Pneumonia Prevalence among Under-Five in Nigeria: A Poisson Regression Model Approach.

Adesupo A. Akinrefon, M.Sc.*; Olateju A. Bamigbala*, B.Sc.; and S.S. Abdulkadir, Ph.D.

Department of Statistics and Operations Research, Modibbo Adama University of Technology, Yola Adamawa State, Nigeria.

E-mail: akinrefon.adesupo@mautech.edu.ng*
olatejubamigbala@yahoo.com

ABSTRACT

Pneumonia deaths occur at high levels in developing countries, and three-quarters take place in just 15 countries. The majority of pneumonia cases is preventable or treatable. Using solid fuels such as, charcoal and crops is associated with increased mortality from pneumonia and other acute lower respiratory diseases among children, as well as increased mortality from chronic obstructive pulmonary disease, lung cancer (where coal is used) and other disease among adults. (WHO,2013).

Data on pneumonia cases among infants under age 5 admitted in the Specialist hospital and Valli Clinic were obtained for the purpose of the research. The variables on which the data were collected included sex, age group, year, and hospital type using Poisson Regression. Various models were considered in order to select the best model, the AIC of model 1 was found to be 235.04 which consider the main effects only (that is age group, sex, year, and hospital type). The model indicated that females are more likely to be infected with pneumonia than male. The study also revealed that the rate of pneumonia among the age groups increases as they advance in age, for hospital type it shown that patient infected with pneumonia visited Valli clinic (Private owned) than Specialist Hospital (Government owned). The rate of being infected with pneumonia decreased with year but an increase was recorded in 2014.

(Keywords: pneumonia, Chi-square, quasi-Poisson, deviance, Poisson regression, negative binomial, odds)

INTRODUCTION

More than half of the world's annual new pneumonia cases are concentrated in just five countries which 44% of the world children aged less than five years live, India (43 million), China (21.1 million), Pakistan (9.8 million), Bangladesh (6.4 million), and Nigeria (6.1 million). Pneumonia contributed to 56-86% of all death attributed to measles (Yaguo and Nte, 2011). It has also been shown that the leading risk factors contributing to Pneumonia incidence are the lack of exclusive breast feeding, under nutrition, indoor air pollution, domestic use of smoke-generating firewood, low birth weight, crowding and lack of measles immunization (Igor, et. al., 2008).

Recent estimate from the United Nations Children's Fund (UNICEF, 2009) has shown that pneumonia continues to be the number one killer of children around the world causing 18% of all child mortality, an estimated 1.3 million child deaths in 2011 alone. Nearly all pneumonia deaths occur in developing countries, and three-quarters take place in just 15 countries. The majority of pneumonia cases are preventable or treatable. In 2009, World Health Organization (WHO) and UNICEF released the Global Action Plan for Prevention and Control of Pneumonia (GAPP), setting out a 90% coverage target by 2015 for three interventions: vaccination, breastfeeding, access to care, and antibiotic treatment. If 90% coverage is reached, these interventions could prevent two thirds of all childhood pneumonia deaths (WHO/UNICEF, 2009). Using solid fuels such as, charcoal and crops is associated with increased mortality from pneumonia and other acute lower respiratory diseases among children, as well as increased mortality from chronic obstructive pulmonary disease, lung cancer (where coal is used) and other disease among adults. (WHO, 2013).

METHODOLOGY

The data for this work were obtained from Specialist Hospital and Valli Clinic, Yola of Adamawa State, Nigeria, the data is on patient under five years infected with pneumonia grouped by age group, gender, year and hospital. Hence, Poisson Regression was applied to model the prevalence using the R program.

Poisson Regression Model

The general method of fitting a Poisson regression model is to use the Poisson model formulation to derive a likelihood function that can be maximized so that parameter estimates, estimates standard errors, maximized likelihood statistic and other information can be produced. Poisson regression analysis goal is to fit the data to a regression equation that will accurately E(Y) or μ as a function of a set of explanatory variables

$X_1, X_2, ...X_p$.

In Poisson regression it is assumed that the dependent variable Y, number of occurrence of an event, has a Poisson distribution given the independent variables $X_1, X_2, ...X_p$.

