Study of Binary Logistic and Poisson Regression Models of Diabetic patients in Nigeria using Dichotomous and Non- Dichotomous predictors

Abstract

The comparative study of the Binary-logistic and Poisson regression models of diabetic patients in Nigeria was presented using R-squared, Adjusted R-squared, Variance Inflated Factors and Akaike Information Criterion for two different data sets of Diabetic Patients known as the dichotomous and the Non-dichotomous data obtained from the University of Port Harcourt Teaching Hospital (UPTH). The results revealed that the Binary logistic regression was better than the Poisson regression for both dichotomous and non-dichotomous data. It was also, observed that, the Binary-logistic regression model was significant in this study with a non-dichotomous data set, while Poisson regression was not significant. The results also showed that both the type 1 and type 2 diabetes have negative effects on the diabetic Patients.

Keywords: Binary-logistic, Poisson regression, Dichotomous data, Non-dichotomous data, diabetes Type 1 diabetes Type 2 diabetes.

1. Introduction

Obviously, regression analysis makes use of the relationships between two or more quantifiable variables, which means that, an output variable can be predicted from the other(s). This is widely applied in behavioral, business, biological and social sciences among other disciplines. (Ijomah et al. 2018, Michael et al.2005). In practice, there are two types of regressions and they are Linear and Nonlinear regressions. While the linear regression can either be a simple or a multiple linear regression (Ijomah et al. 2018 and Nduka, 1999) and the Nonlinear regressions are either log-linear, quadratic, cubic, exponential, Poisson, logistic and power regression. The interest in this research lies on the Poisson and the Logistic regression. Poisson regression is useful for a count data. It is used to estimate rates or counts comparing different exposure groups, similarly, logistic regression is used to estimate odds ratios comparing different exposure groups. But the two models logistic and Poisson regressions are related in the sense that they are used to determine which variables are important and the direction of the effect for each variable. These models allow analysts to take account of the knowledge present in a set of observations between the dependent variable and independent variables (Ijomah et al. 2018). Poisson regression model in general form is comparable to Binary logistic regression and multiple regression models. Poisson distribution can be applied in many fields linked to counting such as; Telecommunication, Biology, Agriculture, Radioactivity. etc. In the same way, Binary logistic regression is an essential model considered for use when the response variable is binary with two possible outcomes, such as financial status of firm (profit or loss), blood pressure: (High or low) etc. Both models are appropriate for analyzing data arising from either observational or experimental studies (Michael et al, 2005 and Ijomah et al. 2018).

This paper looked at Poisson and Binary-logistic regression models because the response outcomes obtained are discrete (or binary response variable).

Ijomah *et al.* (2018) in their work focused on comparing Logistic model with Poisson model using a count data of household utilized not utilized primary health care services and looked at the effects of the predictors on the response for binary count data. But this paper, seeks to examine the Binary-logistic and the Poisson models on a data set arranged in two forms which

are: dichotomous and non-dichotomous data sets. The dichotomous data considers the diabetic patients as the response and male and female as the predictors, while, the non-dichotomous data set looks at the diabetic patients as response and the type 1 and type 2 diabetes as the predictors. The dichotomous data is the binary count data coded as 0 for female and 1 for male, while the non-dichotomous data is not a coded data. In the other words, Ijomah *et al.* (2018) compared these two models using a binary count data, while in this work, the two models were compared using binary coded and binary un-coded data.

This work is aimed at comparing Poisson and Binary logistic regression analysis using two data sets of diabetic Patients in Nigeria known as dichotomous and non-dichotomous data. The dichotomous data considers the Male and Female Patients, if 1, it is Male, if 0, it is Female, while the non-dichotomous data considered the diabetic Patients on the type of diabetes they are suffering.

Why logistic regression in this study? The reasons are as follows,

1) Mathematically, it is an extremely flexible and easily used function.
2) It gives meaningful interpretation clinically.

