International Journal of Innovative Mathematics, Statistics & Energy Policies
9(2):24-39, April-June, 2021

© SEAHI PUBLICATIONS, 2021 www.seahipaj.org ISSN: 2467-852X

Modeling Doctor's Daily Visits to Patients (Ward Round) with Application of Poisson Regression Model

Chibo, Onyesom & Uti, Uju

Department of Statistics
Captain Elechi Amadi Polytechnic, Rumuola Port Harcourt, Nigeria
E-mail:chiboonyesom5@gmail.com

ABSTRACT

The main objective of the research was to apply Poisson Regression model on doctor visit as the dependent variable. The set of explanatory variables under consideration was tested and subsequently the final model was determined. Poisson regression model, which is a generalized linear model, was chosen as a computing model. Using it guarantees consistent results when working with variables with non-normal data distribution (skewed and discrete). Thus OLS estimator cannot work and is replaced by MLE estimator. The research resulted in selecting the Poisson regression model with an estimated and significant parameter. Then we turn to discussing a major problem of Poisson regression known as overdispersion and suggest possible solutions, including negative binomial regression and zero inflated regression. The Stata software package was used in the data analysis and we noted that the negative Binomial fits the model better than the Poisson regression model because of its flexibility.

Keywords: Overdispersion, Explanatory Variables, Zero-inflated, Maximum Likelihood Estimates.

1.0 INTRODUCTION

A visit to doctor, also known as physician office visit or ward round is a meeting between a patient with a physician to get health advice of treatment for a symptom or condition. According to a survey in the in the United State, a physician typically sees between fifty and one hundred patients per week, but the rate of visitation may vary with medical specialty, but differs one little by community size. This means the socio-political context of the patient (family, work, stress, beliefs) should be assessed as it often offers vital clues to the patient's condition and further management. In the analysis of data it is necessary to first comprehend the type of data before deciding the modeling approach to be used in the context of modeling the discrete, non- negative nature count of a dependent variable, the use of least square regression models several methodological limitations and statistical properties (Miaou, 1993; Karlaftis and Tarko,1998; Shankar,1995). The Poisson regression model is a good starting point of count data modeling. Many examples such as visits to doctor, the number of patent awarded to a firm, the number of road accident death, the number of dengue fever cases are restricted to a single digit or integer with quite low number of events (Cameron and Trivedi, 1998, Hausman et al, 1984; Radin et al, 1996). For such feature of data, Poisson regression suffers one potential problem, this is related to the assumption of equality of the mean and variance a property called equidispersion. When this assumption is violated, for instance the variances excess the mean, an overdispersion occurs. Failure to control for overdispersion will lead to inconsistent estimates, biased in standard error and inflated test statistics. One of the approaches to modeling overdispersion is to use quasi likelihood estimation techniques proposed by Wedderbum (1974).

1.1 Statement of Problem

The analysis of data in this study focuses on the use of Poisson regression with application to visits to Doctor.

This is because Poisson regression has more advantages over conventional linear model.

Therefore it appears worthwhile to devote effort in using Poisson regression to modeling doctor's home visit among the aged in Nigeria with a view of evaluating the impact of doctors home visit among the aged.

1.2 Significance of Study

The importance's of this study is the involvement of Poisson regression model in application to visit to doctor.

To researchers and student, the study will be significant as it is an academic exercise that will serve as reference materials for the future research work.

1.3 Aim of the Study

The aim of the study is geared towards Poisson regression model with application to doctor's home visit among the aged in Nigeria.

1. 4 Objective of the Study

1. Treating count data models and understanding of their estimations obtained by Poisson regression model
2. To find out whether gender age, number of illness and income has any differences in visits to doctor.
3. To find whether Poisson regression has more advantages over other conventional linear model.

1.5 Scope of Study: The research work is on Poisson regression model with application to doctor's visit among the age per annual. The range is chosen to ensure availability of date for the analysis in line with the objectives of the study.

1.6 Limitation of Study

Poisson regression suffers one problem. This is related to the assumption of equality of variances and mean. When this assumption is violated that is when the variances observed counts exceeds the mean, an overdispersion occurs and failure to control overdispersion leads to inconsistent estimates, biased in standard error and inflated test statistics.

1.7 Definition of Keywords

The following key words shall be defined in the context of the study. They include:

1. Doctor visit: means meeting between patients and a doctor to give health advice or treatment for symptoms and conditions
2. Regression: is a statistical measurement used in finance, investing and other discipline that attempts to determine the strength of relationship between one dependent variable using denoted by Y and a series of other changing variables.
3. Modeling: this is simply the abstraction or process of making a simple description about a statistical system or a process that can be used to explain such as statistical system.
4. Count data: is a statistical data type, a type of data in which the observations can take only the non-negative integer values (0, 1,

2.0 LITERATURE REVIEW

Jerald (1987) studied negative binomial regression models and examines efficiency and robustness properties of inference procedures based on them the result pointed out one limits of Poisson regression which is the variance of the data is constrained to be equal to the mean.

Alexander (2012) modeled the occurrences and incidence of malaria cases given the age, gender and time in quarters; he modeled the incidences of severe malaria cases given age, gender and time in years and validated the two models using negative binomial regression model and Poisson regression model. Poisson regression model and negative binomial regression models were used in fitting the data. Both models indicated that malaria is independent of gender. The prevalence of malaria and severe malaria cases were found to be prevalent among Children with less than 1 year old, and those under 5 and 70+

years old. The work concluded that malaria still remains high particularly among children under 5 years and those found between 20-34 age groups.

White and Robert (1996) modeled a likelihood- ratio testing frame work based on the negative binomial distribution that tests for the goodness of ft of this distribution to the observed counts. And then test for differences in the means and/ or aggregation of counts among treatment. Interferences about differences in means among treatment as well as the dispersion of the counts are possible. Simulations demonstrated that the statistics power of ANOVA is about the same as the likelihood-ratio testing procedure also provides information on dispersion. Type1 error rates of Poisson regression exceeded the expected 5%, even when corrected for overdispersion. Count data on orange-crowned Warblers are used to demonstrate the procedure.

