model

Zero-inflated models are statistical models based on a zero-inflated probability distribution, i.e., a distribution that allows for frequent zero-valued observations. When the number of zeros is so large that the data do not fit standard distributions (e.g., normal, Poisson, binomial, negative-binomial, and beta), the data is referred to as zero inflated. A dual state crash system is assumed in these models, in which one state is the zero crash state that can be regarded as virtually safe during the observation period, while the other state is the non-zero crash state. Thus, they are two-part models, a logistic model for whether an observation is zero or not, and a count model for the other part. Both models can use the same predictor variables but estimate their coefficients separately. So, the predictors can have different effects on the two processes. The crash data used in our paper has a high percent of zero crashes (62.3%), which requires to be fitted with zero inflated models. The Poisson zero inflated is called (ZIP). The two model components are described as follows [14] [18] [19] [20] [21]:

Novelty Journals

ISSN 2394-9716

International Journal of Novel Research in Interdisciplinary Studies
Vol. 10, Issue 1, pp: (6-18), Month: January – February 2023, Available at: www.noveltyjournals.com

$$\Pr(y_j = 0) = \pi + (1-\pi)e^{-\lambda}$$
$$\Pr(y_j = h_i) = (1-\pi)\frac{\lambda^{h_i}e^{-\lambda}}{h_i!}, \qquad h_i \geq 1$$

Where:

y_j is the response variable,

λi is the expected rate of the Poisson count for the h_i individual, and,

π is the probability of extra zeros.

The mixed probabilities for the ZIP model are expressed as follows [22] [23] [24] [25] [26]:

1- a membership of Always-0 group is a binary outcome that can be predicted by logit or probit model. The probability ψ_i that observation i is in Always-0 group is predicted by the characteristic of observation i, so that can be written as:

$\psi_i = F(z_i'\gamma)$

where z_i is the vector of covariates and γ is the vector of coefficients of logit or probit regression.

2- The probability that observation i is in Not always-0 group becomes (1- ψ_i) . For observations in Not always-0 group, their positive count outcome is predicted by the standard Poisson model, so that can be written as:

$$P(y_i|x_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}$$

where μ_i is the conditional mean.

The overall ZIP model can be mathematically written as:

Zero counts in *Always-0 group*

$$P(Y_i = 0|x_i, z_i) = \psi_i \times 1 = \psi_i$$

Zero counts in *Not Always-0 group*

$$P(Y_i = 0|x_i, z_i) = (1-\psi_i) \times \frac{e^{-\mu_i}\mu_i^0}{0!} = (1-\psi_i)e^{-\mu_i}$$

Non zero counts in *Not Always- group*

$$P(Y_i = y_i|x_i, z_i) = (1-\psi_i)\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}$$

Overall

$$P(Y_i = y_i|x_i, z_i) = \begin{cases} \psi_i + (1-\psi_i)e^{-\mu_i} & if\ y_i = 0 \\ (1-\psi_i)\dfrac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} & if\ y_i > 0 \end{cases}$$

Novelty Journals

ISSN 2394-9716

International Journal of Novel Research in Interdisciplinary Studies
Vol. 10, Issue 1, pp: (6-18), Month: January – February 2023, Available at: www.noveltyjournals.com

3. DISCUSSION AND FINDINGS

First, a multiple linear regression model was fitted to the test data and the results were obtained in R as shown in Table 2.

Table 2: The Outcome of Multiple Linear Regression Model

Coefficients	Estimate	Std. Error	t value	Pr (>|t|)
Intercept	0.0251348	0.246419	0.102	0.919
mi_post	0.0002855	0.000443	0.644	0.52
rd_char	0.4587676	0.031091	14.76	<2e-16 ***
rd_surf	-0.034161	0.041004	-0.833	0.405
aadt	0.0025256	0.00627	0.403	0.687
weather	0.0010893	0.043452	0.025	0.98
light	0.0046781	0.048854	0.096	0.924
drv_ age	-0.055867	0.060864	-0.918	0.359
drv_sex	-0.041855	0.069098	-0.606	0.545
veh_type	-0.056436	0.057471	-0.982	0.327
Residual standard error: 0.7739 on 535 degrees of freedom Multiple R-squared: 0.2974, Adjusted R-squared: 0.2855 F-statistic: 25.16 on 9 and 535 DF, p-value: < 2.2e-16				

