ScienSage

Journal of Advanced Scientific Research

Available online through http://www.sciensage.info/jasr

ISSN
0976-9595
Research Article

APPLICATIONS OF ZERO INFLATED MODELS FOR HEALTH SCIENCES DATA

Wan Muhamad Amir W Ahmad[1], Siti Aisyah Abdullah[2], Kasypi Mokhtar[3], Nor Azlida Aleng*[4], Nurfadhlina Halim[5] and Zalila Ali[6]

[1,2,4,5]*School of Informatics and Applied Mathematics, Universiti Malaysia Terengganu (UMT), 21030 Terengganu, Malaysia*
[3]*School of Maritime Business and Management, Malaysia Terengganu (UMT), 21030 Terengganu, Malaysia*
[6]*School of Mathematics Sciences, Universiti Sains Malaysia (USM), 11800 Pulau Pinang, Malaysia*
**Corresponding author: wmamir@umt.edu.my*

ABSTRACT

Pneumonia is an infection of one or both lungs which is usually caused by bacteria, viruses, or fungi. Each year, pneumonia attack kills about 1.4 million people in the world, especially among children who are also the main sufferers of the disease. The aim of this study was to examine the factors that are associated directly or indirectly in pneumonia patients among the children. In this present paper, we have considered several regressions model to fit the count data that encounter in the field of Health Sciences. We have fitted Poisson, Negative Binomial (NB), Zero-Inflated Poisson (ZIP) and Zero-Inflated Negative Binomial (ZINB) regressions to pneumonia data. To compare the performance of these models, we analysed data with moderate to high percentage of zero counts. Because the variances were almost two times greater than the means, it appeared that both NB and ZINB models performed better than Poisson and ZIP models for the zero inflated and overdispersed count data. From the results of the ZINB regression can overcome overdispersion so it was better than the Poisson regression model.

Keywords: Poisson regression, Negative Binomial Regression, Overdispersion, Zero-Inflated Poisson, Zero-Inflated Negative Binomial.

1. INTRODUCTION

Count data is very communal in several fields such as in biomedical science, public health and marketing. The Poisson distribution has been used to model the count data for a long time. It has an important constraint that the mean and variance are equal. However, many processes in real life are over dispersed (variances are greater than means) and violate the underlying assumption of Poisson distribution. In that cases the negative binomial (NB) distribution is a natural and more flexible extension of the Poisson distribution and allows for over-dispersion compared to Poisson distribution. Several researchers have suggested using the NB regression model as an alternative to the Poisson regression model when the count data are over or under dispersed. Both Poisson and Negative Binomial distribution have been used for predicting the health sciences count frequencies [1-4]. It is noted that most of the health sciences data contain excess number of counts with zero. Unfortunately, the Poisson and NB models do not address the possibility of zero counts and cannot fit the data properly. Then corresponding inflated models, say zero inflated Poisson (ZIP) and zero inflated negative binomial (ZINB) are very useful to describe the zero inflated count data. Both ZIP and ZINB models incorporate extra variation than the corresponding Poisson and NB models [5]. The most appropriate reference for ZIP regression model are [6, 7], ZINB regression model are [8, 9] among others. To select an appropriate inflated model, that is, ZIP model over Poisson or ZINB model over NB [10] statistics is one ofthe popular tests. Besides modeling crash data, these four models have been used in environmental science by [11] in biomedical science [12] among other discipline. The main objective of this paper is to provide a comprehensive review of these four models and discuss how to fit appropriate statistical models for count data using Statistical software, especially for the overdispersed and an excess number of counts in the data.

2. MATERIAL AND METHODS

2.1. Poisson Regression Approach

According to the study researchers, Poisson regression model is a non-linear regression model and widely used in the field of medicine or epidemiology. Regression analysis is a technique used nonlinear regression to model the dependent variable that reflects the data count or discrete [13]. The Poisson distribution is a discrete probability distribution that important when n is large and p is small, and when the independent variables occur within a period [14]. The Poisson probability distribution, $P(y;\mu)=\frac{e^{-\mu}\mu^{y}}{y!}$, has the same mean and variance (equidispersion), $Var(y)=E(y)=\mu$. As the mean of a Poisson distribution increases, the probability of zeros decreases and the distribution approximates a normal distribution [15]. The Poisson distribution also has the strong

