Assessment of Statistical Approaches to Model Low Count Data: An Empirical Application to Youth Delinquency

Taimoor Malik[1], Syed Arif Ali[2], Abdur Rasheed[2,*] and Afaq Ahmed Siddiqui[3]

[1]*Clinical trials Unit, Dow University of Health Sciences, Karachi, Pakistan*

[2]*Department of Research, Dow University of Health Sciences, Karachi, Pakistan*

[3]*Faculty of Pharmacy, University of Karachi, Karachi, Pakistan*

Abstract: *Objectives*: The aim of this study was to identify the risk factors associated with number of crime committed by youth (Youth Delinquency) between ages 10-17, using Ordinary Least Square (OLS), Poisson Regression model (PRM), Negative Binomial Regression model (NBRM)& Zero Inflated Negative Binomial (ZINB) with the aim to choose the most appropriate model for the observed count data.

Methodology: The data in the study was collected from youth whose mothers enrolled in Philadelphia Collaborative Perinatal Project (CPP). School and delinquency record (between ages 10-17) was obtained by the Centre for studies in Criminology and Criminal Law. Literature search suggest that factors associated with child delinquency can be divided into four main factors as Individual, Family, School and Peer. Therefore we included variables in the analysis accordingly.

Result: For OLS scatter plot of residuals versus estimated counts showed definite pattern of heterogeneity (non-constant variance). The likelihood-ratio (LR) test of over dispersion yields the significant p-value, which implied that the outcome variable is overdispersed. The plot of the difference between the actual probabilities and the mean predicted probabilities for each model showed that PRM has poor predictions for low counts (0-2).

Conclusion: NBRM and ZINB both performed well, however fit statistics revealed that NBRM has provided more closed predication as compare ZINB.NB modeling techniques provides much more compelling and accurate results instead of basic PRM or those available through simple linear or log-linear modeling techniques.

Keywords: Count Data, Poisson regression model, Negative Binomial Regression.

1. INTRODUCTION

Count Variables indicate how many times specific event has occurred. A count variable can take only zero or positive integer values as an event cannot occur a negative number of times. There are several multidisciplinary examples including number of doctor visits [1], number of alcohol drink consumed [2], number of road traffic accidents [3], number of publications [4], number of children ever born to a women [5]. Studies that model theses counts and their association with other variables provide information leading to better understanding of the problem under study.

When the mean of the count variable is relatively high, OLS regression techniques provide reasonable results. However, when the mean of the count is low, OLS regression yields Inefficient, inconsistent and biased estimates [4, 6]. In this regard, Poison Regression Model (PRM) has been served as the basic model as it lends itself well with the nature of count data and relatively easy to understand. However Poisson Regression Model (PRM) assumes that the mean and variance of count variable are equal, a property known as equidispersion. Practically, in most of applications actual data may be overdispersed (that is variance exceeds the mean), which has similar consequences as the failure of homoscedasticity (constant variance) assumption in OLS regression [1].

Due to this restriction, Negative Binomial regression model (NB) has been developed which has the capability to account for overdispersion. But sometimes data in hand contains far more zeros (excessive number of zeros) than are allowed [4, 7]. Although overdispersed count model that is NBRM can be used to model the data having zero counts. The problem of excessive zero counts can be handled by using Zero Inflated Negative Binomial Regression Model (ZINB) introduced by lambert [8] and Greene [9] and there may be little advantage in fitting ZINB. In ZINB data are assumed to come from a mixture of two distributions. The structural zeros from a binary distribution are mixed with the non-negative integer outcomes (including zeros) from a count distribution. Logistic regression is usually used to model the structural zeros, and negative binomial regression is used for the count outcomes.

*Address correspondence to this author at the Department of Research, Dow University of Health Sciences, Karachi, Pakistan;
E-mail: abdur.rasheed@duhs.edu.pk

E-ISSN: 1929-6029/15

The aim of this study is to identify the risk factors associated with number of crime committed by youth between ages 10-17 (Youth Delinquency) using, PRM, NB & ZINB with the aim to choose the most appropriate model for the observed data.

METHODOLOGY

The data we used in the study was collected from youth whose mothers enrolled in Philadelphia Collaborative Perinatal Project (CPP) between 1959 to 1974. (http://doi.org/10.3886/ICPSR08928.v2). Upon registration, each mother was administered a series of interviews. Data recorded for each child from birth through age seven were recorded on several variables. School and delinquency record (between ages 10-17) were obtained by the Centre for studies in Criminology and Criminal Law [10].

Recent literature search suggest that factors associated with child delinquency can be divided into four main factors as Individual, Family, School and Peer [11-13]. Therefore we included variables in the analysis accordingly Table **1**.