Adeniyi and Ayo (2008) evaluated a non-linear model to identify fertility determinants and predict fertility using women's background characteristics. Based on the persistent high growth rate is among top ten most populous countries. The researchers used 2008 Nigeria Demography and Health Survey dataset consisting of 33,385 women with 31.4% from urban area. Fertility was measured using children ever born (CEB) and fitted into multi-factors additive Poisson regression models. Respondents mean age was 28.64 +/- 9.59 years, average CEB of 3.13 +/- 3.07 but higher among rural women than urban women (3.42 +/- 3.16 vs 2.53 +/- 2.79). Women aged 20-24years were about twice as likely to have higher CEB as those aged 15-19 years (IRR = 2.06, 95% CI: 1.95-2.18). Model with minimum deviance was selected and was used to predict CEB by the woman.

Monday (2017) based on the Human Immune deficiency Virus (HIV)/Acquired Immune deficiency syndrome (AIDS) epidemic has become one of the greatest challenges to public health among adults in Sub-Saharan African. In Nigeria, HIV/AIDS epidemic remain one of the major causes of death in the general population, particularly among young adult. This research was modeled, with the use Poisson regression model to study the linear trend of annual deaths resulting from HIV/AIDS in Nigeria for the period of 1996 to 2004. The result from the Poisson regression revealed an increase in rate of death resulting from HIV/AIDS in Nigeria. Therefore, there should be increase in the level of awareness of HIV/AIDS and other precautionary measures should also be observed in other to reduce the menace.

Wan 2010 Attempts to model count data have varied from the use of least square regression techniques to methods involving exponential distribution families including Poisson and Negative Binomial (NB) models. Given the nature of discrete, non-negative integer value of count data, the Poisson distribution has been verified to be the best distribution to describe count data. Though the Poisson regression model works well for count data, it still suffers one potential problem. This relates to the assumption of the equality of variance and mean in the use of Poisson model.. Thus, this paper provides a road map of the practical approach for modeling count data and an illustration using doctor visits data.

Kianififard (1995) evaluated the basic methodology of Poisson regression analysis and its application to clinical research. Overdispersion, model diagonistic and sample size issues were discussed. The methodology is illustrated on a data set from a clinical trial for the treatment of bladder cancer.

Shengping and Gilbert (2015) evaluated how risk factors, such as the timimng of corticosteroid treatment are associated with hospital length of stay for pediatric patients who were admitted due to acute asthma exacerbations using Poisson regression.

Amany and Yasmin (2008) evaluated using Poisson regression model for the number of woman participation in labour force I upper Egypt the work identify the determinants that have an impact on women, participate in the labour market.

Ferenc and Hegedus (2014) studied how Poisson regression can be used in studies in which the dependent variable describes the number of occurrences of some rate event such as suicide after pointing out why ordinary linear regression is inappropriate for handing dependent variables of this sort.

3.0 METHODOLOGY

Having reviewed related work on count data model-Poisson regression, this research depicts the methodology and its point of utilization. It also clarifies the research method and the research that should be utilized and the techniques utilized to guarantee the unwavering quality and legitimacy of the research.

3.1 Count Data Model

Count data is a statistical data type in which the observations can take only non- negative integer values {0,1,2, 3,...} and integers arise from counting.

An individual piece of count data is often termed a count variable that is count variable indicates the number of times something happened. When count variable is treated as a random variable, the Poisson distribution is commonly used to represent its distribution.

Although the use of regression model for counts is relatively recent, even a brief survey of recent application shows how common these outcomes are and the importance of the class of models.

Linear regression model is often applied to count outcomes which results in inefficient, inconsistent, and biased estimates. Even though there are situations in which the linear regression model provides reasonable results. We have several count models bur in the project we consider Poisson regression model (PRM), which is one of the foundation of other count models

3.2.1 The Validity of the Poisson models

1. The Poisson distribution may be useful to model events such as:
2. The number of Meteorites greater than 1 meter diameter that strike Earth in a year.

3.2.2 Assumptions And Validity

The Poisson distribution is an appropriate model if the following assumptions are true.

1. kIs the number of times an event occurs in an interval and k can take values 0,1,2,….
2. The occurrences of one event do not affect the Probability that a second event will occur. That is, events occur independently.

If these conditions are true then is a Poisson random variable and the distribution of k is a Poisson distribution.

3.2.3 Probability Of Events For A Poisson Distribution

An event can occur 0,1,2,3,… times in an interval. The average number of events in an interval is designated λ (lambda). λ is the event rate, also called the rate parameter. The Probability of observing K event in an interval is given by the equation

$$P(keventsininterval) = \frac{e^{-\lambda}\lambda^{k}}{K!}$$, where

1. λ is the average number of events per interval.
2. e is the number $2.71828 \ldots.$ (Euler’s number) the base of the natural logarithms.
3. K takes the values$0,1,2,3 \ldots..$
4. $K!$ =$k \times (k-1) \times (k-2) \times \ldots \times 2 \times 1$ is the factorial of K.

3.3.1 Poisson Regression Model

Is $x \in \square^n$ is a vector of independent variable, then the model takes the form

$\log\left(E\left(\frac{y}{x}\right)\right) = \alpha + \beta' x$ where $\alpha \in \square$ and $\beta \in \square^n$. Sometimes this is written more compactly as $\log\left(E\left(\frac{y}{x}\right)\right) = \theta' x$, where x is now an (n +1)- dimensional vector consisting of n independent variables concatenated to a vector of one. Here θ is simply x concatenated toβ.

Thus when given a Poisson regression model θ and an input vector x, the predicted mean of the associated Poisson distribution is given by

$$E\left(\frac{y}{x}\right) = \theta^{'x}.$$

If Y_i are independent observations with corresponding values x_i of the predictor variables, then θ can be estimated by Maximum Likelihood.

