Applications of some discrete regression models for count data

B. M. Golam Kibria
Department of Statistics, Florida International University
University Park, Miami, FL 33199, USA
Email: kibriag@fiu.edu

Abstract

In this paper we have considered several regression models to fit the count data that encounter in the field of Biometrical, Environmental, Social Sciences and Transportation Engineering. We have fitted Poisson (PO), Negative Binomial (NB), Zero-Inflated Poisson (ZIP) and Zero-Inflated Negative Binomial (ZINB) regression models to run-off-road (ROR) crash data which collected on arterial roads in south region (rural) of Florida State. To compare the performance of these models, we analyzed data with moderate to high percentage of zero counts. Because the variances were almost three times greater than the means, it appeared that both NB and ZINB models performed better than PO and ZIP models for the zero inflated and overdispersed count data.

Keywords: AIC; Count Data; GLM; Goodness of fit; Poisson Model; Negative Binomial; Prediction; Zero-Inflated Poisson; Zero-inflated Negative Binomial.

1. Introduction

Many outcomes in traffic accident, clinical medicine, biomedical research that are non-negative and discrete in nature. Thus it may be natural to model these count data with discrete distribution instead of continuous, which is usually being used as normal. The Poisson (PO) distribution has been used to model the count data for a long time. It has an important constraint that the mean and variance are equal. However, many processes in real life are over dispersed (variances are greater than means) and violate the underlying assumption of Poisson (PO) distribution. In that cases the negative binomial (NB) distribution is a natural and more flexible extention of the Poisson distribution and allows for over-dispersion compared to Poisson distribution. Several researchers have suggested to use the NB regression model as an alternative to the PO regression model when the count data are over or under dispersed. Both Poisson and Negative Binomial distribution have been used for predicting the accidents related count frequencies by Miaou (1994), Shankar et al. (1995, 1997), Poch and Mannering (1996), Milton and Mannering (1998) and Lee and Mannering (2002) among others. It is noted that most of the accidents data contain excess number of counts with zero. Unfortunately, the Poisson and NB models do not address the possibility of zero counts and can not fit the data properly. Then corresponding inflated models, say zero inflated Poisson (ZIP) and zero inflated negative binomial (ZINB) are very useful to describe the zero inflated count data. Both ZIP and ZINB models incorporate extra variation than the corresponding Poisson and NB models. The most appropriate reference for ZIP regression model are Lambert (1992) and Lee et al. (2001) and ZINB regression model are Cameron and Trivedi (1998) and Long (1997) among others. To select an appropriate inflated model, that is, ZIP

model over Poisson or ZINB model over NB, the Vuong (1989) statistics is one of the popular test. Beside modeling crash or accidents data, these four models have been used in environmental science by Warton (2005), in biomedical science (Yau et al. 2003) among other discipline. The main objective of this paper is to provide a comprehensive review of these four models and discuss how to fit appropriate statistical models for count data using STATA software, specially for the over dispersed and an excess number of counts in the data.

The organization of this paper is as follows. The statistical methodology and goodness of fit of the models are given in section 2. To compare the performance of the models, an example has been illustrated in section 3. This paper ends up with some concluding remarks in section 4.

2. Methodology

2.1 Regression Models

In dealing with count data, for examples accident, number of ER visits, crashes that are non-negative and discrete in nature, it make more sense to model these count data using PO, NB, ZIP or ZINB distributions. Regardless of whether the assumed model is a PO, NB, ZIP or ZINB, it will be assumed that the occurrences will be independent of each other. The four types of models are described briefly in the following subsections.

2.1.1 Poisson Regression

If the variance of the counts approximately equals the mean of the count, then the Poisson regression model can be expressed as

$$P(y_i|x_i) = \frac{\exp(-\theta_i)\theta_i^{yi}}{y_i!} \quad \text{for} \quad y_i = 0,1,2, \tag{2.1}$$

where y_i is the number of counts (crashes for example) for a particular period or region i, θ_i is the expected number of crashes per period, which can be modeled as

$$\theta_i = \exp(\mathbf{x}_i'\beta),$$

where $\mathbf{x}_i'$ is the vector of explanatory variables and β is the vector of unknown regression parameters. The main constraint in the PO distribution is that the mean and variance are same, that is, $E(Y) = V(Y) = \theta.$ When there is a heterogeneity or over dispersion in the population, the Poisson regression does not work well. The following negative binomial (NB) regression model is a possible candidate as an alternative to PO regression.

2.1.2 Negative Binomial Model

The Negative Binomial (NB) distribution can be obtained from the mixture of Poisson and Gamma distribution and is expressed as

$$P(y_i|x_i)=\frac{\Gamma(y_i+1/\alpha)}{y_i!\Gamma(1/\alpha)}\left[\frac{1}{1+\alpha\theta_i}\right]^{1/\alpha}\left[\frac{\alpha\theta_i}{1+\alpha\theta_i}\right]^{y_i} \quad \text{for} \quad y_i=0,1,2,3, \qquad (2.2)$$

where y_i is the number of crashes for road segment i, θ_i is the expected number of crashes per period, which can be expressed as

$$\theta_i = \exp(\mathbf{x}_i'\beta),$$

The mean and variance of negative binomial distribution are respectively, $E(y_i|x_i)=\theta_i$ and $Var(y_i|x_i)=\theta_i[1+\theta_i\alpha)]>E(y_i|x_i)$. Thus the NB model is also over-dispersed and allows extra variation relative to the traditional PO model. It has more desirable properties than the Poisson model to describe the relationship between ROR crashes and geometric characteristics (Chin and Quddus 2003). The variance of NB is significantly greater than the mean. Here α represents an ancillary or dispersion parameter which indicate the degree of over dispersion. If $\alpha=0,$ the NB regression model reduces to traditional Poisson regression model. Many researchers in different fields have considered both Poisson and Negative Binomial models: Miaou (1994), Karlaftis and Tarko (1998), Hauer (2001), Lee et al. (2002), Byers et al. (2003), Berhanu (2004), Yau et al. (2004) and Lord et al. (2005) to mention a few. However, when excess zero occur, both PO and NB regression models are not that useful to fit the zero inflated models. In that case both ZIP and ZINB models are appropriate choice.

