IOSR Journal of Mathematics (IOSR-JM)
e-ISSN: 2278-5728, p-ISSN:2319-765X. Volume 10, Issue 1 Ver. IV. (Feb. 2014), PP 106-111
www.iosrjournals.org

A Way of Selecting the Right Statistical Model for Handling Overdispersion and Excess Zeroes

Zalina Abdullah[1], Wan Muhamad Amir W Ahmad[2]

[1](Department of Mathematics, School of Informatics and Applied Statistics, Universiti Malaysia Terengganu (UMT), 21030 Kuala Terengganu, Terengganu, Malaysia)

[2](Department of Mathematics, School of Informatics and Applied Statistics, Universiti Malaysia Terengganu (UMT), 21030 Kuala Terengganu, Terengganu Malaysia)

Abstract: *This study aims to find the risk factors of Diabetes Mellitus (DM) and to find the best model among Poisson, Negative Binomial (NB) and Zero-Inflated Negative Binomial (ZINB) regression models. In the count data, the existence of overdispersed data is a common situation for modeling approach. Overdispersion occurs when the data has greater value of variance compared to its mean. The Poisson regression model is a good starting step to model the data but does not account for overdispersion. Hence, NB regression model provide a better way to handle this case. Although it works well, but its inclusion of dispersion parameter seems to increase the probability of zero counts. As a result, we suggest applying ZINB regression model as it capable to handle both overdispersion and excess zeroes. Under the test of Akaike Information Criterion (AIC), Likelihood Ratio (LR), Vuong and Clarke, the ZINB regression model was chosen as the best model. From this analysis, we found that 11 of the parameters were the risk factors that significantly associated with DM at p<0.05 but sex was not as it had the value of p>0.05. Meanwhile, the respiratory system is the major contributor to the problem of DM.*

Keywords: *Diabetes Mellitus (DM), count data, Poisson regression model, Negative Binomial (NB) regression model and Zero-Inflated Negative Binomial (ZINB) regression model,*

I. Introduction

Count data is referred to the number of events that occur over a fixed period of time. It consist only nonnegative integers and discrete values. The examples for events count that recently had been used are the number of road accident deaths, the number of doctor visit patients and the number of dengue fever cases [1]. In the case where the variable of count outcome has small value of variance, the application of Ordinary Least Square (OLS) may lead to bias results for the predictor and the value of standard error will be large [2]. Hence, Poisson regression model provide a better way for modeling the distribution of the count data compared to other linear models [3], [4] as it was developed to satisfy the nature properties of count data. This regression assumes the mean and variance of the count variable are equal. However, it suffers one potential problem where the assumption is violated because the existence of overdispersed data where the variance of count data is larger than the mean. In this case, the single parameter λ is unable to describe event counts in Poisson distribution. There are two possibilities to the occurrence of overdispersion whether from the heterogeneity in the population or due to excess zeroes. Failure to overcome overdispersed data will tend to bias of standard error, inflated test statistics and inconsistent of population estimation.

As an alternative approach, NB regression model was applied. NB regression model is the generalization and extension of Poisson-gamma regression model [1] to handle overdispersed data by including dispersion parameter to allow variance of the observed count exceeds the mean and also accounts for unobserved heterogeneity. But sometimes, count data may contain a greater proportion of zero counts and it cannot be well modeled by using NB regression model. Thus, we use ZINB regression model as proposed by Lambert [5] to fit with overdispersed data and excess zeroes [6]. The ZINB regression model composed of two mixtures of processes that generate an “always zero group” and a “not always zero group”. According to Yesilova, Kaydan and Kaya [7], different ways are used to observe these two groups. To model whether the outcome is from “always zero group” or “not always zero group”, we used logit model with binomial assumption. Then, to determine the outcome in the “not always zero group”, we used NB model for count data.

