INTERNATIONAL JOUNAL OF: SCIENCE RESEARCH AND TECHNOLOGY

TIMBOU-AFRICA ACADEMIC PUBLICATIONS
INTERNATIONAL JOURNAL, MAY, 2021 EDITIONS
VOL. 5 NO. 9 ISSN: 2623-7861

ANALYSIS OF CONTRIBUTION OF SELECTED RISK FACTORS TO CHOLERA USING MULTIPLE AND LOGISTIC REGRESSION MODEL

BOLANLE, A. OSENI[1], JONATHAN, O. ADEDOKUN[2], OLAWALE, B. FADIORA[3]; & BUSUYI, G. OGUNSANYA[4]

[1]Department of Mathematics and Statistics, the Polytechnic, Ibadan, Nigeria. [2]Department of Public Administration, Federal University Oye-Ekiti, Ekiti-State. Nigeria. [3]Deparment of Computer Science, the Polytechnic, Ibadan, Ibadan, Nigeria. [4]Department of Statistics and Mathematics, Moshood Abiola Polytechnic, Abeokuta, Nigeria.

ABSTRACT

Cholera remains a global threat to public health and one of the key indicators of social development. While the disease is no longer an issue in countries where minimum hygiene standard are met, it remains a threat in almost every developing country like Nigeria. This research work examines the contribution of each risk factor for the occurrence of Cholera diseases

Introduction

Medical practitioners have recognized bodily disorder functions and structure as well as mental abstraction by accumulating facts about possible observations of systems on their patients. These diseases are given specific names according to the symptoms observed. A disease arose as a result of dirty environment. It is simply be defined as any alteration of action that occurred to man or any living thing from normal reaction to abnormal or medically a definite pathological process having a characteristics set of sign and symptoms. It may affect the whole body or any of its parts and its eatiology, pathology and prognosis may be known or unknown.

in Ibadan Metropolis. Multiples regression and Logistic Regression model are the statistical tools employed. Test of individual parameter and joint significant of the explanatory variables were carried out to check for the adequacy of the fitted model. The result concluded that the factors are significantly contributed to the prevalence of Cholera disease. Kernel density estimate and Pnorm were plotted to check for normality in sample. Finally, it was discovered that all the explanatory variables can be related to the occurrence of Cholera disease.

Keywords: *Odds-Ratio, Logistic-regression, eatiology, Explanatory variables, Communicable disease.*

Diseases can be classified as commu nicable and non-communicable. Diseases that have cause and are easily transferable to another either in man or animal is termed as communicable diseases while a disease that have cause and cannot be transferable from one to others is called non-communicable diseases. The study is motivated in the sense that cholera has always been so deadly that so many Nigerian intellectuals have lost their life to due to level of illiteracy in urban geographical area. In the outbreak of cholera, many human resources have lost their life in which the Nigeria labour force is declining, many raw material are from urban areas and the people over the places are not taking into consideration as far as health is concern. The aim of this paper is to determine the contribution of each of the risk factors to cholera outbreak in Ibadan Metropolis using Multiple and Logistic regression model and comparison of these aforementioned models via the estimates of Odds-Ratio (OR).

Literature Review

During the 19th century, cholera spread repeatedly from its original reservoir in the Ganges delta in India to the rest of the world before receding to South Asia. Six pandemics were recorded that killed millions of people across Europe, Africa and the Americas. The seventh pandemic which is still ongoing started in 1961 in South Asia reached Africa in 1971 and the Americas in 1991. The disease is

TIJSRAT
ISSN: 2623-7861

now considered to be endemic in many countries and the pathogen causing cholera cannot currently be eliminated from the environment.

World Health Organization (WHO) has confirmed at least 3,315 cholera cases and registered more than 30,000 cases of acute watery diarrhea-which could also prove to be cholera in its more common milder form. The group has also warned that as the weather cools and temperatures become more favourable for transmission-the bacteria could spread further. Cholera is an acute diarrhoeal infection caused by ingestion of the bacterium Vibrio cholera. Transmission occurs through direct faecal-oral contamination or through ingestion of contaminated water and food. The disease is characterized in its most severe form by a sudden onset of acute watery diarrhoea that can lead to death by severe dehydration and kidney failure.

