International Journal of
Science and Engineering Investigations vol. 7, issue 77, June 2018
ISSN: 2251-8843

Application of Poisson and Negative Binomial Regression Models in Modelling Oil Spill Data in the Niger Delta

N. L. Nwakwasi
Department of Civil Engineering SEET, Federal University of Technology Owerri, P.M.B 1526, Imo State, Nigeria
(ngozinwakwasifuto@yahoo.com)

Abstract-Oil Spill is one of the highest causes of environmental pollution in the Niger Delta region and over the years its causes have been attributed to several random factors. Billions of naira are lost to Oil spill on a yearly basis. The ultimate goal of this study is to apply Poisson and Negative binomial regression models to identify the major factors that could aid Oil spill reduction in the Niger Delta and to ascertain the better model suitable for prediction of Oil spill, using a secondary data which was source from Department of Petroleum Resources (DPR) 2007, the collected data span from 1980-2006. Maximum likelihood estimation procedure was used to estimate the parameters of the selected model. With the number of Oil spill as the response variable (Y) and 6 - explanatory variable (*X's*), also applying the forward selection criteria to the 6-explanatory variables, model 3 is best suitable for forecasting the subject under study. The result of the Poisson regression model showed that there was over dispersion in the Oil spill data since the dispersion parameter (2.774) was greater than 1, hence underestimating the standard error and overestimating the coefficients of the explanatory variable, consequently, giving misleading inferences. The result of the assessment criteria for Poisson regression model and Negative binomial regression model revealed that the Negative binomial regression models Oil spill data better in the Niger Delta region as considered in this study. Production Operation (PO), Corrosion of Pipeline (CP) and Flowline Replacement (FR) are the major contributors to Oil spill in the Niger Delta region. With flowline replacement ranking first. A model suitable for prediction of oil spill in the Niger Delta has been developed.

Keywords- *Oil Spills, Poisson Regression Model, Negative Binomial Regression Model, Akaike Information Criteria (AIC)*

I. INTRODUCTION

Environmental concern had been growing over the Niger Delta region, both nationally and internationally. Inspite of the Delta's resources endowment, it immense potential for economic growth and sustainable development, the region is in a terrible state. This under increasing threat from rapid deteriorating economic conditions and social tensions which are not being addressed by current Government Policies and behaviors patterns (Osuji, 2001)

A recent study of this region by the World Bank has warned that "an urgent need exists to implement mechanism to protect the life and health of the regions inhabitants and its ecological systems from further deterioration" (Osuji, et.al, 2006).

In line with this, there are enormous unique challenges facing oil companies in Nigeria, the Shell Petroleum Development Company alone has 86 flow-stations (East & West) and dome 6,200 kilometers of pipeline and flowlines in 31,000 square kilometers of the Niger Delta and one can guess the amount of oil spills expected from such facilities. The six major causes of oil of oil spills include:

a. Production operation and facilities
b. Corrosion of pipelines
c. Flowline replacement
d. Flow station Upgrade
e. Leaking values
f. Tampering/Sabotage.

Available records (Fig 1&2) shows that oil spill incidents have occurred in many part of the Niger Delta since 1970 till date. According to the Directorate of Petroleum Resources (DPR), between 1986 - 2006 (which this study deems to address) a total of 4835 incident s resulted in the spill of approximately 2,446,322 barrels (388940.73m^3) of oil into the environment. Of this quantity, an estimated 1,896,930 barrels (301592.9m^3) comprising 77% were lost to the environment (DPR, 2007).