Testing for significance of a study in logistic regression, it implies that the overall significance is based upon the value of G test-statistic which is usually referred to as the deviance statistic. Usually, interpreting a regression equation comprises relating the predictor variables to the response variable that the equation was developed to answer. Nonetheless, using logistic regression, it is difficult to interpret the relation between the predictor variables and that probability that Y = 1 directly because the logistic regression equation is non-linear. As a result, Statisticians have shown that the relationship can be interpreted indirectly using the odds ratio. The odds in favour of an event occurring is defined as the probability that the event will occur divided by the probability that the event will not occur. In logistic regression, the event of interest is always Y = 1. Logistic regression forms a best fitting equation or function using the maximum likelihood method, which maximizes the probability of classifying the observed data into the appropriate category given the regression coefficients.

Wansu *et al*. (2018) posited that point estimates from the robust Poisson models were unbiased. Under model misspecification, the robust Poisson model was generally preferable because it provided unbiased estimates of risk ratios.

Berk (2003) observed that count data are used commonly in criminological research. When the response variable is a count, one option is to employ Poisson regression as a special case of the generalized linear model. Poisson formulation is somewhat simple to interpret because the right hand side is the familiar linear combination of predictors and because when it is exponential, the regression coefficients are interpreted as multipliers. The applications of Poisson regression have been discussed by a number of researchers, expecially in Criminology; Paternoster and Brame (1997), Sampson and Laub (1997); Osgood (2000).

According to Yanqiu *et al.* (2009), Poisson regression was used to study time trends and regional differences in maternal mortality (RMM) in China from 2000-2005 and found that RMM declined by an average of 5% per year. Here, Poisson regression model is used to examine the incidence of maternal mortality at the hospital. The Poisson model assumes that the variance of the count data is equal to the mean (Agresti, 2007). The coefficients of the Poisson regression model are estimated using the maximum likelihood techniques. The deviance (likelihood ratio) test statistic, G2, is used to assess the adequacy of the fitted model.

Xiu & Engel (2012) observed that neglect of retransforming random errors in the random-effects multinomial logit model results in severely biased longitudinal trajectories of health probabilities as well as overestimated effects of covariates on the probabilities.

2. Materials and Methods

Logistic Regression Model

According to Ijomah et al (2018), Logistic regression model with binary response variables is given as;

$$Y_i = \beta_0 + \beta_1 X_i + \mathcal{E}_i \quad (1)$$

Where Y_i is a binary response, having the values 0 or 1.

$E(Y_i)$ has a special meaning, we have that;

$$E(Y_i) = \beta_0 + \beta_1 X_i \quad (2)$$

Since $E(\mathcal{E}_i) = 0$

Where E represents expectation.

For Y_i to be a discrete random variable for which the probability distribution can be stated as seen in table 1 below.

Table 1 (Probability distributions of binary response variable)

Y_i	Probability

1	$\mathrm{P}(Y_i = 1) = \pi_i$
0	$\mathrm{P}(Y_i = 0) = 1 - \pi_i$

Such that

$\mathrm{P}(Y_i = 1) + \mathrm{P}(Y_i = 0) = \pi_i + 1 - \pi_i = 1$

The above simply means that, if 1 is probability of success and 0 is the probability of failure, then,

$$P(Success) + P(Failure) = 1 \quad (3)$$

$$P(Success) = 1 - P(Failure) \quad (4)$$

In this study, for dichotomous data, if 1, the diabetic patient is male, if 0, it is female.