2.1.3 Zero-Inflated Poisson (ZIP)

The zero-inflated Poisson model has a long history to use in the literature of count data to deal with an excess zeros in data. The ZIP model can be defined as

$$P(y_i|x_i)=\begin{cases}\psi+(1-\psi)\exp(-\theta_i) & \text{for} \quad y_i=0\\ (1-\psi)\dfrac{\exp(-\theta_i)\theta_i^{yi}}{y_i!} & \text{for} \quad y_i>0,\end{cases} \qquad (2.3)$$

where y_i is the number of crashes for road segment $i,$ for a chosen time period, θ_i is the expected number of crashes per period, which can be modeled as

$$\theta_i = \exp(\mathbf{x}_i'\beta),$$

where x_i' is the vector of explanatory variables and β is the vector of parameters, and $\psi(0<\psi<1)$ is the probability of being in the zero crash state, determined by a logit model (Lambert 1992, Long 1997). That is

$$\text{logit}(\psi)=\log_e\left(\frac{\psi}{1-\psi}\right)=\mathbf{z}_i'\gamma,$$

or

$$\psi=\frac{\exp(\mathbf{z}_i'\gamma)}{1+\exp(\mathbf{z}_i'\gamma)},$$

where $\mathbf{z}_i'$ is the vector of explanatory variables and γ is the corresponding vector of parameters. The mean and variance of ZIP model are respectively,

$$E(y_i|x_i) = \theta_i(1-\psi_i) < \theta_i \quad \text{and} \quad Var(y_i|x_i) = E(y_i|x_i)\left[1+\frac{\psi_1}{1-\psi_i}E(y_i|x_i)\right] > E(y_i|x_i).$$

Thus the ZIP model is over-dispersed and allows extra variation relative to the Poisson model. If $\psi = 0,$ the ZIP model reduces to a classical Poisson regression model, otherwise the variance exceeds the mean (Long 1997). Different researchers have used ZIP model for several purposes and times, among them, Mullahy (1986), Lambart (1992), Gupta et al. (1996), Lee et al. (2001), Cheung (2002) and Yau et al. (2003) are notable. In the case of ZIP model the dual-state system exists and can be described by combining the Poisson count model (normal-count state) and the binary process (zero state) for the ZIP model. Testing for overdispersion in Poisson and binomial regression models we refer Dean (1992) among others.

2.1.4 Zero-Inflated Negative Binomial (ZINB)

The zero-inflated negative (ZINB) model can be formulated as section 2.1.3. Following Cheung (2002), the ZINB model can be expressed as

$$P(y_i|x_i) = \begin{cases} \psi + (1-\psi)\left[\dfrac{1}{1+\alpha\theta_i}\right]^{1/\alpha} & \text{for} \quad y_i = 0 \\ (1-\psi)\dfrac{\Gamma(y_i+1/\alpha)}{y_i!\Gamma(1/\alpha)}\left[\dfrac{1}{1+\alpha\theta_i}\right]^{1/\alpha}\left[\dfrac{\alpha\theta_i}{1+\alpha\theta_i}\right]^{y_i} & \text{for} \quad y_i > 0, \end{cases} \tag{2.4}$$

where y_i is the number of crashes for road segment i, θ_i is the expected number of crashes per period, which can be modeled as

$$\theta_i = \exp(\mathbf{x}_i'\beta),$$

where $\mathbf{x}_i'$ is the vector of explanatory variables and β is the vector of parameters, and ψ is the probability of being in the zero crash state, determined by a logit model (Long 1997). That is

$$\text{logit}(\psi) = \log_e\left(\frac{\psi}{1-\psi}\right) = \mathbf{z}_i'\gamma,$$

or

$$\psi = \frac{\exp(\mathbf{z}_i'\gamma)}{1+\exp(\mathbf{z}_i'\gamma)},$$

where $\mathbf{z}_i'$ is the vector of explanatory variables and γ is the corresponding vector of parameters. The mean and variance of ZINB model are respectively,

$$E(y_i|x_i) = \theta_i(1-\psi_i) < \theta_i \quad \text{and} \quad Var(y_i|x_i) = E(y_i|x_i)\left[1+\left(\frac{\psi_1+\alpha}{1-\psi_i}\right)E(y_i|x_i)\right] > E(y_i|x_i).$$

Thus the ZINB model is also over-dispersed and allows extra variation relative to the traditional NB model. If $\psi = 0,$ the ZINB model reduces to a classical NB regression model. For $\alpha = 0,$ the ZINB regression model reduces to ZIP regression model and for $\psi = 0$ and $\alpha = 0,$ it reduces to a classical Poisson regression model. For properties and statistical inference, including the maximum

likelihood estimation of the parameters for ZIP or ZINB models, we refer Gupta et al. (1996), Lambert (1992), Long (1997) among others. The parameters of the models have been estimated by maximum likelihood estimation method using statistical software STATA 9.0.