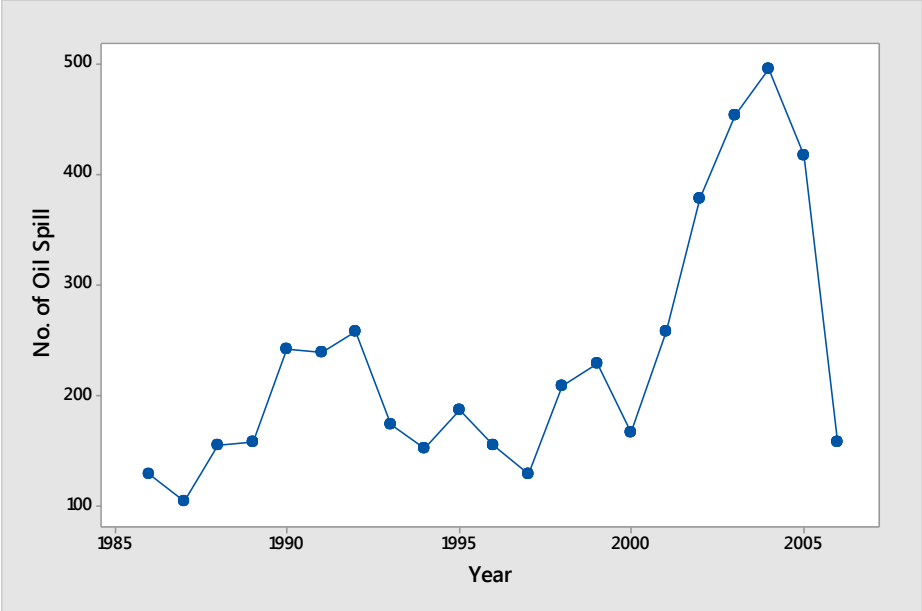

Figure 1. Time series plot of Oil spill data in Niger Delta from 1986 - 2006

One of the major objectives of this study is to provide a model which will be able to predict the major contributors of oil spill in Niger Delta, in the process of oil exploration and exploitation based on the available oil spill data. This will keep oil companies to take safety measures in their operations.

II. METHODS

A. *Poisson Regression Model*

*B*y definition, $Y_{(\text{dependent variable})}$ follows a Poisson distribution with parameter $\lambda > 0$ iff the probability distribution function is given by:

$$P(Y = k) = \frac{\exp(-\lambda)\lambda^k}{k!} \quad (1)$$

For K=0,1,2,… such that $E(Y) = \lambda \; and \; Var(Y) = \lambda$

For n independent random variables $Y_1, Y_2, \dots, Y_n$

$Y_i \sim P(\mu_i)$, and suppose we want to let the mean and the variance depend on the explanatory variables x_i's

$$g(\mu_i) = \mu_i = x_i'\beta = \eta_i, \quad (2)$$

We can consider a generalized linear model with log link as:

$$\log(\mu_i) = x_i'\beta. \quad (3)$$

Where X_i denotes the vector of explanatory variables and β denotes the vector of regression parameters.

B. *Model Specification*

The model for the Poisson regression is given as:

$$\lambda_i = E(Y_i) = \exp\left(\alpha + \sum_{j=1}^{k} x_j'\beta_j\right), \quad (4)$$

Where α is the intercept, Since the mean is equal to the variance the usual assumption of Homoscedasticity would not be appropriate for a Poisson data.

C. *Parameter Estimation-Iterative Reweighted Least Squares*

Estimation of parameters in Poisson regression relies on maximum likelihood estimation (MLE) method.

The log likelihood function is given as:

$$\log L(\beta) = \sum\{y_i \log(\mu_i) - \mu_i\}, \quad (5)$$

To obtain the maximum likelihood for the parameter β_j we employ the chain rule

$$\frac{\partial l}{\partial \beta_j} = U_j = \sum_{i=1}^{N} \frac{\partial l_i}{\partial \beta_j} = \sum_{i=1}^{N} \left(\frac{\partial l_i}{\partial \theta_i} \cdot \frac{\partial \theta_i}{\partial \mu_j} \cdot \frac{\partial \mu_i}{\partial \beta_i}\right) \quad (6)$$