Note that

$$E(Y_i) = 1(\pi_i) + 0(1 - \pi_i) = \pi_i \quad (5)$$

Hence, $E(Y_i) = \pi_i = P(Y_i = 1)$ (6)

Equating (2) and (6), we have

$$E(Y_i) = \beta_0 + \beta_1 X_i = \pi_i \quad (7)$$

The mean response of Logistic regression is given as

$$E(Y_i) = \pi_i = \frac{\exp(\beta_0 + \beta_1 X)}{1 + (\beta_0 + \beta_1 X)} \quad (8)$$

Though, Logistic regression which is one of the generalized linear model is given as seen in Parastoo and Ahmad (2016) as

$$Logit(p) = \ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}, i = 1, \ldots, n \quad (9)$$

The link function of a logit regression is $f(\mu_r) = \ln\left(\frac{p}{1-p}\right)$

Where

$p = p_r(Y_i = 1)$ and

$$p = p_r\left(Y_i = \frac{1}{X}\right) = \frac{e^{\mu+\beta_1 x_{1i}+\cdots+\beta_k x_{ki}}}{1+e^{\mu+\beta_1 x_{1i}+\cdots+\beta_k x_{ki}}} \quad (10)$$

Actually, logistic regression is similar to the linear regression but the difference is that their coefficients are different. As the linear regression minimizes the sum of square errors, while the logistic regression maximizes the probability that an event will occur. We use the F and t-statistics for linear regression, while we use chi-square and Wald for logistic regression.

It is generally given as

$$\log\left(\frac{p}{1-p}\right) = \mu + \sum \beta X \quad (11)$$

X is a vector of independent variables

β is the vector of estimated parameters

P=is the likely outcomes or events.

The dichotomous and non-dichotomous data were applied in equation (1) and (9) using Minitab software.

Poisson Regression Model

The Poisson distributions are count variables, say,
$Y_i = 0, 1, \ldots,$ which always difficult to obtain repeated numbers. It is given as
$f(\mathfrak{r}_i) = \frac{\mu^{\mathfrak{r}}\exp(-\mu)}{\mathfrak{r}!}$; $\mathfrak{r}$=0, 1, . . . (12)
where
$f(\mathfrak{r}_i)$ is the probability that $\mathfrak{r}$ and $\mathfrak{r}!=\mathfrak{r}(\mathfrak{r}-1)$,. . ..3.2.1
the mean and variance of a Poisson distribution are respectively $E(\mathfrak{r}) = \mu$ and $\delta^2(\mathfrak{r}) = \mu$
that is to say the mean and variance of a Poisson distribution are equal for certain. For a certain time, interval t
$f(\mathfrak{r}_i) = \frac{t(\mu^{\mathfrak{r}})\exp(-t\mu)}{\mathfrak{r}!}$ (13)
The Poisson model is given as
$E(\mathfrak{r}_i) + \mathcal{E}_i$; i=0, 1, . . . , (14)

Coefficient of Determination

As seen in Onu and Inamete (2022), the Coefficient of determination is given as

$$R^2 = \frac{\text{SSR}}{\text{SSTotal}} \quad (15)$$

$$=1 - \frac{\text{SSE}}{\text{SSTotal}} \quad (16)$$

It lies between 0 and 1, the more the value is close to one, the better the model fit on the data, while as the value becomes nearer to zero, the inferior the model fit.

Adjusted Coefficient of Determination

We also another test statistic known as adjusted Coefficient of Determination given as seen

$$R^2_{Adjusted=} = (\frac{n-1}{n-p})(\frac{SS_{Error}}{SS_{Total}}) \quad (17)$$

$$= 1 - \frac{\frac{MSE}{SSTotal}}{\frac{}{n-1}} \quad (18)$$

Odd ratio Test

Specifically, Odd ratio is a useful way of interpreting the relationship that exists between a predictor and responses. The odds ratio (q) can be any nonnegative number. The odds ratio = 1 is used to serves as the standard for comparison. If q = 1, it indicates there is no association between the response and predictor. Also, if q > 1 the odds of success are higher for the reference level of the factor, also, if q < 1, then, the odds of success are less for the reference level of the factor or for greater levels of a continuous predictor. higher values farer from 1 represent stronger degrees of association.

The odd increases multiplicatively by the factor e^{β_i} for every one unit increase in X, which is equivalent to $exp(\beta_i)$

Akaike Information Criterion

This criterion was also applied and it is given as

$$AIC = n \ln SSE - n \ln n + 2p \quad (19)$$

Where n is the number of sampled variables, SSE is the Sum of Square Error and p is the number of model parameters.