The aims of this study are to select the right statistical model for handling overdispersion and excess zeroes in count data and to identify the risk factors of DM. In this section, we used three statistical methods: Poisson, NB and ZINB regression models for analyzing the data and it was done with SAS 9.3 statistical software program by using PROC GENMOD. The respondents are a total of 1000 patients diagnosed clinically with DM since 2002 until 2009. The data were collected at the Medical Record Unit in Hospital Universiti Sains Malaysia (HUSM), Kubang Kerian, Kelantan, Malaysia. The disease of DM is a chronic disease resulting from disability of insulin production, insulin action, or both and it is characterized by having high levels of blood glucose [8]. The prevalence of DM increased over the world as well as in Malaysia due to various factors such as growth, population, urbanization, aging and increasing prevalence of physical inactivity and obesity [9]. The dependent variable Y is the number of complication effects among DM disease. Meanwhile, 12 independent variables are age, sex, gastritis and duodenitis (gnd), primary hypertension (hyper), hypertensive heart disease (hyperheart), disease of the urinary system (urinary), chronic obstructive pulmonary disease (obs), renal failure (renal), cellulitis, disease of the respiratory system (respiratory), pneumonia (pneu) and anemia. All the parameters are estimated by using Maximum Likelihood Estimator (MLE). An AIC selection criterion is used to evaluate the goodness of fit of the model. This test indicates that the smallest value of AIC is accepted as the best model [10]. Meanwhile, to find the best regression model, we test the non-nested models (NB and ZINB regression models) by using Vuong test [11] and Clarke test [12]. The significance test statistics and positive value of the tests indicate one of the models is chosen as the best model.

II. Methodology

2.1 Poisson Model

Poisson regression model is suitable for modeling the count data as it fulfills the nature properties of count data. The dependent variable y_i is distributed as Poisson distribution with conditional mean of μ_i on a linear function of independent variables, X_i for case *i*. Thus, the density function of y_i can be written as

$$f\left(y_i \mid X_i\right)=\frac{e^{-\mu}\mu^{y_i}}{y_i!}, \qquad \text{for } y = 0, 1, 2,\ldots \tag{1}$$

where $\mu_i=\exp\left(X_i^{'}\beta\right)$. Poisson regression model with log link function can be expressed as $\ln\mu_i=\left(X_i^{'}\beta\right)$ or $\mu_i=\exp\left(X_i^{'}\beta\right)$. μ_i is a function of explanatory variables. When independent variables of X_i are given, then the log-likelihood function can be as

$$LL(\beta)=\sum_{i=1}^{n}\left[y_i X_i^{'}\beta-\exp\left(X_i^{'}\beta\right)-\ln y_i!\right] \tag{2}$$

where β are unknown parameters and parameter estimation for β is estimated by using maximum log-likelihood function [13].

2.2 Negative Binomial Model

NB regression model is an alternative of Poisson regression model that was generalized by Poisson regression model. This model could account the overdispersion by the inclusion dispersion parameter. It uses log link function to model between dependent and independent variables. Its conditional mean, μ_i for y_i is determined by X_i and heterogeneity component of ε_i unrelated to X_i. The formulation can be written as

$$\begin{aligned}\hat{\mu}_i &=\exp\left(X_i^{'}\beta+\varepsilon_i\right)\\ &=\exp\left(X_i^{'}\beta\right)\exp\left(\varepsilon_i\right)\end{aligned} \tag{3}$$

where $\exp\left(\varepsilon_i\right)\sim Gamma\left(\alpha^{-1},\alpha^{-1}\right)$. The density function of Y can be expressed as

$$f\left(Y=y_i \mid X_i\right)=\frac{\Gamma\left(y_i+\alpha^{-1}\right)}{\Gamma\left(y_i+1\right)\Gamma\left(\alpha^{-1}\right)}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right)^{\alpha^{-1}}\left(\frac{\mu_i}{\alpha^{-1}+\mu_i}\right)^{y_i} \tag{4}$$