Hence the iterative equation would be written as:

$$b^{(m)} = (X^T W X)^{-1} X^T W \quad (7)$$

Where $w_{ii} = \frac{1}{\text{var}(Y_i)}$ and $z_i = y_i$

D. *Negative Binomial Regression (NBR) Model*

NBR is a popular generalization of Poisson regression because it loosens the highly restrictive assumption that the variance is equal to the mean made by the Poisson model. The traditional negative binomial regression model, commonly known as NB2, is based on the Poisson-gamma mixture distribution. This model is popular because it models the Poisson heterogeneity with a gamma distribution. This is given as:

$$\lambda = E(y_i) = \exp(\alpha + \sum \beta_i X_i + \sum \gamma_i D_i) \quad (8)$$

$$P(Y_i = y_i) = \frac{\exp(-\mu_i)\mu_i^{y_i}}{y!}, \quad y_i = 0,1,2, \dots \quad (9)$$

$$Y_i \sim NB(\lambda = \exp(X^T\beta), \psi) \quad (10)$$

Compounding Poisson distribution above and a gamma distribution would give a Negative Binomial distribution

$$\Pr(Y_i = y_i|\mu_i, \alpha) = \frac{\Gamma(y_i+\alpha^{-1})}{\Gamma(y_i+1)\Gamma(\alpha^{-1})}\left(\frac{1}{1+\alpha\mu_i}\right)^{\alpha^{-1}}\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i} \quad (11)$$

Where $E(Y_i) = \mu_i$ while the variance $Var(Y_i) = \mu_i(1 + \alpha^{-1}\mu_i)$ where

$$\alpha = v^{-1} \quad (12)$$

- The regression coefficients are estimated using the method of maximum likelihood.
- The significant regression parameters from (11) will be subjected to a test of hypothesis.
- The AIC would reveal the better model between equation (4) and equation (11) then the rank test would be employed to ascertain the levels of contribution of the various factors under consideration to Oil spill in the Niger Delta.

E. Model Evaluation

In this research work, we would be applying the techniques below for evaluating the models to also ascertain the better model suitable for modelling Oil spill in the Niger Delta from 1976 – 1996:

Deviance: is a goodness of fit statistic for a model that is often used for statistical hypothesis testing and often used to compare two different models. Larger deviance indicates low performance of the model and less deviance indicates better performance of the model.

$$D = 2(l(y, \phi; y) - l(\hat{\mu}, \phi; y)) \quad (13)$$

Where $l(y, \phi; y)$ is the log-likelihood of the full model and $l(\hat{\mu}, \phi; y)$is the log-likelihood of the current model.

AIC: The Akaike would be used to in selecting the best model that fit our data. The model with the smaller AIC is the best.

$$AIC = -2(likelihood - K) \quad (15)$$

$$BIC = -2In(likelihood) - KIn(n) \quad (16)$$

Where n is sample size and K is number of predictors;

Pearson Residual: It is used to check for model fit of each variable (explanatory variables). It is the discrepancy between our observed and fitted values for each observation.

III. RESULTS AND DISCUSSION

In modelling the number of oil spill in Niger Delta, R statistical software version 3.3.3 was used. The Generalized Linear Model (GLM) with Poisson as the fundamental distribution for modelling a count data using the Log link function and the Negative Binomial distribution was latter employed to correct the error of over dispersion in the count data in situation where the result of the Poisson regression model shows over dispersion.

TABLE I. DESCRIPTIVE STATISTICS OF THE RESPONSE VARIABLE AND THE EXPLANATORY VARIABLES

Parameter	Minimum	Maximum	Mean	Std. Deviation
No.of.Spill	104	495	230.24	112.437
Production operation	22	104	48.38	23.604
Corrosion of pipeline	31	149	68.95	33.788
Flowline replacement	12	59	27.52	13.456
Flow station upgrade	9	45	20.86	10.185
Leaking values	5	25	11.52	5.733
Tampering/ sabotage	24	114	52.90	25.873