assumption that events are independent. Thus, this distribution do not fit well if μ differs across observations (heterogeneity) [10]. The Poisson regression model (PRM) incorporates observed heterogeneity into the Poisson distribution function, $Var(y_i \mid x_i) = E(y_i \mid x_i) = \mu_i = \exp(x, \beta)$. As μ increases, the conditional variance of y increases, the proportion of predicted zeros decreases, and the distribution around the expected value becomes approximately normal [10]. The conditional mean of the errors is zero, but the variance of the errors is a function of independent variables, $Var(\varepsilon \mid x) = \exp(x\beta)$. The errors are heteroscedastic. Thus, the (PRM) rarely fits in practice due to overdispersion [16, 17].

Let x be $n \times (p+1)$ matrix of explanatory variables. The relationship between $y_i$and i^{th} row vector of x, $x_{i,}$ linked by $g(\mu_i)$ is the canonical link function given by: $E(y_i) = \mu_i = e^{x_i^T \beta}$, where, $x_i = (x_{i0}, x_{i1}, ..., x_{in})^T$ and $\beta = (\beta_0, \beta_1, ..., \beta_p)^T$. The standard estimator for Poisson model is the maximum likelihood estimator (MLE). To find the MLE, we define the likelihood function as follows:

$$L(\beta) = \prod_{i=1}^{n} \frac{e^{-\mu} \mu_i^{y_i}}{y_i!} = \prod_{i=1}^{n} \frac{e^{-e^{x_i^T \beta}} (e^{x_i^T \beta})^{y_i}}{y_i!} \quad (1)$$

Taking log on both sides and we get,

$$\ell(\beta) = \sum_{i=1}^{n} [y_i \ln \mu_i - \mu_i - \ln y_i!] = \sum_{i=1}^{n} [y_i x_i^T \beta - e^{x_i^T \beta} - \ln y_i!] \quad (2)$$

It can be verified that the first two partial derivatives of the log-likelihood function exists and are given as follows:

$$\frac{\partial \ell(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} (y_i - \mu_i) x_{ij} = \sum_{i=1}^{n} (y_i - e^{x_i^T \beta}) x_{ij}$$

$$\frac{\partial^2 \ell(\beta)}{\partial \beta_j \beta_k} = -\sum_{i=1}^{n} \mu_i x_{ij} x_{ik} = -\sum_{i=1}^{n} e^{x_i^T \beta} x_{ij} x_{ik}$$

Hence, equation (1) or (2) is nonlinear in β so that they need to be solved by using an iterative algorithm. The iterative algorithms commonly used are either Newton-Raphson or Fisher scoring. In practice $\hat{\beta}$ is the solution of the estimating equations obtained by differentiating the likelihood in terms of β and solving them to zero. Therefore, β will be obtained by maximizing (2) using a numerical iterative method [13]. Poisson regression is a powerful analysis tool but as with all statistical methods, it can be used inappropriately if its limitations are not fully understood. There are three problems that might exist in Poisson regression analysis. There are the data might be truncated or censored, the data might contain excess zeroes and the the mean and variance are not equal, as required by poisson distribution. This problem is called overdispersion.

2.2. Negative Binomial Regression

The Negative Binomial (NB) distribution can be obtained from the mixture ofPoisson and Gamma distribution and is expressed as

$$p(y_i \mid x_i) = \frac{\Gamma(y_i + 1/\alpha)}{y_i! \Gamma(1/\alpha)} \left[\frac{1}{1+\alpha\theta_i}\right]^{1/\alpha} \left[\frac{\alpha\theta_i}{1+\alpha\theta_i}\right]^{y_i}$$

for $y_i = 0,1,2,3$

where, $y_i$is the number of crashes for road segment i,q i is the expected numberof crashes per period, which can be expressed as

$$\theta_i = \exp(x'_i \beta),$$

The mean and variance of negative binomial distribution are respectively, $E(y_i \mid x_i) = \theta_i$ and $Var(y_i \mid x_i) = \theta_i[1 + \theta_i \alpha] > E(y_i \mid x_i)$. Thus, the NB model is also overdispersed and allows extra variation relative to the traditional Poisson model. It has more desirable properties than the Poisson model to describe the relationship in health sciences data. The variance of NB is significantly greater than the mean. Here a represents an ancillary or dispersion parameter which indicate the degree of over dispersion. If a = 0, the NB regression model reduces to traditional Poisson regression model. Many researchers in different fields have considered both Poisson and Negative Binomial models [18-25] to mention a few. However, when excess zero occur, both Poisson and NB regression models are not that useful to fit the zero inflated models [26]. In that case both ZIP and ZINB models are appropriate choice.

