2nd Annual International Conference on Qualitative and Quantitative Economics Research (QQE 2012)

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.

Mohd Asrul AA\textsuperscript{a}, Nyi Nyi Naing\textsuperscript{b}

\textsuperscript{a}University Tun Hussien Onn, Jalan Parit Raja, Parit Raja, 86400, Batu Pahat Johor, Malaysia
\textsuperscript{b}Universiti Sains Malaysia, Unit of Biostatistics & Research Methodology, Health Campus,16150, Kota Bharu, Kelantan Malaysia

Abstract

The analysis data with accessing high zero by using the model of Poisson, Negative Binomial Regression (NBR), Zero-Inflated Poisson (ZIP) and Zero-Inflated Negative Binomial (ZINB) is widely used. Deviance and Pearson Chi-Square goodness of fit statistic indicate no over dispersion exists in this study. In the selection of appropriate regression model, Aike Information Criteria (AIC) and Bayesian Information Criteria (BIC) were used. Small value of AIC and BIC of the model accepted as a good model. At the end of these were information criteria, ZINBDR regression was chosen as the best model.

Keywords: Zero Inflated Negative Binomial Death Rate, Standardization Rate, Influential Observation.

Introduction

In many area of interest such as economic fields, agriculture, epidemiology, ecology the dependent or response variable of interest \( (y) \) is a non-negative integer or count which is guess to explain or determine in terms of a set of covariates \( (x) \). Unequal the traditional regression model, the response variable is discrete with a distribution that places probability mass at non negative integer values only. In term of regression models for count, such other limited or discrete dependent variable model as well as the logit and probit, are linear with many condition and special features intimately linked to discreteness and nonlinearity. Thus, NBR is appropriate to replace the PR when absence the over-dispersion [12].
In NBR model, the parameter estimated are converged by considering effect that stems from overdispersion. Basically count observation might have excessive zero than expected. In such case ZIP regression model is appropriate approach to analyze the dependent variable having to much zero observation \[\text{?}\]. ZIP assumes that the population consists of two different type observation whereby one of them is based on count data consist Poisson distribution that can have zero value exists \[\text{?}\]. In such cases, when ZIP existing overdispersion and highly accessing zero such mentioned above, ZINB is alternative method that will used \[\text{?}\]. Like in ZIP regression in ZINB the observation with zero data and those without zero data are modeled in different way. According the discrete model such Pois-

![Diagram of models](image)

Figure 1: The frequently used models in the count data analysis framework

son, NB, ZIP, and ZINB let us consider some examples from microeconomics, begining with samples independent cross-section observations. Such fertility study, frequent modeling number or live births over specified age interval of the mother, with willing in analyzing it variations in terms of like mother schooling, age, and household income\[?\]

1 Methods

1.1 Death Rate

To incorporate into ZINB regression model we employ a death rate function to dependent variable. Rate dependent variable are estimated by requirement as follows. Lets assume mortality rate cases in the \(j^{th}\) observation for \(j = 1, 2, \ldots, n\) a categorical observation age rate death estimation, whereby supposed to be negative binomial distributed with \(d_j\) is the expected death of rate cases. Age death rate normaly was calculated using standard population rate as follows:

\[
d_j = \frac{q_j e_j}{p_j}
\]

where;

- \(q_j\) = Number of death among persons of a given age group.
- \(p_j\) = Population of person in given age group in a standard population
- \(e_j\) = Constant population.
2 Zero Inflated Negative Binomial Death Rate

To incorporate into ZINB regression model we employ a death rate function to dependent variable. Rate dependent variable are estimated by requirement as follows. Let assume mortality rate cases in the \( j \)th observation for \( j = 1, 2, \ldots, n \) a categorical observation age rate death estimation, whereby supposed to be negative binomial distributed with \( d_j \) is the expected death of rate cases. Again, refer the equation (1) substitute count observation \( y_i \) to \( y_{dj} \) death rate observation. Thus the equation ZINBDR as follows:

\[
(Y_{dj}) = \begin{cases} 
\omega_i + (1 - \omega_i)(1 + \psi \theta_i)^{-\psi^{-1}} & y_{dj} = 0 \\
(1 - \omega_i)f(y; \theta, \psi) & y_{dj} > 0 
\end{cases}
\]

and the log-likelihood ZINBDR is;

\[
Lc(y_{dj}; \gamma, \beta, \psi) = \sum_{y_{dj}=0}^{n} \ln[\exp(z'_i\gamma) \\
+ (1 + \psi \exp(x'_i \beta))^{-\psi}] \\
+ \sum_{y_{dj}>0}^{n} \sum_{j=0}^{y_{dj}-1} \ln(j + \psi^{-1}) \\
+ \sum_{y_{dj}>0}^{n} \{ - \ln(y_{dj})! - (y_{dj} + \psi^{-1}) \\
+ \ln(1 + \psi \exp(x'_i \beta)) + y_{dj} \ln(\psi) \\
+ y_{dj}x'_i \beta \} \\
- \sum_{dj=1}^{n} \ln[1 + \exp(z'_i\gamma)]
\]