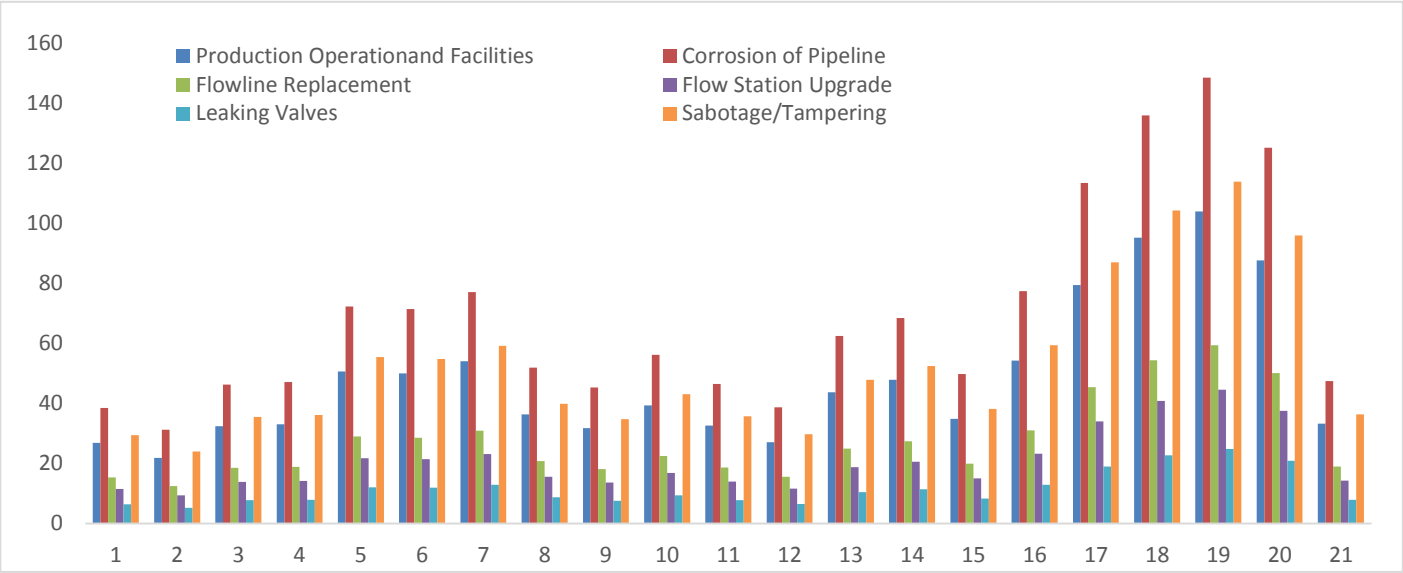

Figure 2. Cause of Oil Spillage (1986 - 2006)

A. Modelling Oil Spill Data in Niger Delta

TABLE II. THE POISSON REGRESSION MODELS FOR THE OIL SPILL DATA IN NIGER DELTA 1986 – 2006 WITH THEIR AIC'S

	Models	AIC
1	$\log(mean\ of\ No\ of\ Oil\ Spill) = \alpha_1 + \beta_i PO$	186.5
2	$\log(mean\ of\ No\ of\ Oil\ Spill) = \alpha_2 + \beta_1 PO + \beta_2 CP$	185.5
3	$\log(mean\ of\ No\ of\ Oil\ Spill) = \alpha_3 + \beta_1 PO + \beta_2 CP + \beta_3 FR$	181***
4	$\log(mean\ of\ No\ of\ Oil\ Spill) = \alpha_4 + \beta_1 PO + \beta_2 CP + \beta_3 FR + \beta_4 FU$	187.2
5	$\log(mean\ of\ No\ of\ Oil\ Spill) = \alpha_5 + \beta_1 PO + \beta_2 CP + \beta_3 FR + \beta_4 FU + \beta_5 LV$	194.4
6	$\log(mean\ of\ No\ of\ Oil\ Spill) = \alpha_6 + \beta_1 PO + \beta_2 CP + \beta_3 FR + \beta_4 FU + \beta_5 LV + \beta_6 TA$	197.4

PO=Production Operation, CP=Corrosion of Pipelines, FR=Flowline Replacement, FU= Flow Station Upgrade, LV=Leaking Values, TA=Tampering/Sabotage.