3.3.3 Maximum Likelihood Based Parameter Estimates

Given a set of parameter θ and an input vector$x$, the mean of the predicted Poisson distribution, as stated earlier is given by

$$P\left(\frac{y}{x};\theta\right) = (\lambda^y/y!)e^{-\lambda} = e^{y\theta'x}e^{e^{-\theta'x}}/y!\,.$$

Now suppose we are given a data set consisting of m vectors$x_i \epsilon \square^{n+1}$, i=1,2,…, m along with a set of m values $y_1, \dots\dots, y_m \in \square$. Then, for a given set of parameterθ, the Probability of attaining this particular set of data is given by

$$P(y_1, \dots., y_m/x_1, \dots., x_m; \theta) = \prod_{i=1}^{m} e^{y_i\theta' x_i} e^{-e^{\theta' x_i}} / y_{i!}$$

By the method of Maximum likelihood, we wish to find the set parameters θ that makes this Probability as large as possible. To do this, the equation is first rewritten as a likelihood function in terms of θ:

$$L\left(\frac{\theta}{X,Y}\right) = \prod_{i=1}^{m} e^{y_i\theta' x_i} e^{-e^{\theta' x_i}} / y_{i!}$$

Note that the expression on the right hand side has not actually changed. A formula in this form is typically difficult to work with; instead one uses the$log - likeli□ood$.

$$\ell\left(\frac{\theta}{X,Y}\right) = \sum_{i=1}^{m} (y_i \theta^i e^{\theta^i x_i}) \ .$$

Notice that the parameter θonly appears in the first two terms of each term in the summation. Therefore, given that we are only interested in finding the best value of θwe may drop the $y_i!$ and simply write

$$\ell(\theta/X,Y) = \sum_{i=1}^{m} y_i \theta' x_i - e^{\theta' x_i}.$$

To find a Maximum likelihood, we need to solve an equation

$$\ell \frac{(\theta/X,Y\ )}{\delta\theta} = 0$$

3.3.4 Overdispersion

A characteristic of Poisson distribution is that its mean is equal to its variances. In certain circumstances, it will be found that the observed variances that the observed variances is the greater than the mean, this is known as overdispersion and indicates the model to be appropriate.

3.4 Negative Binomial Regression Model

In negative binomial regression, the mean of y is determined by the exposure time t and a set of k regressor variables $the\ x's.$ the expression relating these qualities is

$$\mu_i = \exp(\ln(t_i) + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki})$$

Often, $x_1 \equiv 1$, in which case β_1 is called the intercept. The $\beta_1,\ \beta_2,\ \beta_3$ are unknown parameters that are estimated from a set of data, their estimates are symbolized as $b_1, b_2, b_3.$

The fundamental negative binomial regression model for an observation i is given as;

$$P_r\,(Y = y_i/\mu_i, \alpha) = \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(\alpha^{-1}\Gamma(y_i + 1))}\left(\frac{1}{1 + \alpha\mu_i}\right)^{\alpha^{-1}}\left(\frac{\alpha y_i}{1 + \alpha\mu_i}\right)^{y^i}$$

3.4.1 Maximum Likelihood Based Estimation

The regression coefficients are estimated using the method of maximum likelihood. The logarithm of the likelihood function is given as

$$\mathcal{L} = \sum_{i=1}^{n}\{\ln[\Gamma(y_i + \alpha^{-1})] - \ln[\Gamma(\alpha^{-1})] - \ln[\Gamma(y_i + 1)] - \alpha^{-1}\ln(1 + \alpha_i\mu_i) - y_i\ln(1 + \alpha\mu_i) + y_i\ln(\alpha) + y_i\ln(\mu_i)\}$$

.

$$\mathcal{L} = \sum_{i=1}^{n}\{(\sum_{j=0}^{y_i-1}\ln(j + \alpha^{-1})) - \ln(\Gamma(y_i + 1)) - (y_i + \alpha^{-1})\ln(1 + \alpha_i\mu_i) + y_i\ln(\mu_i) + y_i\ln(\alpha)\}$$

the first derivative of $\mathcal{L}$

$\frac{\delta \mathcal{L}}{\delta \beta_j} = \sum_{i=1}^{n} \frac{x_{ij}(y_i - \mu_i)}{1+\alpha\mu_i}$, j=1, 2, 3, ….,$k$

$$\frac{\delta \mathcal{L}}{\delta \alpha} = \sum_{i=1}^{n} \{\alpha^{-2}(\ln(1+\alpha\mu_i) - \sum_{j=1}^{y_{i-1}} \frac{1}{j+\alpha^{-1}}) + \frac{y_i - \mu_i}{\alpha(1+\alpha\mu_i)}\}.$$

$$-\frac{\delta^2 \mathcal{L}}{\delta\beta_r \delta\beta_s} = \sum_{i=1}^{n} \frac{\boldsymbol{\mu_i(1+\alpha\mu_i)x_{ir}x_{i5}}}{(\mathbf{1}+\boldsymbol{\alpha\mu_i})^2}$$

$\boldsymbol{r, s} = 1, \mathbf{2}, \dots, \boldsymbol{k}$

$$-\frac{\boldsymbol{\delta^2 \mathcal{L}}}{\boldsymbol{\delta\beta_r \delta\beta_s}} = \sum_{i=i}^{n} \frac{\boldsymbol{\mu_i}(y_i - \mu_i +)x_{ir}}{(\mathbf{1}+\boldsymbol{\alpha\mu_i})^2}$$

$\boldsymbol{r = 1, 2, \dots., k}$

$$\frac{\boldsymbol{\delta^2 \mathcal{L}}}{\boldsymbol{\delta\alpha^2}} = \sum_{i=1}^{n}\{\sum_{j=0}^{y_{i-1}} \left(\frac{j}{1+\alpha_j}\right)^2 + 2\alpha^{-3}\ln(1+\alpha\mu_i) - \frac{2\alpha^{-2}\mu_i}{1+\alpha\mu_i} - \frac{(y_i+\alpha^{-1})^2}{(1+\alpha\mu_i)^2}\ \mu_i\}$$

