Zero Inflated Poisson and Zero Inflated Negative Binomial Models with Application to Number of Falls in the Elderly

Yusuf OB [1]*, Bello T [2] and Gureje O[2]

[1]*Department of Epidemiology and Medical Statistics, University of Ibadan, Nigeria*

[2]*Department of Psychiatry, University of Ibadan, Nigeria*

Submission: March 15, 2017; **Published:** May 08, 2017

*Corresponding author: Yusuf OB, Department of Epidemiology and Medical Statistics, University of Ibadan, Nigeria,
Email: bidemiyusuf1@gmail.com

Abstract

The presence of excess zeros and the problem of over-dispersion often occur with count data. Few methods have been developed to deal with extra zeros that occur in response count variables. Such methods include zero inflated Poisson (ZIP) and zero inflated negative binomial (ZINB) regression models. This analysis determined the best fitting model when the response variable is a count variable: number of falls in the elderly.

The data set consists of variables on number of falls as well as other covariates collected in a longitudinal study of the Ibadan study of ageing; a community-based survey, conducted in eight states in the south-western and north-central regions in Nigeria. Descriptive statistics were conducted to check for the presence of over-dispersion. Incidence rate ratios (IRR), and 95% CI were determined. Zero inflated Poisson and zero inflated negative binomial regression models were fitted. The best model was selected based on the values of Vuong z-statistic, -2logLL, AIC and BIC selection criteria. The Vuong z-statistic of the ZIP and ZINB models were 7.14 and 0 respectively. The results showed that the ZINB model gave the best fit with AIC of 3281.569 and -2LogLL of 3237.57. The ZINB identified chronic pain has a factor associated with number of falls. This study suggests the Zero inflated negative binomial regression as the best model for predicting number of falls in the presence of excess zeros and over-dispersion.

Keywords: Zero inflated Poisson regression; Zero inflated negative binomial regression; Falls in the elderly; Chronic pain

Background

Generalized linear models (GLMs) are powerful tools for analyzing count data. The starting point for count data is a GLM with Poisson-distributed errors, but not all count data meet the assumptions of the Poisson distribution. Therefore it is important to test if the variance > mean or if the number of zeros is greater than expected.

Usually when zeros occur in data sets, researchers or analysts deal with it as missing data, delete or impute data. At times these zeros have meanings and are important and should be dealt with as such. It may mean that the subject or respondent did not experience the outcome and such data should not be considered as missing. Generally data should not be transformed to fit a linear model and, particularly, count data should not be log-transformed. Another approach commonly used is to make the variable a categorical one by considering the zeros as "outcome absent" and the observations as "outcome present". This is not a very good option either since the variable has been transformed and interest is no more in the number of occurrences.

A standard Poisson model would not distinguish between the two processes causing an excessive number of zeroes, but a zero-inflated model allows for and accommodates this complication. When analyzing a dataset with an excessive number of outcome zeros and two possible processes that arrive at a zero outcome, a zero-inflated model should be considered. A histogram of the response variable is drawn to see if the number of zeros is excessive. If two processes generated the zeroes in the response variable but there is not an excessive number of zeroes, a zero-inflated model may or may not be used.

In dealing with dependent variables with excess zeros, the generalized linear model (GLM) with Poisson distribution can be considered. However, the GLM with Poisson distribution does not deal with over dispersion particularly when it is due

to excessive zeros [1]. However, the zero inflated model with Poisson distribution (ZIP) deals with excessive zero responses in count data. Theory suggests that the excess zeros can be modeled independently. The ZIP model has two parts: a Poisson count model and the logit model for predicting excess zeros. The zero inflated negative binomial regression is also for modeling count data with excessive zeros and it is usually for over dispersed count outcome variables. Again theory also suggests that the excess zeros are generated by a separate process from the count values and that the excess zeros can be modeled independently [2].

In this analysis, we considered the number of falls; a count variable with zero responses. If the respondent didn't experience a fall, the only possible outcome is zero. If the respondent experienced a fall, then it is a count process. Previous studies on falls have examined falls as a categorical variable by examining factors associated with its presence or absence [3-5]. Number of falls is usually a count variable and should be treated as such.

We determined the best fitting model between the zero inflated Poisson and the zero inflated negative binomial regression models in identifying the factors associated with number of falls in the elderly using data from the Ibadan Study for Ageing.