2.3. Zero-Inflated Poisson Model

Consider the ZIP model, which is denoted by $\Pr(Y_i = y_i)$, in which the response variable $Y_i = (1,2,..., n)$ has a probability mass function (pmf) given by

$$\Pr(Y_i = y_i) = \begin{cases} \omega_i + (1-\omega_i) e^{-\theta_i}, & y_i = 0, \\ (1-\omega_i) \dfrac{\theta_i^{y_i} e^{-0_i}}{y_i!}, & y_i > 0, \end{cases} \quad (3)$$

Where $0 \le \omega_i < 1$ and $\theta_i > 0$. The random variable Y_i has a Poisson (θ_i) distribution when $(\omega_i) = 0$. The parameters θ_i and ω_i depend on vectors of covariates X_i and z_i, respectively. The ZIP model is given by $\log(\theta_i) = x_i^t \beta$, $\log(\frac{\omega_i}{1-\omega_i}) = z_i^t \gamma$ and the mean and variance ZIP model are given by $E(Y_i) = (1-\omega_i)\theta_i$, $Var(Y_i) = (1-\omega_i)\theta_i(1+\omega_i\theta_i)$

2.4. Zero-Inflated Negative Binomial Model

For zero-inflated and overdispersed data a frequent modelling choice is the Zero-Inflated Negative Binomial (ZINB) model. The response variable $Y_i (i = 1,2,..., n)$ has a pmf given by

$$Pr(Y_i = y_i) = \begin{cases} \omega_i + (1-\omega_i)(1+\kappa\theta_i^c)^{-\theta_i^{1-c}/\kappa}, \\ y_i = 0, \\ (1-\omega_i) \dfrac{\Gamma(y_i + \theta_i^{1-c}/\kappa)}{y_i! \Gamma(\theta_i^{1-c}/\kappa)} (1+\kappa\theta_i^c)^{-\theta_i^{1-c}/\kappa} (1+\theta_i^{-c}/\kappa)^{-y_i}, \\ y_i > 0 \end{cases} \quad (4)$$

Where $0 \le \omega_i \le 1$ and $\theta_i > 0$, κ is the dispersion parameter with $\kappa > 0$ and $\Gamma(.)$ is the gamma function. The mean and the variance of the model are defined as

$E(Y_i) = (1-\omega_i)\theta_i$ and $Var(Y_i) = (1-\omega_i)\theta_i(1+\theta_i\kappa^{-1}+\omega_i\theta_i)$. The response variable Y_i has a negative binomial distribution with mean θ_i and dispersion parameter κ when $\omega_i = 0$. Ridout, et al. (2001) fitted various models to these data on the basis of the Poisson and negative binomial distributions and their zero-inflated counterpart [15].

2.5. Parameter Estimation and Model Selection

2.5.1. Parameter Estimation

The maximum likelihood method has been considered due to limitation of the application of SAS, which consider the maximum likelihood estimation (mle) technique. To evaluate the model, it is necessary to examine the significance of the variables included in the model. For a better model, the estimated regression coefficients have to be statistically significants. Usually, the *t* test is used to determine the significance of the regression coefficients. Moreover, the intuitive judgment of the experimenters should be considered.

2.5.2. Goodness of Fit Test

After fitting some models to the data, it is essential to check the overall fit as wellas quality of the fit. The quality of the fit between the observed values (y) and predicted values ($\hat{\mu}$) can be measured by various test statistics; however, the one of the useful statistic is called deviance and defined as:

$$D(y:\hat{\mu}) = -2[L(\hat{\mu};y) - L(y;y)]$$

For a better model, one would expect smaller value of the $D(y:\hat{\mu})$.

2.5.3. Selecting Best Model

Akaike's information criterion (AIC, Akaike 1973) [27] was used to compare thedifferent models. The AIC is defined as

$$AIC = -2L + 2k,$$

Where, L is the log-likelihood and $k < p$ is the number of parameters in the model.For the best fitted model one must expect lowest AIC value.

