Zero-Inflated Regression Models with an Application to Vehicle Theft Count Data

Malina Zulkifli[1], Noriszura Ismail[2] & Ahmad Mahir Razali[3]

School of Quantitative Sciences, College of Arts and Science, Universiti Utara Malaysia Kedah, Malaysia[1]
School of Mathematics, Faculty of Technology and Science, Universiti Kebangsaan Malaysia Bangi, Malaysia[2]
School of Mathematics, Faculty of Technology and Science, Universiti Kebangsaan Malaysia Bangi, Malaysia[3]

malina@uum.edu.my[1], ni@ukm.my[2]
& mahir@ukm.my[3]

Abstract

Poisson regression model has been widely used for modeling claim count data in actuarial and insurance literatures. However, in several cases, claim count data often have excessive number of zeros than are expected in the Poisson model. In that case the Poisson regression may underestimate the standard errors and giving misleading inference about the regression parameters. This paper aims to apply the zero-inflated regression models on vehicle theft crime data. These zero-inflation phenomenon is a very specific type of overdispersion and zero-inflated Poisson (ZIP) regression model has been suggested for handling zero-inflated data. If the crime count data continue to suggest additional overdispersion, the alternative models the zero-inflated negative binomial-1 (ZINB-1) and the zero-inflated negative binomial-2 (ZINB-2) will be fitted on the private car theft claim count data. In addition, two different forms of link function will be used in the fitting procedure of the zero-inflated regression models, producing different estimates for each model. The results of this study indicate that the ZINB-2 models is better compared to ZIP regression model for handling zero-inflated and additional overdispersed crime count data.

Keywords: vehicle theft, count data, crime, zero-inflated poisson, zero-inflated negative binomial

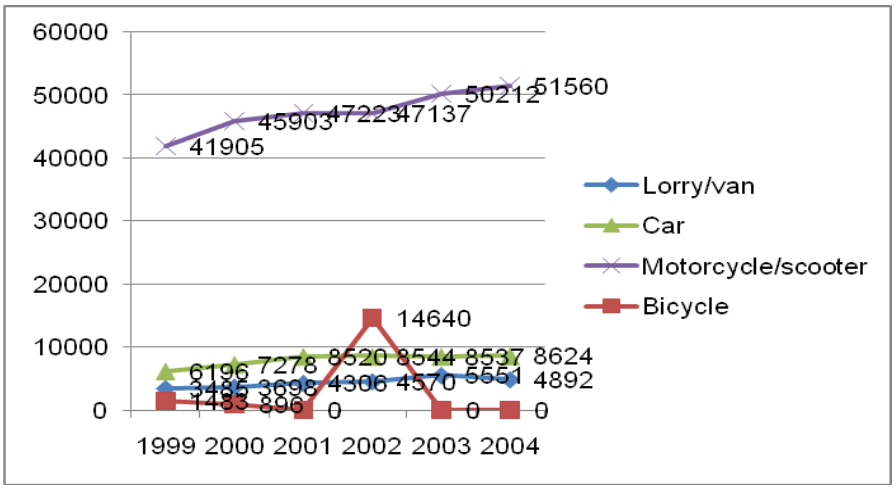
1. Introduction

The level of crime in Malaysia is a serious concern in recent years. Varieties of objectives, targets and strategy conducted by the government with the help of various agencies as well as NGOs, however, have not been able to reduce people's concerns

about crime. Even now people are more concerned with the increase in crime. Crime, whether the violence crime or property crime is very prevalent and can occur anywhere and at any time irrespective of rural or urban areas, in public or private places such as at home, either in the daytime or night, morning or afternoon. Malaysia is as an International Police Community use phrase "Index Crime" to measure crime. Index Crime is defined as:"*Crime that is reported with sufficient regularity and with sufficient significance to be meaningful as an index to the crime situation"* (Sidhu 2005).