Pr(Y=y) = $\frac{e^{-\mu}\mu^y}{y!}$, y = 0,1,2,

Where E(Y) = μ and Var(Y) = μ. This is called the equi-dispersion property of the Poisson distribution.

The log of the mean μ is assumed to be a linear function of the independent variables, that is:

$$In\mu = \beta_o + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

Where Y ~P(μ)

or equivalently,

$$\mu = e^{(\beta_o + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p)}$$

This is the model for analyzing normal count data.

Sometimes, the response may be in the form of events of certain type that occur over time, space or some other index of size. In this situation, it is often relevant to model the data as the rate at which events occur. When a response count Y has index (such as population size) equal to t, the sample rate of occurrence is Y/t. The expected value for rate is μ/t. Thus, for analysis rate data, the model can be written as:

$$In\left(\frac{\mu}{t}\right) = \beta_o + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

This model has equivalent representation as:

$$In\mu - Int = \beta_o + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

The adjustment term $-Int$, on the left-hand side of the equation is called an *offset.*

Estimation of Parameters Using Maximum Likelihood Estimation (MLE)

Estimation of parameters in Poisson regression relies on maximum likelihood estimation (MLE) method. Maximum likelihood estimation seeks on answer question of what values of the regression coefficients are most likely to have given rise to the data.

To discuss the maximum likelihood estimation for Poisson regression. Let μ_i be the mean for the ith response, for i=1,2,…,n. Since the mean response is assumed to be a function of a set of explanatory variables. $X_1, X_2, ..., X_n$, the notation $\mu(X_i, \beta)$ is used to denote the function that relates the mean response μ_i *and* X_i (the values of explanatory variables for case i) and β (the values of regression coefficients). Now consider the Poisson regression model in the following form:

$$\mu_i = \mu(X_i, \beta) = e^{(X_{\cdot}, \beta)} \quad \textbf{(1)}$$

Then, from Poisson distribution:

$$P(Y, \beta) = \frac{[\mu(X_i, \beta)]^Y e^{-\mu(X_i, \beta)}}{Y!} \quad \textbf{(2)}$$

The likelihood function is given by:

$$L(Y; \beta) = \prod_{i=1}^{N} P(Y; \beta)$$

$$L(Y; \beta) = \prod_{i=1}^{N} \frac{[\mu(X_i, \beta)]^Y e^{-[(X_i, \beta)]}}{Y!}$$

$$L(Y; \beta) = \frac{\left\{\prod_{i=1}^{N} [(\mu(X_i, \beta))]^{Y_i}\right\} e^{\left[-\sum_{i=1}^{N} \mu(X_{\cdot}, \beta)\right]}}{\prod_{i=1}^{n} Y_i!} \quad \textbf{(3)}$$

The next thing to do is taking natural log of the above likelihood function. Then, differentiate the equation with respect to β and equate the equation to zero. The log likelihood function is given as:

$$In\, L(Y_i, \beta) = \sum_{i=1}^{N} [Y_i In[\mu(X_i, \beta)] - \mu(X_i, \beta) - In(Y_i!)] \quad \textbf{(4)}$$

$$\frac{\partial}{\partial \beta}[InL(Y; \beta)] = 0 \quad \textbf{(5)}$$

The solution to the set of Maximum Likelihood given above must generally be obtained by iteration procedure. This procedure will estimate the values of β. Maximum likelihood estimation produces Poisson parameters that are consistent, asymptotically normal and asymptotically efficient. To demonstrate estimation of parameters using maximum likelihood estimation, consider the method of scoring in generalized linear model. The method of scoring in generalized linear model simplifies the estimating equation to:

$$b^{(m)} = b^{(m-1)} + \left[I^{(m-1)}\right]^{-1} U^{(m-1)} \quad \textbf{(6)}$$

Where $b^{(m)}$ is the vector of estimates of the parameters $\beta_0, \beta_1, ..., \beta_p$ at the nth iteration.