From table 2 above, we see that the best model which fit the Oil Spill data in Niger delta from 1986 – 2006 is model 3 because it has the smallest Akaike Information Criterion (AIC). The AIC of model 3 was found to be 181 with a null deviance of 984.016 on 20 degree of freedom and a residual deviance of 47.162 on 17 degree of freedom following the chi-square distribution ($\chi^2_{(1)}$).

The parameters and their estimates, the standard errors, the Chi-squares with their associated probability values are presented in table 3 below.

TABLE III. THE PARAMETER ESTIMATES OF THE SELECTED POISSON REGRESSION MODEL

Parameter	Estimate	Std. Error	Z values	Pr(>|Z|)
(Intercept)	4.489684	0.037857	118.597	< 2e-16
Production operation	0.002738	0.055449	0.049	0.960612
Corrosion of pipelines	-0.067475	0.033500	-2.014	0.043990
Flowline replacement	0.195370	0.057321	3.408	0.000654

The Table 3 presents the parameter estimates of the selected model for the number of Oil spilled in Niger delta over 21 years ie from 1986-2006. The Akaike information criterion (AIC) of this model was 181 with a null deviances of 984.016 on 20 degree of freedom and a residual deviance of 47.162 on 17 degrees of freedom following the chi-square distribution (X^2) on one degree of freedom. The dispersion parameter was found to be 2.774 (i.e residual deviance/degree of freedom as seen in table 5) and the Omnibus test flag a P-value equal 0.000 which implies that the model is significant at 5% α-level. However, the assumption of equality of mean and variances in Poisson distribution has been violated since the dispersion parameter is not approximately equal to 1. The dispersion parameter of the above model is 2.774 which is greater than 1, a clear indication of over dispersion in the oil spill data. This further implies that the parameters of the stated model have been over-estimated and the corresponding standard errors have been under estimated consequently giving a misleading inference about the regression parameters. To address this issue, Negative Binomial regression was used to modify the model to nullify the effect of over dispersion in the data and the result is shown in table 4 below.

TABLE IV. NEGATIVE BINOMIAL REGRESSION MODEL PARAMETER ESTIMATES

Parameter	Estimate	Std. Error	Z values	Pr(>|Z|)
(Intercept)	4.489682	0.037858	118.593	<2e -16
Production Operation	0.002738	0.055451	0.049	0.009612
Corrosion of pipelines	-0.067474	0.033501	-2.014	0.044000
Flowline replacement	0.195369	0.057322	3.408	0.000654

B. Interpretation of Coefficients

From table 4 it can be seen that flowline replacement; Corrosion of pipeline and production operation were all statistically significant since (p-value < 0.05).

The variables Production Operation (PO), Corrosion of Pipeline (CP) and Flowline Replacement (FR) where all statistically significant at 5% α level. This implies that Oil Spillage in the Niger delta does not just happens it is pivoted by several factors but in this study Production Operation, Corrosion of Pipeline and Flowline Replacement are the most significant contributors or causes of oil spill in the Niger Delta.

From Table 4 it is observed that the parameter estimates have been reduces and the standard errors have also been increased. The parametric analysis for the comparison between the Poisson and Negative Binomial Regression for goodness of fit of the model is shown in table 5.

TABLE V. ASSESSMENT CRITERIA FOR POISSON AND NEGATIVE BINOMIAL REGRESSION

Assessment Parameter	Poisson Regression Model	Negative Binomial Regression Model
Null Deviances	984.016	983.945
Degree of Freedom	20	20
Residual Deviance	47.162	17.161
Degree of Freedom	17	17
Log Likelihood	-83.980	-167.96
Dispersion Parameter	2.774	1.073
Akaike's Information Criterion (AIC)	181	177.96

From the result presented in table 5, it is clear that the negative binomial regression model is the better model which fits the Oil Spills Data in the Niger Delta because the dispersion parameter has reduced from 2.774 which was given by the Poisson model to 1.073 in the Negative Binomial model.