From Table 2, we can see that the only significant variable is the road characteristics as its p-value is less than 0.001. The other independent variables are insignificant. The multiple R-squared measures the strength of the linear relationship between the response variable and the predictor variables, the higher the value the better the fit. We can see that the R-squared is 0.2974, which is small. Also, the adjusted R squared, which is a modified version of R-squared that has been adjusted for the number of predictors in the model is only 0.2855. The assumption of normality was already checked by plotting a histogram for the dependent variable, and it was found to be skewed as shown in Figure 1. Therefore, the normality assumption is violated. The linearity assumption was already checked using the matrix scatterplot of all variables as shown in Figure 2. The matrix scatterplot showed almost linear relationships between the dependent and independent variables. The independence of observations was checked by determining the correlation matrix of all variables, which showed very small correlation for almost all explanatory variables as shown in Figure 3. The assumption of homoscedasticity is checked by generating the plot of residuals vs fitted values as shown in Figure 5. We can see from Figure 5 that the red lines representing the mean of the residuals are all basically horizontal and centered around zero. This means there are no outliers or biases in the data that would make a linear regression invalid. In the Normal Q-Q plot, we can see that the residuals are almost perfect one-to-one line. Based on these residuals, we can say that our model meets the assumption of homoscedasticity. However, since the normality assumption is violated, and almost all variables are insignificant in the model, this suggests using a different model to better fit our data.

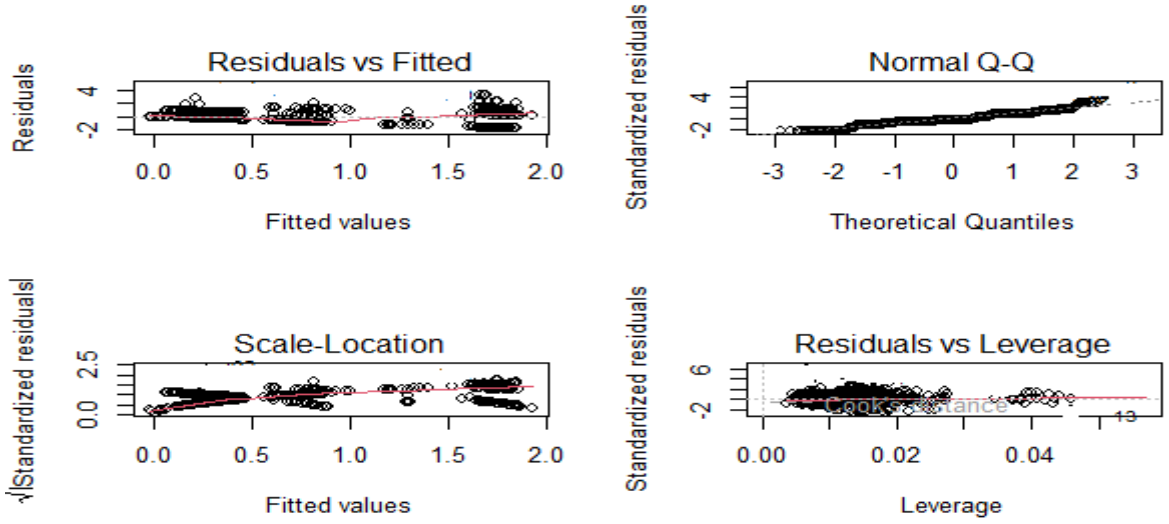

Figure 5: Plot of Residuals vs Fitted for the Multiple Linear Regression Model

Next, a Poisson regression model was fitted to the test data, and the results were obtained in R as shown in Table 3.