$\alpha>0$ shows the level of overdispersion. If $\alpha=0$, NB regression model will reduce to Poisson regression model. The conditional mean and variance are $E\left(Y_i \mid X_i\right)=\mu=\exp\left(X_i^{'}\beta\right)$ and $Var\left(Y_i \mid X_i\right)=\mu_i+\alpha\mu_i^2$ [3]. Then, the log-likelihood function of NB regression model is [1]

$$LL(\beta)=\sum_{i=1}^{N}\left\{\sum_{j=0}^{y_i-1}\ln\left(j+\alpha^{-1}\right)-\ln\left(y_i!\right)-\left(y_i+\alpha^{-1}\right)\ln\left(1+\alpha\exp\left(X_i^{'}\beta\right)\right)+y_i\ln(\alpha)+y_iX_i^{'}\beta\right\} \qquad (5)$$

where $\sum_{j=0}^{y_i-1}\ln\left(j+\alpha^{-1}\right)=\ln\left(\frac{\Gamma\left(y_i+\alpha^{-1}\right)}{\Gamma\left(\alpha^{-1}\right)}\right)$

2.3 Zero-Inflated Negative Binomial Model

To model the count data with excess zeroes, the ZINB regression model was used as proposed by Lambert [5]. It has a mixture of two processes to identify count outcome from always zero group and not always zero group. For not always zero group, NB model will be used to model the outcome of zero counts and other than zero. For process 1 that generates always zero counts having ω_i probability and process 2 (generates counts from NB model) having $1-\omega_i$ probability. Thus,

$$Y_i\sim\begin{cases}0 & \textit{with probability}\ \omega_i\\ g\left(y_i\right) & \textit{with probability}\ 1-\omega_i\end{cases} \qquad (6)$$

Then, the probability of Y_i can be written as

$$\begin{aligned}&P\left(Y_i=0\mid X_i\right)=\omega_i+\left(1-\omega_i\right)g(0)\\ &P\left(Y_i\mid X_i\right)=\left(1-\omega_i\right)g\left(y_i\right), \qquad y_i>0\end{aligned} \qquad (7)$$

where $g\left(y_i\right)$ follows NB distribution. Meanwhile, the log-likelihood function for ZINB is [14]

$$\begin{aligned}LL(\mu,\alpha,\omega;Y)&=\sum_i\left(\begin{array}{l}I_{y_i=0}\log\left(\omega_i+\left(1-\omega_i\right)\left(1+\alpha\mu_i\right)^{-\alpha^{-1}}\right)\\ +I_{y_i>0}\log\left(\left(1-\omega_i\right)\dfrac{\Gamma\left(y_i+\frac{1}{\alpha}\right)}{y_i!\,\Gamma\left(\frac{1}{\alpha}\right)}\dfrac{\left(\alpha\mu_i\right)^{y_i}}{\left(1+\alpha\mu_i\right)^{y_i+\frac{1}{\alpha}}}\right)\end{array}\right)\\ &=\sum_i\left(\begin{array}{l}I_{y_i=0}\log\left(\omega_i+\left(1-\omega_i\right)\left(1+\alpha\mu_i\right)^{-\alpha^{-1}}\right)\\ +I_{y_i>0}\left(\begin{array}{l}\log\left(1-\omega_i\right)-\frac{1}{\alpha}\log\left(1+\alpha\mu_i\right)-y_i\log\left(1+\frac{1}{\alpha\mu_i}\right)\\ +\log\Gamma\left(y_i+\frac{1}{k}\right)-\log\Gamma\left(\frac{1}{\alpha}\right)-\log y_i!\end{array}\right)\end{array}\right)\end{aligned} \qquad (8)$$

where $I(.)$ is the indicator function in the specified event. The description that has been proposed is given as $\log(\mu)=X\beta$ and $\log\left(\frac{\omega}{1-\omega}\right)=G\gamma$ where $X(n\times p)$ and $G(n\times q)$ are the covariate matrixes. β with $(p\times1)$ dimension and γ with $(q\times1)$ dimension are unknown parameter vectors. The EM algorithm was used to estimate the maximum likelihood for β, α and γ.