Methods

The Zero Inflated Poisson (ZIP) Regression Model

In Zero Inflated Poisson regression, the response ($Y = Y_1$, Y_2, ...Y_n) is independent. The assumption of this model is that with probability p the only possible observation is 0, and with probability (1-p), a Poisson (λ) random variable is observed in Y. Therefore the occurrence of Y_i follows this distribution:

$$Y_i = \begin{cases} 0, \text{ with probability } p_i + (1-p_i)e^{-\lambda_i} \\ k, \text{ with probability } (1-p_i)\dfrac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i}, k = 1, 2, \cdots \end{cases} \quad (1)$$

The mean and variance of ZIP distribution are respectively:

$$E(Y_i) = (1-p_i)\lambda_i \quad (2)$$

$$V(Y_i) = (1-p_i)(\lambda_i + \lambda_i^2) - ((1-p_i)\lambda_i)^2 \quad (3)$$

The Zero Inflated Negative Binomial (ZINB) Regression Model

The Zero-Inflated Negative Binomial (ZINB) regression model assumes that there are two distinct data generation processes. The result of a Bernoulli trial is used to determine which of the two processes is used. For observation i, with probability π_i the only possible response of the first process is zero counts, and with probability of $1-\pi_i$, the response of the second process is governed by a negative binomial with mean λ_i. The zero counts are generated from the first and second processes, where a probability is estimated for whether zero

counts are from the first or the second process. The overall probability of zero counts is the combined probability of zeros from the two processes. Following Greene and Yau et al., a ZINB model for the response Y_i is written as:

$$Y_i = 0, \text{ with probability } \pi_i \quad (4)$$

$$Y_i \sim \text{negative binomial } (\lambda_i, k), \text{ with probability } (1-\pi_i) \quad (5)$$

So that,

$$\Pr(Y_i = 0) = \pi_i + (1-\pi_i)(1+k\lambda_i)^{-\frac{1}{k}} \quad (6)$$

$$\Pr(Y_i = y_i) = (1-\pi_i)\frac{\Gamma(y_i + \frac{1}{k})(k\lambda_i)^{y_i}}{\Gamma(y_i+1)\Gamma(\frac{1}{k})(1+k\lambda_i)^{y_i+\frac{1}{k}}}, y_i = 1, 2, \cdots \quad (7)$$

Where the mean and variance of the Y_i are:

$$E(Y_i) = (1-\pi_i)\lambda_i \quad (8)$$

$$V(Y_i) = (1-\pi_i)\lambda_i(1-\lambda(\pi_i + k)) \quad (9)$$

where λ_i is the mean of the underlying negative binomial distribution and k is the over-dispersion parameter.

Y_i = the number of falls in the elderly

$X_1, \ldots \ldots X_n$ = selected independent variables

Application: Number of falls in the elderly

Data Description and Exploratory data Analysis: The data set consists of variables on number of falls as well as other covariates collected in a longitudinal study of the Ibadan study of ageing, a community-based longitudinal survey, conducted in eight states in the south-western and north-central regions in Nigeria (Lagos, Ogun, Osun, Oyo, Ondo, Ekiti, Kogi and Kwara) over a four year period (2003/2004, 2007-2009). Respondents were elderly people aged 65 years or above who resided in the selected states. A multistage cluster sampling of households within enumeration areas (geographical units demarcated by the National Population Commission) within each state was employed. Only one respondent, who had provided consent, mostly verbal due to illiteracy or by choice, was selected per household. Information was collected using a questionnaire that contained socio-demographic, social engagement and health characteristics as well as questions on falls. At baseline in 2003/2004, 2149 respondents completed the study. Respondents were subsequently followed up yearly from 2007 to 2009. Only the baseline data were used for this analysis. The data were weighted to account for the clustering.