2.2 Selecting Appropriate Models

2.2.1 Vuong Statistic: Selecting Inflated Model over Traditional

A number of tests for example, likelihood ratio test, the Wald test and the score tests are available for testing the zero inflation in the model (for example, see van den Broek 1995 and Lee at al., 2004 among others). For our convenience we will consider Vuong statistic, which is available in STATA. To define the Vuong (V) statistic, suppose $f_1(y_i|x_i)$ and $f_2(y_i|x_i)$ denote the probability density function of zero-inflated model (ZIP or ZINB) and parent or traditional model (PO or NB) respectively and $F_1(y_i|x_i)$ and $F_2(y_i|x_i)$ denote their corresponding cumulative distribution functions. We want to test the following hypotheses

H_0 : Two distribution functions are equivalent
H_a : Two distribution functions are different (two tailed test (two sided))
H_a : $F_1(y_i|x_i)$ is better than $F_2(y_i|x_i)$ (upper tailed test (one sided))
H_a : $F_1(y_i|x_i)$ is worse than $F_2(y_i|x_i)$ (lower tailed test (one sided)) (2.5)

Now we define,

$$m_i = \log\left[\frac{\hat{f}_1(y_i|x_i)}{\hat{f}_2(y_i|x_i)}\right],$$

where $\hat{f}_1(y_i|x_i)$ and $\hat{f}_2(y_i|x_i)$ are predicted probabilities of the corresponding models $f_1(y_i|x_i)$ and $f_2(y_i|x_i)$ respectively. Let $\overline{m} = \frac{1}{n}\sum_{i=1}^{n} m_i$ and $s_m = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(m_i - \overline{m})^2}$ denote the mean and standard deviation of the measurements m_i . Then the Vuong statistic is defined as

$$V = \frac{\overline{m}}{S_m / \sqrt{n}}$$

For large sample size and under the null hypothesis the statistic *V* has the asymptotic standard normal distribution. Note that Shankar et al. (1997), Carson and Mannering (2001) and Lee and Mannering (2002) among others have defined that *V* statistic has a *t* distribution instead of approximate standard normal. This is not a correct statement as *t* statistic is developed based on the assumption that data are from normal distribution. In the context of count data, the parent population is discrete and for large sample size, *V* has asymptotic normal distribution. The critical values of *t* statistic depend on its degrees of freedoms (*df*). For small degrees of freedoms the *t* distribution is leptokurtic. However, as the number of degrees of freedom increases, the *t* distribution

approaches the standard normal distribution. Thus to make any decision about the null hypothesis it is reasonable to compare the observed value of the test statistic with the critical value from standard normal distribution.

2.2.2 Selecting Over Dispersed Model

To test the whether data are over dispersed or not, we test the following hypothesis,

$$H_0 : \alpha = 0$$
$$H_a = \alpha \neq 0$$

The corresponding test statistic is

$$z = \frac{\hat{\alpha}}{SE(\hat{\alpha})}.$$

Under H_0 and for large sample size, the *z* has approximate standard normal distribution. Reject H_0 at α level of significance if

$$|z| > Z_{\alpha/2}$$

If we reject the null hypothesis we should accept that NB and ZINB models are more appropriate compare to PO and ZIP models.

2.3 Parameter Estimation and Model Selection

2.3.1 Parameter Estimation

The maximum likelihood method has been considered due to limitation of the application of STATA, which consider the maximum likelihood estimation (mle) technique. To evaluate the model, it is necessary to examine the significance of the variables included in the model. For a better model, the estimated regression coefficients have to be statistically significants. Usually, the *t* test is used to determine the significance of the regression coefficients. Moreover, the intuitive judgment of the experimenters should be considered.

2.3.2 Goodness of fit

After fitting some models to the data, it is essential to check the overall fit as well as quality of the fit. The quality of the fit between the observed values (*y*) and predicted values $(\hat{\mu})$ can be measured by various test statistics, however, the one of the useful statistic is called deviance and defined as:

$$D(y:\hat{\mu}) = -2\left[L(\hat{\mu};y) - L(y;y)\right]$$

For a better model, one would expect smaller value of the $D(y:\hat{\mu})$. For detailed about the fitting of a generalized linear model (GLM)s readers are refer to McCullah and Nelder (1987) and Agresti (2002) among others.

2.3.3 AIC: Selecting Best Model

Akaike's information criterion (AIC, Akaike 1973) was used to compare the different models. The AIC is defined as

$$AIC = -2L + 2k,$$

where L is the log-likelihood and $k < p$ is the number of parameters in the model. For the best fitted model one must expect lowest AIC value.

2.4 Model Assessment: Prediction

Our main objective is to compare the models (PO, ZIP, NB and ZINB) in the sense of better prediction. We will fit the models and then predict the number of crashes. For a better model, one would expect the predicted frequencies should be close to corresponding observed frequencies. Since the theoretical comparison is hard to make, a numerical comparison using a real data set are given in the following section.

