

Modeling Zero – Inflated Regression of Road Accidents at Johor Federal Road F001

Joewono Prasetijo^{1,a} and W. Zahidah Musa¹

¹*Faculty of Civil and Environmental Engineering, University Tun Hussein Onn Malaysia, 86400 Batu Pahat, Johor, Malaysia*

Abstract. This study focused on the Poisson regression with excess zero outcomes on the response variable. A generalized linear modelling technique such as Poisson regression model and Negative Binomial model was found to be insignificant in explaining and handle over dispersion which due to high amount of zeros thus Zero Inflated model was introduced to overcome the problem. The application work on the number of road accidents on F001 Jalan Jb – Air Hitam. Data on road accident were collected for five-year period from 2010 through 2014. The result from analysis show that ZINB model performed best, in terms of the comparative criteria based on the P value less than 0.05.

1 Introduction

Malaysia has experienced a remarkable period of economic expansion and growth in population, economy, industrialisation and motorisation. Since then, the increase in population and motorisation led to a consequent increase in the number of road traffic accidents [1]. According to the survey obtained by iRAP (International Road Assessment Programme) on Malaysian road networks in 2007, the casualties for traffic accident include over 6,000 fatalities in a year. Over 60% of these fatal accidents happen on rural roads. Nearly half of all fatalities occur on federal roads and over a quarter on state roads. It was estimated about 2% of Gross Domestic Product (GDP), approximately RM 9 billion, to Malaysia's economy. In addition, traffic accident has become a major social-economic problem in Malaysia, and it is considered as main cause of death. Due to road accidents, result in annual economic cost of 9 billion ringgit to the nation economy. Numbers of fatalities reach over 6,000 people per year in road accidents nationwide. The amount has been increasing since 2003 and reached 6,872 for year 2010 while 21,397 suffer injuries. Traditionally, the simple risk measures such as observed accident counts, rates, costs, and other similar methods were used to target locations with a number of accidents exceeding a chosen threshold as hot spots [2]. Unfortunately, these measures have several shortcomings as discussed in a number of studies [2]. Alternatively, statistical modelling approach relate to the traffic safety are used for identification of hazardous locations. This approach is extensively used in traffic safety studies for identifying major contributing factors to crashes and injuries, establishing proper relationships between crashes and explanatory variables, and predicting crash frequency and injury severity. A generalized Linear Modelling technique such as Poisson

^a Corresponding author : joewono@uthm.edu.my

distribution was used and was found to be insignificant in explaining the effect of explanatory variables related to intersection geometry with accident count.

Generally Poisson regression analysis is the primary model which is use to find out the relation between independent variables and independent variable. The dependent variable must distribute as Poisson distribution. One of the assumptions of Poisson regression is the mean and variance must be equal, but actually most of the data will have the larger variance or over dispersion. The negative binomial regression is a method which can solve the over dispersion problem [3]. The problem outcome from the Poisson regression is when there are many zeros for dependent variable, the accuracy of the Poisson regression will be reduce, Zero-inflated Poisson regression (ZIP) is introduce to solve this problem [3]. The dependent variable in this study was accident data which occur to be excess zero. Thus, the zeros dependent variable depends on the other independent variables. Zeros dependent variable comes from the zero data that produce from data crashes. The motivation for the ZIP model is the model are comparable to those from fitting a NB regression model for general over-dispersion [4], the potential under-reporting of vehicle accidents especially minor injury and property damage accidents [5], and [6] proposed to identifying effective factors on crash occurrence probability and prove the efficiency of Zero-inflated regression models against traditional Poisson and Negative binomials regression models. At the end of likelihood ratio test, it seems that the ZINBDR model gave better results than NBDR model [7].

2 Methodology

2.1 Poisson regression model, negative binomial model and zero inflation model

The methodology, on which this research proposal is based, firstly treats the mean accident frequency at a site as an unknown quantity. Due to a random, discrete, and positive nature of crash data, count data modelling techniques such as Poisson and negative binomial (NB) regression models are used in such cases. The Poisson regression model has been traditionally considered as the starting point in modelling crash data, with assumption of the mean of crash counts being equal to its variance (that is, equal-dispersion) [8]. In the Poisson regression model, the probability of n_i vehicle crashes occurring at a given road *section* i , $Pr(n_i|\mu_i)$, can be estimated by

$$Pr = \binom{n_i}{\mu_i} = \frac{\exp(-\mu_i) \mu_i^{n_i}}{n_i} \tag{1}$$

$$\mu_i = \exp\left(\sum_{j=1}^M \beta_j X_{ij}\right) \tag{2}$$

where μ_i is the expected number of vehicle crashes at the given section, X_j is a vector of covariates (road geometric, environmental and traffic information) and β_j is a vector of estimable regression coefficients.