Equating the gradients to zero gives the following set of likelihood equations

$$\sum_{i=1}^{n} \frac{x_{ij}(y_i - \mu)}{1+\alpha\mu_i} = 0 \quad j = 1, 2, .., k$$

$$\sum_{i=1}^{n} \{\alpha^{-2}(\ln(1+\alpha\mu_i) - \sum_{j=0}^{y_{i-1}} \frac{1}{j+\alpha^{-1}}) + \frac{y_i + \mu_i}{\alpha(1+\alpha_i\mu_i)}\} = 0.$$

3.5 Zero Inflated Models

Zero inflated models are a statistical model based on a zero-inflated probability distribution that is a distribution that allows for frequent zero-valued observations.

3.5.1 Zero Inflated Poisson (ZIP)

One well known zero-inflated model is Diane Lambert's zero inflated Poisson model, which concerns a random event containing excess zero-count data in unit time.

The zero inflated Poisson (ZIP) model employs two components that correspond to two zero generating processes. The first process is governed by a binary distribution that generates structural zeros. The second process is governed by a Poisson distribution that generates counts some of which may be zero. The two model components are describes as follows

$$P_r(y_i = 0) = \pi + (1 + \pi)e^{-\lambda}$$

$$P_r(y_i = h_i) = (1 - \pi)\frac{\lambda h_i e^{-\lambda}}{h_i\,!}$$

Where the outcome variable y_i has any non-negative integer value, λ is the expected Poisson count for i^{th} individual; π is the probability of extra zeros. The mean is $(I - \pi)\lambda$ and the variance is $\lambda(I - \pi)(I + \pi\lambda)$.

3.5.2 Estimates Of Zero Inflated Poisson

The method of moments estimator are given by

$$\hat{\lambda}_{m_0} = \frac{s^2 + m^2}{m} \quad \text{-1,}$$

$$\hat{\Pi}_{m_0} = \frac{s^2 - m}{s^2 + m^2 - m}$$

Where m is the sample mean and s^2 is the sample variance. The maximum likelihood estimator is derived from the following equation

$$\bar{x}\left(1-e^{\hat{\lambda}_{ml}}\right)=\hat{\lambda}_{ml}\left(1-\frac{n_0}{n}\right).$$

Where $\bar{x}$ is the sample mean, and $\frac{n_0}{n}$ is the observed proportion of zeros. This can be solved by iteration and the maximum likelihood estimator for π is given by $\widehat{\Pi}_{ml}=1-\frac{\bar{x}}{\hat{\lambda}_{ml}}$.

3.5.3 Zero Inflated Negative Binomial (ZINB) Regression Model

The zero-inflated negative binomial (ZINB) regression model is used for count data that exhibit overdispersion and excess zeros. The data distribution combines the negative binomial distribution and the logit distribution. The positive value of Y are the nonnegative integers:0, 1, 2, and so on.

The probability distribution of the ZINB random variable y_i can be written as;

$$P_r(y_i=j)=\begin{cases}\pi_i+(1+\pi_i)g(y_i=0) & if\ j=0\\ (1-\pi_i)g(y_i) & if\ j>0\end{cases}$$

Where $\pi_i$is the logistic link function defined below and $g(y_i)$ is the negative binomial distribution given by ;

$$g(y_i)=P_r(Y=\mu_i,\alpha)=\frac{\Gamma(y_i+\alpha^{-1})}{\Gamma(\alpha^{-1})\Gamma(y_i+1)}\left(\frac{1}{1+\alpha\mu_i}\right)^{\alpha^{-1}}\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i}$$

Exposure time t can be included in the negative binomial component and a set of k regressor variables ($the\ x's$) the expression relating the equation is

$$\mu_i=\exp(ln(t_i))+\beta_1x_{1i}+\beta_2x_{2i}+\cdots.+\beta_kx_{ki}$$

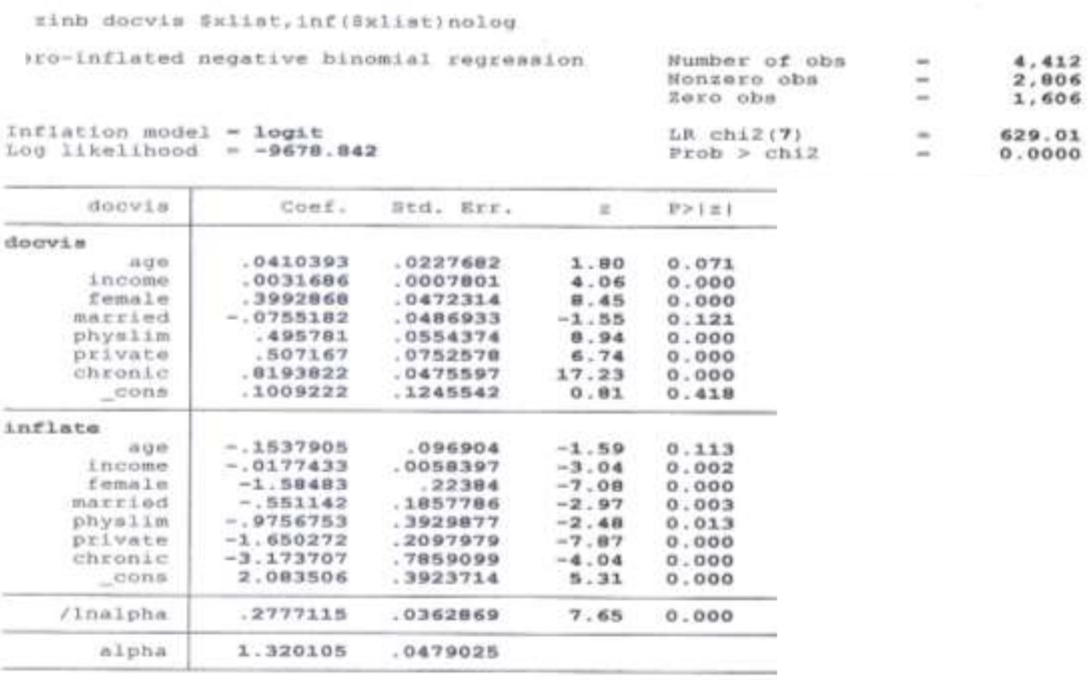
Often, $x_1\equiv 1,$ the logistic link function π_i given by;