2.6. Conceptual Framework

In NBR model, the parameter estimated are converged by considering the effect that stems from overdispersion. Basically count observation might have excessive zero than expected. In such case ZIP regression model is an appropriate approach to analyze the dependent variable having to much zero observation [28]. 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 [29]. In such cases, when ZIP existing overdispersion and highly accessing zero such mentioned above, ZINB is an alternative method that will be used.

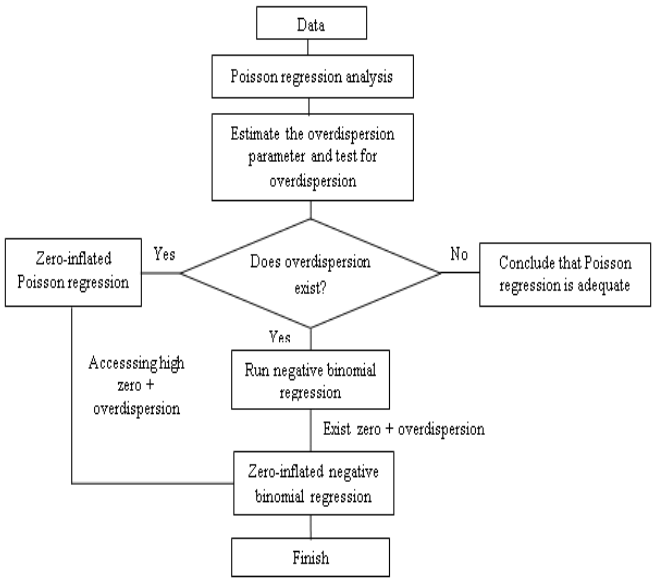

Fig. 1: The frequently used models in the count data analysis framework

Table 1: Explanation of the Variables

Variable Name	Code	Explanation of the variables	Categorical
Diagnosed	Y	The count of disease	
Age	X_1	Age of patients in year	
Sex	X_2	Gender	0=Female, 1=Male
DM	X_3	Diabetes Mellitus	No ,1= Yes
Influenza	X_4	Influenza status	0=No ,1=Yes
Septicemia	X_5	Septicemia status	0=No ,1=Yes
Diarrhea	X_6	Diarrhea status	0=No ,1=Yes
Asthma	X_7	Asthma status	0=No ,1=Yes
Anemia	X_8	Anemia status	0=No ,1=Yes
Tuberculosis	X_9	Tuberculosis status	0=No ,1=Yes
Blood fever	X_{10}	Blood fever status	0=No ,1=Yes
Acute tonsillitis	X_{11}	Acute tonsillitis status	0=No ,1=Yes
Streptococcus Pneumonia	X_{12}	Pathogen Streptococcus Pneumonia status	0=No ,1=Yes
Acute Bronchitis	X_{13}	Acute Bronchitis status	0=No ,1=Yes

2.7. Study Design And Location

We conducted a retrospective, cohort study at medical record Hospital Putrajaya, Malaysia. It consists 1252 patients beginning 2005 untill 2009. The ethics committee at Hospital Putrajaya approved access to patient records. Patients confidentially have been maintained.

2.8. Patients And Data Collection

We studied patients with pneumonia disease patients who attended the Family Health Clinic in Hospital Putrajaya, Malaysia. Data from 1252 respondents (patients) were collected between ages 0 until 12 years old and diagnosed to

have pneumonia. Material of this study is a hypothetical sample which is composed of eleven variations. Namely variables are as in Table 1.

A total of 12500 registered pneumonia patients were registered from 2005 to 2009, among which 1252 met the inclusion /exclusion criteria in Table 2.

Table 2: Inclusion and exclusion Criteria

S. No.	Inclusion Criteria
1	The subject is diagnosed with pneumonia
2	From Malaysia population
3	Other disease diagnosed: diabetes, influenza, septicemia, diarrhea, asthma, anemia, tuberculosis, health services, acute tonsillitis, streptococcus pneumonia, heart disease
4	Age range: 0-12 years
	Exclusion Criteria
1	Age 13 years and above
2	Accident patients
3	Any other conditions as evaluate by the physician