Unfortunately, much of the cases, the crash data produce the variance which is greater than the mean, well known as over-dispersion. The over-dispersion is a result of extra variation in crash means across sites which can be caused by various factors such as model misspecification, omission of important covariates, and excess zero counts. In this case, applying a Poisson regression model for vehicles crash data would result in underestimation of standard error of the regression parameters. Therefore, the negative binomial model will be introduced with

$$\mu_i = \exp(\beta X_i + \varepsilon_i) \tag{3}$$

where the $exp(\varepsilon_i)$ is gamma distributed with mean 1 and variance α (the *over dispersion* parameter). By taking the unobserved heterogeneity of the Poisson mean to be counted, the following variance differ from the mean is introduced.

$$Var[y_i] = E[y_i][1 + \alpha E[y_i]] = E[y_i] + \alpha E[y_i]^2 \tag{4}$$

and the maximum likelihood function of the binomial regression model is shown as:

$$L(\mu_i) = \prod_i^N \frac{\Gamma(\theta + n_i)}{\Gamma(\theta)n_i} \left[\frac{\theta}{\theta + n_i} \right]^\theta \left[\frac{\mu_i}{\theta + \mu_i} \right]^{n_i} \tag{5}$$

where θ is inverse dispersion parameter ($1/\alpha$), $\Gamma(\cdot)$ is a value of gamma distribution, and N is the total number of road sections.

In some cases, excess zeros in crash data exist and considered as a result of over dispersion. In this case, the NB model cannot be used to handle the over dispersion which is due to high amount of zeros. To do this, zero-inflation (ZI) models including Zero Inflated Poisson (ZIP) and Zero Inflated Negative Binomial (ZINB) models can be alternatively used. Both the ZIP and ZINB models assume that all zeros count come from two different processes: the process generating excess zero count derived from a binary model, and the process generating non-negative counts for vehicle crashes including zero values. As description of ZIP model, let P_i as the probability of being excess zero for section i , and $(1 - P_i)$ be the probability of crash counts derived from the Poisson distribution as it is given as

$$P(Y = y_i) = \begin{cases} P_i + (1 - P_i)\theta^{\mu_i} & y_i = 0 \\ (1 - P_i) \frac{\theta^{\mu_i} \mu_i^{y_i}}{y_i!} & y_i > 0 \end{cases} \tag{6}$$

where y is the number of vehicle crashes for section i and μ_i is the expected crash frequency in section i as a function of road section covariates, $\mu_i = exp(X_i\beta)$. The probability of being in the zero-crash-state, P_i , is often fitted using logistic regression model, as follows:

$$logit(P_i) = \ln\left(\frac{P_i}{1 - P_i}\right) = \gamma_0 + \gamma_1 Z_i + \dots + \gamma_N Z_N \tag{7}$$

where $Z = (Z_1, Z_2, \dots, Z_N)$ is a function of some explanatory variables and $\gamma = (\gamma_1, \gamma_2, \dots, \gamma_N)$ is the estimable coefficients.

Similar to ZIP model, the probability density function for the ZINB given mean μ_i , and dispersion parameter, α , is given by:

$$P(Y = y_i) = \begin{cases} P_i + (1 - P_i) \frac{1}{(1 + \alpha\mu_i)^{1/\alpha}} & y_i = 0 \\ (1 - p) \frac{\Gamma\left(y_i + \frac{1}{\alpha}\right)(\alpha\mu_i)^{y_i}}{\Gamma(y_i + 1)\Gamma\left(\frac{1}{\alpha}\right)(1 + \alpha\mu_i)^{y_i + \frac{1}{\alpha}}} & y_i > 0 \end{cases} \tag{8}$$

The maximum likelihood method is used to estimate the parameters of the ZI models. In general, the probability of observing zero counts in ZI models is the sum of the probability of observing an excess zero in the first process and probability of observing a zero in the second process.

2.2 Study area and data collection

Study area of this research work was selected and considered from the south of Peninsular Malaysia after consulting with Malaysia road authorities such as Public Work Dept. (JKR), Malaysian Institute of Road Safety Research (MIROS), and Malaysia Royal Police. This study focused on F001 roadway which is Jalan Jb-Air Hitam. They will include those routes where high traffic volumes are present, large numbers of crashes occur and where a range of different road conditions are evident. The R software was used to analyze the data.