3 NB - Dependent Death Rate (NBDR)

Similar with NBDR, substitute equation (7) \( y_i \) count observation to \( y_{dj} \) death rate observation. Thus, the equation dependent death rate negative binomial is expressing such;

\[
P(Y_{dj} = y_{dj}) = \frac{\Gamma(y_{dj} + 1/\psi)}{y_{dj}!\Gamma(1/\psi)} \left[ \frac{1}{1 + \psi \theta_i} \right]^{1/\psi} \left[ \frac{\psi \theta_i}{1 + \psi \theta_i} \right]^{y_{dj}}
\]

for \( y_{dj} > 0 \) and \( y_{dj} \) is the death of rate by age categorical followed by, \( \theta_i \) is the expected rate of death per year. To incorporate covariate, assume that \( \theta_i = \exp(x'_i \beta) \) where \( \beta \) is a \((P + 1) \times 1\) vector of covariates and intercept of \( \beta_0 \), the coefficient for regression \((\beta_0, \beta_1, \beta_2, \ldots, \beta_p)\). Taking the exponential of \( x'_i \beta \) ensure that the mean parameter \( \theta_i \) is nonnegative. Thus, the log-likelihood NBDR as follows;
\[ L_c(y_{dj}; \psi, \theta) = \sum_{i=1}^{n} \log \left( \frac{\Gamma(y_{dj} + \psi - 1)}{y_{dj}! \Gamma(\psi - 1)} \right) \]

\[ - (Y_{dj} + \psi - 1) \log (1 + \psi \theta) \]

\[ + (Y_{dj} \log (\psi \theta)) \]

3.1 Data

We used a secondary data death of (AIDS) Kota Bharu, Kelantan Malaysia. The data consisted (n=945) measurement of gender, national, race, marital status, occupation, and mode transmission. Table below describe of the covariates used;

| Variable          | Description         |
|-------------------|---------------------|
| Gender            | 0=Female 1=Male     |
| National          | 0=Non Malaysian 1=Malaysian |
| Race              | 0=Non Malay 1=Malay |
| Marital status    | 0=Single 1=Married 2=Divorce/Widow |
| Occupation        | 0=unemployment 1=selfemployment 2=government 3= non government 4= housewife 5=retired 6=student |
| Mode transmission | 0= IVDU 1=sexual transmission 2=unknown |

The data were collected for 2000 until 2008 in Kelantan area and the dependent variable in each model is the rate of death for AIDS patients by using aged group (categorized as) 20-24, 25-29,...,65-69. The independent variables moduled as table above.
4 Model Selection

Model goodness of fit was examined by the loglikelihood using the Akaike Information Criteria (AIC) and the Bayesian Information Criteria (BIC). The likelihood ratio test was used to compared the Poisson model and NB model. Mento-Carlo simulation indicate that AIC and BIC selection criteria need to be used together [?]. The equation of AIC and BIC described as follows:

\[
AIC = -2LL + 2r
\]

and

\[
BIC = -2LL + r\ln(n)
\]

where \( LL \) is a log likelihood value, \( r \) indicates number of parameter and \( n \) is a sample size.

5 Results

Descriptive statistics for the variable rate of death age, and gender, nation, race, status, occupation and transmission used in the present study are given in table 2 below. The 945 sample of observation values belonging each variable were used in the study. While the smallest values mean for the rate of death by age categorical was 0, the highest values detected as 281.07.

The almost 60\% observation values out of 945 observation used in the study were zero valued among the variable used. The number of variables given in following in figure 2.