I is the information matrix with elements I_{jk} given by:

$$I_{jk} = \sum_{i=1}^{N} \frac{X_{ij} X_{jk}}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \quad \textbf{(7)}$$

And U is the vector of elements given by:

$$U_j = \sum_{i=1}^{N}\left[\frac{(Y_i-\mu_i)}{Var(Y_i)}X_{ij}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)\right] \quad \textbf{(8)}$$

U is called the score function.

If both sides of equation (6) are multiplied by $I^{(m-1)}$ it will become:

$$I^{(m-1)}b^{(m)} = I^{(m-1)}b^{(m-1)} + U^{(m-1)} \quad \textbf{(9)}$$

From (8), ***I*** can be written as:

$$I = X^T W X \quad \textbf{(10)}$$

Where W is the N x N diagonal matrix with elements:

$$w_{ii} = \frac{1}{Var(Y_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2 \quad \textbf{(11)}$$

The expression on the right-hand side of equation (9) is the vector with elements:

$$\sum_{k=o}^{p}\sum_{i=1}^{N}\frac{X_{ij}X_{jk}}{Var(Y_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2 b_k^{(m-1)} + \sum_{i=1}^{N}\left[\frac{(Y_i-\mu_i)}{Var(Y_i)}X_{ij}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)\right]$$

Evaluated at $b^{(m-1)}$. Thus, the right-hand side of equation (7) can be written as:

$$X^T W z \quad \textbf{(12)}$$

Where z has elements:

$$z_i = \sum_{k=0}^{p} X_{ik}b_k^{(m-1)} + (Y_i-\mu_i)\left(\frac{\partial\eta_i}{\partial\mu_i}\right) \quad \textbf{(13)}$$

Note that z is N x 1 matrix. Hence, finally, the iterative equation for parameter estimation can be written as:

$$\left(X^T W X\right)^{(m-1)} b^{(m)} = \left(X^T W z\right)^{(m-1)} \quad \textbf{(14)}$$

This equation has to be solved iteratively because, in general, **z** and **W** depend on **b** .This iterative method is known as iteratively reweighted least squares method (IRWLS).

Now, consider a set of Poisson regression data, $Y_1, Y_2, ..., Y_N$ satisfying the properties of generalized linear model. Parameters β_o *and* β_1 (let's just consider these two) are related to the Y_i'*s* through $E(Y_i) = \mu_i$ *and* $g(\mu_i) = \eta_i = In(\mu_i) = \beta_o + \beta_1 X_1$

From equation (9), the following equation is obtained:

$$w_{ii} = \frac{1}{\mu_i}(\mu_i)^2 = \mu_i$$

Using the estimate $\begin{bmatrix} \beta_o \\ \beta_1 \end{bmatrix}$ for β, equation (13) becomes:

$$z_i = \beta_o + \beta_1 X_1 + \frac{(Y_i - \mu_i)}{\mu_i}$$

Essentially, to find the formula for estimating equation, the following matrices must be obtained:

$$X = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ . & . \\ . & . \\ . & . \\ 1 & X_N \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_o \\ \beta_1 \end{bmatrix}$$

$$W = \begin{bmatrix} \mu_1 & & & \\ & \mu_2 & & 0 \\ & & . & \\ & & & . \\ & 0 & & . \\ & & & & \mu_N \end{bmatrix} \qquad Z = \begin{bmatrix} \beta_o + \beta_1 X_1 + \frac{(Y_1 - \mu_1)}{\mu_1} \\ \beta_o + \beta_1 X_1 + \frac{(Y_2 - \mu_2)}{\mu_2} \\ . \\ . \\ . \\ \beta_o + \beta_1 X_1 + \frac{(Y_N - \mu_N)}{\mu_N} \end{bmatrix}$$