III. Results

Table 1: Diagnosis count of DM patients

Diagnosis	Frequency	Percent (%)
0	818	81.8
1	82	8.2
2	29	2.9
3	32	3.2
4	24	2.4
5	15	1.5
Total	1000	100.0

Based on Table 1, a total 81.8% of the patients did not suffer from any complication effects to their health. This situation indicating the existence of high proportion of zero counts in the data. A part from that, 8.2% only had one complication effect to their health. The remaining 2.9% only had two complication effects and 3.2% had three complication effects. The rest 2.4% had four complication effects and the lowest percentage with 1.5% who was got five complication effects.

As the starting point, we had analyzed the Poisson regression model as the baseline model for the count data. The result showed the possibility of overdispersion due to larger value of variance (1.080) compared to its mean (0.41). Hence, the equality assumption of Poisson distribution had been violated. Other than that, the result from Pearson Chi-square was higher than one (1.2898) proved the existence of overdispersion. As an alternative approach to handle this case, we used NB regression model. From this analysis, the dispersion parameter was 0.0030. This model could collapse into Poisson regression model if the dispersion parameter equal to 0. Due to overdispersion and excess zeroes come together in the data, ZINB regression model provides a way to solve this situation.

Table 2: Model selection criteria for Poisson, NB and ZINB

Models	Log-likelihood	AIC	Vuong Test	Clarke Test
Poisson	-283.8228	1044.6920		
NB	-283.7052	1046.4560	Not preferred model	Not preferred model
ZINB	**-507.2194**	**1046.4389**	**Preferred model**	**Preferred model**

From the Table 2, it was seen that Poisson regression model had the smallest value of AIC compared to NB and ZINB regression models. Since Poisson regression model and NB regression model are nested models, the Likelihood Ratio (LR) test was done to make a comparison. The LR is given as $-2\left(LL_{Poisson} - LL_{NB}\right) = -2\left(-283.8228 - \left(-283.7052\right)\right) = 0.2352$, which showed that NB regression model most preferred as it significantly higher than Poisson regression model. Whereas, to compare non-nested model between NB and ZINB regression models, the Vuong statistics test as well as Clarke test had been computed by using SAS macro. The result showed that ZINB regression model was preferred as it had positive value and being significant at $p < 0.0001$. Next, the ML parameter estimations and standard error for the Poisson regression model were obtained in Table 3.

Table 3: Parameter estimations and standard error for Poisson regression model

Parameters	estimate	Standard Error	95% Confidence Limits		*p*-value
Intercept	-2.4649	0.2191	-2.8943	-2.0355	<.0001
Age	0.4310	0.2090	0.0215	0.8406	0.0391*
Sex	0.1243	0.1140	-0.0991	0.3477	0.2755
Gnd	0.4290	0.1315	0.1714	0.6867	0.0011*
Hyper	1.1734	0.1155	0.9471	1.3997	<.0001*
Hyperheart	0.8893	0.1479	0.5994	1.1793	<.0001*
Urinary	0.5493	0.1507	0.2539	0.8446	0.0003*
Obs	0.5440	0.1609	0.2285	0.8594	0.0007*
Renal	0.8759	0.1626	0.5572	1.1947	<.0001*
Cellulitis	1.0745	0.1705	0.7404	1.4086	<.0001*
Respiratory	1.4049	0.1479	1.1150	1.6949	<.0001*
Pneu	0.8807	0.1571	0.5728	1.1886	<.0001*
Anemia	0.7331	0.1389	0.4609	1.0052	<.0001*

*significant at p<0.05

Then, the ML parameter estimations and standard error for the NB regression model were obtained in Table 4.

