# Comparison between Negative Binomial and Poisson Death Rate Regression Analysis: AIDS Mortality Co-Infection Patients

Mohd Asrul AA[1], Nyi Nyi Naing[2]
[1]*Universiti Tun Hussien Onn,Jalan Parit Raja,86400,Batu Pahat, Johor, M alaysia.*
[2]*Universiti Sains M alaysia, Health Campus,16150, Kota Bharu, Kelantan, Malaysia*

***Abstract:*** *Analysis of count data is widely used in medical studies, epidemiology, ecology and many research of interest. Basically negative binomial (NB) will used when Poisson data lead to heterogeneity or over dispersion. Thus, when a Poisson data proof of over dispersion phenomenon exists NB will be replaced for that purpose. We modeled categorical age of death rate cases as the dependent variable comparing a NB and Poisson regression. To con- duct this purpose; SAS was used by using PROC GENMODE procedure. To estimate parameter according categorical age of death rate, we did standardization of rate via NB and Poisson distribution. The objective of this study was to compare Negative Binomial Death Rate (NBDR) and Poisson Death Rate (PDR).*

***Key words:*** *negative binomial, Poisson death rate, count data*

## I. Introduction

Mortality rate among AIDS co-infection should be given a full intention in our country. The WHO recorded that highest figure about this problem. There are several factors that affected of mortality AIDS co-infection patients. Regression technique was used to estimate the relationship between exposure death rate. The usual regression method models, widely use for mortality data in environment epidemiology is the Poisson regression modeling (McCullagh.P. and Nelder.J.A, 1989). While existing over dispersion is a common problem with poisson regression when conditional variance is greater than conditional mean in the observed count data. When absence of over dispersion in Poisson regression, negative binomial has been proven able to some situations when the Poisson model is poor fit (Cameron and Trevedi, 1998).This paper considered a standardization of death rate via negative binomial and Poisson regression to compare of the models. Thus, there are several methods considered a new approaches in environmental epidemiology such as generalized linear mixed effect model was used the penalized splinnes as the smoothing methods(Chunag.K.J. et al., 2007) or generalized additive model that use natural cubic splines as the smoothing method (Morgan.G. et al., 1998; Michelozzi.P. et al., 1998) with Poisson and negative binomial regression. To model this approaches in negative binomial and Poisson regression we used via substituted standardization of rate to count observation based on age categories whereby samples population mortality AIDS begin 2000-2008 in Kelantan areas.

## II. Data

Considering part of analysis data, we used a secondary data death of (AIDS) Kota Bharu, Kelantan Malaysia. The data 945 population sample size measurement of gender, national, race, marital status, occupation, and mode of transmission. The data presented mortality rate consist 945 patients were death year 2000 until 2008 according age categorical.

Hence, in figure 1, age between 25-29 and 30-34 shows the highest recorded death rate followed by 35-39. Why this age group the highest? According to figure 1, there has been a sharp rise in the number of young men dying over

Table 1: Summary of variables used in the analysis of AIDS mortality data

| Variable | Description |
|---|---|
| Gender | 0=Female<br>1=Male |
| National | 0=Non Malaysian<br>1=Malaysian |
| Race | 0=Non Malay |

| | |
|---|---|
| | 1=Malay |
| Marital Status | 0=Single |
| | 1=Married |
| | 2= Divorce/Widow |
| Occupation | 0= unemployment |
| | 1=self employment |
| | 2=government |
| | 3= non government |
| | 4= housewife |
| | 5=retired |
| | 6=student |
| Mode transmission | 0= IVDU |
| | 1=sexual transmission |
| | 2=unknown |

the last decade. Besides that, record death rate among second generation for Irish migrants living in England and Wales are 20 percent higher than for the rest of the population. Addition, mental disorders which include some drugs and alcohol related deaths, approximations 6 percent under group of age 25 to 39 in England and Wales both for men and women. Vice- versa, for United State older individual with HIV or AIDS also report more chronic medical condition and limitation in physical functioning (Bureau, 2009). Thus, recorded in 1999, 62 percent of the population was aged 20-29 years both males and females. The main causes of death for men aged 25 to 39 were homicide and accident, while for women they were accident and malignant neoplasm and liver diseases (Rico.P, 1999). Hence, age under 25 to 39 can be categorized a younger adults a higher for AIDS, because several causes such demography and epidemiology basically influence families under that age.