About 75% of people infected with cholera do not develop any symptoms and has extremely short incubation period of two hours to five days-enhances the potentially explosive pattern of outbreaks as the number of cases can rise very quickly. Cholera is an extremely virulent disease that affects both children and adults. Unlike other diarrhoeal diseases, it can kill healthy adults within hours. Individuals with lower immunity such as malnourished children or people living with HIV are at greater risk of death if infected by cholera. Since 2005, the re-emergence of cholera has been noted in parallel with the ever-increasing size of vulnerable populations living in unsanitary conditions.

Cholera remains a global threat to public health and one of the key indicators of social development. While the disease is no longer an issue in countries where minimum hygiene standard are met, it remains a threat in almost every developing country like Nigeria. The number of cholera cases reported to WHO during 2006 rose dramatically, reaching the level of the late 1990s. A total of 236,896 cases were notified from 52 countries including 6,311 deaths an overall increase of 79% compared with the number of cases reported in 2005.

Several studies of this type have been made in Tuberculosis control by Waaler et al (1962); Brogger (1967) and ReVelle et al (1967). Similar applications have also been utilized for a number of other bacterial diseases such as typhoid fever, tetanus and cholera in the works of

Cvjetanovic et al (1971, 1972 and 1978) amongst others. Furthermore, a model of risk factors in a non-infectious disease and skin cancer has been constructed using Logistic regression by Vitaliano (1978) and Ajiboye (2014). Log-linear models were used to analyze data from a cohort study of acute respiratory illness in the works of Melia et al (1979) and Florey et al (1979).

Other studies using Logistic regression include Stavraky et al (1983) and Lugosi (1985) amongst others and those using Log- linear modeling include McGlynn et al (1985) and Perillo et al (1986). The rest of work is structured as follow: Section 2 introduces some existing literature review. Section 3 discusses materials and methods for analysis of cholera epidemic. Section 4 provides results and discussion while Conclusion and recommendations are presented in Section 5.

Materials and Methods

Data description and sources

The data sets used in this work contains the data on the number of cholera cases recorded in Jericho Nursing Home, Jericho Ibadan at the cholera unit for 683 patients. Four risk factors of cholera were examined viz; water, food, dirty environment and overcrowding. Yearly total record of cholera used as response variable and explanatory variables are $x_1 = water, x_2 = food, x_3 = dirty\ environment$ and $x_4 = overcrowding$ in this study.

Logistic Regression Model

In logistic regression model, the relationship between a response variable and one or more explanatory variables are examined. The outcome variable is discrete. The outcome variable can take two or more possible values depending on the nature of the subject under study. The logistic regression is very useful to find the best fitting to describe the relationship between an outcome variable and a set of explanatory variables. Chao-Ying et al (2002) in their work opined that the main difference between the logistic regression model and linear regression model is that the outcome variable in logistic regression is well suitable for describing and testing hypotheses in categorical outcome variable

and one or more categorical or continuous predictor variables. In some studies, the response variable is not a numerical value. For example in Bernoulli trial study, the two possible outcomes can be classified as success or failure, yes or no, go and no go basis, presence or absence, true or false, alive or dead Bayaya (2010).

Assume Y represent the response variable and x_i the explanatory variable. The expected value of $Y \mid x_i$ denoted by the quantity π_i is defined by

$$\pi(x_i) = \frac{e^{\beta_0 + \sum_{i=1}^{4} \beta_i x_i}}{e^{\beta_0 + \sum_{i=1}^{4} \beta_i x_i} + 1} \qquad (1)$$

where β_0 is the regression constant; $x_i's$ are the risk factors and $\beta_i's$ are the regression coefficient.