2.9. Statistical Analyses

To show the utility of the developed approach, we applied the poisson regression model, negative binomial regression model, ZIP and ZINB real data set in underlying Pneumonia disease among children. Pneumonia is a lung infection which is usually caused by bacteria, viruses, or fungi. Each year, pneumonia attack kills about 1.4 million people in the world, especially among children who are also the main sufferers of the disease. This study involved a sample of 1251 patients with Pneumonia, representative for children age between 0-12 years. Consider a dataset on children; we count how many diagnosis children aged 0-12 years had while suffering from Pneumonia. Patient's data with Pneumonia disease were collected at the Medical Record Unit in Hospital Putrajaya in Malaysia for year 2005 to 2009. For the purpose of this study, we have employed the count data. The diagnosis is considered as the response variable, and the selected variables are: diabetes, influenza, septicemia, diarrhoea, asthma, anemia, tuberculosis, health services, acute tonsilitis, streptoccocus pneumonia, and heart disease. The main objective of this paper is to handling overdispersion in a dataset, it is a one-sided test and the level of significance was set as $\alpha = 0.05$. The analysis of these data was performed in SAS 9.3, by using *proc genmod*. In counting the number of response to an exposure, the patient may have no diagnosis response because of their immunity or resistence to the disease.

3. RESULTS AND DISCUSSION

Descriptive statistics for the variable diagnosis, diabetes, influenza, septicemia, diarrhoea, asthma, anemia, tuberculosis, blood fever, acute tonsilitis, streptoccocus pneumonia, and heart disease used in the present study are given in table below. The almost 62.3% observation out of 1252 observation used in the study were zero valued among the variable used. The overdispesion might exist due to many excess zeros for the case when $y = 0$ because 62.3% of observed counts are zeros. Frequency (percent) of patients who received a different types of diagnosis is a total of 283 (22.6%), while patients who received two different types of diagnosis the frequency (percent) was 142 (11.3%). for patients who received three different types of diagnosis was 31 (2.5%) and patients who received four different types of diagnosis was 16 (1.3%).The data has too many zeros and over-dispersed which led us to apply the zero-inflation model [30].

Table 3: Diagnosis count from 1252 Pneumonia Patient in Hospital Putrajaya

Diagnosis	Patient	
	Frequency	Percent (%)
0	780	62.3
1	283	22.6
2	142	11.3
3	31	2.5
4	16	1.3
Total	1252	100.0

To make comparison among two models, two methods can be used. One is by using the deviance statistic and the otheris using the Wald statistic. However, only deviance statistic method will be discussed here. Table 5 summarizes log likelihood and deviance for Poisson regression and ZIP.

Table 5: Log likelihood and deviance for Poisson and ZIP

Model	Degrees of Freedom	Log likelihood	Deviance
Poisson	1251	-1120.5393	1568.5737
ZIP	1236	-611.5493	1804.1870

The following hypothesis need to be tested:

$$H_0 : F(y) = F^*(y) \text{ (Poisson regression model)}$$

$$H_1 : F(y) \neq F^*(y) \text{ (ZIP regression model)}$$

The following statistic is used:

$$d = D_0 - D_1$$

Where D_0 indicates the deviance of the less inclusive model while D_1 indicates the deviance of the more inclusive model. This statistic has an approximate chi-squared distribution with degress of freedom equal to the difference between the numbers of unknown parameters in the two models. Thus H_0 will be rejected if $d > \chi^2_{0.05}(df)$.

For comparing Poisson and ZIP,

$d = 1804.1870 - 1568.5737 = 228.8157$ and

$df = 1251 - 1236 = 15$

From statistical table, $\chi^2_{0.05}(15) = 25.0$. Since 228.8157 >25.0, H_0 is rejected at significance level $\alpha = 0.05$. Therefore, ZIP model is the best model to pneumonia data.

Vuong test

We want to test the following hypothesis:

H_0 : Two distributions functions are equivalent

H_1 : Two distributions are different

Table 6: Vuong test between NB and ZINB

Vuong Statistics	*Z*	*p*-value	Preferred model
Unadjusted	1.31	0.10	ZINB
Akaike Adjusted	1.31	0.10	ZINB
Schwarz Adjusted	1.31	0.10	ZINB

Clarke Sign Test

H_0 : Models are equally close to the true model

H_1 : One of the models is close to the true model

Table 7: Clarke sign test between NB and ZINB

Clarke Statistics	*M*	*p*-value	Preferred Model
Unadjusted	276.0	<.0001	ZINB
Akaike Adjusted	276.0	<.0001	ZINB
Schwarz Adjusted	276.0	<.0001	ZINB

The output above shows the Vuong test followed by the Clarke Sign Test. This test compares the zero-inflated negative binomial model to a standard negative binomial model [10]. Because the z-value is significant and *p*-value is small, the Vuong test and Clarke Sign Test shows that the zero-inflated negative binomial is a better fit than the standard negative binomial.