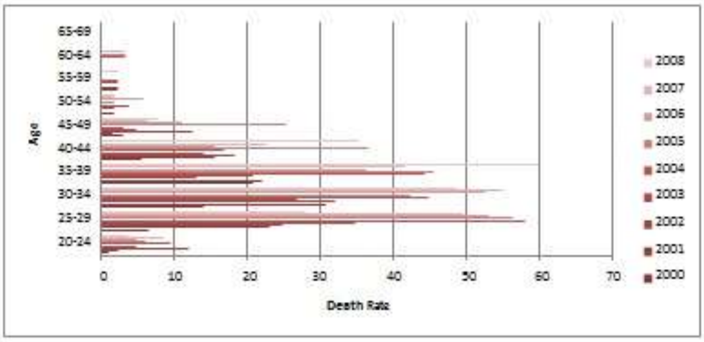


Figure 1: Death of rate 2000-2008

## III. Negative Binomial and Poisson Regression

### 3.1 Standardization of rate

Generally, to estimation of rates in various discipline whether of percentage, crude birth rate, frailty rate and act. The population data are used to study and comparability statistical analysis on disease incidence, prevalence, and mortality. Rates can be calculated by dividing the total number of events such death, incident cases, in the population in a specified year. It usually indicates per 1000 or 100,000 population whereby consider the actual experience of population and should always be examined when assessing the mortality of a population. Considering in this research, age specific rate are estimated as the number of event over a given time period in a specified age group divided by the population in that age group over the same time period typically express as a per 1000 or 100,000 population per year.

While, the comparison of age specific rate is the most extended and reliable method of comparing rates over time or between populations groups ( Bains.N, 2009). To shown a calculation, estimation, presentation, and interpretation of rates for each group of results in a large numbers of likeness which is heavy. Connecting with discrete distribution model for discrete method of age whereby specific rates of the study population such as population of interest are utilized to a randomly chosen standard population. This represent the expected number of events whereby producing the numbers of event that would be expended in the standard population, if it had experienced the same age specific event as a study population. Death rate is obtaining by divided the expected numbers of events by the total standard population. To state the requirement for calculation of rates are (Bains.N, 2009);

1. Number of events such as death, incident cases for each group in the study.
2. Population for each group in the standard population.
3. Constant population 1000 or 100,000.

To incorporate into negative binomial regression model we employ a death rate function to dependent variable. Rate dependent variables are estimated by requirement as follows. Let assume mortality rate cases in the $j^{th}$ observation for $j = 1,2,\dots,n$ a categorical observation age rate death estimation, whereby supposed to be negative binomial distributed. To calculated mortality rate, assuming $d_j$ is the expected death rate cases. Age death rate normally was calculated using standard population rate (Bains.N, 2009). Thus, $d_j$ formula describe as follows;

$$d_j = \frac{q_j e_j}{p_j} \qquad (1)$$

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.

Addition, in a count data $y_i$ is a count observation instead of $i = 1,2,\dots,n$ . Here, we replace $i$ with $d_j$ observation to shown that $d_j$ is a rate observation represent the calculation death rate based on age categorical. This function generated by SAS version 9.2 via PROC GENMODE procedure.

### 3.2 Negative Binomial Dependent Death Rate (NBDR)

Similar with NBDR, substitute equation (1) for $y_i$ observation to $y_{dj}$ death rate observation. Thus, the equation dependent death rate negative binomial is expressing such;

$$P(Y_{dj} = y_{dj}) = \frac{\Gamma\left(y_{dj} + {}^{1}/_{\psi}\right)}{y_{dj}!\left({}^{1}/_{\psi}\right)}\left[\frac{1}{1+\psi\theta_i}\right]^{1/\psi}\left[\frac{1}{1+\psi\theta_i}\right]^{y_{dj}} y_{dj>0} \qquad (2)$$

where $y_{dj}$ 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 $\theta_i = exp\ (x'_\beta)$ where $\beta$ is a $(P+1)\times 1$ vector of covariates and intercept of $\beta_0$, the coefficient for regression $\beta_0, \beta_1, \beta_2, \beta_3\ \dots,\ \beta_p$ . Taking the exponential of $x'_\beta$ ensure that the mean parameter $\theta_i$ is nonnegative.