$$\pi_i = \frac{\lambda_i}{1+\lambda_i} \ \ where\ \lambda_i = \exp(\ln(t_i)) + Y_1 Z_{1i} + Y_2 Z_{2i} + \cdots . + Y_m Z_{im})$$

Where $(the\ z's)$ are the regressor variables and t is the exposure time.

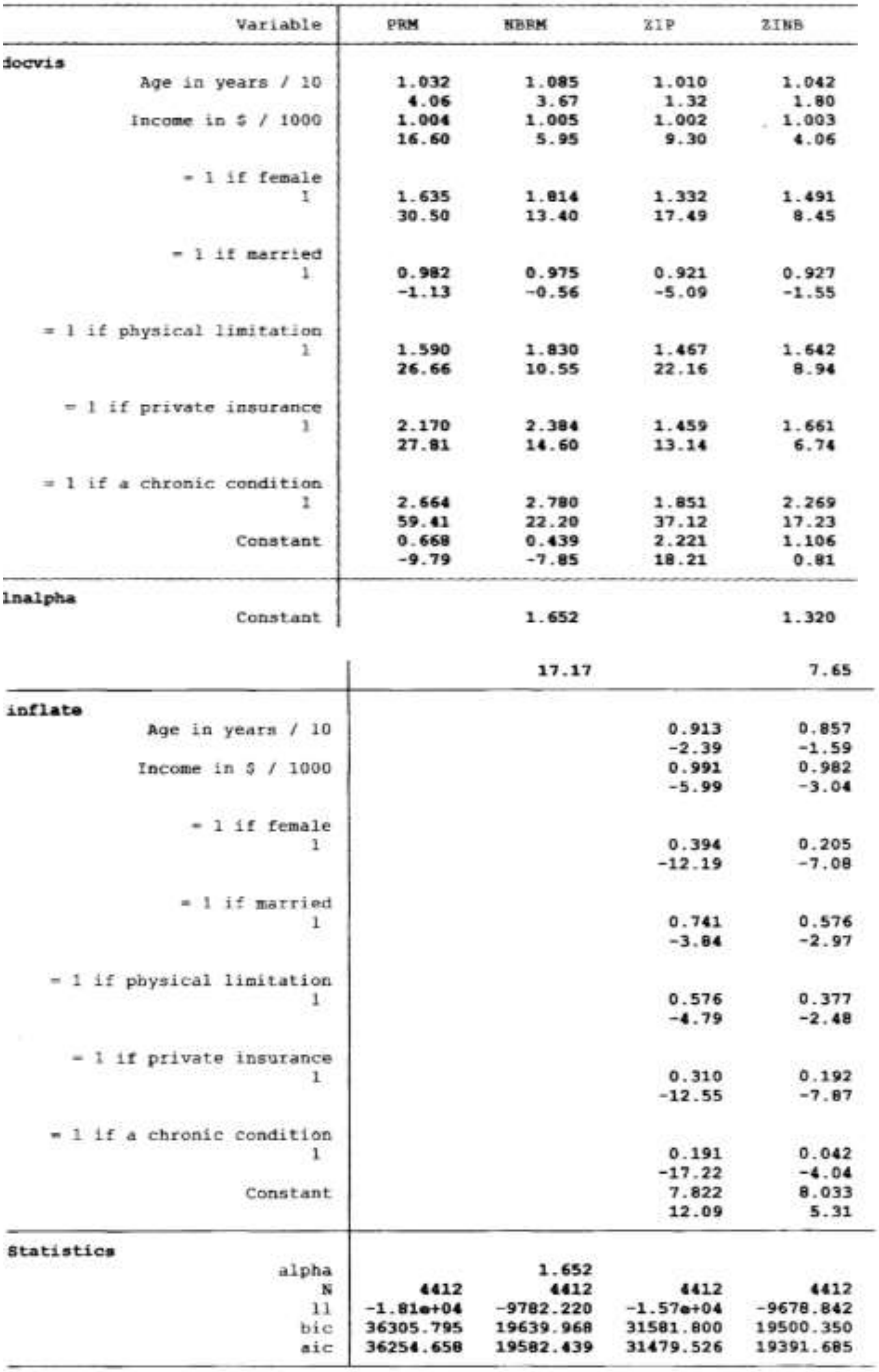
4.0 RESULTS

Table 4.1 zero inflated negative binomial

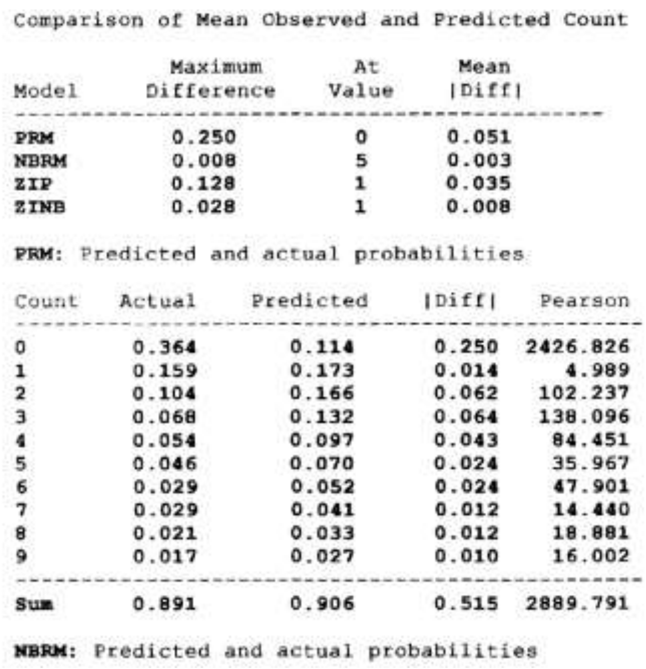


```
  zinb docvis $xlist,inf($xlist)nolog

)ro-inflated negative binomial regression        Number of obs     =      4,412
                                                 Nonzero obs       =      2,806
                                                 Zero obs          =      1,606

Inflation model = logit                          LR chi2(7)        =     629.01
Log likelihood  = -9678.842                      Prob > chi2       =     0.0000

------------------------------------------------------------------
      docvis |      Coef.   Std. Err.      z    P>|z|
-------------+----------------------------------------------------
docvis       |
         age |   .0410393   .0227682     1.80   0.071
      income |   .0031686   .0007801     4.06   0.000
      female |   .3992868   .0472314     8.45   0.000
     married |  -.0755182   .0486933    -1.55   0.121
     physlim |    .495781   .0554374     8.94   0.000
     private |    .507167   .0752578     6.74   0.000
     chronic |   .8193822   .0475597    17.23   0.000
       _cons |   .1009222   .1245542     0.81   0.418
-------------+----------------------------------------------------
inflate      |
         age |  -.1537905    .096904    -1.59   0.113
      income |  -.0177433   .0058397    -3.04   0.002
      female |   -1.58483     .22384    -7.08   0.000
     married |   -.551142   .1857786    -2.97   0.003
     physlim |  -.9756753   .3929877    -2.48   0.013
     private |  -1.650272   .2097979    -7.87   0.000
     chronic |  -3.173707   .7859099    -4.04   0.000
       _cons |   2.083506   .3923714     5.31   0.000
-------------+----------------------------------------------------
    /lnalpha |   .2777115   .0362869     7.65   0.000
-------------+----------------------------------------------------
       alpha |   1.320105   .0479025
------------------------------------------------------------------
```