The logit transformation is used to express $\pi(x_i)$ as a linear function of $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$. The logit transformation is given by

$$h(x_i) = \ln\left(\frac{\pi(x_i)}{e^{\beta_0 + \sum_{i=1}^{4} \beta_i x_i}}\right) = \beta_0 + \sum_{i=1}^{4} \beta_i x_i \qquad (2)$$

The expression in equation (2) is a linear in its parameters may be continuous and may range from $-\infty$ to $+\infty$ depending on the range of x (Hosmer and Lemeshow, 2000). The probability of Y occurring is denoted by $E(Y \mid x_i)$ and it is equal to $\pi(x_i)$. This expected value follows the binomial distribution with the probabilities $\pi(x_i)$ and $1-\pi(x_i)$ for Y takes only the values 1 or 0 respectively.

Maximum likelihood estimation was used to estimates the regression parameters. The parameter of interest in the logistic regression is the computation of odds-ratio which is estimated as

$$OR = e^{\beta_i}, \quad i = 1, 2, 3, 4 \qquad (3)$$

However, the inferences about the odds-ratio are based on the distribution of $\ln(OR) = \hat{\beta}_i$ which tends to follow a normal distribution.

Multiple Regression Model

TIJSRAT
ISSN: 2623-7861

In a simple linear regression model, if the objective were to predict fairly accurately the value of a dependent variable. In most practical situations, the value of response variable is affected by several related variables. The more information we have the better our predictions should be. The researchers often think of several explanatory variables that might be related to a response variable. To know the strong predictors of the response variable, a multiple regression analysis is used to eliminate those predictors that are not effective. In matrix notation, multiple regression models are of the form:

$$y = X\beta + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2) \qquad (4)$$

where

y is an $n \times 1$ vectors of the response variables and deterministic, ε is an $n \times 1$ vector of disturbance term. Our task is in the estimation of the $k \times 1$ vector of the coefficients β and the variance of the error term σ^2. Hence, we writes down the likelihood function of the model in equation (4) as;

$$L(y \mid \beta, \sigma^2) = \left(\frac{1}{\sigma\sqrt{2\pi}} \right)^n \exp\{-\frac{1}{2\sigma^2} \sum (y - X\beta)\} \qquad (5)$$

The estimates $\hat{\beta}$ and σ^2 are obtained by maximizing the likelihood function in equation (5). The maximum likelihood estimator of β is given by

$$\hat{\beta} = (X^T X)^{-1} X^T y. \qquad (6)$$

This result gives the familiar Ordinary Least Squares (OLS) estimator for the coefficients. We are interested in the following estimated regression model as:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{\beta}_4 X_{4i} \qquad (7)$$

since there are four explanatory variables in the study.

Results and Discussion

The study focuses of the data on the number of cholera cases recorded in Jericho Nursing Home, Jericho Ibadan as earlier mentioned in section 3.

Interpretation

TIJSRAT
ISSN: 2623-7861

All the regression model parameters are Significant since their p-values less than 0.05, the Null hypothesis was rejected in favour of the alternative hypothesis and the result concluded that the factors are significantly contributed to the prevalence of cholera disease. The estimated model is

$$\hat{Y} = 0.783 + 0.046X_1 + 0.039X_2 + 0.044X_3 + 0.035X_4$$

and shown in the table table1 below:

Table1. Test of individual parameter

Variable	Coef. (β_i)	Std. Error	t-test	p-value	Decision	Conclusion
CONSTANT	0.783	0.017	47.408	0.000	Reject	Significant
FOOD	0.046	0.004	11.193	0.000	Reject	Significant
WATER	0.039	0.007	5.415	0.000	Reject	Significant
DIRTY ENVIRONMENT	0.044	0.007	6.091	0.000	Reject	Significant
OVER CROWDING	0.035	0.004	7.979	0.000	Reject	Significant

Also the regression model parameters are jointly significant since the p – value is less than 0.05, the Null hypothesis was rejected in favour of the alternative hypothesis and the result concluded that the factors are significantly contributed to the prevalence of cholera disease. This is shown in table2 via Anova table.

Table2 Joint Significant Test (Anova)

Model	Sum of Squares	df	Mean Square	F	Sig.
Regression	120.046	4	30.011	576.073	.000[b]
Residual	35.322	678	.052		
Total	155.367	682			

Null Hypothesis: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ (The Factors are not jointly significant)

We also carried out post estimation test of normality and homoscedasticity assumption in the study. The figures below illustrate

Kernel density estimate plot and Pnorm plot for check normality in sample together with homoscedasticity check.