From the above matrices

$$X^T W X = \begin{bmatrix} 1 & 1 & . & . & . & 1 \\ X_1 & X_2 & . & . & . & X_N \end{bmatrix} \begin{bmatrix} \mu_1 & & & & \\ & \mu_2 & & 0 & \\ & & . & & \\ & & & . & \\ & 0 & & . & \\ & & & & \mu_N \end{bmatrix} \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ . & . \\ . & . \\ . & . \\ 1 & X_N \end{bmatrix}$$

$$X^T Wz = \begin{bmatrix} 1 & 1 & . & . & . & 1 \\ X_1 & X_2 & . & . & . & X_N \end{bmatrix} \begin{bmatrix} \mu_1 & & & 0 \\ & \mu_2 & & \\ & & . & \\ & & & . \\ & 0 & & . \\ & & & \mu_N \end{bmatrix} \begin{bmatrix} \beta_o + \beta_1 X_1 + \dfrac{(Y_1 - \mu_1)}{\mu_1} \\ \beta_o + \beta_1 X_1 + \dfrac{(Y_2 - \mu_2)}{\mu_2} \\ . \\ . \\ . \\ \beta_o + \beta_1 X_1 + \dfrac{(Y_N - \mu_N)}{\mu_N} \end{bmatrix}$$

Which then give:

$$X^T WX = \begin{bmatrix} \sum_{i=1}^{N} \mu_i & \sum_{i=1}^{N} \mu_i X_i \\ \sum_{i=1}^{N} \mu_i X_i & \sum_{i=1}^{N} \mu_i X_i^2 \end{bmatrix} \quad \textbf{(15)}$$

And

$$X^T Wz = \begin{bmatrix} \sum_{i=1}^{N} \mu_i \left(\beta_o + \beta_1 X_i + \dfrac{(Y_i - \mu_i)}{\mu_i} \right) \\ \sum_{i=1}^{N} \mu_i X_i \left(\beta_o + \beta_1 X_i + \dfrac{(Y_i - \mu_i)}{\mu_i} \right) \end{bmatrix} \quad \textbf{(16)}$$

Since $In(\mu_i) = \beta_o + \beta_1 X_i, \quad thus, \quad \mu_i = e^{(\beta_o + \beta_1 X_i)}$

Therefore, (15) and (16) become:

$$X^T WX = \begin{bmatrix} \sum_{i=1}^{N} e^{(\beta_o + \beta_1 X_i)} & \sum_{i=1}^{N} X_i e^{(\beta_o + \beta_1 X_i)} \\ \sum_{i=1}^{N} X_i e^{(\beta_o + \beta_1 X_i)} & \sum_{i=1}^{N} X_i^2 e^{(\beta_o + \beta_1 X_i)} \end{bmatrix} \quad \textbf{(17)}$$

$$X^T Wz = \begin{bmatrix} \sum_{i=1}^{N} e^{(\beta_o + \beta_1 X_i)} \left(\beta_o + \beta_1 X_i + \dfrac{Y_i}{e^{(\beta_o + \beta_1 X_i)}} - 1 \right) \\ \sum_{i=1}^{N} X_i e^{(\beta_o + \beta_1 X_i)} \left(\beta_o + \beta_1 X_i + \dfrac{Y_i}{e^{(\beta_o + \beta_1 X_i)}} - 1 \right) \end{bmatrix}$$

The maximum likelihood estimates are obtained iteratively using Equation (14), initial values can be obtained by applying the link to the data, that is taking the natural log of the response, and regressing it on the predictors (or explanatory variables) using ordinary least square (OLS) method given by:

$$\hat{\beta} = \left(X^T X \right)^{-1} X^T Y$$

Goodness of Fit Test and Model Selection

Naturally, the modeling process yield a number of models out of which the 'best' model is then selected. Thus, a number of selection strategies were used; the Pearson Chi-Square, the Akaike Information Criterion (AIC). R-Programming was used in the modeling.

Akaike Information Criterion

Bozdogan (2000) described AIC as one of the criterion in determining the best model, as follows:

$$AIC = -2InL(\theta) + 2k$$

Where $L(\theta)$ is likelihood valued, and k the number of parameters. The best model is the model that has the smallest AIC value.

Model Checking Using Pearson Chi-Squares and Deviance

The popular measures of the adequacy of the model fit the Pearson Chi-Squares and deviance. To check the goodness of fit of the model, the following hypotheses are required.

