MODEGLING PREVALANCE OF MALARIA IN NIGERIA USING LOGISTIC REGRESSION

ABSTRACT

Malaria continues to pose a significant health threat in many parts of the world, particularly in regions where transmission rates are high. Early detection and prediction of malaria prevalence are essential for effective public health interventions and resource allocation. This study focuses on the development of a logistic regression model to predict the prevalence of malaria based on a comprehensive analysis of key demographic and environmental factors. The model incorporates variables such as the age of children, the source of drinking water, maternal education levels, gender, regional factors, and household wealth. Through the application of logistic regression, this study identifies the relative influence of each predictor on malaria prevalence. The results reveal that certain factors, such as the age of children and the source of drinking water, have a more significant impact on malaria risk. The model's performance is evaluated using (ROC) curves, which demonstrate the model's ability to distinguish between high and low-risk individuals with a reasonable degree of accuracy. The ROC curve analysis suggests that the model is effective, with an AUC value of **0.689490** indicating a fair level of discriminatory power.By identifying high-risk groups and areas, health authorities can prioritize interventions such as the distribution of insecticide-treated nets, indoor residual spraying, and public education campaigns. Moreover, the model's predictive capacity can be used to inform the deployment of limited medical resources, ensuring that they are directed to the areas where they are most needed. This project emphasize the importance of data-driven approaches in the ongoing fight against malaria and provides a foundation for future research and model refinement. This Analysis was carried out completely with R package

Key Words: Receiver Operating Characteristic ,Area Under the Curve

INTRODUCTION

Malaria is a mosquito-borne infectious disease caused by Plasmodium parasites. These parasites are transmitted to humans through the bites of infected female Anopheles mosquitoes. It is one of the common and major health issues in Africa at large. It has been one of the factors of Mortality rates in Nigeria especially among children of ages (0-59) months. According to WHO (2023), African Region continues to carry a disproportionately high share of the global malaria burden. In 2022, the Region was home to about 94% of all malaria cases and 95% of deaths. Children under 5 years of age accounted for about 78% of all malaria deaths in the Region.

Over the years in Nigeria, a lot of lives have been lost due to lack of education, ignorance and illiteracy in regards to Malaria. Malaria has been one amongst other the major factors causing child Mortality in Nigeria especially in rural areas due to of lack of awareness, bad living condition, lack of finances etc. Reports have shown that 30% of under 5 years old mortality and 25% of infant mortality is caused by Malaria.

Despite significant efforts to control malaria in Nigeria, it remains a major public health concern, causing substantial morbidity and mortality. Traditional approaches to understanding malaria prevalence often rely on epidemiological data (population health data), which lack precision and fail to capture the complex may interplay of environmental, socio-economic, and demographic factors.

To address this issue, this study tend to develop a logistic regression model to predict the prevalence of malaria in Nigeria.

Malaria has been analyzed by various researchers using geostatistical modelling approach, machine learning classifiers and various types of logistic regression. Little or no effort has been made using binomial logistic regression to model prevalence of malaria among children with the age (0-59) months in Nigeria. This study aims to determine the relationship among factors causing malaria, fit a model for the prevalence of malaria, examine the adequacy of the model and Predict future occurrence of malaria cases in Nigeria.

Logistic Regression model is a predictor model variable and a categorical response variable .It is used when the research method is focused on whether or not an event occurred, rather than when it occurred. It is a part of the generalized Linear Models that predicts the values of one dependent variable from one or more predicting variables when the dependent variable is dichotomous. For Example, we could use logistic regression to model the relationship between various measurement of a manufactured specimen to predict if the item is defective or not (a binary variable either Yes or No).

It helps us to estimate a probability of falling into a certain level of the categorical response given a set of predictors. Some types of logistic regression include; Binary Logistic Regression, Ordinal Logistic Regression, Multinomial Logistic Regression and Nominal Logistic Regression. In this research work, we explore Binomial Logistic Regression for the data set on prevalence of malaria.

METHODOLOGY

The methodology adopted in this study is binomial logistic regression. For the purpose of this study, a secondary data was obtained from Demographic and Health Survey (DHS). A general exploratory data analysis (EDA) was performed on the data. We then further by fitting a model for the data and examining the adequacy of the model.

Logistic Regression Model

The Multiple Binary Logistic Regression Model is given as follows;

$$\boldsymbol{logit(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + E_i}$$

p =Malaria occurrence/prevalence (yes or no).

$\boldsymbol{X_1}$ =Source of drinking water.

$\boldsymbol{X_2}$ =Sex of the children.

$\boldsymbol{X_3}$ =Mother's highest educational level.