3. Results and Discussion

Illustration 1 For a non-dichotomous data using Binary logistic regression

Binary Logistic Regression: gender versus tupe 1, type 2

Method

Link function Logit

Rows used 24

Response Information

Variable	Value	Count	
gender	1	12	(Event)
	0	12	
	Total	24	

Table 2: Deviance Table

Source	DF	Adj Dev	Adj Mean	Chi-Square	P-Value
Regression	2	6.776	3.388	6.78	0.034
tupe 1	1	1.616	1.616	1.62	0.204
type 2	1	1.932	1.932	1.93	0.165
Error	21	26.495	1.262		
Total	23	33.271			

Table 3: Model Summary

Deviance R-Sq	Deviance R-Sq(adj)	AIC
20.37%	14.35%	32.50

Table 4: Coefficients

Term	Coef	SE Coef	VIF
Constant	2.68	1.34	
tupe 1	-0.0598	0.0561	1.12
type 2	-0.0716	0.0548	1.12

Table 5: Odds Ratios for Continuous Predictors

	Odds Ratio	95% CI
tupe 1	0.9420	(0.8440, 1.0513)
type 2	0.9309	(0.8361, 1.0364)

Table 6: Regression Equation

P(1) = exp(Y')/(1 + exp(Y'))

Y' = 2.68 - 0.0598 tupe 1 - 0.0716 type 2

Table 7: Goodness-of-Fit Tests

Test	DF	Chi-Square	P-Value
Deviance	21	26.50	0.188
Pearson	21	23.82	0.302
Hosmer-Lemeshow	8	5.54	0.698

Illustration 2 For a non-dichotomous data using Poisson regression

Poisson Regression Analysis: gender versus tupe 1, type 2

Method

Link function Natural log

Rows used 24

Table 8: Deviance Table

Source	DF	Adj Dev	Adj Mean	Chi-Square	P-Value
Regression	2	3.3876	1.6938	3.39	0.184
tupe 1	1	0.7964	0.7964	0.80	0.372
type 2	1	0.6622	0.6622	0.66	0.416
Error	21	13.2480	0.6309		
Total	23	16.6355			

Table 9: Model Summary

Deviance R-Sq	Deviance R-Sq(adj)	AIC
20.36%	8.34%	43.25

Table 10: Coefficients

Term	Coef	SE Coef	VIF
Constant	0.400	0.641	
tupe 1	-0.0325	0.0411	1.44
type 2	-0.0277	0.0350	1.44

Regression Equation

gender = exp(Y')

Y' = 0.400 - 0.0325 tupe 1 - 0.0277 type 2

Goodness-of-Fit Tests

Test	DF	Estimate	Mean	Chi-Square	P-Value
Deviance	21	13.24796	0.63086	13.25	0.900
Pearson	21	11.84354	0.56398	11.84	0.944

Illustration 3 For a Dichotomous data using Poisson regression

Poisson Regression Analysis: Diabetes versus Gender

Method

Link function Natural log

Rows used 48

Deviance Table

Source	DF	Seq Dev	Seq Mean	Chi-Square	P-Value
Regression	1	133.9	133.925	133.92	0.000
Gender	1	133.9	133.925	133.92	0.000
Error	46	458.6	9.971		
Total	47	592.6			