3 Example

3.1 Data Description

To demonstrate the performance of the models, we consider two years of run-off-road (ROR) crash data (2000-2001) which encompassing 588 centerline miles in the rural south region of Florida State Highway System classified as Principal Arterial only. The data was collected from two data sources. The first is the crash data which were extracted from the Florida Department of Transportation (FDOT) Crash Analysis Reporting System (CARS). This database contains a great deal of data regarding the drivers or pedestrians, conditions of the vehicles, contributing causes, weather condition, lighting condition, etc. The second source of data includes all of the roadway geometric and traffic related features, which were extracted from the Roadway Characteristics Inventory (RCI) database also maintained by the FDOT. The Civil and Environmental Engineering Department at the Florida International University (FIU) has developed a Dynamic Segmentation (DYSEG) program. The DYSEG combines both databases: Crash Analysis Reporting System (CARS) and Roadway Characteristics Inventory (RCI). This is a database program that allows the user to filter necessary variables. One of the advantages of this program was that it allowed for the user to segment the entire State Highway System by section in equal lengths of similar geometric and traffic characteristic. The study area includes all of the counties in Florida that encompass the rural south region. There are several segment lengths are available (one mile, one and half and two). However, this paper consider only one and half segment length to accomplish the objective of the paper. The choice of one and half mile segment give us the excessive segments with zero ROR crashes. There are several variables to be considered, however, to compare the models, we have chosen only endmilepost (end mile post), lanewidth (width of the travel lane measured in feet), medwidth (the width of the median measured in feet), nolanes (no of lanes), pavecond (rating of the pavement condition ranging from 1, very poor to 5, very good), surwidth (the width of the entire pave roadway surface measured in feet), adt (the average annual daily traffic is total traffic volume on a roadway segment for one year divide by 365 days. This is measured vehicles per day). These variables include both roadway related variables and traffic related variables. For detailed about

the data and their properties we refer Gonzalez (2004). After fitting several models, we found that these variables are statistically significant to predict the number of crashes. We have created different data sets by deleting zero ROR crashes from 392 to 250. The summary statistics of the four set of data (number of crashes) are given in table 3.1. From the summary table it appears that all data sets are overdispersed. Thus one might expect that both NB and ZINB would possibly be better models to predict the ROR crashes.

Table 3.1: Summary Statistics

	Data 250	Data 275	Data 325	Data 392
% of zero	10.0	18.2	31.0	43.0
Mean	2.33	2.12	1.79	1.48
Variance	6.66	6.50	6.08	5.49
Skewness	3.03	3.04	3.14	3.35
Kurtosis	16.28	16.55	17.62	19.58

3.2 Model Fitting

There are four sets of data and we have fitted 4 different models for each data set. To save the space of the paper, the STATA output have not provided here, however, they are available from the author upon request. The total 16 possible models and the summary of statistical analysis have been given in Table 3.2.

3.2.1 Model fitting to Data Set 250

There are 10% zero ROR crashes in this data set. From Table 3.2, we observed that all models (PO, ZIP, NB and ZINB) are approximately and equivalently significant. In PO regression model, there are five explanatory variables (lanewidth, nolanes, pavecond, surwidth and adt) which have significant effect on the ROR crashes. For ZIP model, four explanatory variables (maxspeed, pavecond, surwidth and adt) are statistically significant for PO part and three explanatory variables (lanewidth, pavecond and adt) are statistically significant for logit part. Four explanatory variables (nolanes, pavecond, surwidth and adt) are statistically significant for the NB regression model. For ZINB model, four explanatory variables (nolanes, pavecond, surwidth and adt) are statistically significant for NB part and three explanatory variables (nolanes, pavecond and adt) are statistically significant for logit part. Vuong statistics suggested for traditional models instead of zero inflated models. The over dispersion parameters are statistically significant which indicated that over dispersion in the data. The deviance and AIC supported for both NB and ZINB models to fit the ROR crash data. The differences of observed and the mean predicted proportions for number of crashes is presented in Figure 3.1. This figure supported for both PO and NB models. Thus when the data have smaller amount of zeros, it make more sense to use PO or NB regression models. However, since the data is over dispersed, we have suggested for NB regression model.

Table 3.2: Summary of the fitted Models

Data	Characteristics	Model			
		Poisson	ZIP	NB	ZINB
250	Log-Likelihood	-443.87	-447.00	-493.55	-435.90
	Deviance	887.35	894.0	879.09	871.79
	DF	244	241	244	240
	ML R^2	0.56	0.57	0.36	0.36
	AIC	899.74	912.00	891.09	891.79
	Over Dispersion			$\hat{\alpha}$ =0.128(0.007)	α =0.114(0.010)
	Vuong Statistics		z =0.47(0.32)		z =1.31(0.10)
275	Log-Likelihood	-472.17	-470.43	-463.67	-457.83
	Deviance	944.24	940.86	927.33	915.66
	DF	270	267	269	264
	ML R^2	0.62	0.61	0.30	0.42
	AIC	954.34	956.86	939.33	937.66
	Over Dispersion			α =0.165(0.002)	$\hat{\alpha}$ =0.136=(0.006)
	Vuong Statistics		z =1.57(0.06)		z =1.65(0.05)
325	Log-Likelihood	-546.66	-534.98	-526.53	-526.65
	Deviance	1093.33	1069.96	1053.06	1053.31
	DF	319	313	319	318
	ML R^2	0.62	0.57	0.33	0.33
	AIC	1105.33	1093.96	1065.06	1067.31
	Over Dispersion			$\hat{\alpha}$ =0.338(0.000)	$\hat{\alpha}$ =0.329(0.000)
	Vuong Statistics		z =1.72(0.04)		z =0.23(0.41)
392	Log-Likelihood	-627.98	-597.11	-588.07	-581.35
	Deviance	1255.97	1194.23	1176.14	1162.70
	DF	385	382	384	383
	ML R^2	0.61	0.51	0.28	0.31
	AIC	1269.97	1214.23	1192.14	1180.70
	Over Dispersion			$\hat{\alpha}$ =0.587(0.000)	$\hat{\alpha}$ =0.405(0.000)
	Vuong Statistics		z =3.24(0.00)		z =2.79(0.003)