Table 3: The Outcome of Poisson Regression Model

Coefficients	Estimate	Std. Error	z value	Pr (> |z|)
Intercept	-1.38562	0.428308	-3.235	0.00122 **
mi_post	0.00048	0.000753	0.634	0.52578
rd_char	0.58693	0.04265	13.76	< 2e-16 ***
rd_surf	-0.05984	0.070023	-0.855	0.39278
aadt	0.00247	0.010732	0.23	0.81817
weather	0.00974	0.076799	0.127	0.89903
light	0.01	0.081079	0.123	0.90188
drv_age	-0.0952	0.108811	-0.875	0.38163
drv_sex	-0.06581	0.120209	-0.547	0.58406
veh_type	-0.13072	0.117035	-1.117	0.26401
Null deviance: 710.64 on 544 degrees of freedom Residual deviance: 522.74 on 535 degrees of freedom AIC: 994.62				

From Table 3, we can notice that the only significant independent variable is the road characteristics as its p-value is less than 0.001. The other independent variables are insignificant. The coefficient for road characteristics is 0.586. This means that the expected log count for a one-unit increase in rd_char is 0.586, which will positively increase the mean number of crash_frq by 0.586. Likewise, the coefficient for (aadt) for example is 0.0024, which indicates that the expected log count for (aadt) is 0.0024. This variable is statistically insignificant ($p = 0.818 > 0.05$). The coefficient for road surface is - 0.059. This means that the expected log count for a one-unit increase in rd_surf is – 0.059, which will negatively affect the crash frequency and decrease the mean number of crash_frq by 0.059. The null deviance tells us how well the response variable can be predicted by a Poisson model with only an intercept term. The residual deviance tells us how well the response variable can be predicted by a Poisson model that include all independent variables. Since the Poisson regression model is a form of the Generalized Linear Models (GLMs), there are many goodness of fit measures can be used for estimating how well the model fits the data, such as the residual deviance, and Pearson Chi square test. If the model fits the data well, the ratio of the residual deviance to the degrees of freedom should be close to one. From the model outcome, we can see that the ratio of deviance/df = 522.74/535 = 0.977 which is very close to 1. In addition, we conducted a Chi-Square goodness of fit test in R, and we got the p-value for this test as (p-value = 0.639 > 0.05), which suggests that the data fits the model reasonably well. Another important aspect related to Poisson regression is the overdispersion that is often exist in count data. Poisson regression of overdispersed data leads to a deflated standard errors and insufficient test statistics. In R, overdispersion can be analyzed using the "qcc" package. If the dispersion ratio is larger than one, this will indicate overdispersion in the data. After conducting the test in R, we got the overdispersion ratio of the test as (1.0596), which is slightly greater than 1, suggesting an overdispersion in the data. Overdispersion can be fixed by choosing a different distributional family, such as negative binomial regression. So, we will fit a negative binomial model using the 'glm.nb' function in the 'MASS' package in R.

Next, a Negative Binomial regression model was fitted to the test data, and the results were obtained in R as shown in Table 4.

Table 4: The Outcome of Negative Binomial Regression Model

Coefficients	Estimate	Std. Error	z value	Pr (> |z|)
Intercept	-1.3841882	0.4296004	-3.222	0.00127 **
mi_post	0.0004798	0.0007558	0.635	0.52554
rd_char	0.5868364	0.0427853	13.716	< 2e-16 ***
rd_surf	-0.0601328	0.0702482	-0.856	0.39200
aadt	0.0024895	0.0107674	0.231	0.81715
weather	0.0098412	0.0770335	0.128	0.89834

light	0.0098125	0.0813588	0.121	00.90400
drv_age	-0.0952326	0.1091446	-0.873	0.38292
drv_sex	-0.0666828	0.1205998	- 0.553	0.58031
veh_type	-0.1304260	0.1172604	-1.112	0.26602
Null deviance: 707.78 on 544 degrees of freedom Residual deviance: 520.80 on 535 degrees of freedom AIC: 990.72				