Table 11: Model Summary

Deviance R-Sq	Deviance R-Sq(adj)	AIC
22.60%	22.43%	713.49

Table 12: Coefficients

Term	Coef	SE Coef	VIF
Constant	3.8094	0.0304	
Gender	-0.5740	0.0506	1.00

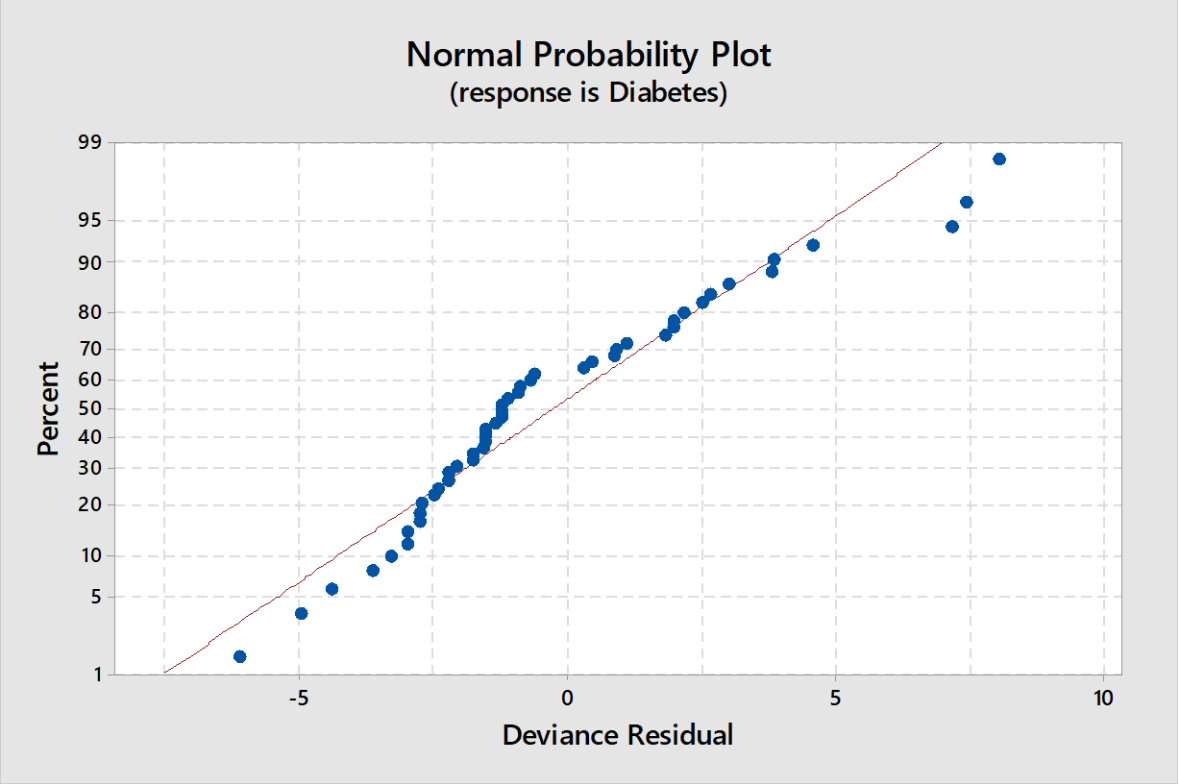
Table 13: Regression Equation

Diabetes = exp(Y')

Y' = 3.8094 - 0.5740 Gender

Table 14: Goodness-of-Fit Tests

Test	DF	Estimate	Mean	Chi-Square	P-Value
Deviance	46	458.64494	9.97054	458.64	0.000
Pearson	46	506.26877	11.00584	506.27	0.000



ng

Figure 1: Normal Probability plot

Illustration 4 For a Dichotomous data using Binary logistic regression

Binary Logistic Regression: Diabetes versus Gender

Method

Link function Logit

Rows used 48

Table 15: Response Information

Variable	Value	Count	
Diabetes1	1	31	(Event)
	0	17	
	Total	48	

Table 16: Deviance Table

Source	DF	Adj Dev	Adj Mean	Chi-Square	P-Value
Regression	1	16.88	16.8757	16.88	0.000
Gender	1	16.88	16.8757	16.88	0.000
Error	46	45.52	0.9896		
Total	47	62.40			