Statistical Analysis

The adequacy of OLS model was examined by scatter plot of residuals versus estimated counts. To compare count models, the difference between the observed probabilities for each count and the average predicted probabilities for each model were plotted on the same graph. All regression analyses were performed using Stata version 11.1.

RESULTS

The results of the estimated Regression Models presented in Table **2**. For OLS scatter plot of residuals versus estimated counts showed definite pattern of heterogeneity (non-constant variance), indicating OLS regression model was not appropriate choice.

To choose between PRM and NBRM, the likelihood-ratio (LR) test of over dispersion yields the significant p-value, chibar2 (01) =183.72, p =<.01, which implied that the outcome variable is overdispersed and hence NBRM is the better choice.

Table 1: Description of Variables Used in the Analyses

Variable	Description
Gender	0= Male 1= Female
Income	Family income at the time of Registration in CPP
Remedial Disciplinary Codes In School	Number of times an individual enrolled in a disciplinary program during High School
Nursery School Attendance	0=Yes 1=No
Subscales of CAT[1]	
CAT_1	Vocabulary Score
CAT_2	Comprehension Score
CAT_3	Reading Score
CAT_4	Computation Score
CAT_5	Concept& Problem Score
CAT_6	Total Maths Score
CAT_7	Mechanics Score
Subscales of WISC[2]	
WISC-1	Information Score
WISC-2	Comprehension Score
WISC-3	Vocabulary Score
WISC-4	Digit Span Score
WISC-5	Block Design Score
WISC-6	Coding Performance Score
WISC-7	Verbal I.Q SCORE
WISC-8	Performance I.Q Score

Table 2: Estimated Models to Analyze Youth Delinquency

Regression Model	Poisson				Negative Binomial				Zero Inflated Negative Binomial			
Variable	**Coeff**	**SE**	**e^b**	**P-Value**	**Coeff**	**SE**	**e^b**	**P-Value**	**Coeff**	**SE**	**e^b**	**P-Value**
Gender												
Female	-0.90944	0.1056669	0.4027	<0.01	-0.9658702	0.1822669	0.3807	<0.01	-0.939789	0.1799274	0.3907	<0.01
Family Income	-0.0001202	0.0000259	0.9999	<0.01	-0.0001161	0.0000475	0.9999	0.014	-0.0001213	0.0000461	0.9999	0.008
Remedial Disciplinary Codes In School	0.2475586	0.0235893	1.2809	<0.01	0.3250835	0.1318478	1.3841	0.014	0.2728055	0.1117078	1.3136	0.015
Retarded	-0.0866352	0.0504951	0.917	0.086	-0.0650381	0.0855598	0.937	0.447	-0.0520493	0.0938061	0.9493	0.579
Subscales of CAT[1]												
Vocabulary Score	-0.0017094	0.0029647	0.9983	0.564	-0.0048391	0.0057505	0.9952	0.4	-0.0042933	0.005579	0.9957	0.442
Comprehension Score	-0.0094624	0.0038414	0.9906	0.014	-0.0084006	0.0072804	0.9916	0.249	-0.0050991	0.0070953	0.9949	0.472
Reading Score	-0.0222458	0.0040609	0.978	<0.01	-0.0215281	0.0068799	0.9787	0.0020	-0.024581	0.0066807	0.9757	<0.01
Computation Score	0.0103111	0.0028505	1.0104	<0.01	0.012072	0.0058244	1.0121	0.038	0.0117496	0.0056746	1.0118	0.038
Concept& Problem Score	-0.0054292	0.0031216	0.9946	0.082	-0.005009	0.0055582	0.995	0.367	-0.0040867	0.0053547	0.9959	0.445
Total Maths Score	0.0174498	0.0035963	1.0176	<0.01	0.0106158	0.0069551	1.0107	0.127	0.0105066	0.0068229	1.0106	0.124
Mechanics Score	-0.0047984	0.0044883	0.9952	0.285	-0.0003341	0.0079657	0.9997	0.967	0.0008428	0.0077938	1.0008	0.914
Subscales of WISC[2]												
Information Score	-0.1550568	0.1535085	0.8564	0.312	0.0500873	0.3296517	1.0514	0.879	-0.0383882	0.3266869	0.99	0.906
Comprehension Score	-0.2002367	0.1540842	0.8185	0.194	-0.0177752	0.3293874	0.9824	0.957	-0.0818141	0.3264187	0.99	0.802
Vocabulary Score	-0.1788257	0.1551708	0.8363	0.249	0.0073434	0.3310648	1.0074	0.982	-0.2109335	0.3293144	0.99	0.522
Digit Span Score	-0.2133848	0.1547352	0.8078	0.168	-0.0187186	0.3312071	0.9815	0.955	-0.0853319	0.3279963	0.99	0.795
Block Design Score	-0.0054404	0.029305	0.9946	0.853	-0.0299404	0.0614853	0.9705	0.626	-0.0184781	0.0597649	0.99	0.757
Coding Performance Score	0.0292024	0.0256732	1.0296	0.255	0.0263197	0.0509468	1.0267	0.605	0.0341352	0.0492585	0.99	0.488
Verbal I.Q SCORE	0.11941	0.0976037	1.1268	0.221	-0.0006731	0.2083913	0.9993	0.997	0.0426289	0.2062347	0.99	0.836
Performance I.Q Score	-0.0142703	0.0083109	0.9858	0.086	-0.0067547	0.0166896	0.9933	0.686	-0.0100017	0.016097	0.99	0.534