Crime rate in Malaysia measured to every one thousand people. According to Malaysian Quality of Life (2004) crime rate increase from 3.8 cases in year 1990 to 6.2 cases in year 2002 which more than four-fifths criminal cases involving property crime. Property crimes are criminal acts related to property such as burglary, robbery, stolen vehicles etc. It is worth to note that the motor vehicle theft makes up 49% of the total index crime in Malaysia (Sidhu 2005). Common places of vehicle theft are in the public parking area, shopping complex and hotels, residential areas and on roadsides. Figure 1 shows vehicle theft crime which reported to police according to type of vehicle.

Figure 1. Vehicle theft crime reported to police according to type of vehicle



The above figure shows that vehicle theft crimes increased for all types of vehicles in the past six years except for bikes. This is because the case involving the theft of a bicycle under-reported due to consider that it is a vehicle that is cheap and does not need to insurance coverage.

In several cases, count data often have excessive number of zero outcomes than are expected in the Poisson. This paper aims to apply the zero-inflated regression models on vehicle theft crime data. The first model that has been suggested for handling these zero-inflated data is zero-inflated Poisson (ZIP). The use of ZIP becomes more widespread since the publication of Lambert (1992) who applied the ZIP on manufacturing defects data. Jansakul and Hinde (2002) has implemented a score test to compare Poisson regression model with ZIP by allowing the zero probability to depend on covariates, and its application have been shown in two sets of data, the apple shoot propagation data and the HIV data. If the count data continue to suggest overdispersion, then we can consider the zero-inflated negative binomial-1 (ZINB-1) and zero-inflated negative binomial-2 (ZINB-2) regression model as alternatives. The applications of ZINB-1 and ZINB-2 can be found in Ridout *et al.* (2001) in which these regression models were fitted to apple shoot propagation data and also provided score statistics for testing the ZIP against the ZINB-1 and the ZINB-2.

2. Materials and Methods

2.1. Data

In this study, vehicle theft crime data for private cars are compiled from ten insurance companies in Malaysia. The data, which was based on 1.2 million private car policies for a period of three years, was provided by the Insurance Services Malaysia (ISM). The exposure was expressed in terms of units of the car-year and incurred claims consist of claims that already paid as well as outstanding.

Table 1 shows the rating factor and class for the exposure and the claims incurred. By excluding zero exposures, we have a total of 1059 count data to be fitted to the ZIP, ZINB-1 and ZINB-2 regression model. The zero-inflated regression models are considered for this data because the data consists of 740 of zero claims, which contributes about 70% of the data.

Table 1. Rating factors and rating classes for Malaysia vehicle theft data

Rating factors	Rating classes
Coverage	Comprehensive
	Non-comprehensive
Vehicle year	0-1
	2-3
	4-5
	6-7
	8+
Vehicle c.c.	0-1000
	1001-1300
	1301-1500
	1501-1800
	1801+
Vehicle make	Local type 1
	Local type 2
	Foreign type 1
	Foreign type 2
	Foreign type 3

2.2 Zero-inflated Poisson (ZIP) Model

The ZIP regression model has been used by researchers for handling purely zero-inflated count data. The p.m.f. of the ZIP is given by,

$$\Pr(Y_i = y_i \mid \mu_i, \omega_i) = \begin{cases} \omega_i + (1-\omega_i)\exp(-\mu_i), & y_i = 0 \\ (1-\omega_i)\dfrac{\mu_i^{y_i}}{y_i!}\exp(-\mu_i), & y_i > 0 \end{cases} \quad (1)$$

where $0 \le \omega_i < 1$ and $\mu_i > 0$, with mean $E(Y_i) = (1-\omega_i)\mu_i$, and variance $Var(Y_i) = (1-\omega_i)\mu_i(1+\omega_i\mu_i)$.