From Table 4, we can see again that the only significant independent variable is the road characteristics as its p-value is less than 0.001. The other independent variables are insignificant. There are positive and negative coefficients. The positive coefficients indicate that the expected log count for a one-unit increase will positively increase the mean number of crash_frq by the coefficient value. The negative coefficients indicate that the expected log count for a one-unit increase will decrease the mean number of crash_frq by the coefficient value. Therefore, the positive coefficients of (mi_post, rd_char, aadt, weather, light) will positively increase the mean of crash frequency. But the negative coefficients of (rd_surf, drv_age, drv_sex, veh_type) will negatively decrease the mean of crash frequency. From the model outcome, we can see that the ratio of the residual deviance/df = 520.8/535 = 0.988, which is very close to 1. So, we can say that the data fits the model well and better than the Poisson regression. In addition, we conducted an overdispersion ratio test in R, and we got an overdispersion ratio of (1.0309), which is slightly greater than 1, but better than what we got from the Poisson model. The AIC of the negative binomial is slightly lower than the AIC of Poisson model, which indicates a better fit as well. However, since most independent variables are still insignificant, and our data consists of 62% zero crash sections, we will fit a zero inflated Poisson model.

Next, a Zero Inflated Poisson (ZIP) regression model was fitted to the test data, and the results were obtained in R as shown in Table 5.

Table 5: The Outcome of Zero-Inflated Poisson Regression Model

Count model coefficients (Poisson with log link)				
Coefficients	Estimate	Std. Error	z value	Pr (> |z|)
Intercept	-1.4519528	0.3544320 -	-4.097	-4.19e-05 ***
mi_post	0.0005170	0.0006146	0.841	< 2e-16 ***
rd_char	0.5910766	0.0342967	17.234	< 2e-16 ***
rd_surf	0.0624270	0.0571916	1.092	< 2e-16 ***
aadt	0.0120272	0.0085381	1.409	5.79e-10 ***
weather	0.0288467	0.0620952	0.465	< 2e-16 ***
light	0.0389533	0.0677771	0.575	< 2e-16 ***
drv_age	0.0334547	0.0864880	0.387	< 2e-16 ***
drv_sex	-0.0837287	0.0964907	- 0.868	0.3855
veh_type	0.1708939	0.0959439	1.781	< 2e-16 ***
Zero-inflation model coefficients (binomial with logit link)				
Coefficients	Estimate	Std. Error	z value	Pr (> |z|)
Intercept	-1.5343	0.3755	-4.086 4	4.38e-05 ***
mi_post	1.646	0.0923781	19.314	< 2e-16 ***
rd_char	160.681	0.4184956	6.196	5.79e-10 ***
rd_surf	2.9428092	0.8523669	3.453	0.000555 ***
aadt	15.902	0.0002064	3.735	0.000188 ***
weather	105.894	0.0390319	9.589	< 2e-16 ***
light	1.7772101	0.3233166	5.497 3	3.87e-08 ***
drv_age	1130.471	0.8360020	3.659	0.000253 ***
drv_sex	-177.016	0.3226731	5.502	0.09834
veh_type	501.927	0.0005307	6.020	1.75e-09 ***
Number of iterations in BFGS optimization: 109 Log-likelihood: -467.1 on 20 Df				

From Table 5, we can find two model parts; the first contains a Poisson regression coefficients for each of the variables along with standard errors, z-scores, and p-values for the coefficients. The second part corresponds to the inflation model. This includes logit coefficients for predicting excess zeros along with their standard errors, z-scores, and p-values. We can see that all the independent variables become statistically significant in both the count model and the zero inflation model as their p-values are now much less than 0.001, except the variable of driver sex, which turned to be insignificant as its p-value is greater than 0.05. Moreover, we can notice that all the independent variables having now positive coefficients in both the count model and the zero inflation model, except the driver sex, which has a negative coefficient. The positive coefficients of all explanatory variables suggest that the expected log count for a one-unit increase will positively increase the mean number of crash_frq by the coefficient value. Therefore, all variables included in the model (except the drv_sex) will positively increase the mean of crash frequency in an amount that corresponds to each coefficient. We can check if our ZIP model fits the data significantly better than the null model, i.e., the intercept-only model. To show that this is the case, we can compare the current model to a null model without predictors using chi-squared test on the difference of log likelihoods in R, and we got a very small p-value (0.00297). Therefore, we can be confident to say that our ZIP model is statistically significant and fits the data very well. In addition, we generated the plot of the residuals vs predicted and the Q-Q plot of the ZIP model as shown in Figure 6.