Table 8: Summary of different four methods, Poisson, NB and Zero-Inflated Model Fit

Variable	Estimate (Standard Error)			
	Poisson	NB	ZIP	ZINB
Diarrhoea	0.9740* (0.4261)	1.0325*(0.4228)	0.0371 (0.2474)	0.1060 (0.2472)
Influenza	1.1733* (0.1067)	1.2085*(0.1067)	1.1598* (0.0984)	1.1038* (0.1010)
Septicemia	1.1436* (0.1505)	1.1793*(0.1503)	1.1318* (0.1401)	1.0883* (0.1409)
Diabetes	0.9984* (0.2046)	1.0380* (0.231)	1.2557* (0.1737)	1.1777* (0.1764)
Asthma	0.9410* (0.0969)	0.9354* (0.0971)	1.0159*(0.0863)	0.9635* (0.0883)
Anaemia	0.7006* (0.1535)	0.7564* (0.1511)	0.7532* (0.1460)	0.7595* (0.1460)
Tuberculosis	0.7067* (0.1207)	0.6803* (0.1194)	0.5608* (0.1239)	0.5482* (0.1236)
Blood fever	0.7766* (0.0926)	0.7936* (0.0929)	0.6744* (0.0890)	0.6386* (0.0903)
Tonsillitis	1.0989* (0.1182)	1.0413* (0.1161)	1.0684* (0.1130)	0.9974* (0.1165)
Meningitis	1.0415* (0.0845)	1.0316* (0.0848)	1.0683* (0.0789)	1.0738* (0.0790)
Bronchitis	1.0796* (0.0885)	1.0113* (0.0865)	0.9738* (0.0866)	1.0164* (0.0884)
alpha		1.05367E-8		1.0543383E-8
Log likelihood	-1120.5393	-1090.9977	-1084.1983	-815.9125
AIC[a]	2652.7191	2595.6455	2582.0373	1691.8249
BIC[b]	2657.8516	2605.9009	2592.3023	1845.7998

* *Significantly (p<0.05); a = Akaike's information criterion; b = Bayesian information criterion*

Finally, we summarize the performance of each of the four fitted models when fitted to each of the four types of generated data. The data generated by the Poisson distribution can be predicted equally well by each of the four models that we consider. The data generated by the zero-inflated Poisson can be predicted most accurately using either a zero-inflated Poisson or a zero-inflated negative binomial model. The negative binomial model performs next best. The Poisson model fares the worst: it significantly underpredicts the number of zeros and overpredicts the number of ones. The data generated by the negative binomial process can be predicted equally well by either a negative binomial or a zero-inflated negative binomial model. These models are followed by the zero-inflated Poisson and the Poisson. The data generated by the zero-inflated negative binomial model can be predicted best by a zero-inflated negative binomial, followed by a negative binomial, a zero-inflated Poisson and a Poisson. Notice that the Poisson model provides the worst fit in all cases other than in the case of Poisson-generated data. Thus, a Poisson model should be used only in cases where there is strong evidence that it is the correct specification. As long as data sample is reasonably large, a slight loss of efficiency is, on average, more preferable compared to model misspecification.

4. CONCLUSIONS

Through a comparison of the model, it was found that the ZINB model is preferred to use instead of the Poisson regression model, Negative Binomial regression and ZIP because the model has a value of AIC is the smallest. AIC values for the Poisson regression model, NB, ZIP and ZINB respectively are 2652.7191, 2595.6455, 1836.1870 and 1691.8249.Also*p*-value for the Vuong test is positive and statistically significant. Therefore, ZINB model more suitable for modeling the patient data pneumonia in children. The estimate ZINB regression is given by the following expression:

$$\ln[E(\text{diagnosis}_1)] = \beta_0 + \beta_1 \text{influenza} + \beta_2 \textit{septicemia} + \beta_3 \textit{diabetes} + \beta_4 \text{asthma} + \beta_5 \textit{anemia} + \beta_6 \textit{tuberkulosis} + \beta_7 \text{blood fever} + \beta_8 \textit{tonsilitis} + \beta_9 \textit{meningitis} + \beta_{10} \textit{bronchitis}$$

$$\mu = \exp(-1.4117 + 1.1038X_1 + 1.0883X_2 + 1.1777X_3 + 0.9635X_4 + 0.7595X_5 + 0.5482X_6 + 0.6386X_7 + 0.9974X_8 + 1.0738X_9 + 1.0164X_{10})$$