Table 4: Parameter estimations and standard error for NB regression model

Parameters	estimate	Standard Error	95% Confidence Limits		*p*-value
Intercept	-2.4713	0.2383	-2.9382	-2.0043	<.0001
Age	0.4344	0.2752	0.0123	0.8566	0.0437*
Sex	0.1250	0.1147	-0.0999	0.3499	0.2760
Gnd	0.4294	0.1323	0.1700	0.6888	0.0012*
Hyper	1.1785	0.1376	0.9088	1.4482	<.0001*
Hyperheart	0.8936	0.1614	0.5772	1.2101	<.0001*
Urinary	0.5494	0.1515	0.2525	0.8463	0.0003*
Obs	0.5457	0.1637	0.2249	0.8665	0.0009*
Renal	0.8762	0.1632	0.5564	1.1960	<.0001*
Cellulitis	1.0761	0.1726	0.7378	1.4143	<.0001*
Respiratory	1.4072	0.1523	1.1087	1.7058	<.0001*
Pneu	0.8827	0.1603	0.5685	1.1969	<.0001*
Anemia	0.7344	0.1409	0.4582	1.0106	<.0001*

*significant at p<0.05

Lastly, the ML parameter estimations and standard error for the ZINB regression model were obtained in Table 5.

Table 5: Parameter estimations and standard error for ZINB regression model

Parameters	estimate	Standard Error	95% Confidence Limits		*p*-value
Intercept	-2.6576	0.2605	-3.1683	-2.1469	<.0001
Age	0.5234	0.2379	0.0572	0.9897	0.0278*
Sex	0.1361	0.1264	-0.1117	0.3839	0.2818
Gnd	0.4391	0.1634	0.1189	0.7593	0.0072*
Hyper	1.3727	0.1667	1.0460	1.6993	<.0001*
Hyperheart	1.0298	0.1949	0.6478	1.4118	<.0001*
Urinary	0.5805	0.1847	0.2184	0.9425	0.0017*
Obs	0.6561	0.1971	0.2343	1.0070	0.0016*
Renal	0.6207	0.1848	0.5178	1.2420	<.0001*
Cellulitis	1.1145	0.1947	0.7328	1.4961	<.0001*
Respiratory	1.4520	0.1757	1.1077	1.7964	<.0001*
Pneu	0.9555	0.1852	0.5924	1.3185	<.0001*
Anemia	0.7858	0.1714	0.4499	1.1217	<.0001*

*significant at p<0.05

To summarize, the ML parameter estimations and standard error for all the regression models were obtained in Table 6.

Table 6: Parameter estimations and standard error for all models

Parameters	Poisson	NB	ZINB
Intercept	-2.4649 (0.2191)	-2.4713 (0.2383)	-2.6576 (0.2605)
Age	0.4310* (0.2090)	0.4344* (0.2752)	0.5234* (0.2379)
Sex	0.1243 (0.1140)	0.1250 (0.1147)	0.1361 (0.1264)
Gnd	0.4290* (0.1315)	0.4294* (0.1323)	0.4391* (0.1634)
Hyper	1.1734* (0.1155)	1.1785* (0.1376)	1.3727* (0.1667)
Hyperheart	0.8893* (0.1479)	0.8936* (0.1614)	1.0298* (0.1949)
Urinary	0.5493* (0.1507)	0.5494* (0.1515)	0.5805* (0.1847)
Obs	0.5440* (0.1609)	0.5457* (0.1637)	0.6561* (0.1971)
Renal	0.8759* (0.1626)	0.8762* (0.1632)	0.6207* (0.1848)
Cellulitis	1.0745* (0.1705)	1.0761* (0.1726)	1.1145* (0.1947)
Respiratory	1.4049* (0.1479)	1.4072* (0.1523)	1.4520* (0.1757)
Pneu	0.8807* (0.1571)	0.8827* (0.1603)	0.9555* (0.1852)
Anemia	0.7331* (0.1389)	0.7344* (0.1409)	0.7858* (0.1714)

*significant at p<0.05

In short, according to Table 6, as Poisson, NB and ZINB regression models had different specifications, but they shared the same significant value of *p* such as age, gnd, hyper, hyperheart, urinary, obs, renal, cellulitis, respiratory, pneu and anemia but sex was not. Hence, sex was not categorized as the risk factor of DM. Meanwhile, the disease of the respiratory system gave the major contributor to the problem of DM as it had highest value of parameter estimation.