### 3.3 Poisson Death Rate (PDR)

Poisson Regression is a famous technique in statistical filed to analysis a count data distribution the mostly use in many setting in areas of interest. Poisson distribution interested in modeling rates of this event such as infection diseases. Poisson regression modeling providing a formal way to evaluate possible association between dependent rate of age $(Y)$ and a factor affected $(X)$ on mortality rate AIDS co-infection patients. To account Poisson based on the equation (1), again substitute equation (1) for $y_i$ count observation to $y_{dj}$ death rate observation. Thus, the equation dependent death Poisson is expressing such;

$$P\left(Y_{dj} = y_{dj}\right) = \frac{(\exp(-\mu_i)\,(\mu_t)^{y_{dj}})}{y_{dj}!} \qquad y_{dj>0} \quad (3)$$

Basically the main assumption under the Poisson model is that expected value of the covariates age categorical death rate for subject ($d_j$) equal to mean and variance $\mu = E\left(Y_{dj}\right) = Var\left(Y_{dj}\right)$ . To incorporate covariate via the Poisson death rate according the equation express $\mu_i \;=\; exp(\beta_0 \;+\; \sum_{j=1}^{n} X_{ij}\,\beta_j)$ where $\mu_i$ is the number of mortality death rate to be expected $X_{i1}\,X_{i2}\,X_{i3}\,\ldots X_{in}$ are the values of the covariates during that time period and the $\beta_0, \beta_1, \beta_2, \beta_3\;\ldots\;,\;\beta_n$ are the coefficient to be estimated by the modeling.

## IV. Analysis and Results

Using PROC GENMODE in SAS procedure can fit wide range in generalized linear model (Castelloe.J.M, 2000). To asses of adequate model, (McCullagh.P. and Nelder.J.A, 1989) express deviance residual and Pearson chi- square $\chi^2$ residual is a popular technique choosing appropriate model was used. Consider both of tools, statistical model is correct than both quantities are asymptotically distributed as $\chi^2$ statistic with $(n - p)$ degrees of freedom when a sample size and $p$ (the number of fitted parameter including the intercept)(White.G.C. and Bennetts.R.E, 1996; Jain.G.C. and Consul.P.C, 1971). Refer that assumption, if the model is adequate the expected values of both deviance and Pearson chi-square $\chi^2$ is equal or closed to $(n - p)$ ( the scaled deviance closed to 1) or scaled Pearson chi-square which is $\chi^2/_{df}$ is approximation to 1.The deviance residuals express as follows;

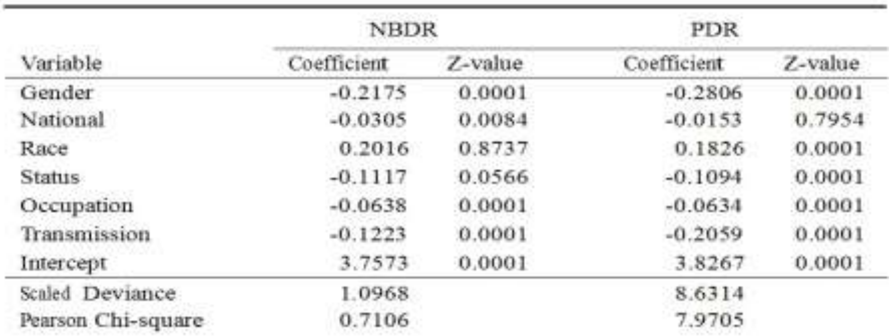
$$r_i^D = sig(y_{dj} - \mu_i)\sqrt{d_i} \qquad (4)$$

where $d_j$ is the contribution of the $i^{th}$ subject to deviance with total deviance given by $D = \sum\ (r_i^D\ )^2$ corresponding to Pearson residuals, such as;

$$r_i^D = \frac{(y_{dj} - \widehat{\mu_i})}{\sqrt{V(\widehat{\mu_i})}} \qquad (5)$$

Thus, so that $\chi^2 = \sum\ (r_i^D\ )^2$. Subsequently ,the deviance residual are more commonly used because their distribution tendency to be closer to normal than that of the Pearson residuals. Table (2) show there result negative binomial and Poisson death rate for AIDS co-infection patients.