$\boldsymbol{\beta_0}$ = intercept.

$\boldsymbol{\beta_1, \beta_2, \beta_3}$= slope of their respective explanatory variable.

$\boldsymbol{E_i}$ = error.

Parameter Estimation of Logistic Regression Model

The Multiple Binary Logistic Regression Model is given as follows;

$$\boldsymbol{logit(p) = \beta_0 + \beta_1 X_1 + \cdots \ldots \ldots .. + \beta_P X_P + E_i} \qquad \textbf{(1)}$$

$$\mathbf{log}\left(\frac{\boldsymbol{p}}{\mathbf{1} - \boldsymbol{p}}\right) = \boldsymbol{\beta^T X}$$

Take the log of both sides

$$\frac{p}{1-p} = e^{\beta^T X}$$

Collect like terms and find the equation for p

$$p = (1-p)e^{\beta^T X}$$

$$p = e^{\beta^T X} - pe^{\beta^T X}$$

$$p + pe^{\beta^T X} = e^{\beta^T X}$$

$$p(1 - e^{\beta^T X}) = e^{\beta^T X}$$

$$p = \frac{e^{\beta^T X}}{(1 + e^{\beta^T X}\ )}$$

$$p = \frac{1}{1 + e^{\frac{1}{\beta^T}} X}$$

$$P = \frac{1}{1+e^{-\beta^T} X}$$
(2)

$\boldsymbol{\beta_0}$ is the intercept term, $\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \ldots\ldots\ldots, \boldsymbol{\beta_p}$ are the coefficient for the dependent variables $\boldsymbol{X_1}, \boldsymbol{X_2}, \ldots\ldots\ldots\ldots, \boldsymbol{X_P}$. In equation **(2)**, P is a logistic model. Logistic Regression uses Likelihood Ratio to test null hypothesis that any subset of the β's is equal to 0. The number of the β's in the full model is **K+1** while the number of β's in the reduced model is r+1. Thus, the number of $\boldsymbol{\beta}$'s being tested is **(K+1)-(r+1) =K-r**. The likelihood ratio test is given by

$$\lambda = -2(e(\beta^{(0)}) - e(\beta))$$
(3)

Where

$e(\boldsymbol{\beta})$ Is the log likelihood of the fitted (full) model

$e(\boldsymbol{\beta}^{(0)})$ Is the log likelihood of the (reduced) model specified by the null hypothesis evaluated at the maximum likelihood estimate of that reduced model.

This test statistics has a χ^2 distribution with **K-r** degrees of freedom. Statistical software often presents results for this test of "deviance," Which is defined as -2 times log likelihood. The notation used for the test statistics is typically $\boldsymbol{G^2}$=deviance (reduced) –deviance (full).

Interpreting the odds ratio

Suppose we write:$\frac{p}{1-p} = e^{\beta^T X} = e^{\beta_0+\beta_1 X_1+\cdots\ldots\ldots+\beta_P X_P}$ **(4)**

$\frac{p}{1-p}$is called the odd ratio. we can see that increasing $\boldsymbol{X_j}$ by one unit while keeping all other predictors fixed or constant, multiplies the odds by $\boldsymbol{e^{B_j}}$**.** Alternatively, we could write **log** $(\frac{p}{1-p}) = \boldsymbol{\beta^T xX}$

In this case, holding all other variable constant, a unit increase of $\boldsymbol{X_j}$ will increase the log of the odds by $\boldsymbol{\beta_j}$

Assumptions of Logistic Regression
Homoscedasticity (constant variance).
The dependent variable is binary.
There must be little or no multicolinearity between the explanatory variable.
It requires sufficient large sample size.

Diagnostic check tools for Logistic Regression
Residual Analysis.
ROC Curve and AOC Value.
Deviance and Pearson Goodness-of-Fit Tests.

Variance Inflation Factor(VIF) or Correlation matrices.

Data Description

The data used for this study is a secondary data gotten Demographic Health Survey (DHS). This data consist on malaria test results of children of age 0-59 months and factors that might affect the prevalence of malaria, it was collected from 70428 house old across all 36 states in Nigeria , out of the 70428 household, children of 6592 households were found negative,4105 were found positive,131 were not present during the survey 246 of them refused to be tested while 59354 of them did not have children around the age of 0-59 months

Analysis and Results

Here we present the results obtained from m odeling the prevalence of malaria in Nigeria