Table 17: Model Summary

Deviance R-Sq	Deviance R-Sq(adj)	AIC
27.04%	25.44%	49.52

Table 18: Coefficients

Term	Coef	SE Coef	VIF
Constant	2.398	0.739	
Gender	-2.909	0.850	1.00

Table 19: Odds Ratios for Continuous Predictors

	Odds Ratio	95% CI
Gender	0.0545	(0.0103, 0.2888)

Table 20: Regression Equation

P(1) = exp(Y')/(1 + exp(Y'))

Y' = 2.398 - 2.909 Gender

Table 21: Goodness-of-Fit Tests

Test	DF	Chi-Square	P-Value
Deviance	46	45.52	0.492
Pearson	46	48.00	0.392
Hosmer-Lemeshow	0	0.00	*

Table 22: Fits and Diagnostics for Unusual Observations

Obs	Observed Probability	Fit	Resid	Std Resid	
41	0.0000	0.9167	-2.2293	-2.28	R
43	0.0000	0.9167	-2.2293	-2.28	R

R Large residua

Discussion of Results

Illustration 1(Non-dichotomous data with Binary Logistic regression)

The study reveals that the Binary logistic regression is significant, which means that the model fits the data as evident in the P-value of 0.034 less than the 0.05 significant level. The R-square and adjusted R-square were obtained respectively as 20.37% and 14.35% and the Akaike Information Criterion value of 32.50. The Variance Inflated Factor was found to be 1.12 for both type 1 and type 2 diabetes. The odd ratio and confidence interval were respectively 0.9480 and (0.8140, 1.0513).

The regression equation shows that both type 1 and type 2 diabetes have negative effect on the diabetic Patients.

Illustration 2 (Non-dichotomous data with Poisson regression)

It was revealed that the Poisson regression is not significant in the analysis of the diabetic Patients as affected by the type 1 and type 2 cases. The R-square was found to be 20.36%, while adjusted R-square was 8.34% with the Akaike Information Criterion obtained as 43.25. The Variance Inflated Factor for type 1 and type 2 were found to be equal, with the value of 1.44. The regression equation shows that, the type 1 and type 2 cases both have negative effects on the diabetic Patients.

Comparing Illustration 1 and 2 (Binary logistic vs Possion regressions) for non-dichotomous data

It was observed that, the Binary logistic regression model was significant in this study with a non-dichotomous data set, while Poisson regression was not significant. It was also revealed that for the coefficient of determination, the Binary logistic model has a small value of coefficient of determination and the adjusted, but its value was greater than the value obtained in the Poisson regression with a percentage difference of 0.01 R-square, but the percentage difference between the adjuster-squared for Binary logistic and the Poisson regressions was 6.01. The AIC value for Binary logistic regression was found to be smaller than that of the Poisson regression by 10.75. The Variance Inflated Factor for Binary logistic and Possion regression differed by 0.32. All the above discussions show that Binary logistic regression is better than the Poisson regression for a non-dichotomous data. The two regressions reveal that type 1 and 2 diabetes have negative effects on the diabetic Patients.

Illustration 3 (dichotomous data for Poisson regression)

The study reveals that both the Poisson regression and Gender are significant at 5% level of significant level. The R-square was found to be 22.60% and the adjusted R-square was 22.43% with AIC value of 713.49 and VIF value of 1.0.

Illustration 4 (dichotomous data for Binary logistic regression)

The study reveals that both the regression and Gender are significant at 0.05 level. The R-square was found to be 27.04% and adjusted R-square was 25.44% and the AIC value was found to be 49.52.

Comparing Illustration 3 and 4 (Binary logistic vs Possion regressions) for dichotomous data

It was observed that Binary logistic regression was still better than the Poisson regression as revealed by the R-square, R-square adjusted and AIC. The two regressions have equal value of VIF.

Comparing Illustration 1 and 4 Binary logistic (non-dichotomous vs dichotomous data)

The comparison reveals that Binary logistic for a non-dichotomous data, is favoured by the AIC criterion, as it has the smallest value as compared with the Binary logistic for dichotomous data, but Binary logistic for dichotomous data was favoured by the R-squared and adjusted R-squared statistics.