NB: The P-values of the tests are presented within parenthesis

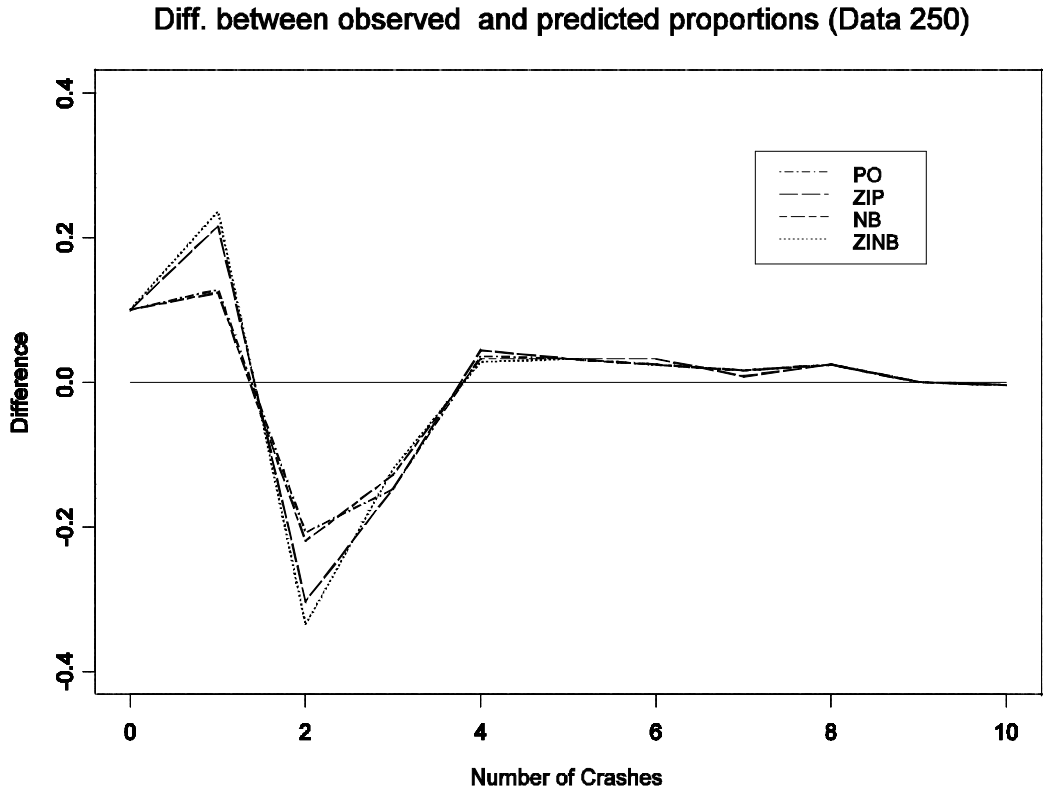

Figure 3.1: Difference between observed and predicted proportions for number of crashes

3.2.2 Model fitting to Data Set 275

There are 18% zero ROR crashes in this data set. In PO regression model, there are four explanatory variables (nolanes, pavecond, surwidth and adt) which have significant effect on the ROR crashes. For ZIP model, four explanatory variables (maxspeed, pavecond, surwidth and adt) are statistically significant for PO part and two explanatory variables (lanewidth, pavecond) are statistically significant for logit part. Four explanatory variables (nolanes, pavecond, surwidth and adt) are statistically significant for the NB regression model. For ZINB model, four explanatory variables (nolanes, pavecond, surwidth and adt) are statistically significant for NB part and four explanatory variables (nolanes, pavecond, surwidth and adt) are statistically significant for logit part. Both Vuong statistic and the scale parameter of 0.129 (significant) suggested for ZINB model. The deviance and AIC supported for ZINB model only to fit the ROR crash data. The differences of observed and the mean predicted proportions for number of crashes is presented in Figure 3.2. This figure is also supported for ZINB model.