Discussion

Table 1 and shows the summary statistics of malaria prevalence and its factors which reveals that approximately 38% of cases are positive for malaria, with children's ages ranging from 6 to 59 month, mothers generally have low education levels, and the population exhibits significant economic disparity, with the majority having negative or low wealth. The sources of drinking water and regional distribution are relatively even, while the gender distribution is nearly balanced, with a slight skew towards females. Figure 1 shows the pie chart which reveals that 61.6% of the observation taken were malaria- negative and 38.4% were positive .Figure2-5 shows the histogram of each factors likely to cause malaria against malaria prevalence, from the histogram we can see that age significantly affect the prevalence of malaria while the other factors appears to be evenly distributed

Variable	Min	1st QU	median	mean	3rd QU	Max	class
Result of malaria	0.0000	0.0000	0.000	0.3838	1.0000	1.0000	N:6592
Age of children	6.0000	21.0000	36.000	34.0000	49.00	59.00	P: 4105
Source of drinkin g water	11.0000	21.0000	31.000	30.2800	32.0000	96.0000	
Mothers level of e ducation	0.0000	0.0000	2.0000	2.28000	2.0000	9.0000	
Sex of child	1.0000	1.0000	1.0000	1.4870	2.0000	2.0000	
region	1.0000	8.0000	15.000	16.6300	26.0000	37.0000	
Wealth	-194845	-104270	-49532	-24668	48602	207282	

.Table 1: Descriptive Statistics

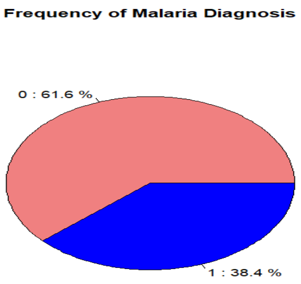

FIGURE1: Pie chart Representation of Class

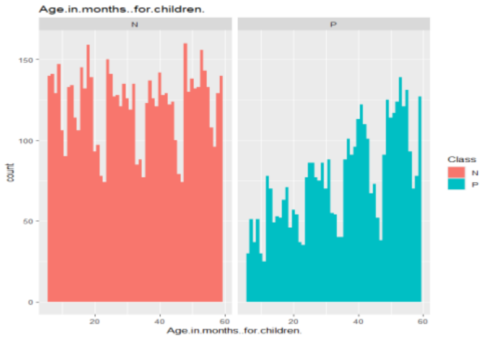

Figure2: Histogram visualizing age in months of children

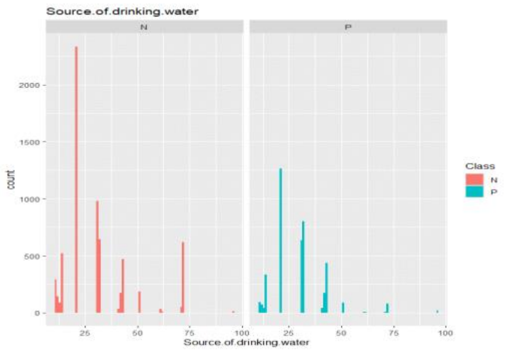

Figure3: Histogram visualizing source of drinking water

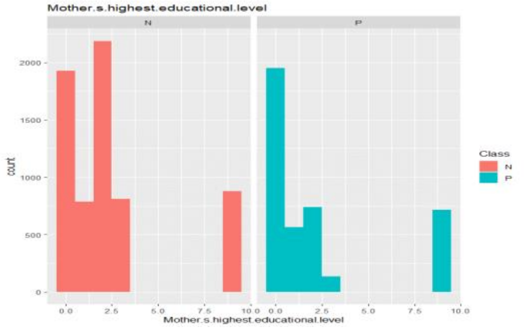

Figure 4: Histogram visualizing mother's highest educational level

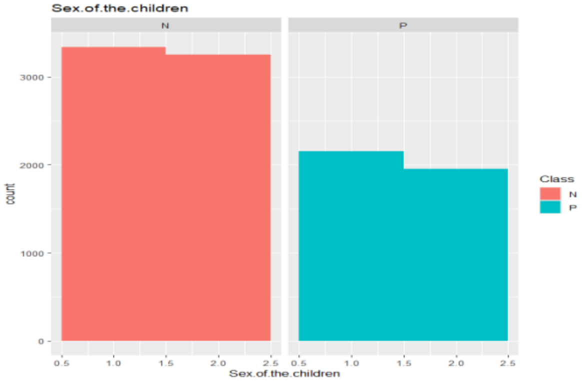

Figure4.4: Histogam visualizing sex of child

.