The outcome variable was number of falls in the last 12 months, while independent variables were socio demographic characteristics: age, sex, years of education, wealth index developed using principal component analysis to create 5 wealth quintiles: poorest, poorer, middle, richer, richest. The wealth index was constructed using the asset approach whereby all household possessions are included as much as possible. Items used for the wealth index include: wall clock, bucket, radio, cell

phone, motor bike, fan, gas or electric stove, video cassette, bicycle, television, air conditioner, livestock, personal computer, clothes iron, deep freezer, refrigerator, and motor vehicle. The reliability, precision and accuracy of the wealth index increases when many items are used for its computation [6]. Other explanatory variables considered were presence of chronic medical conditions such as arthritis, insominia, and presence of chronic pain. A total of 2118 respondents answered the question: how many times have you fallen in the last 12 months? Of these, 475 had fallen 1 time, 91 had fallen 2 times, 63 had fallen 3 times, etc., We checked for the number of zeros {zeros =1643 (77.57%)}, and then examined the means (mean=0.4759, variance =1.304) and confirmed that variable (number of falls) was over dispersed. Features of data suggest the use of zero inflated models. Figure 1 shows the number of falls. The ZIP and ZINB models were fitted. Model selection and comparison were done using the Akaike information criteria (AIC), Bayesian Information criteria (BIC),-2logL as well as the Voung test.
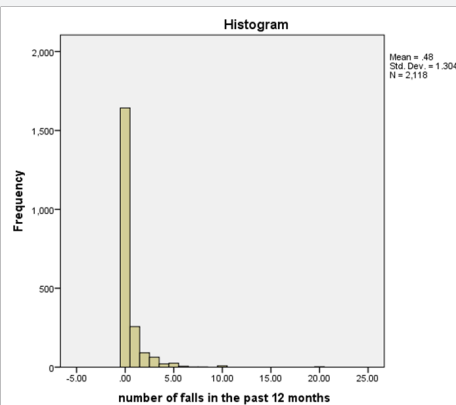

Figure 1: Histogram showing number of falls.

Results

A total of 475 (22.4%) respondents reported having 1 or more fall while 1643 had zero falls. About a third (30.6%) were above 80years, slightly above half (53.4%) were females, 25.8% lived in the urban areas and about half (50.3%) were married. Majority (70.8%) had no medical condition, 79.1% had chronic pain and 25.6% were visually impaired. Table 1 shows the characteristics of the respondents.

Table 1: Characteristics of Respondents.

Variable	Frequency	Percentage
Age Group		
65-69	702	32.7
70-74	495	23
75-79	295	13.7
80+	657	30.6
Sex		
Male	1001	46.6
Female	1148	53.4
Location		
Urban	555	25.8
Semi urban	870	40.5
Rural	724	33.7
Marital Status		
Currently married	1080	50.3
Never married	9	0.4
Divorced	32	1.5
Separated	43	2
Widowed	985	45.8
Occupation		
Skilled/higher level	221	11.5
Elementary	794	41.2
Trade	914	47.4
Medical Condition		
Yes	628	29.2
No	1521	70.8
Chronic Pain		
Yes	1700	79.1
No	449	20.9
Wealth Index		
Poorest	535	24.9
Poorer	523	24.3
Middle	426	19.8
Richer	339	15.8
Richest	326	15.2
Vision Impairment		
Yes	517	25.6
No	1503	74.4

Results from the ZIP model

The number of falls was approximately lower by 23.2% among female respondents compared to male respondents (IRR = 0.768, 95% CI: 0.594, 0.991). Number of falls among respondents from rural areas was higher by 34.0% (IRR = 1.340, 95% CI: 1.025, 1.751) compared to respondents from urban areas. Respondents who had a chronic medical condition had a 19.1% decrease in number of falls (IRR = 0.809, 95% CI: 0.668, 0.981) compared to respondents who had no medical condition. The risk of experiencing a fall increased by 28.8% among respondents who were visually impaired (IRR = 1.288, 95% CI: 1.076, 1.541) compared to respondents who were not visually impaired, however, this was not significant.

For the absolute zero group, the risk of being in the absolute zero group among female respondents decreased by 44.3% (IRR = 0.557, 95% CI: 0.379, 0.818) compared to male respondents. In addition, the risk of being in the absolute zero group among respondents with chronic pain decreased by 39.2% (0.608, 95% CI: 0.421, 0.877) compared to respondents with no chronic pain. The risk of being in the absolute zero group for respondents in semi-urban areas decreased by 21.5% (0.785, 95% CI: 0.544, 1.134) compared to respondents in urban areas. However the risk of being in the absolute zero group for respondents in rural areas increased by 5.2% (1.052, 95% CI: 0.710, 1.558) compared to respondents in urban areas. Location was found to be non-significant. The risk of being in the absolute zero group for respondents that were visually impaired increased by 59.7% (1.597, 95% CI: 0.786, 1.398) compared to respondents who were not visually impaired, but this result was not significant. Table 2 shows the parameter estimates from the zero inflated Poisson regression.