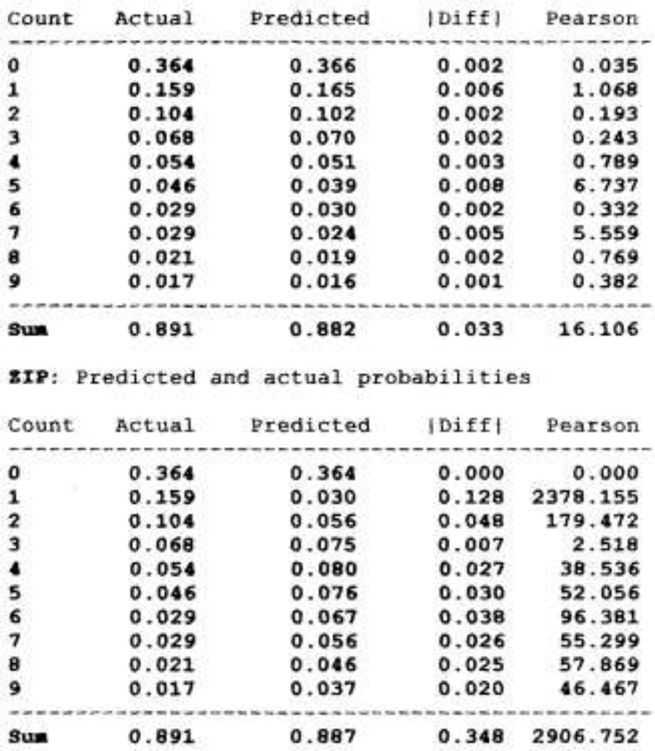
Table 4.2 Comparison of the model



Variable	PRM	NBRM	ZIP	ZINB
docvis				
Age in years / 10	1.032	1.085	1.010	1.042
	4.06	3.67	1.32	1.80
Income in $ / 1000	1.004	1.005	1.002	1.003
	16.60	5.95	9.30	4.06
= 1 if female				
1	1.635	1.814	1.332	1.491
	30.50	13.40	17.49	8.45
= 1 if married				
1	0.982	0.975	0.921	0.927
	-1.13	-0.56	-5.09	-1.55
= 1 if physical limitation				
1	1.590	1.830	1.467	1.642
	26.66	10.55	22.16	8.94
= 1 if private insurance				
1	2.170	2.384	1.459	1.661
	27.81	14.60	13.14	6.74
= 1 if a chronic condition				
1	2.664	2.780	1.851	2.269
	59.41	22.20	37.12	17.23
Constant	0.668	0.439	2.221	1.106
	-9.79	-7.85	18.21	0.81
lnalpha				
Constant		1.652		1.320
		17.17		7.65
inflate				
Age in years / 10			0.913	0.857
			-2.39	-1.59
Income in $ / 1000			0.991	0.982
			-5.99	-3.04
= 1 if female				
1			0.394	0.205
			-12.19	-7.08
= 1 if married				
1			0.741	0.576
			-3.84	-2.97
= 1 if physical limitation				
1			0.576	0.377
			-4.79	-2.48
= 1 if private insurance				
1			0.310	0.192
			-12.55	-7.87
= 1 if a chronic condition				
1			0.191	0.042
			-17.22	-4.04
Constant			7.822	8.033
			12.09	5.31
Statistics				
alpha		1.652		
N	4412	4412	4412	4412
ll	-1.81e+04	-9782.220	-1.57e+04	-9678.842
bic	36305.795	19639.968	31581.800	19500.350
aic	36254.658	19582.439	31479.526	19391.685



```
Comparison of Mean Observed and Predicted Count

            Maximum       At       Mean
Model      Difference    Value    |Diff|
---------------------------------------------
PRM          0.250         0       0.051
NBRM         0.008         5       0.003
ZIP          0.128         1       0.035
ZINB         0.028         1       0.008

PRM: Predicted and actual probabilities

Count    Actual     Predicted     |Diff|    Pearson
------------------------------------------------------
0        0.364       0.114        0.250   2426.826
1        0.159       0.173        0.014      4.989
2        0.104       0.166        0.062    102.237
3        0.068       0.132        0.064    138.096
4        0.054       0.097        0.043     84.451
5        0.046       0.070        0.024     35.967
6        0.029       0.052        0.024     47.901
7        0.029       0.041        0.012     14.440
8        0.021       0.033        0.012     18.881
9        0.017       0.027        0.010     16.002
------------------------------------------------------
Sum      0.891       0.906        0.515   2889.791

NBRM: Predicted and actual probabilities
```



```
Count    Actual     Predicted     |Diff|    Pearson
------------------------------------------------------
0        0.364       0.366        0.002      0.035
1        0.159       0.165        0.006      1.068
2        0.104       0.102        0.002      0.193
3        0.068       0.070        0.002      0.243
4        0.054       0.051        0.003      0.789
5        0.046       0.039        0.008      6.737
6        0.029       0.030        0.002      0.332
7        0.029       0.024        0.005      5.559
8        0.021       0.019        0.002      0.769
9        0.017       0.016        0.001      0.382
------------------------------------------------------
Sum      0.891       0.882        0.033     16.106