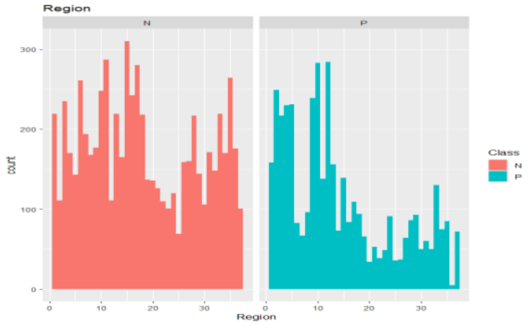

Figure4.5: Histogram visualizing region

Coefficients	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.099e+00	1.006e-01	-10.924	< 2e-16
Age in months for children	1.999e-02	1.378e-03	14.508	< 2e-16
Source of drinking water	-1.484e-03	1.522e-03	-0.975	0.32950
Mothers highest educational level	4.757e-03	7.037e-03	0.676	0.49906
Sex of the children	-1.087e-01	4.205e-02	-2.586	0.00971
Region	-6.053e-03	2.388e-03	-2.535	0.01126
Wealth	-6.786e-06	2.886e-07	-2.535	< 2e-16
Signif. Codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				
(Dispersion parameter for binomial family taken to be 1)				
Null deviance: 14246 on 10696 degrees of freedom				
Residual deviance: 13075 on 10690 degrees of freedom				
AIC: 13089				
Number of Fisher Scoring iterations: 4				

Table 2 logit Model

Table 2 shows the result from fitting of the model which is written as

$$\mathrm{logit(p)} = -1.099 + 0.01999\beta_1 - 0.001484\beta_2 + 0.004757\beta_3 - 0.1087\beta_4 - 0.006053\beta_5 - 6.786\mathrm{e} -06\beta_6$$

Where$\beta_1 =$ Age in months for children $\beta_4 =$ Sex of the children

$\beta_2 =$ Source of drinking water $\beta_5 =$ Region

$\beta_2 =$ Mothers highest educational level $\beta_6 =$ Wealth

The above models shows:
Intercept (-1.099): This is the baseline log-odds when all predictor variables are zero.
0.01999β1: For every additional month in a child's age, the log-odds of the outcome increase by 0.01999, holding all other variables constant.i.e as a child gets older the likelihood of them having malaria slightly increases.
0.001484β2:For each unit increase in the variable representing the source of drinking water, the log-odds of the outcome decrease by 0.001484.i.e the type of drinking water a child has access to can affect their likelihood of having malaria. In this model, certain types of drinking water sources slightly decrease the chance of malaria.
0.004757β3: For each unit increase in the mother's educational level, the log-odds of the outcome increase by 0.004757.i.eThe higher the education level of the child's mother, the more likely it is that the child might have malaria, though this increase is very small.
0.1087β4: If the child is of a specific sex (which is coded as 1 for male and 2 dor female), the log-odds decrease by 0.1087 0.006053β5:For each unit increase in the region variable, the log-odds decrease by 0.006053.i.e The region where the child lives plays a role in determining their risk of malaria. Some regions have a slightly lower or higher risk. 6.78610β6: For each unit increase in wealth, the logodds decrease by 0.000006786.

Lastly table 3 and figures 6 and figure 7 are the diagonistic check tools used in this study to check if the data agrees with the assumption of lack of multicolinearity normality, and homogeneity of variance we can see from the plots that this assumptions were satisfied respectively while figure 8 and AUC Value of **0.689490084315836** were used to check the fitness of the model which indicates that the model's ability to distinguish between positive and negative outcomes is moderate.

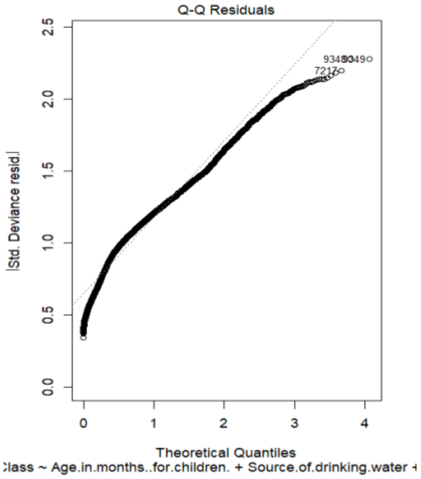

Figure6: Q-Q Plot

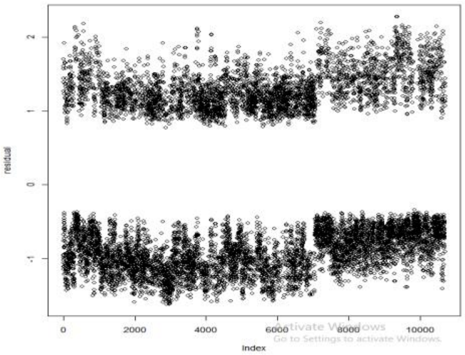

Figure 7 residual plot

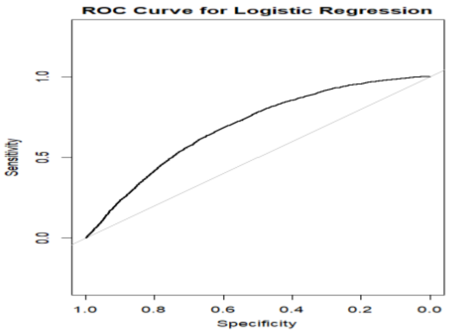

Figure7 ROC Curve

.