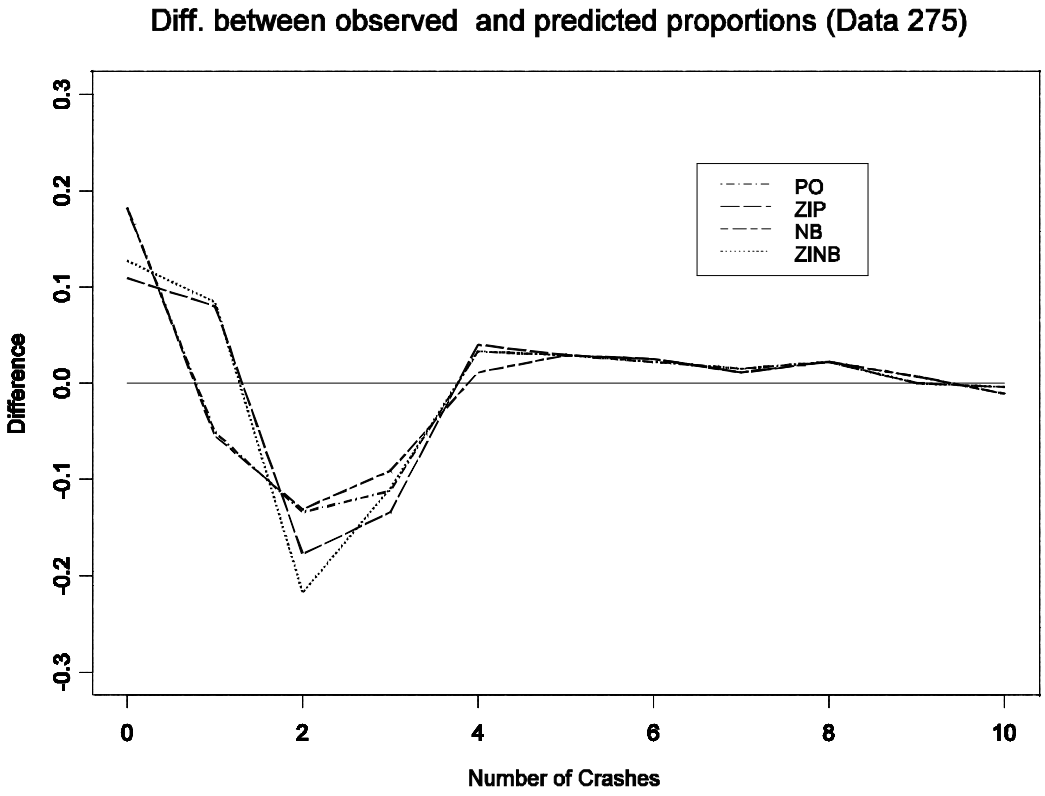

Figure 3.2: Difference between observed and predicted proportions for number of crashes

3.2.3 Model fitting to Data Set 325

There are 31% zero ROR crashes in this data set. In PO regression model, there are five explanatory variables (lanewidth, nolanes, pavecond, surwidth and adt) which have significant effect on the ROR crashes. For ZIP model, seven explanatory variables (endmilepost, lanewidth, medwidth, nolanes, pavecond, surwidth and adt) are statistically significant for PO part and three explanatory variables (maxspeed, medwidth, pavecond) are statistically significant for logit part. Four explanatory variables (nolanes, pavecond, surwidth and adt) are statistically significant for the NB regression model. For ZINB model, three explanatory variables (maxspeed, pavecond, and adt) are statistically significant for NB part and only pavecond is statistically significant for logit part. Vuong statistic suggested for ZIP and NB models. However, the scale parameter of 0.338 is statistically significant and indicating substantial overdispersion in the zero-counts. Thus we might select the NB model. The deviance and AIC also supported for NB model to fit the ROR crash data. The differences of observed and the mean predicted proportions for number of crashes is presented in Figure 3.3. This figure also supported for NB model.

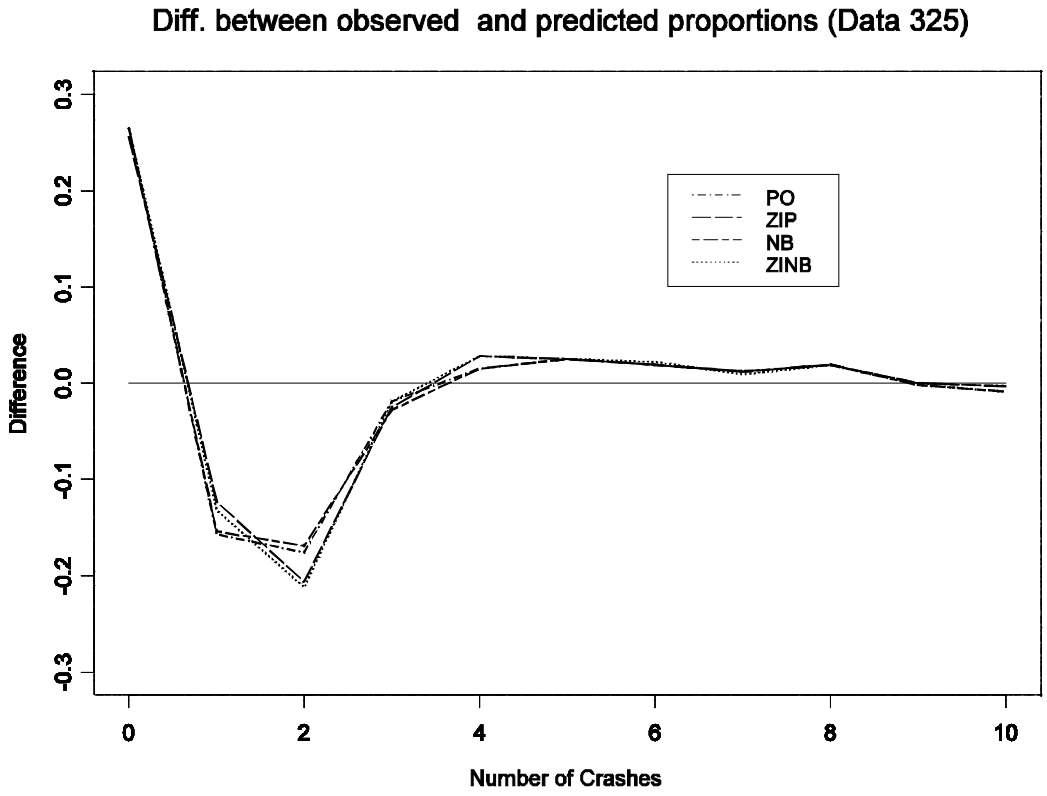

Figure 3.3: Difference between observed and predicted proportions for number of crashes

