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, Nyi Nyi Naing

*A University Tun Hussien Onn, Jalan Parit Raja, Parit Raja, 86400, Batu Pahat Johor, Malaysia
bUniversiti 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 [?]. ZIP assumes that the population consists of two different type observation whereby one of them is based on count data consists Poisson distribution that can have zero value exists [?]. In such cases, when ZIP existing overdispersion and highly accessing zero such mentioned above, ZINB is alternative method that will used [?]. 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-

![Figure 1: The frequently used models in the count data analysis framework](image)

son, NB, ZIP, and ZINB let us consider some examples from microeconomics, beginning with samples independent cross-section observations. Such fertility study, frequent modeling number of 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 normally 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;

\[
L_c(y_{dj}; \gamma, \beta, \psi) = \sum_{y_{dj} = 0}^{n} \ln[\exp(z_i'\gamma) \\
+ (1 + \psi \exp(x_i'\beta))^{-\psi}] \\
+ \sum_{y_{dj} > 0}^{n} \sum_{j=0}^{y_{dj}-1} \ln(j + \psi^{-1}) \\
+ \sum_{y_{dj} > 0}^{n} \{- \ln(y_{dj})! - (y_{dj} + \psi^{-1}) \\
+ \ln(1 + \psi \exp(x_i'\beta)) + y_{dj} \ln(\psi) \\
+ y_{dj} x_i'\beta \} \\
- \sum_{d_j=1}^{n} \ln[1 + \exp(z_i'\gamma)]
\]

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;
\[ L_c(y_{dj}; \psi, \theta) = \sum_{i=1}^{n} \log \left( \frac{\Gamma(y_{dj} + \psi^{-1})}{y_{dj}! \Gamma(\psi^{-1})} \right) \]

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

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

### 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;

<table>
<thead>
<tr>
<th>Table 1: Summary of variables used in the analysis of AIDS mortality data</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Variable</strong></td>
</tr>
<tr>
<td>Gender</td>
</tr>
<tr>
<td>National</td>
</tr>
<tr>
<td>Race</td>
</tr>
<tr>
<td>Marital status</td>
</tr>
<tr>
<td>Occupation</td>
</tr>
<tr>
<td>Mode transmission</td>
</tr>
</tbody>
</table>

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 modulated as table above.
4 Model Selection

Model goodness of fit was examined by the loglikelihood using the Aikake 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:

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

(3)

and

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

(4)

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

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rate</td>
<td>945</td>
<td>35.84</td>
<td>0</td>
<td>281.06</td>
</tr>
<tr>
<td>Gender</td>
<td>945</td>
<td>0.93</td>
<td>0</td>
<td>0.06</td>
</tr>
<tr>
<td>Nation</td>
<td>945</td>
<td>0.98</td>
<td>0</td>
<td>0.01</td>
</tr>
<tr>
<td>Race</td>
<td>945</td>
<td>0.96</td>
<td>0</td>
<td>0.03</td>
</tr>
<tr>
<td>Status</td>
<td>945</td>
<td>0.47</td>
<td>0</td>
<td>0.46</td>
</tr>
<tr>
<td>Occupation</td>
<td>945</td>
<td>1.23</td>
<td>0</td>
<td>1.77</td>
</tr>
<tr>
<td>Transmission</td>
<td>945</td>
<td>0.25</td>
<td>0</td>
<td>5.33</td>
</tr>
</tbody>
</table>

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 ZINBDR. Therefore ZINBDR model shown in Table 3, with a bold letters was choosen 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

<table>
<thead>
<tr>
<th>Models</th>
<th>Log-likelihood</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZINBDR</td>
<td>8062.4</td>
<td>8092.4</td>
<td>8165.1</td>
</tr>
<tr>
<td>NBDR</td>
<td>8079.7</td>
<td>8095.7</td>
<td>8174.5</td>
</tr>
</tbody>
</table>

6 Discussion

Determination goodness of fit via model selection basically based on some criteria information 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. Besides 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 Hussien Om Malaysia, were supported this research. Thanks to medical record Hospital Universiti Sains Malaysia provided the data through ethical approved on 15 Jun 2011 (USMKK/PPP/JEPeM [238.4(1.12)])

**References**