Table 2: Parameter estimates in the Zero inflated Poisson model.

Covariate	IRR	SE	p-value	95% CI	
				Lower	Upper
Intercept	1.557	0.378	0.068	0.967	2.507
Age group					
65-69*					
70-74	0.826	0.103	0.126	0.656	1.056
75-79	0.824	0.119	0.179	0.621	1.093
80+	0.957	0.105	0.685	0.771	1.186
Sex					
Female	0.768	0.1	0.043	0.594	0.991
Male*					
Location					
urban*					
semi-urban	1.03	0.133	0.817	0.8	1.327
rural	1.34	0.183	0.032	1.025	1.751
Marital Status					
Currently married*					
Never married	0.606	0.619	0.624	0.082	4.49
Divorced	0.942	0.369	0.879	0.438	2.029
Separated	3.862	0.841	<0.001	2.519	5.919
Widowed	1.295	0.162	0.039	1.013	1.656
Occupation					
Elementary occupation	1.161	0.185	0.348	0.85	1.586
Trade	0.872	0.151	0.429	0.622	1.224
Skilled/ Higher level occupation*					
Medical Condition					
Yes	0.809	0.079	0.031	0.668	0.981
No*					
Constant pain					
Yes	1.168	0.153	0.236	0.903	1.509
No*					
wealth index					
Poorest*					
Poorer	1.003	0.118	0.973	0.798	1.263
Middle	0.801	0.097	0.068	0.631	1.016
Richer	0.719	0.114	0.038	0.526	0.982
Richest	0.952	0.143	0.741	0.709	1.277
Any Visual Impairment					
Yes	1.288	0.118	0.006	1.076	1.541
No*					
Inflation Variables					
Intercept	5.071	0.336	<0.001	2.623	9.804
Age group					
65-69*					
70-74	0.951	0.188	0.789	0.657	1.375
75-79	0.837	0.219	0.414	0.545	1.284
80+	0.883	0.17	0.461	0.633	1.23
Sex					
Female	0.557	0.196	0.003	0.379	0.818
Male*					
Location					
urban*					
semi-urban	0.785	0.188	0.197	0.544	1.134
rural	1.052	0.201	0.802	0.71	1.558
Marital Status					
Currently married*					
Never married	0.48	1.655	0.657	0.019	12.304
Divorced	1.191	0.585	0.766	0.378	3.751
Separated	2.277	0.468	0.079	0.91	5.695
Widowed	1.064	0.188	0.741	0.736	1.54
Occupation					
Elementary occupation	1.188	0.235	0.465	0.749	1.883

Trade	1.316	0.251	0.275	0.804	2.154
Skilled/ Higher level occupation*					
Medical Condition					
Yes	0.769	0.15	0.08	0.574	1.032
No*					
Chronic Pain					
Yes	0.608	0.187	0.008	0.421	0.877
No*					
wealth index					
Poorest*					
Poorer	1.178	0.185	0.377	0.819	1.694
3	0.744	0.195	0.129	0.508	1.09
Middle	1.016	0.235	0.947	0.641	1.611
Richer	1.126	0.228	0.602	0.721	1.76
Richest					
Any Visual Impairment					
Yes	1.597	0.147	0.75	0.786	1.398
No*					
-2logL	3386.994				
AIC	3466.944				
BIC	3688.992				

*reference category, Voung test of ZIP vs. Standard Poisson z =7.14

Results from the ZINB model

Respondents who had chronic pain had a 67.0% increase in number of falls (IRR = 1.670, 95% CI: 1.217, 2.292) compared to respondents who had no pain. Respondents who were visually impaired had a 36.4% increase in number of falls (IRR = 1.364, 95% CI: 1.042, 1.786) compared to respondents that were not visually impaired. Number of falls was approximately higher by 35.4% among female respondents compared to male respondents (IRR = 1.354, 95% CI: 0.961, 1.907). Number of falls among respondents from semi-urban and rural areas were higher by 26.7% (IRR = 1.267, 95% CI: 0.931, 1.725) and 29.4% (IRR = 1.294, 95% CI: 0.917, 1.825) respectively compared to respondents from urban areas. Respondents who had elementary occupation had an 18.1% increase in number of falls (IRR = 1.181, 95% CI: 0.775, 1.800) compared to respondents who had skilled/higher level occupation. However, respondents who were traders had a 21.1% decrease in number of falls (IRR = 0.789, 95% CI: 0.506, 1.231) compared to respondents who had skilled/higher level occupation. Sex, location, marital status and occupation were not significantly associated with number of falls. Table 3 shows the parameter estimates from the zero inflated negative binomial regression.