Based on the p.m.f. shown in (1), the ZIP exhibits overdispersion when $\omega_i > 0$. The parameters, μ_i and ω_i, depend on the covariate vectors of $\mathbf{x}_i$ and $\mathbf{z}_i$, whereby the link functions can be written as,

$$\log(\mu_i) = \mathbf{x}_i^T\boldsymbol{\beta} \quad \text{and} \quad \log\left(\frac{\omega_i}{1-\omega_i}\right) = \mathbf{z}_i^T\boldsymbol{\gamma}. \tag{2}$$

If $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are estimated by the maximum likelihood method, the log likelihood for the ZIP regression model is,

$$\begin{aligned}\log L(\boldsymbol{\beta},\boldsymbol{\gamma}) = &\sum_{y_i=0}\log(\omega_i + (1-\omega_i)\exp(-\mu_i)) \\ &+ \sum_{y_{i>0}}\log(1-\omega_i) + y_i\log(\mu_i) - \log(y_i!) - \mu_i\end{aligned}, \tag{3}$$

Therefore, the maximum likelihood estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ can be obtained by maximizing $\log L(\boldsymbol{\beta},\boldsymbol{\gamma})$ with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$.

2.3 Zero-inflated Negative Binomial-1 (ZINB-1) Model

The ZINB-1 and the ZINB-2 regression models can be used to handle both zero-inflation and additional overdispersion in count data. The p.m.f. of the ZINB-1 regression model is (Ridout *et al.* 2001, Yang *et al.* 2009),

$$\Pr(Y_i = y_i \mid \mu_i, \omega_i, a) = \begin{cases} \omega_i + (1-\omega_i)(1+a)^{-\mu_i a^{-1}}, & y_i = 0 \\ (1-\omega_i)\dfrac{\Gamma(y_i + \mu_i a^{-1})}{y_i!\,\Gamma(\mu_i a^{-1})}(1+a)^{-\mu_i a^{-1}}(1+a^{-1})^{-y_i}, & y_i > 0 \end{cases} \tag{4}$$

where $0 \le \omega_i < 1$ and $\mu_i > 0$, with mean $E(Y_i) = (1-\omega_i)\mu_i$ and variance $Var(Y_i) = (1-\omega_i)\mu_i(1+a+\omega_i\mu_i)$.

The p.m.f. shown in (4) indicates that the ZINB-1 reduces to the ZIP when $a = 0$. The variance, $Var(Y_i) = E(Y_i)(1+a+\omega_i\mu_i)$, indicates that the ZINB-1 exhibits overdispersion when $a > 0$ and $\omega_i > 0$, implying that the model can be used for handling both zero-inflation and additional overdispersion in count data. The

parameters, μ_i and ω_i, depend on the covariate vectors of $\mathbf{x}_i$ and $\mathbf{z}_i$, whereby the link functions can be written as (2). If $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and a are estimated by the maximum likelihood method, the log likelihood for the ZINB-1 regression model can be written as,

$$\begin{aligned}\ln L(\boldsymbol{\beta},\boldsymbol{\omega},a) = &\sum_{y_i=0}\log(\omega_i+(1-\omega_i)(1+a)^{-\mu_i a^{-1}}) \\ &+\sum_{y_i>0}(\log(1-\omega_i)+\log(\Gamma(y_i+\mu_i a^{-1}))-\log(y_i!) \\ &\quad -\log(\Gamma(\mu_i a^{-1}))-\mu_i a^{-1}\log(1+a)-y_i\log(1+a^{-1}))\end{aligned} \tag{5}$$

The maximum likelihood estimates of $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and a can be obtained by maximizing the log likelihood with respect to $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and a.

2.4 Zero-inflated Negative Binomial-2 (ZINB-2) Model

The p.m.f. of the ZINB-2 regression model is given by (Ridout *et al.* 2001, Yang *et al.* 2009),

$$\Pr(Y_i=y_i\mid\mu_i,\omega_i,a)=\begin{cases}\omega_i+(1-\omega_i)(1+a\mu_i)^{-a^{-1}}, & y_i=0\\ (1-\omega_i)\dfrac{\Gamma(y_i+a^{-1})}{y_i!\Gamma(a^{-1})}(1+a\mu_i)^{-a^{-1}}(1+a^{-1}\mu_i^{-1})^{-y_i}, & y_i>0\end{cases} \tag{6}$$

where $0\le\omega_i<1$ and $\mu_i>0$, with mean $E(Y_i)=(1-\omega_i)\mu_i$ and variance $Var(Y_i)=(1-\omega_i)\mu_i(1+a\mu_i+\omega_i\mu_i)$.