H_o : the model has a good fit

H_1 : the model has lack of good fit

Pearson Chi-Squares

Let Y_i be the observed count and $\hat{\mu}_i$ be the fitted mean value. Then, for Poisson regression analysis, the Pearson Chi-Squares statistics is given by:

$$X^2 = \sum \frac{\left(Y_i - \hat{\mu}_i\right)^2}{\hat{\mu}}$$

Its degree of freedom is equal to the number of response counts. N minus the number of parameters in the model, p. H_o will be rejected if $X^2 > \chi^2_{\alpha(N-p)}$ indicating lack of fit of the model.

Deviance

The deviance is given by:

$$D = 2\sum Y_i In\left(\frac{Y_i}{\hat{\mu}_i}\right)$$

For large samples, the deviance also has the approximate chi-square distribution with (N-p) degrees of freedom. Similar to Pearson Chi-Squares, H_o will be rejected if $D > \chi^2_{\alpha(N-p)}$ indicating lack of fit of the model.

Over-Dispersion Test

In the Poisson regression analysis using, there is an assumption of equi-dispersion that the mean value of the variance must meet. However, this assumption is rarely fulfilled so that it lead to the case of over dispersion. Values for detecting the over dispersion, can be seen from the value of deviance/df or Pearson/df. If the value of deviance/df or Pearson/df is greater that 1, It can be said to be a case of over dispersion whereas if it is less that 1 then there is under dispersion.

RESULTS

The number of infants infected with pneumonia was modelled using Poisson regression and the results are presented in the Table 1 below. The table contains various models generated and their AIC's.

From Table 1 above shows that the best model which from 2013 to 2015 is Model 1 which is the main effect only because it has the smallest AIC. The AIC of Model 1 was found to be 235.04.

To check the goodness of fit of the model, the following hypotheses are required.

H_o : the model has a good fit

versus

H_1 : the model has lack of fit

Table 1: The Poisson Regression Models for the Number of Infants Infected with Pneumonia from 2013 to 2015 with their AIC's.

Model	Possible Models Generated	Equations	AIC's
1.	Main Effect Only	$In(y) = \alpha + \beta_1 Age + \beta_2 Gender + \beta_3 Hosp + \beta_4 Year$	235.04
2.	Possible Interaction	$In(y) = \alpha + \beta_1 Age + \beta_2 Gender + \beta_3 Hosp + \beta_4 Year + \beta_5 (Age * Gender)$	236.12
3	Possible Interaction	$In(y) = \alpha + \beta_1 Age + \beta_2 Gender + \beta_3 Hosp + \beta_4 Year + \beta_5 (Hosp * Year)$	239.01
4	Possible Interaction	$In(y) = \alpha + \beta_1 Age + \beta_2 Gender + \beta_3 Hosp + \beta_4 Year + \beta_5 (Age * Year)$	241.80
5.	Saturated Model	$In(y) = \alpha + \beta_1 Age + \beta_2 Gender + \beta_3 Hosp + \beta_4 Year + \beta_5 (Age * Gender) + \beta_6 (Age * Year) + \beta_7 (Gender * Year) + \beta_8 (Age * Gender * Year)$	260.41

Table 2 (a): The Parameter Estimate of Selected Poisson Regression Model for Number of Infants Infected with Pneumonia from 2013 to 2015.

Predictor	Estimate	Standard Error	Z-value	P-Value	Odd Ratio
Intercept	-1.371037	0.041790	-32.808	2e-16***	0.2538
Male (Reference)					
Female	0.003618	0.041204	0.088	0.93002	1.0036
Age <1 Year (Reference)					
1 – 3 year	0.013198	0.043717	0.302	0.76273	1.0133
4 – 5 years	0.192946	0.140690	1.371	0.17024	1.2128
Valli (Ref)					
Specialist Hosp.	-0.320619	0.118218	-2.712	0.00669**	0.7257
2013 (Reference)					
2014	-0.236317	0.047514	-4.974	6.57e-07***	0.7895
2015	-0.212099	0.050936	-4.164	3.13e-05***	0.8089

Significant code: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 2(b): Summary Result of Infants Infected with Pneumonia Using Poisson Regression.