![Figure 2: Frequency zeroes values in the model](image)

In NBDR analysis, Deviance and Pearson Chi square goodness of statistics indicating no overdispersion was obtained 1.10 and 0.71 respectively. Being higher than (1) of the
Table 2: Descriptive statistic for variables

| Variable     | N   | Mean | Min | Max  |
|--------------|-----|------|-----|------|
| Rate         | 945 | 35.84| 0   | 281.06|
| Gender       | 945 | 0.93 | 0   | 0.06 |
| Nation       | 945 | 0.98 | 0   | 0.01 |
| Race         | 945 | 0.96 | 0   | 0.03 |
| Status       | 945 | 0.47 | 0   | 0.46 |
| Occupation   | 945 | 1.23 | 0   | 1.77 |
| Transmission | 945 | 0.25 | 0   | 5.33 |

mentioned goodness of fit statistic represents that were was an overdispersion in data set. AIC and BIC selection criteria for the model of NBDR and ZINBDR are given in table 3. The model selection criteria given in table 3 found extremely different from each other. It was found out that ZINBDR selection criteria were low as to NBDR. The model with a smallest AIC and BIC was ZINDR. Therefore ZINBDR model shown in table 3, with a bold letters was chosen as the best model. All independent variables analyses programming was done using PROC NLMIXED in SAS 9.2. Statistical significant was set at $\alpha = 0.05$ and 95% confidence interval.

Table 3: Model selection criteria for ZINBDR and NBDR

| Models | Log-likelihood | AIC    | BIC    |
|--------|----------------|--------|--------|
| ZINBDR | 8062.4         | 8092.4 | 8165.1 |
| NBDR   | 8079.7         | 8095.7 | 8174.5 |

6 Discussion

Determination goodness of fit via model selection basically based on some criteria informaton theoretical procedure. This theory was developed in the 1950’s and was quantified with Akaike Information Criterion (AIC) in 1970. An extended summary of information theoretical criteria involving model closeness and practical uses of the model inference. In general, the regression model which has the smallest AIC and BIC values is regarded as the best model [?]. In this case value AIC ZINBDR smallest than NBDR but value BIC NBDR smallest than ZINBDR. Beside that, at the end of likelihood ratio test, it seems that ZINBDR model gave better results than NBDR model.

Goodness of statistic (Deviance and Pearson Chi Square), determining whether regression method such as negative binomial and logistic part were applicable very essential [?]. In this analysis of study, values both of goodness of statistic were obtained 1.10 and
0.71 respectively and indicating no overdispersion exist. Besides, if overdispersion absent with the high value, it might had effect in two different regression model goodness of criteria and parameter estimated values in the model. NB regression model was preferred to PR model in classical approach as well.

We choose to use the ZINBDR model, thus it was possible that all assumption for this model were not met especially in regard for the underlying dual-state distribution.

Competing interest
The author(s) declare that they have no competing interests.

Author’s contributions
MAAA outlined the paper, performed the analyses and wrote the manuscript. NNN edited the manuscript for intellectual content and supervised the work and helped conceive the paper. All authors read and approved the final manuscript.

Acknowledgements
We would like to thank the Ministry of Higher Education and Universiti Tun Husseing Om Malaysia, were supported this research. Thanks to medical record Hospital University Sains Malaysia provided the data through ethical approved on 15 Jun 2011 (USMKK/PPP/JEPeM [238.4(1.12)])

References

[1] Lambert.D, Zero-Inflated Poisson Regression with an application to detection in manufacturing, Technometrics.,1-14,34,(1992).

[2] Agresti.A, Categorical Data Analysis,John and Wiley and Sons Incorporation,(1997).

[3] Cameron.A.C and Trivedi.P.K, Analysis of Count Data,New York, Cambridge University Press,(1998).

[4] Frome.E.D and Kutner.M.H and Beauchamp.J.J, Regression Analysis of Poisson Distributed Data, Journal of American Statistical Association,68(344):935-940(1973).

[5] Stokes.M.E and Davis.C.S and Koch.G.G, Categorical Data Analysis Using SAS System,John and Wiley and Sons Incorporation,(2000).

[6] Yesilova.A and Kaki.B and Kasap.I, Regression methods used in modelling of dependent variable obtained based on Zero Inflated Count Data, Journal of Statistical Research.,5:1-9.,(2007).

[7] Long.J.S and Freese.J, Regression Model for Categorical Dependent Variable Using Stata, A Stata Press Publication,USA.,(2001).
[8] McCullagh.P and Nelder.J, *Generalized Linear Model*, Second Edition, Chapman and Hall., London, (1989).

[9] Cox.R, *Some Remarks on Overdispersion*, Biometrika., 70:269-274., (1983).

[10] SAS, *SAS/Stat. Software*, Hangen and Enhanced, USA, Institute Incorporation., (2007)

[11] Wang.P and Puterman.M.L and Cockburn.I.M and Le.N, *Mixed Poisson Regression Models with Covariate Dependent Rates*, Biometrics., 52:381-400., (1996).