Table 1. A summary of some predictor variables.

Variable	Name	Variable	Name
Dependent	Accident Crashes	Independent	Median width
Independent	AADT	Independent	No. of lanes
Independent	Traffic flow	Independent	Land use
Independent	Lane width	Independent	Access density
Independent	Shoulder width	Independent	Horizontal alignment
Independent	Roadside hazard rating	Independent	Vertical alignment

3 Results and Discussion

The empirical distribution of crash counts is plotted in Figure 1. The figures indicate that a peak of extra zeros which representing crash free exists. It confirms the presence of excess zeros in road accident crashes. This pattern is typical for most type of vehicle crash data. Moreover, the ratio of the variance to mean for fatal is 4.2 indicating the crash data is over-dispersed implying the poisson model do not adequately fit the overdispersed crash data. Then as alternative we use the better regression model which is Zero-inflated poisson and Zero-inflated negative binomial regression model.

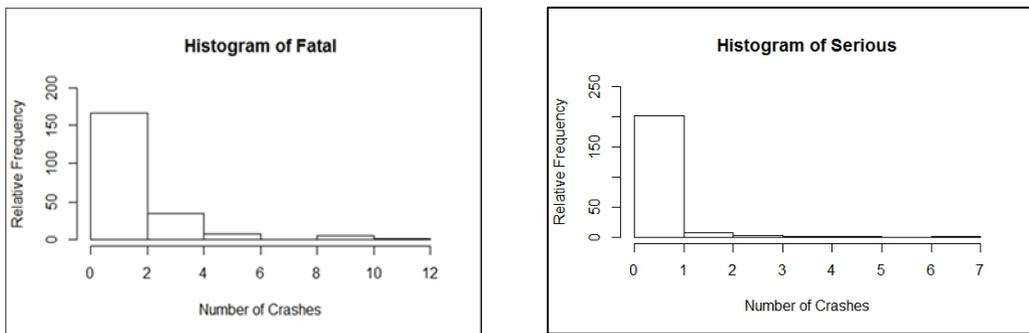


Figure 1. Histogram of observed accident crashes for F001 Jln Jb-Air Hitam during 2010-2014.

From Table 3, it shows the Poisson model coefficient. The less AIC (Akaike Information Criteria) the greater the model be more fitted. Hence from Table 3, AIC values for Poisson was 10154 which were high value compare to the AIC value for ZINB (727.2). The Poisson model assumes that the

variance of the dependent variable is equal to mean but in many applications, count data were found to show the over-dispersion [9]. The Poisson regression model is fit but when consider the raw data, there are many zeros which the ZIP will be approach.

Table 2. Result for showing existence of over dispersion in crash data.

Variables	Standard deviation [SD]	Variance	Mean	Over dispersion=[var>mean]
Fatal	2.057139	4.23181917	1.526	Over dispersion
Serious	0.8486408	0.72019126	0.3116	Over dispersion
Minor	3.475564	12.079547924	1.558	Over dispersion
Wreckage	125.4524	15738.29633704	70.24	Over dispersion

Table 3. Poisson model coefficient.

Poisson				
Pearson residuals:				
Min				
1Q			-13.2799	
Median			-5.8067	
3Q			-0.6374	
Max			5.1534	
AIC			9.4189	
			10154	
	Estimate	Std. Error	Z value	Pr(> z)
Intercept	4.904e+00	3.244e-02	151.160	< 2e-16
Fatal	-3.382e-02	3.685e-03	-9.178	< 2e-16
Serious	-7.509e-03	8.825e-03	-0.851	0.394826
Speed	-3.742e-03	5.447e-04	-6.870	6.41e-12
Volume	2.015e-04	5.288e-05	3.810	0.000139

Table 4. Zero inflated Poisson model coefficients.