Since pneumonia is a disease that has a high correlation with other diseases except for diarrhoea disease, it is suggested that intensive treatment should be given prominence to the babies since the beginning of the birth of the risks of the disease can be minimized. Next risk of infant mortality can be reduced. Factors obtained can be analyzed and verified with other procedures such as multivariate analysis, logistic regression, structured modeling (SEM) and so on.

5. REFERENCES

1. Shankar V, Mannering F, Barfield W. *Accident Anal. And Prevention*, 1995; **27(3)**:371-389.
2. Poch M, Mannering F. *Journal of Transportation Engineering,* 1996; **122 (2)**: 105-113.
3. 3. Milton J, Mannering F. *Transportation,* 1998; **25 (4)**: 395-413.
4. Lee J, Mannering F. *Accident Analysis and Prevention,* 2002; **34 (2)**: 149-161.
5. Mullahy J. *Journal of Econometrics,* 1986; **33(3)**:341-365.
6. Lambert D. *Technometrics*, 1992; **34**:1-14.
7. Lee AH, Wang K, Yau KKW. *Biometrical Journal,* 2001; **43(8)**: 963-975.
8. Cameron AC, Trivedi PK. Regression Analysis of Count Data. New York, Cambridge University Pres. 1998.
9. Long JS. Regression Models for Categorical and Limited Dependent Variables, *T*housand Oaks, CA: Sage Publications, 1997.
10. Vuong QH. *Econometrica,* 1989; **57**:307-334.
11. Warton DI. *Environmetrics,* 2005; **16 (2)**:275-289.
12. Yau KKW, Wang K. Lee AH. *Biom. J.*, 2003; **45**: 437-452.
13. McCullagh P, Nelder JA. Generalized Linear Models. 2nd Edition. London: Chapman & Hall, 1989.
14. Long JS, Jeremy F. Regression Models for Categorical Dependent Variables Using STATA 2nd ed. College Station, TX: STATA Press, 2006.
15. Ridout M, Hinde J, Demetrio CGB. *Biometrics*, 2003; **57**:219-233.
16. Jansakul N. *Proceeding 20th International Workshop on Statistical Modelling, Sydney, Australia* 2005; 277-284.
17. Long JS, Freese J. Regression Model for Categorical Dependent Variable Using Stata, A Stata Press Publication, USA, 2005; 277-284.
18. Miaou SP. *Accidents Analysis and Prevention,* 1994; **26**:471-482.
19. Karlaftis MG Tarko AP. *Accidents Analysis and Prevention,* 1998; **30:** 425-433.
20. Hauer E. *Accidents Analysis and Prevention,* 2001; **33**:799-808.
21. Lee AH, Stevenson MR, Wang K, Yau KKW, et al. *Accidents Analysis and Prevention,* 2002; **34:**515-521.
22. Byers AL, Allore H, Gill TM, Peduzzi P, et al. *Journal of Clinical Epidemiology,* 2003; **56:**559-564.
23. Berhanu G. *Accidents Analysis and Prevention,* 2004; **36:**697-704.
24. Yau KKW, Lee AH, Carrivick PJW. *Computer Methods and Programs in Biomedicine,* 2004; **74:**47-52.
25. Lord D, Washington SP, Ivan JN. *Accident Analysis and Prevention,* 2005; **37(1)**:35-46.
26. Yau KKW, Wang K, Lee AH. *Biometrical Journal,* 2013; **45**:437-452.
27. Akaike H. The second international symposium on information theory, edited by Petrov BV, Csaki BF, Academical Kiado, 1973.
28. Asrul MAA, Naing NN. 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, 2012.
29. Frome ED, Kutner MH, Beauchamp J. *Journal of American Statistical Association*, 1973; **68(344)**:935-940.
30. Nur Syabiha Zafakali, Wan Muhamad Amir W Ahmad *Journal of Modern Applied Statistical Methods*, 2013; **12(1)**:255-260.