The p.m.f. shown in (6) indicates that the ZINB-2 reduces to the ZIP when $a=0$. The variance, $Var(Y_i)=E(Y_i)(1+a\mu_i+\omega_i\mu_i)$, indicates that the ZINB-2 exhibits overdispersion when $a>0$ and $\omega_i>0$, also implying that the model can be used for handling both zero-inflation and additional overdispersion in count data. The parameters, μ_i and ω_i, depend on the covariate vectors of $\mathbf{x}_i$ and $\mathbf{z}_i$, whereby the link functions can be written as (2). If $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and a are estimated by the maximum likelihood method, the log likelihood for the ZINB-2 regression model can be written as,

$$\ln L(\boldsymbol{\beta},\boldsymbol{\omega},a) = \sum_{y_i=0}\log(\omega_i + (1-\omega_i)(1+a\mu_i)^{-a^{-1}}) + \sum_{y_i>0}(\log(1-\omega_i) + \log(\Gamma(y_i+a^{-1})) - \log(y_i!) - \log(\Gamma(a^{-1})) - a^{-1}\log(1+a\mu_i) - y_i\log(1+a^{-1}\mu_i^{-1})) \quad (7)$$

The maximum likelihood estimates of $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and a can be obtained by maximizing the log likelihood with respect to $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and a.

It should be noted that the covariates related to the parameters, μ_i and ω_i, in the ZIP, the ZINB-1 and the ZINB-2 regression models may or may not be the same. Famoye and Singh (2006) considered the case where the same covariates are applied, and wrote the link functions as,

$$\log(\mu_i) = \mathbf{x}_i^T\boldsymbol{\beta} \quad \text{and} \log\left(\frac{\omega_i}{1-\omega_i}\right) = -\tau\mathbf{x}_i^T\boldsymbol{\beta}, \quad (8)$$

where τ is a parameter which relates μ_i and ω_i, and can be estimated via the estimation of the regression coefficients.

If μ_i and ω_i are related through the link functions shown in (8), the log likelihood for the ZIP regression model can be rewritten as,

$$\ln L(\boldsymbol{\beta},\tau) = -\sum_{y_i}\log(1+\mu_i^{-\tau}) + \sum_{y_i=0}\log(\mu_i^{-\tau} + e^{-\mu_i}) + \sum_{y_i>0} y_i\log(\mu_i) - \log(y_i!) - \mu_i \quad (9)$$

Similarly, the log likelihood for the ZINB-1 and the ZINB-2 regression model can be rewritten as,

$$\ln L(\boldsymbol{\beta},a,\tau) = -\sum_{y_i}\log(1+\mu_i^{-\tau}) + \sum_{y_i=0}\log(\mu_i^{-\tau} + (1+a)^{-\mu_i a^{-1}}) + \sum_{y_i>0}\log(\Gamma(y_i+\mu_i a^{-1})) - \log(y_i!) - \log(\Gamma(\mu_i a^{-1})) - \mu_i a^{-1}\log(1+a) - y_i\log(1+a^{-1}) \quad (10)$$

$$\ln L(\boldsymbol{\beta},a,\tau) = -\sum_{y_i}\log(1+\mu_i^{-\tau}) + \sum_{y_i=0}\log(\mu_i^{-\tau} + (1+a\mu_i)^{-a^{-1}}) + \sum_{y_i>0}\log(\Gamma(y_i+a^{-1})) - \log(y_i!) - \log(\Gamma(a^{-1})) - a^{-1}\log(1+a\mu_i) - y_i\log(1+a^{-1}\mu_i^{-1}) \quad (11)$$

The fitting procedure for the ZIP, ZINB-1 and the ZINB-2 can be performed by using *R programming* based on the *nlm* function. For faster convergence, the estimated parameters obtained from fitting the ZIP are used as initial values.