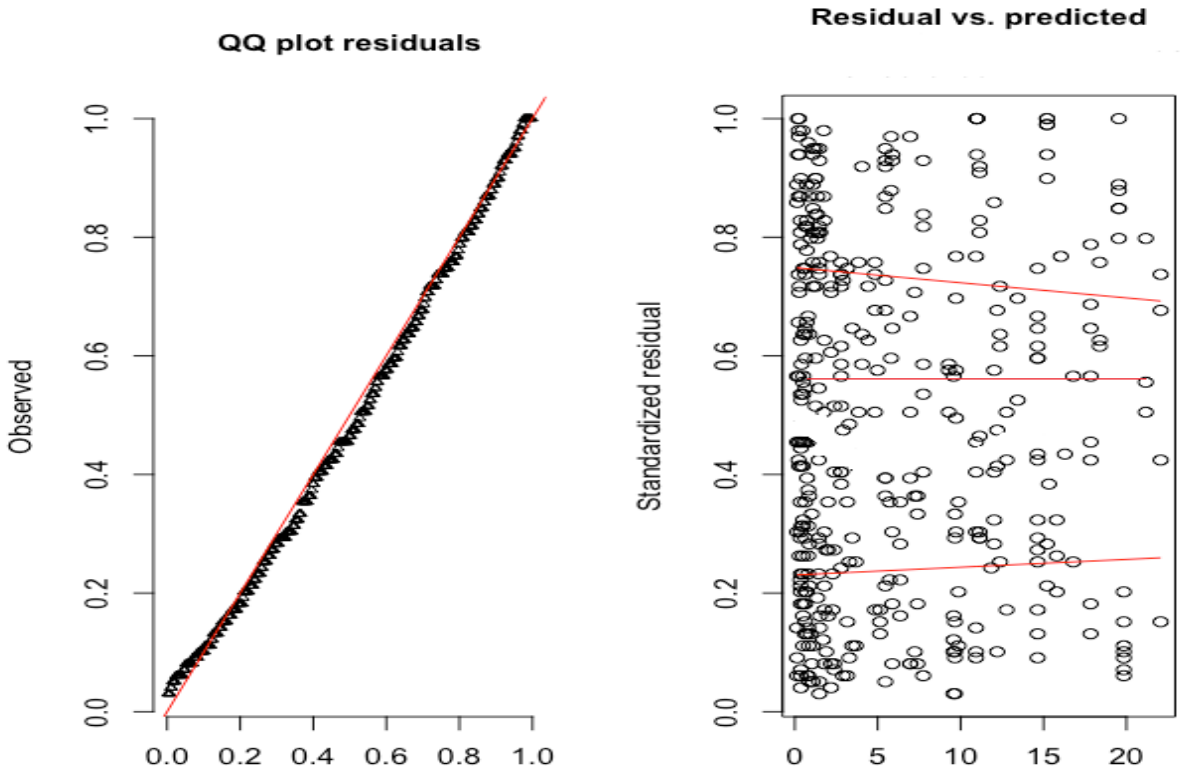

Figure 6: Plot of Residuals vs Predicted for the ZIP model

We can see from Figure 6 that the quantile-quantile plot (Q-Q) looks almost fine, and the standardized residuals show close pattern with expectations, since the red lines are nearly horizontal, which indicate a very good model fit to the data. So, we can conclude that our zero inflated Poisson regression (ZIP) model has captured almost all the independent variables in the model and fit the data much better than the multiple linear regression, Poisson regression, and the negative binomial regression models. Therefore, our ZIP model is considered as the optimal regression model to predict the crash frequency in this paper.

Next, the predicted crash frequencies for each crash rate at the I-94 in MN were obtained using our ZIP model and the test data as shown in Table 6. The predicted zero crash sections are 119 (89% prediction), the predicted one crash sections are 54 (85% prediction), and the predicted two crash sections are 16 (76% prediction). The achieved prediction percent by the ZIP model looks very good.