3.2.4 Model fitting to Data Set 392

There are 43% zero ROR crashes in this data set. In PO regression model, there are six explanatory variables (endmilepost, medwidth, nolanes, pavecond, surwidth and adt) which have significant effect on the ROR crashes. For ZIP model, six explanatory variables (lanewidth, medwidth, nolanes, pavecond, surwidth and adt) are statistically significant for PO part and two explanatory variables (lanewidth and adt) are statistically significant for logit part. Six explanatory variables (endmilepost, medwidth, nolanes, pavecond, surwidth and adt) are statistically significant for the NB regression model. For ZINB model, three explanatory variables (pavecond, surwidth, and adt) are statistically significant for NB part and three variables (nolanes, surwidth, adt) are statistically significant for logit part. Both Vuong statistics have suggested for ZIP and ZINB models. However, the scale parameters of 0.587 and 0.405 are statistically significant and indicating substantial overdispersion in the zero-counts. Thus we might select the ZINB model. The deviance and AIC are also supported for ZINB model only to fit the ROR crash data. The differences of observed and the mean predicted proportions for number of crashes is presented in Figure 3.4. This

figure also indicated to select ZINB model. For this data one might consider ZINB model than the ZIP model in the sense of better zero prediction.

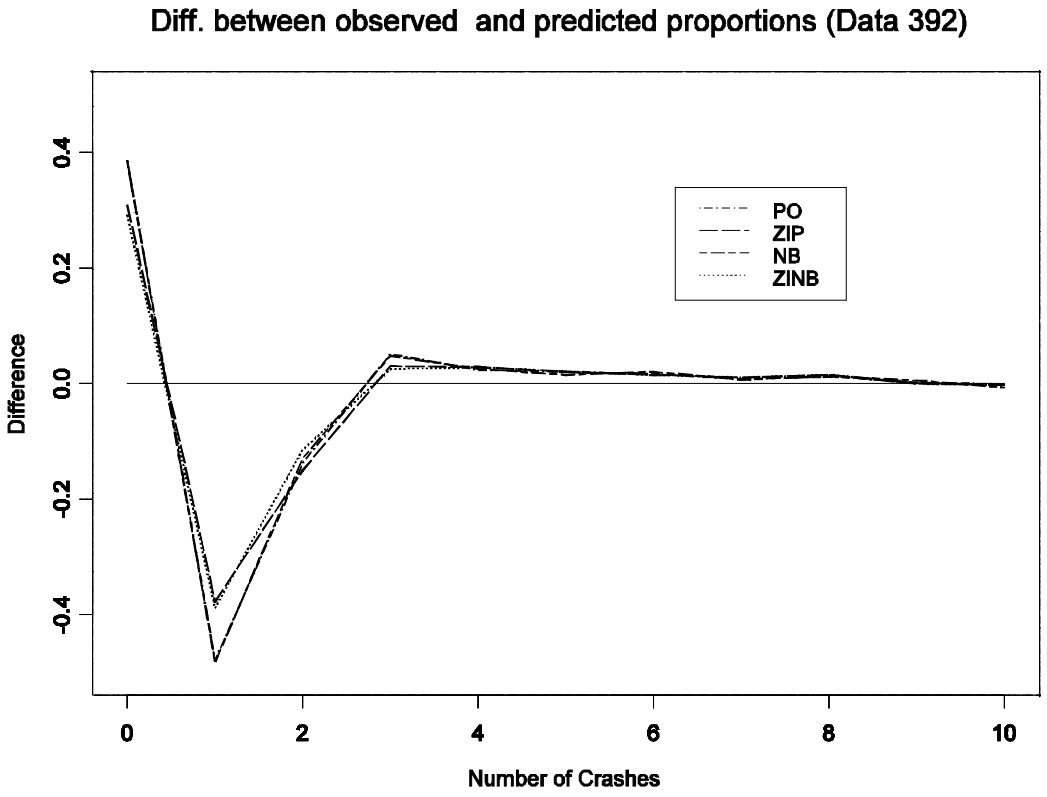

Figure 3.4: Difference between observed and predicted proportions for number of crashes

4 Concluding Remarks

This paper provides both methodological and empirical analysis of ROR (run-off-road) crash data which collected on arterial roads in the rural area of south region of the Florida State. We have fitted several popular regression models, Poisson (PO), Negative Binomial (NB), Zero-Inflated Poisson (ZIP) and Zero-Inflated Negative Binomial (ZINB) to predict the ROR crashes. We consider moderate (10%) to high (43%) number of zeros in the models. A total of Sixteen different statistical models were fitted in this paper. All fitted models include significant explanatory variables. Based on deviance and AIC, it appeared that both NB and ZINB models performed better than PO and ZIP models respectively. The empirical study of this paper revealed that if the over-dispersion and zero-inflation of ROR crashes is found to be moderate to high, both NB and ZINB models are potential alternatives to PO and ZIP regression models. Poisson regression models serve well under nearly homogeneous condition,

while NB models serve better while data are over dispersed. However, for an excess number of zero counts, one might consider both ZIP and ZINB regression models. It is important to note that the same set of variables or same model may not necessarily be statistically significant for another data sets with the same set of variables. For an applied research, it is advisable to fit data and then conclude based on the findings of the analysis. For a definite statement about the best fitted model, one needs more data and more analysis. Hopefully the present analysis can provides some insights to model other kind of count data, for instance, ER visit, clinical epidemiology, biometrical and environmental data.

Acknowledgments

This paper has been written based on the results of the summer 2005 project. The author is grateful to the Dean of the College of Arts and Sciences of Florida International University for providing him with summer research funding 2005. He is also thankful to the Lehman Center at FIU for using their data. The help of Javier Gonzalez with the STATA program and providing data is highly appreciated.