	ZIP				ZINB			
Pearson residuals:								
Min								
1Q								
Median								
3Q								
Max								
Log-like hood:								
	Estimate	Std. Error	Z value	Pr(> z)	Estimate	Std. Error	Z value	Pr(> z)
Intercept	0.068389	1.988734	0.034	0.973	6.68847	0.00	0.00	0.00
Speed	-	0.017027	-	0.144	-0.03952	0.00	0.00	0.00
	0.024876		1.461					
Volume	0.001118	0.001710	0.654	0.513	0.01719	0.04404	0.39	0.696
SW	-	0.655635	-	0.561	-11.47283	0.00	0.00	0.00
	0.381369		0.582					
LW	0.271939	0.5223228	0.521	0.603	-0.34308	0.00	0.00	0.00

The absolute value of V is less than a threshold value such as 1.96 for 0.95 confidence level. From Table 5 the result for vuong was -1.93 which indicated that model 2 (ZINB) was preferred. Vuong test can indicate in any way if zero-inflated model is an improvement over standard a Negative Binomial regression [10]. It can compare between two model regressions. The vuong tests compare the zero-inflated model with an ordinary Negative Binomial model. In this analysis we can see that our test

statistic is significant, indicating that zero-inflated model is superior to the standard Negative Binomial model. The less AIC (Akaike Information Criteria) the greater the model be more fitted. Hence in this analysis, Zero inflated Negative binomial was the most fitted model between the others due to the value of the AIC was 727.2. While for Zero inflated Poisson was 773.2, Negative Binomial was 734.16 and lastly Poisson was 10154. Good P-value for each independent variable was when P value was less than 0.05. The result from Table 4 show that most of the p value for ZINB was less than 0.05 which is 0.00.

Table 5. Vuong non-nested hypothesis Test-statistic.

	Vuong z-statistic	H_A	p-value
Raw	0.8474188	model1 > model2	0.198381
AIC-corrected	-1.9370687	model2 > model1	0.026368
BIC-corrected	-6.6298185	model2 > model1	1.6805e-11

4 Conclusion

This study is focusing on the Poisson regression and negative binomial model with many zero outcomes on the response variable. The zero-inflated Poisson regression model is more effective for many zero outcomes than Poisson regression. While, the zero-inflated negative binomial regression model is more effective for many zero outcomes than negative binomial regression. The application is work on the number of road accidents on federal road in the Johor, south of Malaysia. The zero-inflated negative binomial regression was used to find the relationship between dependent and independent variables when there are many zeros value in the dependent variable, where the relationship is the mixture between Negative binomial model and logistic model. This paper aimed to focus on developing a model which can properly handle stochastic nature of crash event. It also implies to find out the most relevant contributory factors at the hazardous sections and type of vehicle crashes. Lastly, this research can develop a statistical approach to estimate the safety performance of a road network in Johor federal roads. There are many further work to be done in this research (as the study is still ongoing) such as comparing the ZIP and ZINB model in terms of nested models, full probit and logit models, examine the coefficients of variables in different models and examine the diagnostic statistics of models to make it best performed to incorporate the models into traffic management systems.

References

- [1] U. Radin, Updates of road safety status in Malaysia, *IATSS Research*, **29**, 78-80, (2005).
- [2] K. Geurts and G. Wets, Black spot analysis method: literature review, **7**, 7-13, (2003).
- [3] C. Pudprommarat, M. Khamkong and P. Bookkamana, Zero-inflated Poisson regression in road accidents on major road in the north of Thailand, *IRCMSA Proc.*, 323-330, (2005).
- [4] A.H. Lee, M.R. Stevenson, K.Wang and K.K.W.Yau, Modeling young driver motor vehicle crashes: data with extra zeros, *Accident Analysis and Prevention*, **34**, 515-521, (2002).
- [5] S.P. Miaou, The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions, *Accident Analysis and Prevention*, **26**, 471-482, (1994).
- [6] E. Ayati and E. Abbasi, Modeling accidents on mashad urban highways, *Open J. of Safety Science and Technology*, **4**, 22-35, (2014).
- [7] M. Asrul and N.N. Naing, Analysis death rate of age model with excess zeros using zero-inflated negative binomial and negative binomial death rate: mortality AIDS co-infection patients, Kelantan, Malaysia, *Procedia Economics and Finance*, **2**, 275-283, (2012).

- [8] M.H. Pour, J. Prasetijo, A.S. Yahaya and S.M.R. Ghadiri, Modeling vehicle-pedestrian crashes with excess zero along Malaysia federal roads, *Procedia Social and Behavioral Sciences*, 1218-1227, (2012).
- [9] J. Garber, Stochastic Models Relating Crash Probabilities with geometric and corresponding traffic characteristics data, Research Report No. UVACTS-5-15-74, (2001).
- [10] Q.H. Vuong, Likelihood ratio test for model selection and non-nested hypotheses, *Econometrica*, **57**, 307-333, (1989).