The Akaike Information Criteria (AIC) of the Poisson Regression model also reduced from 181 to 177.96 in the negative binomial model, we can see more supporting claims from the Null Deviances; Residual Deviance; Log Likelihood all of which also satisfy the condition that the Negative Binomial Model is the better model.

C. *The Model for Predicting Amount of Oil Spill in Niger Delta*

For negative binomial regression, the model for the Oil spill data is obtained as:

$$\log(\textit{mean of No of Oil Spill}) = 4.4897 + 0.002738\,\mathbf{(PO)} - 0.067474\,\mathbf{(CP)} + 0.195369\,\mathbf{(FR)} + e$$

TABLE VI. RANKING THE PROBABILITY VALUE OF THE ESTIMATES OF NEGATIVE BINOMIAL REGRESSION MODEL

s/n	Parameter	P-values	Ranks
1	Production Operation (PO)	0.009612	2nd
2	Corrosion of Pipeline (CP)	0.044000	3rd
3	Flowline replacement (FR)	0.000654	1st

From table 6 which ranks the P-values of the better model (Negative binomial Regression) to identify which parameter was more significant also which factor contributed more than the other. Flowline replacement was rank in the first position. Implication is that flowline replacement factor is the most statistically significant factor to contributing to Oil spills in the Nige delta followed by Production Operation and lastly Corrosion of pipeline.

IV. CONCLUSION

Negative binomial regression model is the better model suitable for modeling Oil Spillage in the Niger delta region as considered in this study. Production Operation (PO), Corrosion of Pipeline (CP) and Flowline Replacement (FR) are the major factors that contribute to the high number of Oil Spillage in the Niger Delta. The model which can be used to forecast future oil spill in the region keeping the factors considered in this study constant is shown below:

$$\log(\textit{mean of No of Oil Spill}) = 4.4897 + 0.002738\,\mathbf{(PO)} - 0.067474\,\mathbf{(CP)} + 0.195369\,\mathbf{(FR)} + e$$

REFERENCES

[1] DPR (2007) Environmental guidelines and standards for the Petroleum Resources, Ministry of Petroleum and Mineral Resources, Lagos Nigeria, pp 38-49

[2] Osuji, L.C (2001). Total Hydrocarbon content of soils, fifteen months after Eneka and Isiokpo oil spills. J. Appl. Sci. Environ Mgt. 5(2): 35-38

[3] Osuji L.C, Iniobong, D.C & Ojinnaka C.M (2006). Preliminary Investigation of Mgbede zo oil polluted site in Niger Delta, chen. Bioduv. 3:568-577

[4] NDES (1995), Niger Delta Environmental Survey, Steering Committee Report, Briefing Note 1, pp 2-4

[5] Cameron, A.C., and Trivedi, P.K. (1998). *Regression Analysis of Count Data*. Cambridge University Press. Cambridge, UK.

[6] Hinkelmann, K.; Kempthorne, O. (1994). Design and analysis of experiments. New York: Wiley-Interscience, v.1, 495p.

[7] Heinzl, H. and Mittlbock, M. 2003. Pseudo R-squared Measures for Poisson Regression Models with Over or Under-dispersion. Computational Statistics & Data Analysis. 44: 253 – 271.

[8] McCullagh P., Nelder J. A. (1989). Generalized Linear Models (Second edn). New York: Chapman and Hall.

[9] Nelder J. A., Wedderburn, R. W. M. (1972). Generalized Linear Models. Journal of the Royal Statistical Society, Series A 135(3), 370–384.

[10] White, G.C. & Bennetts, R.E. (1996) Analysis of frequency count data