References

1. Agresti, A. (2002). Categorical Data Analysis. New York, Wiley.
2. Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle, In the second international symposium on information theory, edited by B. V. Petrov and B. F. Csaki, Academical Kiado.
3. Berhanu, G. (2004). Models relating traffic safety with road environment and traffic flows on arterial roads in Addis Abba. Accidents Analysis and Prevention, 36, 697- 704.
4. Byers, A. L., Allore, H., Gill, T. M. and Peduzzi, P. (2003). Application of negative binomial modeling for discrete outcomes: A case study in aging research. Journal of Clinical Epidemiology, 56, 559-564.
5. Cameron, C. and Trivedi, P. (1998). Regression analysis of count data. New York: Cambridge University Press.
6. Carson, J. and Mannering, F. (2001). The effect of ice warning signs on ice-accident frequencies and severities. Accidents Analysis and Prevention, 33, 99-109.
7. Cheung, Y. B. (2002). Zero-inflated models for regression analysis of a count data: a study of growth and development. Statistics in Medicine, 21, 1481-1469.
8. Chin, H. C. and Quddus, M. A. (2003). Modeling count data with excess zeros: An empirical application to traffic accidents. Sociological Methods and Research, 32, 90-116.
9. Dean, C. B. (1992). Testing for overdispersion in Poisson and binomial regression models. Journal of the American Statistical Association, 87, 451-457.
10. Gonzalez, J. S., Kibria, B. M. G. and Gan, A. (2005). Some discrete regression models to determine the effects of roadway geometric features on run-off-road crashes. Submitted to Accidents Analysis and Prevention.
11. Gupta, P. L., Gupta, R. C. and Tripathi, R. C. (1996). Analysis of zero-adjusted count data. Computational Statistics and Data Analysis, 23, 207-218.

12. Hauer, E. (2001). Overdispersion in modeling accidents on road sections and in empirical Bayes estimation. Accidents Analysis and Prevention, 33, 799-808.
13. Javier S. Gonzalez (2004): Effects of roadway geometric features on run-off-road crashes on the Florida State highway system. Unpublished Ph. D. thesis, Department of Civil and Environmental Engineering, Florida International University, Miami, USA.
14. Karlaftis, M. G. and Tarko, A. P. (1998). Heterogenity considerations in accident modeling. Accidents Analysis and Prevention, 30, 425-433.
15. Lambert, D. (1992). Zero-Inflated Poisson regression, with an application to defects in manufacturing. Technometrics, 34, 1-14.
16. Lawless, J. F. (1987). Negative binomial and mixed Poisson regression. Canadian Journal of Statistics, 15, 209-225.
17. Lee, A. A., Wang, K. and Yau, K. K. W. (2001). Analysis of zero-inflated Poisson data incorporating extent of exposure. Biometrical Journal, 43, 8, 963-975.
18. Lee, A. A., Xiang, L. and Fung, W. K. (2004). Sensitivity of score tests for zero-inflation in count data. Statistics in Medicine, 23, 2757-2769.
19. Lee, A. H., Stevenson, M. R., Wang, K. and Yau K. K. W. (2002). Modeling young driver motor vehicle crashes: data with extra zeros. Accidents Analysis and Prevention, 34, 515-521.
20. Lee, J. and Mannering, F. (2002). Impact of roadside features on the frequency and severity of run-off-roadway accidents: an empirical analysis. Accidents Analysis and Prevention, 34, 149-161.
21. Long, J. S. (1997). Regression Models for Categorical and Limited Dependent Variables. Thousand Oaks, CA: Sage.
22. 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, 37(1), 35-46.
23. McCullagh, P. and Nelder, J. A. (1987). Generalized linear Models, Chapman and Hall, London.
24. Miaou, S. P. (1994). The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. Accidents Analysis and Prevention, 26, 471 - 482.
25. Milton, J. and Mannering (1998). The relationship among highway geometries, traffic-related elements and motor -vehicle accident frequencies. Transportations, 25, 395-413.
26. Mullahy, J. (1986). Specifications and testing of some modified count data models. J. of Econometrics, 33 (3), 341-365.
27. Poch, M and Mannering, F. L. (1996). Negative binomial analysis of intersection accident frequency. Journal of Transportation Engineering, 122, 105-113.
28. Shankar, V., Mannering, F., and Barfield, W. (1995). Effect of roadway geometric and environmental factors on rural freeway accidents frequencies. Accidents Analysis and Prevention, 27, 371-389.
29. Shankar, V., Milton, J. and Mannering, F. (1997). Modeling accident frequencies as zero-altered probability processes: An empirical inquiry. Accidents Analysis and Prevention, 29, 829-837.
30. Stata Corporation (1999). STATA Release 9.0. College Station, Texas.
31. Van Den Broek, J. (1995). A score test for zero-inflation in a Poisson distribution. Biometrics, 51, 738-743.

32. Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica, 57, 307-334.
33. Warton, D. I. (2005). Many zeros does not mean zero inflation: comparing the goodness-of-fit of parametric models to multivariate abundance data. Environmetrics.
34. Wood, G. R. (2002). Generalized linear accident models and goodness of fit testing. Accident Analysis & Prevention, 34, 417-427.
35. Yau, K. K. W, Wang, K. and Lee, A. H. (2003). Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. Biometrical Journal, 45, 437-452.
36. Yau, K. K. W., Lee, A. H. and Carrivick, P. J. W. (2004). Modeling zero-inflated count series with application to occupational health. Computer Methods and Programs in Biomedicine, 74, 47-52.