Criterion	Value	Df	Value/df
Pearson Chi-Square	34.76787	29	1.1989
Deviance	32.73079	29	1.1286

Table 2(b) above show Pearson Chi-square of 34.76787 and deviance of 32.73079 both on 29 degree of freedom following the critical value of chi-square distribution (χ^2) with 29 degree of freedom with 42.557 at 5% level of significance which Pearson Chi-Square and deviance value are both less than the critical value of 42.557 indicating good fit. The dispersion parameter was found to be 1.1989 and 1.1286 which is not far from 1, which indicate there is no over dispersion in the data and Poisson regression is best fit model for infant infected with pneumonia from 2013 to 2014.

Interpretation of Coefficient

From Table 2(a), it is noted that the risk factors Specialist hospital, year 2014 and 2015 are significant at α=0.05 with their respective significance values equal to 0.00669, 6.57e-07 and 3.13 e-05. Also, female, age one-three years and four-five years are not statistically significant.

The odds that female children will suffer from pneumonia is about 1.0036 times that of males. Children under one year seem less likely to suffer from the disease than those between one-thee years and four-five years are 1.0133 and 1.2128 times as likely to have pneumonia. Clearly, the chance of being affected with pneumonia increases as the babies advance in age. For Valli Clinic compared to Specialist Hospital has odds of 0.7257, which indicate sthat patient who visited the clinic infected with pneumonia in Specialist Hospital are less than that of Valli Clinic. For the year 2013 compared to 2014 and 2015 recording an odds of 0.7895 and 0.8089 which are less than 1 indicating that patient with pneumonia decrease every year but rose in 2015 compared to 2014.

CONCLUSIONS

The Pearson Chi-Square and deviance statistic was invoked to check the goodness of fit of the model, the model fitted Poisson regression. The findings suggest that Pneumonia is still prevalent amidst under five in Adamawa State with reference to a specialist hospital and Valli Clinic. Pneumonia occurred more commonly in females. We suggest that educating mothers about the mechanism of environmentally related disease transmission.

REFERENCES

1. Bozdogan, H. 2000. “Akaike Information Criterion and Recent Development in Information Complexity”. *Journal of Mathematical Psychology*, 44: 62-91.

2. WHO/UNICEF. 2009. “Global Action Plan for Prevention and Control of Pneumonia (GAPP)”. WHO (Geneva). Nov. (Cited:2011 Oct.). Available from http://www.unicef.org/media/files/GAPP3-web.pdf.

3. WHO/UNICEF. 2013. “Progress on Sanitation and Drinking Water: 2013 Update”. World Health Organization: Geneva, Switzerland.

4. Yaguo Ide, L.E. and A.R. Nte. 2011. “Pneumonia and Under Five Morbidity and Mortality”. *The Nigeria Health Journal*.11(3, July-September).

5. Igor, R., P.C. Boschi, Z. Biloglar, K. Mulhaland, and H. Combell. 2008. “Epidemiology and Etiology of Childhood Pneumonia”. *Bulletin of World Health Organization*. 86:408-416.

ABOUT THE AUTHORS

Adesupo Akinrefon, is a Lecturer in the Department of Statistics and Operations Research, Modibbo Adama University of Technology, Yola. He holds a Master of Science Degree in Statistics, (Ph.D. in view) from the University of Ilorin, Nigeria. His research interests include categorical data analysis, stochastic processes, and modelling.

Olateju A. Bamgbala, is a Graduate of Statistics from the Department of Statistics and Operations Research, Modibbo Adama University of Technology, Yola. He holds a Bachelor of Technology Degree in Statistics.

Saidu S. Abdulkadir, is a Lecturer in the Department of Statistics and Operations Research, Modibbo Adama University of Technology, Yola. He holds a Ph.D. in Statistics from the University of Ilorin, Nigeria. His research interests include categorical data analysis and biostatistics.

SUGGESTED CITATION

Akinrefon, A.A., O.A. Bamigbala, and S.S. Abdulkadir. 2016. “Pneumonia Prevalence among Under-Five in Nigeria: A Poisson Regression Model Approach”. *Pacific Journal of Science and Technology*. 17(2):300-309.