1. California Achievement Test.
2. Wechsler Intelligence Scale for Children.

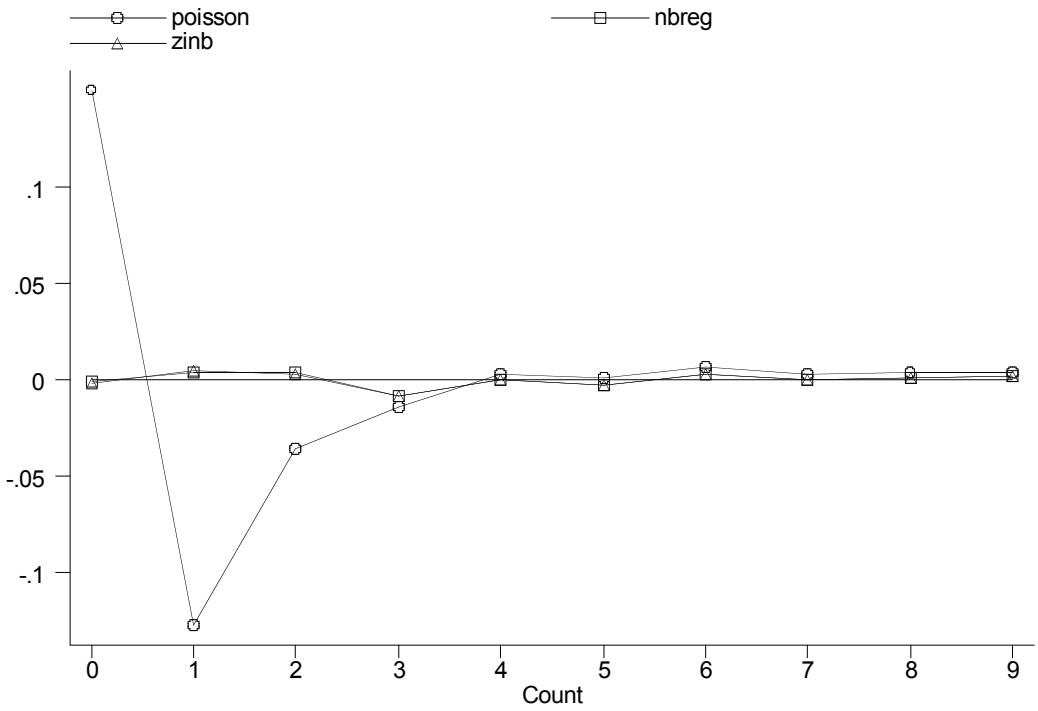

Figure 1: Difference between the observed probabilities and the average predicted probabilities.

Additionally, the results of Voung test, a test to determine whether ZINB is statistically preferred over NBRM, yields the insignificant p-value of 0.105, supporting NRBM as compare to ZINB.

Figure **1** is the plot of the difference between the actual probabilities and the mean predicted probabilities for each model. It is apparent that PRM has poor predictions for low counts (0-2). Among NBRM and ZINB, both performed well, however investigation based on fit statistics revealed that NBRM has provided more closed predication as compare ZINB.

Gender, Income, evidence of disciplinary problems at school and two subscales of CAT were found to be the statistically significant risk factors associated with youth delinquency in all models. Additionally, one subscale of CAT was significant only in PRM and one subscale was significant both in OLS and PRM. Since, the results obtained from OLS and PRM could lead to incorrect inference and also as Figure **1** suggest that the NBRM is the most appropriate model, therefore, we focused only on NB Model.

Being a male youth, increases the expectation of committing a crime. Income at registration in CPP was negatively associated with delinquency. Similarly youth with low achievement score are more opt to have an offense record. Whereas, evidence of disciplinary problems at school, presented in the model as the number of times an individual enrolled in a disciplinary program was positively associated with youth delinquency (Table **2**).

Estimated Coefficients of NB and ZINB can be interpreted in the same way as in PRM. For example, results of NB regression revealed that being a female youth decreases the expected number of crimes by a factor of 0.38. Results can also be expressed in term of percentage as being a female youth decreases the expected number of crimes by 62% (Table **2**).