2.5 Likelihood Ratio Test and Wald test

2.5.1 LRT

If we are interested in the adequacy of the ZIP against the ZINB-1 or the ZINB-2, a two-sided test of likelihood ratio can be applied where the hypothesis can be stated as $H_0 : a = 0$ vs. $H_1 : a \neq 0$. The likelihood ratio is $T = 2(\ln L_1 - \ln L_0)$, where $\ln L_1$ and $\ln L_0$ are the model's log likelihood under respective hypothesis. T has an asymptotic chi-square distribution with one degree of freedom.

If we are interested in testing overdispersion in the ZIP against the ZINB-1 or the ZINB-2, i.e. $H_0 : a = 0$ vs. $H_1 : a > 0$, a one-sided likelihood ratio test can be implemented. The standard asymptotic theory states that under H_0, $\text{sgn}(a)\sqrt{T} = \text{sgn}(a)\sqrt{2(\ln L_1 - \ln L_0)}$ follows an asymptotic standard Normal distribution, where $\text{sgn}(.)$ is the sign function, so that when $a > 0$ it takes the value of 1, otherwise it takes the value of -1.

2.5.2 Wald

In addition to the likelihood ratio, the test of overdispersion in the ZIP versus the ZINB-1 or the ZINB-2 can be performed by using a Wald statistic which is defined as a ratio of the estimated overdispersion parameter to its standard error, $\frac{\hat{a}}{\sqrt{Var(\hat{a})}}$, where asymptotically, the statistic follows a standard Normal distribution.

3. Results and Discussion

Table 2 shows the parameters, the log likelihood, the AIC and the SBC for the ZIP and the ZINB-2 models using the link functions shown in (8). Unfortunately, the fitting procedure does not provide converged solutions for the ZINB-1. The results show that the regression parameters for the ZIP and the ZINB-2 models have similar estimates, and as expected, many of the *t*-ratios in the ZINB-2 are smaller than the ZIP. For testing overdispersion in the ZIP versus the ZINB-2, i.e. $H_0 : a = 0$ vs. $H_1 : a > 0$, the likelihood ratio and the Wald statistic respectively are 2,025.38 and

7.99, indicating that the null hypothesis is rejected and the ZINB-2 is more adequate for fitting the theft claim data.

Table 2. Estimated parameters for ZIP and ZINB-2 using link functions (8)

Parameter	ZIP		ZINB-2	
	Est.	t-ratio	Est.	t-ratio
Intercept	-7.56	-43.36	-7.39	-18.28
Coverage: Non-comprehensive	0.42	6.49	1.52	6.72
Vehicle year: 2-3	0.35	4.16	0.33	1.14
4-5	0.42	5.03	0.21	0.74
6-7	-0.22	-2.25	-0.75	-2.58
8+	1.01	13.73	0.79	2.57
Vehicle cc:1001-1300	0.41	2.59	0.83	2.34
1301-1500	0.68	4.00	1.71	4.79
1501-1800	1.77	10.56	2.00	5.91
1801+	2.35	13.81	2.35	7.15
Vehicle make: Local type 2	0.80	5.23	1.69	5.61
Foreign type 1	-0.91	-14.50	-0.67	-2.94
Foreign type 2	-0.50	-5.72	0.53	2.14
Foreign type 3	-2.11	-13.73	-1.20	-3.15
τ	0.09	1.92	0.32	4.29
a	-	-	2.35	7.99
Log likelihood	-2,382.59		-1,369.90	
AIC	4,795.18		2,771.81	
SBC	4,869.66		2,851.25	

Table 3 shows the parameters, the log likelihood, the AIC and the SBC for the ZIP and the ZINB-1 models using the link functions shown in (2). Unfortunately, the fitting procedure does not provide converged solutions for the ZINB-2. The results show that the regression parameters for the ZIP and the ZINB-1 models also have similar estimates, and as expected, many of the *t*-ratios in the ZINB-1 are smaller than the ZIP. For testing overdispersion in the ZIP versus the ZINB-1, i.e. $H_0 : a = 0$ vs. $H_1 : a > 0$, the likelihood ratio and the Wald statistic respectively are 1993.20 and 8.86, indicating that the null hypothesis is rejected and the ZINB-1 is more adequate for fitting the theft claim data.