ZIP: Predicted and actual probabilities

Count    Actual     Predicted     |Diff|    Pearson
------------------------------------------------------
0        0.364       0.364        0.000      0.000
1        0.159       0.030        0.128   2378.155
2        0.104       0.056        0.048    179.472
3        0.068       0.075        0.007      2.518
4        0.054       0.080        0.027     38.536
5        0.046       0.076        0.030     52.056
6        0.029       0.067        0.038     96.381
7        0.029       0.056        0.026     55.299
8        0.021       0.046        0.025     57.869
9        0.017       0.037        0.020     46.467
------------------------------------------------------
Sum      0.891       0.887        0.348   2906.752
```

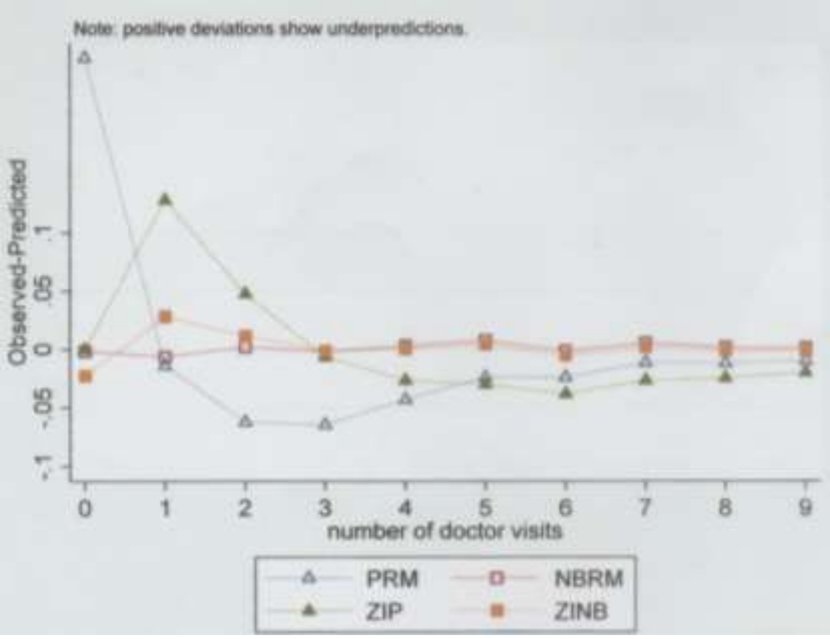


Fig. 4.1: A Graph that connect Observed-Predicted with Number of Doctor visits

5.0 DISCUSSION OF RESULTS

In line with our stated aim and objectives, this research work has been able to achieve the following;

1. It used data obtained from http://www.stata-press.com to model docvis among age, income, female, black, Hispanic, married, physlim, private, chronic.
2. It has analyzed the data as fitted by Poisson regression model (PRM), negative binomial regression model(NBRM), zero-inflated Poisson(ZIP), zero-inflated negative binomial.
3. Finally it has compared between the four model to discover which model fits perfectly.

The interpretation of table 4.1 gives a clear description of the data. Table 4.2 is the overall summary of the count data, the mean number of docvis is approximately 3.957 and the variance is $(3.957)^2$ = 15.06, which is substantially more than the mean this leads to our first regression. Coef. - the Poisson coefficient can be interpreted as follows; for a one unit change in the predictor variable, the difference in logs of the expected counts is expected to change by the respective regression coefficient, given the other predictor variables in the model are held constant. Age- this is the Poisson regression estimate for a one unit increase in age, given the other variables are held constant in the model. If a patient were to increase docvis by one unit point, the difference in the logs of expected counts would be expected to increase by 0.032 unit, while holding the other variables constant. Income - this is the Poisson regression estimate for one unit increase in income, given the other variables are held constant in the model. If a patient were to increase docvis by one unit point, the difference in the logs of expected counts would be expected to increase by 0.004 units, while holding the other variables constant. Female - this is the Poisson regression estimate for a one unit increase in number of female, given the other variables are held constant in the model. If a female patient were to increase docvis by one unit point, the difference in the logs of expected counts would be expected to increase by 0.49 units, while holding the other variables constant. Married - this is the Poisson regression estimate for a one unit increase in number of married patient, given the other variables are held constant in the model.