DISCUSSION

In this study we have compared modern approaches of analyzing count outcome by using empirical data of Youth delinquency. The study showed that gender, lower family income, lower achievements & disciplinary problems at school were the strongest predictors of youth delinquency.

While PRM is the good starting point due to its simplicity, it rarely explains the data in hand [3, 7]. The NBRM that induces overdispersion may be sometimes more suitable. When there is high frequency of zero count in dataset, ZINB may be more appropriate [8, 9].

In this study NBRM completely outperformed the basic PRM. On the other hand although the proportion

of zero observations in dataset was approximately 25% but still NBRM was the best fitted model which suggests that when fitting a series of models without any theoretical justification, it is easy to overfit the data [14, 15]. The fit of count models was assessed by comparing the predicted average proportion of each count outcome to observed proportion [16, 17].

The results of our study suggest that serious over dispersion can results into inflated test statistics which could eventually lead to incorrect statistical inference .For example two variables two CAT variables were not significant in both PRM and NBRM attained statistical significance in PRM. Due to its structure, the Negative Binomial modeling techniques provides much more compelling and accurate results instead of basic PRM or those available through simple linear or log-linear modeling techniques.

CONCLUSION

NBRM has provided more closed predication as compare ZINB.NB modeling techniques provides much more compelling and accurate results instead of basic PRM or those available through simple linear or log-linear modeling techniques.

REFERENCES

[1] Cameron AC, Trivedi PK. Econometric models based on count data. Comparisons and applications of some estimators and tests. Journal of Applied Econometrics 1986; 1(1): 29-53. http://dx.doi.org/10.1002/jae.3950010104

[2] Armeli S, Mohr C, Todd M, Maltby N, Tennen H, Carney MA, *et al*. Daily evaluation of anticipated outcomes from alcohol use among college students. Journal of Social and Clinical Psychology 2005; 24(6): 767-92. http://dx.doi.org/10.1521/jscp.2005.24.6.767

[3] Chin HC, Quddus MA. Modeling count data with excess zeroes an empirical application to traffic accidents. Sociological Methods & Research 2003; 32(1): 90-116. http://dx.doi.org/10.1177/0049124103253459

[4] Scott Long J. Regression models for categorical and limited dependent variables. Advanced Quantitative Techniques in the Social Sciences 1997; 7.

[5] Poston Jr DL, McKibben SL. Using zero-inflated count regression models to estimate the fertility of US women. Journal of Modern Applied Statistical Methods 2003; 2(2): 10.

[6] Malik T, Khan M, Sheikh Z. Models of association between demographics and the hospital visits by patients with type 2 diabetes mellitus. International Journal of Diabetes in Developing Countries 2014: 1-5.

[7] Cameron AC, Trivedi PK. Regression analysis of count data: Cambridge university press 2013. http://dx.doi.org/10.1017/cbo9781139013567

[8] Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics 1992; 34(1): 1-14. http://dx.doi.org/10.2307/1269547

[9] Greene WH. Accounting for excess zeros and sample selection in Poisson and negative binomial regression models 1994.

[10] Denno DW. Sociological and human developmental explanations of crime: Conflict or consensus?* Criminology 1985; 23(4): 711-41. http://dx.doi.org/10.1111/j.1745-9125.1985.tb00371.x

[11] Wasserman GA, Keenan K, Tremblay RE, Coie JD, Herrenkohl TI, Loeber R, *et al*. Risk and protective factors of child delinquency: US Department of Justice, Office of Justice Programs, Office of Juvenile Justice and Delinquency Prevention Washington 2003. http://dx.doi.org/10.1037/e501772006-001

[12] Shader M. Risk factors for delinquency: An overview: US Department of Justice, Office of Justice Programs, Office of Juvenile Justice and Delinquency Prevention 2001.

[13] Belknap J, Holsinger K. The gendered nature of risk factors for delinquency. Feminist Criminology 2006; 1(1): 48-71. http://dx.doi.org/10.1177/1557085105282897

[14] Fekedulegn D, Andrew M, Violanti J, Hartley T, Charles L, Burchfiel C. Comparison of statistical approaches to evaluate factors associated with metabolic syndrome. The Journal of Clinical Hypertension 2010; 12(5): 365-73. http://dx.doi.org/10.1111/j.1751-7176.2010.00264.x

[15] Gardner W, Mulvey EP, Shaw EC. Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models. Psychological Bulletin 1995; 118(3): 392. http://dx.doi.org/10.1037/0033-2909.118.3.392

[16] Long JS, Freese J. Regression models for categorical dependent variables using Stata: Stata press 2006.

[17] Long JS, Freese J. Predicted probabilities for count models. Stata Journal 2001; 1(1): 51-7.

Received on 13-06-2015 Accepted on 26-07-2015 Published on 19-08-2015

http://dx.doi.org/10.6000/1929-6029.2015.04.03.6