Table 3. Estimated parameters for ZIP and ZINB-1 using link functions (2)

		ZIP		ZINB-1	
		Est.	t-ratio	Est.	t-ratio
Covariates for μ_i	Intercept	-7.09	-44.67	-6.95	-20.89
	Coverage: Non-comprehensive	0.42	6.47	0.76	4.64
	Vehicle year: 2-3	0.31	3.69	0.40	2.14
	4-5	0.40	4.70	0.33	1.79
	6-7	-0.29	-2.81	-0.23	-1.13
	8+	1.07	14.48	0.22	1.08
	Vehicle cc:1001-1300	-0.04	-0.30	0.00	0.00
	1301-1500	0.14	0.90	0.41	1.29
	1501-1800	1.24	8.31	1.05	3.28
	1801+	1.85	12.20	1.57	4.84
	Vehicle make: Local type 2	0.32	2.38	0.13	0.50
	Foreign type 1	-0.95	-15.13	-0.62	-4.07
	Foreign type 2	-0.61	-6.99	0.26	1.52
	Foreign type 3	-2.21	-14.03	-1.15	-4.17
Covariates for ω_i	Intercept	-7.16	-6.09	-8.70	-5.63
	Coverage: Non-comprehensive	4.73	9.01	5.43	5.51
	Vehicle year:2-3	-1.42	-1.85	-1.85	-1.01
	4-5	-1.88	-2.29	-2.28	-1.47
	6-7	-1.61	-1.67	-1.85	-1.31
	8+	-1.08	-1.67	-0.77	-0.73
	Vehicle cc:1001-1300	-1.80	-1.48	-1.45	-1.01
	1301-1500	-4.34	-3.39	-4.75	-2.98
	1501-1800	-2.63	-2.24	-2.72	-1.92
	1801+	-2.04	-1.78	-2.58	-1.82
	Vehicle make: Local type 2	-1.26	-0.98	-1.10	-0.71
	Foreign type 1	-2.04	-2.63	-1.81	-1.88
	Foreign type 2	-0.04	-0.06	0.62	0.70
	Foreign type 3	-0.48	-0.38	-0.01	0.00
a		-	-	6.56	8.86
Log likelihood		-2,317.53		-1,320.93	
AIC		4,691.05		2,699.85	
SBC		4,830.07		2,843.84	

4. Conclusions

The ZINB-1 does not provide converged solutions if the link functions shown in (8) are applied. The likelihood ratio and the Wald tests indicate that the ZINB-2 is a better model compared to the ZIP. If the link functions shown in (2) is used instead, the

ZINB-2 does not provide converge solutions. The likelihood ratio and Wald tests also imply that the ZINB-1 is a better model compared to the ZIP. Since the crime data that were used contained more than 70% of zero claims, the result proved that the proposed assessment measures were useful and reliable to examine the predictive performance of the whole model as well as the realization of individual observations in the data include 'zero' occurrence in crime data. This study was limited to vehicle theft data collected in Malaysia and the covariates are only focusing on the vehicles. For future research, it is recommended to include the demographic factor as covariates in which the vehicles are stolen.

References

Famoye, F., & Singh, K.P. (2006). Zero-inflated Generalized Poisson Regression Model with an Application to Domestic Violence Data. *Journal of Data Science*. 4, 117-130.

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

Lambert, D. (1992). Zero-inflated Poisson Regression, with an Application to Defects in Manufacturing. *Technometrics*. 34, 1-14.

Malaysian Quality of Life.(2004). Public Safety.

Ridout, M.S., Hinde, J.P., & Demetrio, C.G.B. (2001). A Score Test for Testing a Zero-inflated Poisson Regression Model against Zero-inflated Negative Binomial Alternatives. *Biometrics*. 57, 219-223.

Sidhu, A.S. (2005). The Rise of Crime in Malaysia. An Academic and Statistical Analysis. *Journal of the Kuala Lumpur Royal Malaysia Police College.* 4, 1-28.

Yang, Z., Hardin, J.W., & Addy, C.L. (2009). Testing Overdispersion in the Zero-inflated Poisson Model. *Journal of Statistical Planning and Inference*. 139, 3340-3353.