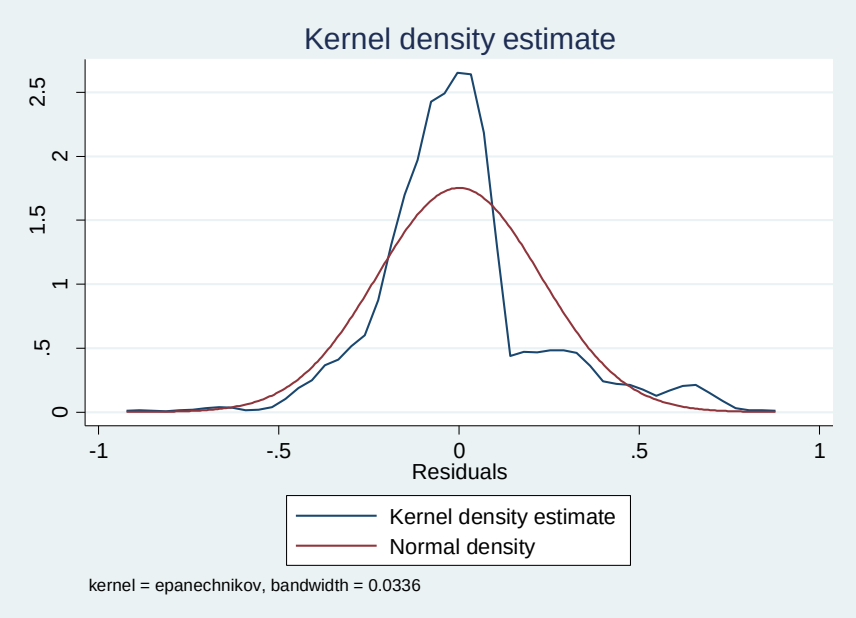

Fig 1: Kernel density estimate plot to check normality in sample

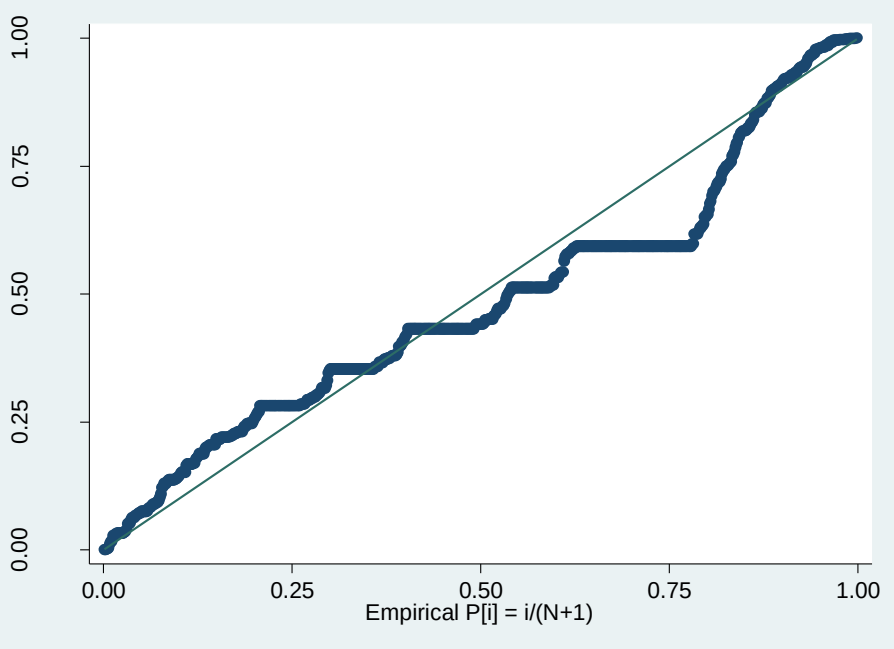

Fig 2: Pnorm plot to check normality in sample

In figures 1 and 2 above, the normal curve on the Kernel density plot was skewed left which mean that the samples are not normally distributed. Normally distributed sample plot assumes a well formed bell shape and non-skewed plot. Also few points were plotted on the straight line of Pnorm plot. A normally distributed sample Pnorm plot used to have almost all the point plotted on the central line.