[12] Jansakul.N, *Zero-Inflated Negative Binomial via R*, In Proceeding 20th International Workshop on Statistical Modelling., Sydney., Australia, 277-284 (2005).

[13] Ridout.M and Hinde.J and Demetrio.C.G.B, *A Score Test for Zero Inflated Poisson Regression Model Against Zero Inflated Negative Binomial Alternatives*, Biometrics., 57., 219-233., (2003)

[14] Lawles.J.F, *Negative Binomial and Mixed Poisson Regression*, The Canadian Journal of Statistic., 15(3)., 1-13, (1987)

[15] Bohning.D, *Zero Inflated Poisson Models and C.A.MAN. A Toturial Collection of Evidence*, Biometrical Journal., 40(7):833-834 (1998)

[16] Chuen.Y.B, *A Study of Growth and Development*, Statistic in Medicine., 21:1461-1469 (2002)

[17] Lee.A.H and Wang.K and Yau.K.K.W, *Analysis of Zero Inflated Poisson Data Incorporating Extent of Exposure*, Biometrical Journal., 43 (8):963-975., (2001)

[18] Bohning.D and Dietz.E and Schlattmann.P, *The Zero-Inflated Poisson Model and the Decayed Missing and Filled Teeth Index in Dental Epidemiology*, Journal of Royal Statistic Society A., 162:195-209 (1999).

[19] Yau.K.K.W and Lee.A.H, *Zero Inflated Poisson Regression with Random Effect to Evaluate an Occupational Injury Prevention Programme*, Statistic in Medicine., 20:2907-2920 (2001).

[20] Rose.C.E and Martin.S.W and Wannemuehler.K.A and Plikaytis, *On the Zero-inflated and Hurdle Models for Modeling Vaccine Adverse event count Data*, Journal of Biopharmaceutical Statistic., 16:463-481 (2006).

[21] Winkelmann.R, *Duration Dependence and Dispersion in Count Data Models*, Journal of Business and Economic Statistic., 13., 467-474.

[22] Rose.N, *Profitability and Product Quality*, Economic Determinants of Airline Safety Performance., Journal of Political Economy., 98., 944-964.

[23] Gurmu.S and Travedi.P.K, *Excess Zeros in Count Model for Recreational Trips*, Journal of Business and Economic Statistic., 14., 469-477
[24] Cameron.A.C and Trevidi.P.K and Piggot.J, *A Microeconometric Model of the Demand for Health Care and Health Insurance Care in Australia*, Review of Economic Studies.,55,85-106.

[25] Davutyan.N, *Bank Failures as Poisson Variates*, Economic Letters.,29.,333-338.

[26] Hausman.J.A and Hall.B.H and Griliches.Z, *Econometric Model for Count Data With an Application to the Patents-R and D relationship*, Econometrica.,52.,909-938

[27] Dalrymple.M.L and Hudson.I.L and Ford.J.J, *Finite Mixture, Zero Inflated Poisson and Hurdle Models with Application to SID*, Computational Statistics & Data Analysis.,68(344):935-940(1973).

[28] Bains.N, *Standardization of rates*, Associtaion of public health epidemiologists in Ontario(APHEO).,2009.,March.

[29] Anderson.N. and Resenbery.H.M, *Age standardization of death rate implementation of the year 2000 standard*, National vitae statistic report.,47.,(3).,1993.

[30] Choi.B.C.K and deGuia.N.A and Walsh.P, *Look before you leep: stratify before you standardize*, American Journal of Epidemiology.,149(12).,1087-1095.

[31] Kitagawa.E.M, *Standardization comparisons in population research*, Demography.,1.,296-315.

[32] Rosenberg.H.M and Curtin.L.R and Maurer.J, *Choosing a standard population: Some Statistical Consideration*, Chapter 5, Workshop Proceeding. Vital and Health Statistics.,Series 4,(29).,93-1466

[33] Minami.M. and Cody.C.E.L. and Goa.W. and Verdesato.M.R, *Modeling shark by-catch: The zero-inflated negative binomial regression model with smoothing*, Journal of Fisheries Research.,210-221.,(84).,(2006)

[34] Hall.D and Berenhaut.K.S, *Score test for heterogeneity and overdispersion in Zero-inflated Poisson and Binomial Regression Models*, The Canadian Journal of Statistics.,1-15,30,(3),(2002)

[35] Belsey.D.A and Kuh.E and Welsch.R.E, *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, John Wiley & Son.,(1980)

[36] Joseph.M.Hilbe, *Negative Binomial Regression*, Cambridge University Press.,New York.,(2011)