Table 3: Parameter estimates in the Zero Inflated negative binomial model.

Covariate	IRR	SE	p-value	95% CI	
				Lower	Upper
Intercept	0.23	0.682	<0.001	0.129	0.411
Age Group					
65-69*					
70-74	0.834	0.139	0.275	0.602	1.155
75-79	0.908	0.176	0.619	0.621	1.328
80+	1.016	0.159	0.919	0.747	1.381
Sex					
Female	1.354	0.237	0.083	0.961	1.907
Male*					
Location					
urban*					
semi-urban	1.267	0.199	0.132	0.931	1.725
rural	1.294	0.227	0.142	0.917	1.825
Marital Status					
Currently married*					
Never married	1.074	1.014	0.939	0.169	6.836
Divorced	0.863	0.44	0.773	0.318	2.346
Separated	1.712	0.693	0.184	0.774	3.786
Widowed	1.1	0.18	0.563	0.797	1.517
Occupation					
Elementary occupation	1.181	0.254	0.439	0.775	1.8
Trade	0.789	0.179	0.296	0.506	1.231
Skilled/ Higher level occupation*					
Medical Condition					
Yes	0.966	0.126	0.789	0.748	1.248
No*					
Chronic Pain					
Yes	1.67	0.27	0.001	1.217	2.292
No*					
wealth index					
1					
2	0.901	0.153	0.539	0.645	1.257

3	1.014	0.178	0.935	0.72	1.43
4	0.683	0.134	0.052	0.464	1.004
5	0.824	0.168	0.344	0.552	1.23
Any Visual Impairment					
Yes	1.364	0.188	0.024	1.042	1.786
No*					
Inflation variables					
Intercept	-12.8103	391.963	0.974	-781.044	755.4237
-2logL	3237.57				
AIC	3281.569				
BIC	3403.695				

Model Comparison and Selection

We used the AIC, BIC, -2log LL and Voung statistic for model comparison and selection. For the ZIP model, AIC was 3466.944 while for the ZINB, AIC was 3281.569. In addition, the -2logL was 3386.994 and 3237.57 for the ZIP and ZINB models respectively (Table 4).

Table 4: Model Comparison.

	ZIP	ZINB
AIC	3466.944	3281.569
BIC	3688.992	3403.695
-2logL	3386.994	3237.57
Voung statistic	7.14	0

Discussion

In this study we identified and quantified the relationships between frequency of falls in the elderly and selected variables in the presence of over dispersion by using zero inflated regression models. The number of falls was over dispersed and the zero inflated negative binomial models were the preferred model in identifying the factors associated with falls. Mixed probability distributions were used to model the zero and non zero counts. The estimated results are efficient and non - biased. Our results showed that the standard errors from the zero inflated Poisson model were smaller than those of the zero inflated negative binomial model which could have led to some insignificant regression parameters. These large standard errors in the ZINB model suggest that in the presence of over dispersion, the ZIP model overstates the significance of the regression parameters and the significance of the evaluation factors. This is compatible with findings from other studies [7-10]. The ZINB was the best model selected based on the values of the model selection test/criteria utilized.

About a quarter of the respondents reported experiencing 1 or more fall. This low rate of fall had been discussed extensively by Bekibele and Gureje [3]. In addition, this rate is comparable with reports from Japan [11], China [12], but low when compared to Australia [13], Spain [14], Italy [15], Brazil and Iran [16-19]. The ZIP model showed that female sex, being separated, or widowed, having chronic medical condition, fourth wealth quintile were significantly associated with number of falls. However, the ZINB identified chronic medical pain as the only predictor of falls in the elderly. Previous studies in which fall in the elderly was treated as a categorical variable have identified regular medication, depression, sufficient exercise, wearing slippery shoes, as significant factors of falls in the elderly [5]. However, in another analysis where falls had been used as a categorical variable, chronic pain was the only significant factor associated with falls using the binary logistic regression [3]. Similarly, our findings using the ZINB model also identified presence of chronic pain as the only significant factor associated with number of falls in the elderly. We however note that in carrying out the ZINB model, we used the option "constant" for the inflate part in the analysis procedures in stata; and we reported the modeling of the non zero counts. We didn't include the covariates in the inflation part of the ZINB model because it disrupted the model from converging; hence the option was deleted.