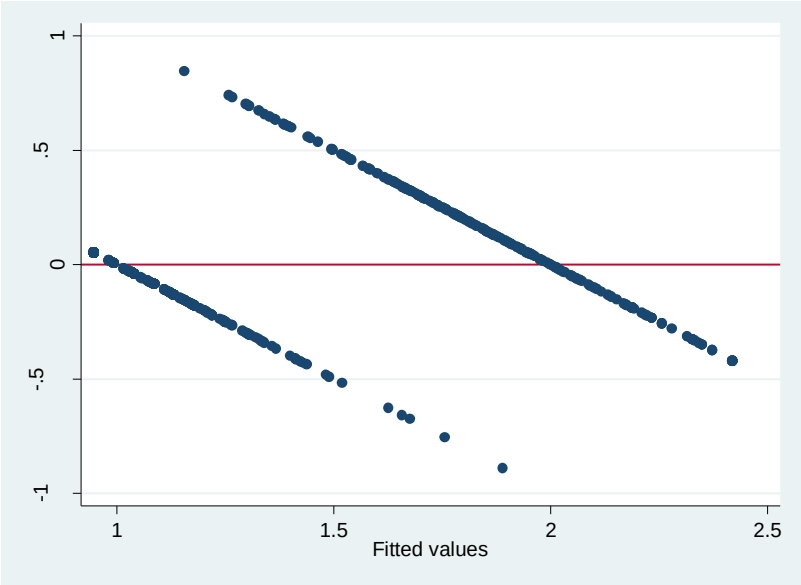

Fig 3: Model estimated residual vs Fitted plot to check homoscedasticity of estimated errors

Figure 3 shows the residual vs fitted values plot shows linear pattern i.e there should be no pattern on the plot. The residual vs fitted plot of heteroscedastics error sample pattern of the data points used to be narrower towards the right end. The white's test Chi-square value 154.06 is significant with p-value of 0.0000 which less than 0.05 significant level, we reject the null hypothesis and concluded that the estimated residuals of the model are heteroscedastics (Non-Homogeneous).

The model has good predictive power since about 77.3% of the total variation of dependent variable was explained by the regression plane.

The test of individual parameter of the logistic regression model was based on the Wald test. The table below shows the parameters value,

TIJSRAT
ISSN: 2623-7861

the standard error of each parameter, the normal (Z) statistic with their corresponding degree of freedom, the probability of accepting null hypothesis (P-value) and the exponential value of the parameters (the odd value of individual parameter).

Table 3. Logistic Regression estimates and Odd-Ratio

Variable	Par. (β)	S.E	(Odd Ratio)	Wald (Z value)	df	P-value	Decision
FOOD	0.646	0.114	1.908	32.275	1	0.000	Reject H_0
WATER	0.351	0.175	1.421	4.045	1	0.044	Reject H_0
DIRTYENVIRONMENT	0.720	0.181	2.055	15.858	1	0.000	Reject H_0
OVERCROWDING	0.442	0.105	1.556	17.577	1	0.000	Reject H_0
Constant	-8.446	0.816	0.000	107.083	1	0.000	Reject H_0

From the Wald test above using the normal (Z) value and the P-values, we concluded that all the estimated parameters are significant to the model.

Conclusion and Recommendations

This paper investigates the contribution of each of the factors that cause cholera disease in Ibadan Metropolis of 683 cholera Cholera patients reported at Jericho Nursing Homes, Jericho, Ibadan. The results and analyses show that food, water, dirty-environment and overcrowding influences the occurrence of Cholera disease. The binary logistic model indicates that all the explanatory variables turns out to be a good fit. This means we can determine the odds ratio of Cholera from knowing the aetiology of the disease. The risk of dirty-environment is highest risk factor to Cholera (2.055) which means that it is 2.055 more likely to cholera occurrence in the area where the environment is dirty, followed by poor food, overcrowding and water respectively.