Comparing Illustration 1 and 3 Poisson (non-dichotomous vs dichotomous data)

The Poisson regression has superior (better) AIC value for non-dichotomous data than the Poisson for dichotomous data, while, the Poisson for dichotomous data has superior R-squared and adjusted R-squared.

Conclusion

The study concludes that the Binary-logistic regression was better than the Poisson regression for both dichotomous and non-dichotomous data, and that, the Binary-logistic regression model was significant in this study with a non-dichotomous data set, while Poisson regression was not significant. Finally, it concludes that both the type 1 and type 2 diabetes have negative effects on the diabetic Patients.

Recommendation

It is recommended in this research that, Binary logistic regression is better than the Poisson regression for both dichotomous and non-dichotomous data.

Contribution to knowledge

This study has revealed that, in any of Binary logistic or Poisson regressions, the type 1 and 2 diabetes have negative effects on the diabetic patients. Also, the Binary-logistic regression gives smaller value of Aikaike Information Criterion than the Poisson. Comparing the Poisson regression for dichotomous and non-dichotomous data, the poisson regression gives better AIC for non-dichotomous data, while that of dichotomous data gives better R-square and R-squared adjusted.

References

Agresti, A. (2007), An Introduction to Categorical Data Analysis, Second Edition, *Wiley, Inc., New York.*

Berk, R. A. (2003), Regression Analysis: A Constructive Critique, C A: *Sage Publications.*

Ijomah, M. A., Biu, E. O., & Mgbeahurike, C. (2018). Assessing Logistic and Poisson Regression Model in Analyzing Count Data. *International Journal of Applied Science and Mathematical Theory 4(1), 42-68.*

Liu, X., & Engel, C. (2012). Predicting longitudinal trajectories of health probabilities with random-effects multinomial logit regression.

Michael, H. K.; Christopher, J. N., John N., and William. L (2005). "Applied Linear Statistical Model", fifth Ed.; 555-623., *McGraw Hill International.*

Nduka, E.C., (1999), Principles of Applied Statistics 1, Crystal Publishers, Okigwe.

Onu, O. H. & Inamete, E. N. H. (2022). Computational and Comparative study of the Impact of corporate Governance on financial performances of quoted Insurance Companies in Nigeria. *International Journal of Science Academic Research,* 3(2), 3475-3482.

Osgood, W. (2000), "Poisson-based Regression Analysis of Aggregate crime Rates". Journal of Quantitative Criminology 16, 21 – 43.

Paternoster R, and Brame R. (1997), "Multiple routes to delinquency: A test of developmental and general theories of crime", Journal of Criminology 35, 45-84.

Sampson, R. J. and Laub, J. H. (1997), A Life-Course Theory of Cumulative Disadvantage and the Stability of Delinquency, in Developmental Theories of Crime and Delinquency. (Advances in Criminological Theory, 7, 133-161.

Wansu C., Lei Q., Jiaxiao S., and Meredith F. (2018). Comparing performance between Log-binomial and robust Poisson regression models for estimating risk ratios under model specification. *BMC Medical Research Methodology*, *63*(2).

Yanqiu, G., Ronsmans, C. and Lin, A. (2009), Time Trends and Regional Differences in Maternal Mortality in China from 2000 to 2005, *Peking University Health Science Center, Beijing, China, Bull World Health Organ* 87:913–920.

Armstrong, J. S. (2012), Illusions in Regression Analysis, International Journal of forecasting. 28 (3), 689.

Henry, G. A. (2010). Comparison of Akaike Information Criterion (AIC) and Bayesian Criterion (BIC) in Selection of an Asymmetric Price Relationship. Journal of Development and Agricultural Economics, Vol 2 (1), page 001-006.

Hosmer, D. W., and Lemeshow, S. (2000), Applied Logistic Regression, Second Edition, John

Wiley and Sons Inc.