This study demonstrated that the ZINB model is the best model to determine the factors that predict the frequency of falls in the elderly when there is an indication of the presence of excess zeros and over dispersion. It is recommended that objective criteria should be used to select appropriate statistical models for analysing count data in the presence of over dispersion.

References

1. Hinde J, Demetrio C (1998) Overdispersion models and estimation. Computational Statistics and Data Analysis 27(2): 151-170.

2. Lambert D (1992) Zero-Inflated Poisson Regression, with application to defects in manufacturing. Technometrics 34(1): 1-14.

3. Bekibele CO, Gureje O (2010) Fall Incidence in a Population of Elderly Persons in Nigeria. Gerontology 56(3): 278-283.

4. Takahiro, Kondo K, Suzuki K, Yamada M, Matsumoto D (2014) Factors Associated with Falls in Community-Dwelling Older People with Focus on Participation in Sport Organizations: The Japan Gerontological Evaluation Study Project. Bio Med Research International 2014 (2014): 1-10.

5. Piyathida K, Prasert P, Bangonsri J (2013) Factors related to falls among community Dwelling elderly. South east Asian J Trop Med Public Health 44(5): 906-915.

6. Rutstein SO, Staveteig S (2013) Making the Demographic and Health Surveys Wealth Index Comparable. In: IUSSP Int Popul Conf 26-31 BEXCO, Busan, Korea, p. 1-59.

7. Piza BEL (2013) Using Poisson and Negative Binomial Regression Models to Measure the Influence of Risk on Crime Incident Counts 2012.

8. Ismail N, Jemain AA, (2007) Handling overdispersion with negative binomial and generalized poisson regression models. Casualty Actuarial Society Forum Winter pp. 103-158.

9. Sadia F (2013) Performance of generalized Poisson regression model and negative binomial regression model in case of over-dispersion count data. IASIR pp. 558-563.

10. Islam MM, Alam M, Tariquzaman M, Kabir MA, Pervin R, et al. (2013) Predictors of the number of under-five malnourished children in Bangladesh: application of the generalized poisson regression model. BMC Public Health 13: 11.

11. Yoshida H, Kim H (2006) Frequency of falls and their prevention. Clin Calcium 16(9): 1444-1450.

12. Chu LW, Chiu AY, Chi I (2008) Falls and subsequent health service utilization in community-dwelling Chinese older adults. Arch Gerontol Geriatr 46(2): 125-135.

13. Gill T, Taylor AW, Pengelly A (2005) A population-based survey of factors relating to the prevalence of falls in older people. Gerontology 51(5): 340-345.

14. Varas-Fabra F, Castro Martin E, Pérula de Torres LA, Fernández Fernández MJ, Ruiz Moral R, et al. (2006) Falls in the elderly in the community: prevalence, consequences, and associated factors. Aten Primaria 38(8): 450-455.

15. Mancini C, Williamson D, Binkin N, Michieletto F, DeGiacomi GV, et al. (2005) Epidemiology of falls among the elderly. Ig Sanita Pubbl 61(2): 117-132.

16. Perracini MR, Ramos LR (2002) Fall related factors in a cohort of elderly community residents. Rev Saude Publica 36(6): 709-716.

17. Siqueira FV, Facchini LA, Piccini RX, Tomasi E, Thumé E, et al. (2007) Prevalence of falls and associated factors in the elderly. Rev Saude Publica 41(5): 749-756.

18. Abolhassani F, Moayyeri A, Naghavi M, Soltani A, Larijani B, et al. (2006) Incidence and characteristics of falls leading to hip fracture in Iranian population. Bone 39(2): 408-413.

19. Piza BEL (2012) Using Poisson and Negative Binomial Regression Models to Measure the Influence of Risk on Crime Incident Counts. 2012.

Your next submission with Juniper Publishers will reach you the below assets

- Quality Editorial service
- Swift Peer Review
- Reprints availability
- E-prints Service
- Manuscript Podcast for convenient understanding
- Global attainment for your research
- Manuscript accessibility in different formats
  (Pdf, E-pub, Full Text, Audio)
- Unceasing customer service

Track the below URL for one-step submission
https://juniperpublishers.com/online-submission.php