The majority of patients- up to 80% can be treated adequately through the administration of Oral rehydration salts (WHO/UNICEF) standard sachet .Government should set up cholera treatment centres among the affected populations in order to ensure timely access to treatment and also provide adequate clean water supply if not the masses should ensure their water is treated before its domestic uses. Finally, masses should feed on hygienic food and sanitize their environment for control of outbreak of not only cholera but other diseases.

References

Ajiboye, A. S (2014). A Study of Accident Fatality on Akure-Owo High Way using Logistic Regression. Journal of the Nigerian Statistical Association Vol. 26, 2014, 1 – 10.

Bayaga, A (2010). Multinomial Logistic Regression Usage and Application in Risk Analysis. Journal of Applied Quantitative Methods, 5(2), 288 – 297. .

Brogger, S. (1967). Systems Analysis in Tuberculosis: A Model. American Review of Respiratory Diseases, 95, 421 – 434.

Chao-Ying, J.P., Kuk, L.L. and G,M. Ingersoll (2002). An Introduction to Logistic Regression Analysis and Reporting. The Journal of Education Research, 6 (1), 3 – 14.

Cvjetanovic, B., Grab, B. and Uemura, K. (1971). Epidemiological Model of Typhoid amd its use in the Planning and Evaluation of Antityphoid Immunization and Sanitation Programmes, Bulletin, World Health Organisation, 45, 53 – 75.

Cvjetanovic, B., Grab, B., Uemura, K. and Bytchenko, B. (1972). Epidemiological Model of Tetanus and its use in the Planning of Immunization Programmes, International Journal of Epidemiology, 1, 2, 125 – 137.

Cvjetanovic, B., Grab, B. and Uemura, K. (1978). Dynamics of Acute Bacterial Diseases, Bulletin, World Health Organization, 56, Supplement 1.

Florey, C. du V. Melia, R. J. W. and Chinn, S, et al. (1979). The Relation Between Respiratory Illness in Primary School Children and the Use of Gas for Cooking III: Nitrogen Dioxide, Respiratory Illness and Lung Infection, International Journal of Epidemiology, 8, 347 – 353.

Hosmer, D, W. and S. Lemeshow (2000). *Applied Logistic Regression*, 2nd Ed. New York, John Wiley.

Lugosi, L. (1985). Trends in Childhood Tuberculosis in Hungary 1958 – 1983: Qualitative Methods for Evaluation of BCG Policy, I ternational Journal of Epidemiology, 14, 304 – 312.

Melia, R. J. W., Florey, C. du V. and Chinn, S. (1979). The Relation Between Respiratory Illness in Primary School Children and the Use of Gas for Cooking I: Results from a National Survey, International Journal of Epidemiology, 8, 333 – 338.

McGlynn, K. A., Lulstbader, E. D. and London, W. T.(1985). Immune Responses to Hepatitis B Virus and Tuberculosis Infections in Southeast Asian Refugees, American Journal of Epidemiology, 122, 1032 – 1036.

Perillo, R. P., Strang, S. and Lowry, O. H. (1986). Different Operating Conditions Affect Risk of Hepatitis B Virus Infection at Two Residential Institutions for the Mentally Disabled, American Journal of Epidemiology, 123, 690- 698.

ReVelle, C., Lynn, W. R. and Feldmann, F. (1967). Mathematical Models for the Economic Allocation of Tuberculosis Control Activities in Developing Countries, Amer. Rev. Resp. Dis., 96. 893 – 909.

Stavraky, K. M., Rawls, W. E. and Chiavetta, J. et al. (1983). Sexual and Socio-economic Factors Affecting the Risk of Past Infections with Herpes Simplex Virus Type 2, American Journal of Epidemiology, 118, 109 – 121.

Waaler, H. T. Geser, A. and Andersen, S. (1962). The Use of Mathematical Models in the Study of the Epidemiology of Tuberculosis, American Journal of Public Health, 52, 1002 – 1013.