Table 6: Observed vs Predicted Crashes at I-94 by the ZIP model

Crash Frequency	Observed Crashes	Predicted Crashes	% Prediction
0	133	119	89
1	63	54	85
2	21	16	76
3	9	6	66
4	5	3	60
5	1	1	100

4. CONCLUSION

Road traffic crashes are considered a leading cause of death in the United States and worldwide. Modeling vehicle crash frequency can provide a clear understanding of the significant risk factors contributing to the vehicle crashes. This paper used crash data to model the crash frequency on the interstate I-94 in the State of Minnesota for the years 2013-2015. Different risk factors that could contribute to the crash occurrence were included in the research, such as the road characteristics (i.e., straight sections, upgrades, curves), the road surface conditions (i.e., dry, wet), the weather conditions (i.e., clear, rainy), the annual average daily traffic (AADT) of the road sections, the light conditions (i.e., day light, dark), the driver's age, the driver's sex, and the vehicle type (i.e., passenger car, truck). In order to find the optimal model, varieties of regression models were used to predict the crash frequency, including; the multiple linear regression, the Poisson regression, the negative binomial regression, and the zero-inflated Poisson (ZIP) regression. The multiple linear regression model identified only one significant risk factor (road characteristics) and failed to fulfill the normality assumption, because the crash data was skewed. The Poisson regression model determined only one significant variable (road characteristics) and couldn't consider the overdispersion that existed in the crash data. Although the negative binomial regression model better fitted the data than the Poisson regression, however, it also failed to handle the overdispersion in the crash data. As the crash data contained a big number of zero crash road sections, the zero-inflated Poisson regression (ZIP) model was used to handle the excess zeros of crash sections. The ZIP model turned to be very effective in modeling the crash frequency by identifying a large number of significant risk factors contributing to crash occurrence and fitting the data reasonably well. All variables in the research were captured by the ZIP model to be significant (except the driver's sex variable). These significant variables include; the mile post, the road characteristics, the road surface conditions, the weather conditions, the light conditions, the traffic volume (AADT), the driver's age, and the vehicle type. For future research, we can include the interaction between the independent variables as additional risk factors in modeling the crash frequency.

REFERENCES

[1] NHTSA. Traffic Safety Facts. 2021. National Highway Traffic Safety Administration (NHTSA). NHTSA's 2021 Estimate of Traffic Deaths Shows 16-Year High. Retrieved February 5, 2023.

[2] CDC. MOTOR VEHICLE SAFETY AT WORK, crash facts. 2022. Centers for Disease Control and Prevention (CDC). https://www.cdc.gov/niosh/motorvehicle/resources/crashdata/facts.html. Retrieved February 5, 2023.

[3] Lao, Y., Wu, Y., Corey, J. and Wang, Y. 2011. Modeling Animal-Vehicle Collisions Using Diagonal Inflated Bivariate Poisson Regression. Accident Analysis and Prevention, 43, 220-227. http://dx.doi.org/10.1016/j.aap.2010.08.013

[4] Lord, D. and Mannering, F. 2010. The Statistical Analysis of Crash-Frequency Data: A Review and Assessment of Methodological Alternatives. Transportation Research Part A, 44, 291-305. http://dx.doi.org/10.1016/j.tra.2010.02.001

[5] Caliendo, C., Guida, M. and Parisi, A. 2007. A Crash-Prediction Model for Multilane Roads. Accident Analysis and Prevention, 39, 657-670. http://dx.doi.org/10.1016/j.aap.2006.10.012

[6] Abdulhafedh, A. 2016. Crash Frequency Analysis. Journal of Transportation Technologies, **6**, 169-180. http://dx.doi.org/10.4236/jtts.2016.64017

[7] Park, E.-S. and Lord, D. 2007. Multivariate Poisson-Lognormal Models for Jointly Modeling Crash Frequency by Severity. *Transportation Research Record*, **2019**, 1-6. http://dx.doi.org/10.3141/2019-01