The z-test statistics testing the slope for age on docvis is zero given the other variables are in the model, is (0.031/0.008) -4.06 with an associated P-value of < 0.0001 with alpha been set as 0.05, we would reject the null hypothesis and conclude that Poisson regression coefficient for age is statistically different from zero0 at the bottom is the test of overdispersion parameter alpha. When the overdispersion parameter is equal to zero the test statistics is -2[-18119.329-(-9782.2197) = 16674.22 with an associated p-value 0f <0.0001. the large test statistic would suggest that the response variable is overdispersed and is not sufficiently described by the simpler poisson distribution. Predicted number of docvis among income, female, physlim, private, chronic are significant except married. The bottom half, contains logit coefficients for the factor change in the odds of being in the always zero group. The predicted number of docvis among the variables is statistically significant except age..The Pearson statistics which is calculated as $N(|diff|)^{2/}/predicted$, where N is the number of observations in the dataset. Looking closely at the sum of the Pearson columns gives us a sense of how close the predicted proportions were to the actual proportions using this method to compare, the NBRM and ZINB appears better than the PRM. Finally in the next the result suggests which model is most preferred by the given comparison strength of the evidences supporting this preference. When we compare the four models using BIC and AIC the NBRM and ZINB is preferred over PRM and ZIP. Fig 4.1 is a graph that plots the residuals from the tested models, the models with lines closest to zero should be considered for our data, at the zero and one count NBRM and ZINB appears better than the PRM and ZIP models.

6.1 CONCLUSION

Categorically, this work can be summarized without any fear of contradiction. The Poisson regression model (PRM), negative binomial regression model (NBRM) which is the base for the other regression model zero-inflated Poisson (ZIP) and zero-inflated (ZINB) respectively fits this work, except the fact that some models fits better than the others. The PRM doesn't fit reasonably well because if it's strict conditions of equal conditional mean and equal conditional variances E(x) = Var (x), as a result leading to under predictions of zeros. While NBRM because of its flexibility fits reasonable well because it allows the variances to be greater than the mean called over-dispersion. The zero-inflated model assumes two groups, one has no chance of going beyond zeros. The other group may have zero count but the probability of having a positive count is non-zero. In conclusion the ZINB and NBRM fits better than the other two models.

6.2 RECOMMENDATION

Recommendations are hereby presented;

1. NBRM often maybe good enough for the modeling of count data so the need of zero inflated models might be questioned.
2. Because of PRM strict conditions it makes its result to be inconsistent and biased.
3. This work can furthered for verification purposes or contribution by any researcher who picks interest in it.

REFERENCES

Agresti, A. (2002*). Categorical data analysis. 2nd edition*. New York, Wiley.

Aiken, L. S. and West, S. G. (1991). *Multiple Regression: testing and interpreting interactions,* Newbury Park, CA: Sage.

Bailer A.J. and L.T. Stayner (1997). Modeling fatal injury rates using poisson regression. *Journal of safety research;28:177-186.*

Bair, H. (2013). Poisson regression: Lack of fit ≠ Overdispersion, StatNews #86, Cornell University, http://www.cscu.cornell.edu/news/ststnews/stnews86.pdf .

Cameron, A. C. and Trivedi, P. K. (2009). *Microeconometrics Using Stata*. CollegeStation, TX: Stata Press.

Cameron, A. C. (2008). *Advances in Count Data Regression Talk for the Applied Statistics Workshop,March* 28, 2009. http://cameron.econ.ucdavis.edu/racd/count.html.

Dobson,A.J. (2002). *An introduction to generalized linear model,2nd ed.*, New York: Chapman & Hall/CRC.

*Famoye, F., Wulu, J. T. Jr.,*and *K. P. Singh.* (*2004)*. On the generalized Poisson regression model with an application to accident data. *Journal of Data Science 2*: *287–295*.

Ferenc M., Rita Hegedus (2014). The use of poisson regression in the sociological study of suicide. Journal of sociology and social policy;Vol.5,No2

Gardner, W., Mulvey, E.P. and Shaw E. C. (1995). Regression analyses of counts and rates: poisson, overdispersed poisson and negative binomial models. *Psychological Bulletin*;*118: 392-404.*

Greene, W. H. (1994). *Econometrics analysis*. New York: Stern School of Business, New York University, Department of Economics.

Hall, D. B. and Zhengang, Z. (2004). Marginal models for zero inflated clustered data. *Statistical modeling, 4:161-180.*

Hilbe, J.M. (2011). *Negative binomial regression*. Cambridge University Press, Cambridge

Joseph M. Hilbe (2014). *Modeling count data*. Cambridge University press, Cambridge.

Jerald F. Lawless (1987). Negative binomial and mixed poisson regression. *Journal of statistics; 15: https://doi.org/10.2307/3314912*

Long, J. S. (1997). *Regression Models for Categorical and Limited Dependent Variables.* Thousand Oaks, CA: Sage Publications.

Monday O. Adenomon (2017). Fitting a Poisson regression model to reported deaths from HIV/AIDS in Nigeria. *Journal of statistical distributions and applications; 3(3):56-60.*

Mc Cullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models, 2nd ed*. London; Chapman and Hall.

Sileshi, G., G. Hailu,and G. I. Nyadzi. (*2009)*. Traditional occupancy–abundance models are inadequate for zero-inflated ecological count data. *Ecological Modeling 220*: *1764–1775*.

White, G. C.,and R. E. Bennetts. 1996. Analysis of frequency count data using the negative binomial distribution. *Ecology 77*: *2549–2557*.

Wan F. (2011). Applying fixed effects panel count model to examine road accident occurrence. Journal of applied-science;11:1185-1191.

Ver Hoef, J. and Boveng, P. L. (2007). *Quasi-poisson* vs *Negative Binomial Regression: How should we model overdispersed count data?* Ecology88:2766-2772. http://doi.org/10.1890/07-0043.1

Zamani, H., Ismail, N. (2013). Score test for testing zero-inflated Poisson regression against zero-inflated generalized poisson alternatives. *Journal of Applied Statistics* 40(9):2056-2068