Table II shows the results of negative binomial (NBDR) and Poisson (PDR) death rate model and estimate coefficient of all two models together with related statistic are listed. While NBDR regression provides a baseline model, and the NBDR is demonstrate the better fit than the PDR. Thus, it is interesting to see that all although NBDR and PDR have a different assumption and specification, they all indicate that the information related all covariates in the results above. The results indicate that the values Pearson chi-square $\chi^2$ of NBDR less than PDR (0.7106 and 7.9705) respectively. According this values, we can indicate that NBDR is appropriate than PDR. While corresponding for over-dispersion, NBDR is less than PDR and lead to summarize a NBDR is adequate regression model.



| | NBDR | | PDR | |
|---|---|---|---|---|
| Variable | Coefficient | Z-value | Coefficient | Z-value |
| Gender | -0.2175 | 0.0001 | -0.2806 | 0.0001 |
| National | -0.0305 | 0.0084 | -0.0153 | 0.7954 |
| Race | 0.2016 | 0.8737 | 0.1826 | 0.0001 |
| Status | -0.1117 | 0.0566 | -0.1094 | 0.0001 |
| Occupation | -0.0638 | 0.0001 | -0.0634 | 0.0001 |
| Transmission | -0.1223 | 0.0001 | -0.2059 | 0.0001 |
| Intercept | 3.7573 | 0.0001 | 3.8267 | 0.0001 |
| Scaled Deviance | 1.0968 | | 8.6314 | |
| Pearson Chi-square | 0.7106 | | 7.9705 | |

Table 2: Results of NBDR and PDR model on death of age rate (AIDS).

## V. Discussion

Negative Binomial and Poisson regression model are closed distribution used when absence of over-dispersion. To handle over-dispersion or heterogeneity there are several approaches. Existing over dispersion in poison distribution a popular choice is to use the negative binomial; assuming a gamma density for an observed weakness factor. This method does not concede an irregular distribution such as that with an excess amount with zeroes. A better way is to excess for with extra zeros by addendum the frailty distribution with a probability mass at the lower end point and treat as a nuisance parameter. The over-dispersion in the analysis would lead to the underestimation of the standard error and consequently will affect the level of significant in the hypothesis testing. Thus, to choose appropriate model for count data observation could change a statistical inference(Alex.P, 1998).However, model selection approaches with spatial emphasis on information criteria such (AIC,BIC and others). Selection of suitable models has become a common approach to interpretation of complex results.

To conduct this issue with suitable normalization for follow up time, it is possible to apply mixture models with count date to analysis tumor return clinically; all cancer patient will actually suffer a return. Follow a view like this, a hurdle model would be more appropriate existing approach. When a model fit the data very well it seduces to recommend that the concepts with disease mapping process behind with model are correctly.

**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 Ministery of Higher Education and Universiti Tun Hussien Onn Malaysia, were supported this research. Thanks to medical record Hospital Universiti Sains Malaysia provided the data throught ethical approved on 15 Jun 2011 (USMKK/PPP/JEPeM [238.4(1.12)])

## References

[1] Alex.P. Analysis of count data using the sas system. *Paper Presented at the SUGI conference,Long Beach, Carlifornia*, 1998.

[2] Bains.N. Standardization of rates. Technical report, Associtaion of public health epidemologists in Ontario (APHEO), March 2009.

[3] Population Reference Bureau. Hiv/aids and older adults in the United States. *Today's Reserach on Aging*, 18:1–7, 2009.

[4] Cameron and Trevedi. *Regression Analysis for Count Data*. Cambridge Uni- versity Press,New York, 1998.

[5] Castelloe.J.M. Sample size computations and power analysis with sas system, 2000.

[6] Chunag.K.J., Chan.C.C., Lee.C.T., and Tang.C.S. The effect of urban air pollution on inflammation, oxidative stress, coagulation and autonomic dysfunction in young adults. *AM Journal Respiration Care Medicine*, 176: 370–376, 2007.

[7] Jain.G.C. and Consul.P.C.A generalized negative binomial distribution.

[8] *Siam Journal Application Mathematics*, 21(4):501–513, 1971.

[9] McCullagh.P. and Nelder.J.A. *Generalized Linear Models*. Chapman and Hall, 2nd edition edition, 1989.

[10] Michelozzi.P., Forastiere.F., D.Perucci.CA., Ostro.B., and Ancona.C. Air pollution and daily mortality in rome, italy. *Occupation Environment Medicine*, 55(9):605–610, 1998.

[11] Morgan.G., Corbett.S., Wlodarczyk.J., and Lewis.P. Air pollution and daily mortality in sydney,australia, 1989 through 1993. *American Journal of Public Health*, 88(5):759–764, 1998.

[12] Rico.P. Who, health situation analysis and trend summary. *Demography*, 1: 296–315, 1999.

[13] White.G.C. and Bennetts.R.E. Analysis of frequency count data using the negative binomial distribution. *Journal the ecological society of america*, 77(8):2549–2557, Dec 1996.