[8] Ma, J., Kockelman, K.M. and Damien, P. 2008. A Multivariate Poisson-Lognormal Regression Model for Prediction of Crash Counts by Severity, Using Bayesian Methods. *Accident Analysis and Prevention*, **40**, 964-975. http://dx.doi.org/10.1016/j.aap.2007.11.002

[9] El-Basyouny, K. and Sayed, T. 2009. Collision Prediction Models Using Multivariate Poisson-Lognormal Regression. *Accident Analysis and Prevention*, **41**, 820-828. http://dx.doi.org/10.1016/j.aap.2009.04.005

[10] Abdulhafedh, A.2017. Road Crash Prediction Models: Different Statistical Modeling Approaches. Journal of Transportation Technologies, 7,190-205

[11] Abdulhafedh, A.2022. Modeling Vehicle Crash Frequency When Multicollinearity Exists in Vehicle Crash Data: Ridge Regression versus Ordinary Least Squares Linear Regression. Open Access Library Journal, 9: e8873.

[12] El-Basyouny, K. and Sayed, T. 2006. Comparison of Two Negative Binomial Regression Techniques in Developing Accident Prediction Models. *Transportation Research Record*, **1950**, 9-16. http://dx.doi.org/10.3141/1950-02

[13] Abdulhafedh, A.2017. Road Traffic Crash Data: An Over-view on Sources, Problems, and Collection Methods. Journal of Transportation Technologies, 7, 206-219

[14] Abdulhafedh, A.2017. Road Crash Prediction Models: Different Statistical Modeling Approaches. Journal of Transportation Technologies, 7,190-205.

[15] Lord, D., Washington, S.P. and Ivan, J.N. 2007. Further Notes on the Application of Zero Inflated Models in Highway Safety. *Accident Analysis and Prevention*, **39**, 53-57. http://dx.doi.org/10.1016/j.aap.2006.06.004

[16] Xie, Y. and Zhang, Y. 2008. Crash Frequency Analysis with Generalized Additive Models. *Transportation Research Record*, **2061**, 39-45. http://dx.doi.org/10.3141/2061-05

[17] Abdulhafedh, A.2022. Comparison between Common Statistical Modeling Techniques Used in Research, Including: Discriminant Analysis vs Logistic Regression, Ridge Regression vs LASSO, and Decision Tree vs Random Forest. Open Access Library Journal, 9: e8414.

[18] Washington, S.P., Karlaftis, M.G. and Mannering, F. 2010. Statistical and Econometric Methods for Transportation Data Analysis. 2nd Edition, Chapman Hall/CRC, Boca Raton.

[19] Lord, D., Washington, S.P. and Ivan, J.N. 2007. Further Notes on the Application of Zero Inflated Models in Highway Safety. Accident Analysis and Prevention, 39, 53-57. http://dx.doi.org/10.1016/j.aap.2006.06.004.

[20] Lord D., Washington S. P., and Ivan J. N. 2005. Poisson, Poisson-Gamma, and Zero-Inflated Regression Models of Motor Vehicle Crashes: Balancing Statistical Fit and Theory. Accident Analysis & Prevention, Vol. 37, No. 1, 2005, pp. 35–46.

[21] Shankar V., Milton J., and Mannering F. 1997. Modeling Accident Frequencies as Zero-Altered Probability Processes: An Empirical Inquiry. Accident Analysis & Prevention, Vol. 29, No. 6, pp. 829–837.

[22] Wood S. N. 2006. Generalized Additive Models: An Introduction with R. Chapman and Hall/CRC, Boca Raton, Fla.

[23] Greene W. H. 2000. Econometric Analysis, 4th ed. Prentice Hall, Upper Saddle River, N.J.

[24] Cameron, A.C. and Trivedi, P.K. 1998. Regression Analysis of Count Data. Cam-bridge University Press, Cambridge, UK.

[25] Hilbe, J. 2014. Modeling Count Data. Cambridge University Press, London.

[26] Lauridsen, J. and Mur, J. 2006. Multicollinearity in Cross-Sectional Regressions. Journal of Geographical Systems, 8, 317-333