IV. Conclusion

In this paper, we use the DM data to examine the risk factors of DM by using three modeling approach such as Poisson, NB and ZINB regression models. We also have compared these three regression models to find the best applicable model. Poisson regression model may act as a good starting step to model the data but due to its restrictive assumption (the equality of mean and variance), this model is not suitable for handling overdispersion. Ignoring overdispersion can cause underestimation of standard error and affects the significance level of hypothesis testing [15]. Hence, NB regression model provide a better way to account overdispersion by including its dispersion parameter but because of that, the probabilities of zero counts may increase. So, we suggest applying the ZINB regression model as it capable to handle both excess zeroes and overdispersion in the data set.

From the analysis, we found that the ZINB regression model is the best model among Poisson and NB regression models under the test of AIC, LR, Vuong and Clarke. According to Jansakul [6], this model could account for overdispersion and excess zeroes at the same time. It can be concluded that ZINB regression model is very suitable to find the risk factors of DM. Thus, the result showed that the risk factors of DM that positively were age, gnd, hyper, hyperheart, urinary, obs, renal, cellulitis, respiratory, pneu and anemia. Meanwhile, sex was not the risk factor of DM as it insignificant associated with DM. Among these significant factors, we should pay more attention to the disease of the respiratory system as it had major contributor to the problem of DM. More awareness program should be done by the parties concerned to reduce the rate of DM disease.

References

[1]. A. C. Cameron and P. K. Trivedi, *Regression analysis of count data* (Cambridge University Press, 1998).

[2]. J. Cohen, P. Cohen, S. G. West, and L. S. Aiken, *Applied multiple regression/correlation analysis for the behavioral sciences 3rd ed* (Mahwah, NJ: Lawrence Erlbaum Associates, 2003).

[3]. H. C. Chin, and M. A. Quddus, Applying the random effect negative binomial model to examine traffic accident occurrence at signalized intersections, *Accident Analysis and Prevention, 35(2),* 2003, 253-259.

[4]. V. N. Shankar, F. Mannering, and W. Barfield, Effect of roadway geometric and environmental factors on rural freeway accident frequencies, *Accident Analysis and Prevention, 27(3),* 1995, 371-389.

[5]. D. Lambert, Zero-inflated Poisson regression, with an application to defects in manufacturing, *Journal of Techonometrics, 34(1),* 1992, 1-14.

[6]. N. Jansakul, Fitting a zero-inflated negative binomial model via R, *Proc. 20th International Workshop on Statistical Modelling*, Sidney, Australia, 2005, 277-284.

[7]. A. Yesilova, M. B. Kaydan, and Y. Kaya, Modeling insect-egg data with excess zeros using zero-inflated regression models, *Journal of Mathematics and Statistics, 39(2),* 2010, 273-282.

[8]. National Diabetes Fact Sheet, *General information and national estimates on diabetes in the United States, 2003* (Department of Health and Human Services, Centers for Disease Control and Prevention, U. S, 2004).

[9]. S. Wild, G. Roglic, A. Green, R. Sicree, and H. King, Global prevalence of diabetes: Estimates for the year 2000 and projections for 2030, *Diabetes Care, 27,* 2004, 1047-53.

[10]. J. B. Schreiber, F. K. Stage, J. King, A. Nora & E. A. Barlow, Reporting Structural Equation Modeling and Confirmatory Factor Analysis Results: A Review, *The Journal of Educational Research, 51(1),* 2006, 53-66.

[11]. Q. H. Vuong, Likelihood ratio tests for model selection and non-nested hypotheses, *Econometrica, 57,* 1989, 307-333.

[12]. K. A. Clarke, Nonparametric model discrimination in international relations, *Journal of Conflict Resolution, 47,* 2003, 72–93.

[13]. SAS, *SAS/Statistics Software* (USA: Hangen and Enhanced, 2007).

[14]. Z. Yau, *Score tests for generalization and zero-inflation in count data modeling*, Unpublished Ph. D. Dissertation, University of South Caroline, Columbia, 2006.

[15]. P. Alex, *Analysis of count data using the SAS system,* Paper presented at the SUGI Conference, Long Beach, California, 2001.