	Age in months (for children)	Source of drinking water	Mother's highest educational level	Sex of the children	Region	Wealth
Age in months (for children)	1	-0.003395366	0.130145744	0.00124	0.035124	0.010250924
Source of drinking water	-0.003395366	1	0.027338741	-0.0024	0.178147	0.149214862
Mother's highest educational level	0.130145744	0.027338741	1	0.00433	0.179339	0.177057467
Sex of the children	0.001243676	-0.002372097	0.004325322	1	-0.0189	-0.02063096
Region	0.035124442	0.178146911	0.179339485	-0.0189	1	0.536349282
Wealth	0.010250924	0.149214862	0.177057467	-0.0206	0.536349	1

Table 3 Correlation Matrix

Recommendations

Based on the findings of this research, the following recommendations are suggested:

1. Transforming the response variable or predictors to improve model fit.
2. Other modeling techniques can explored, such as generalized additive models (GAMs) or other non-linear models, which might better capture the data patterns.
3. Based the significant predictors defined in this study, targeted interventions can be designed to improve living condition in the specific regions affected by malaria For example free medical healthcare for children of 0-59months with malaria for people with low wealth status

CONCLUSION

The logistic regression model identified significant predictors of the outcome, including age of children, sex of children, region, and wealth. However, source of drinking water and mother's highest educational level were not significant. The residual analysis indicated deviations from normality and potential outliers, suggesting that the model may not fully capture the underlying data structure. However it is essential to acknowledge that this research is an ongoing process and further improvements are required.

Further Research

Investigating additional variables that might influence the outcome but were not included in this model and considering a more comprehensive data collection process to capture these variables, by addressing the recommendations, the model's predictive power and reliability would be improved, providing more accurate insights for decision-making and policy development.

REFERENCES

1.Adugna, F., Wale, M., and Nibret, E. (2022). Prevalence of malaria and its risk factors in Lake Tana and surrounding areas, northwest Ethiopia.

2.White, N. J., Pukrittayakamee, S., Hien, T. T., Faiz, M. A., and Mokuolu, O. a., and Dondorp, AM (2014).

3.Lehrer, S., and Pavličić, M. (1982). *Vitezi medicine*. Zavod za izdavačku delatnost" Filip Višnjić".

4.Dugacki, V. (2005). Dr. Rudolf Battara operation in Nin in 1902, the first systematic battle attempt against malaria in Croatia.

5.Jovic, S. (1998). Istorija Medicine i Zdravstvene Kulture na tlu Dansnje Vojvodine 1718–1849 II. dio. *Matica srpska, Srpska akademija nauka i umetnosti: Novi Sad, Serbia*

6.Tan, S. Y., & Ahana, A. (2009). Charles Laveran (1845-1922): Nobel laureate pioneer of malaria.

7.Cartwright, F. F. (1972). Disease and history.

8.Nas, F. S., Yahaya, A., & Ali, M. (2017). Prevalence of malaria with respect to age, gender and socio-economic status of fever related patients in Kano City, Nigeria.

9.Ugwu, C. L. J., & Zewotir, T. T. (2018). Using mixed effects logistic regression models for complex survey data on malaria rapid diagnostic test results. .

10.Obasohan, P. E., Walters, S. J., Jacques, R., & Khatab, K. (2021). Individual and contextual factors associated with malaria among children 6–59 months in nigeria: A multilevel mixed effect logistic model approach

11.Onyiri, N. (2015). Estimating malaria burden in Nigeria: a geostatistical modelling approach

12.Ukwajunor, E. E., Akarawak, E. E. E., Abiala, I. O., & Adebayo, S. B. (2020). Weighted Logistic Regression Modelling of Prevalence and Associated Risk Factors of Malaria in Nigeria.

13.Adigun, A. B., Gajere, E. N., Oresanya, O., & Vounatsou, P. (2015). Malaria risk in Nigeria: Bayesian geostatistical modelling of 2010 malaria indicator survey data.

14.Oguoma, V. M., Anyasodor, A. E., Adeleye, A. O., Eneanya, O. A., & Mbanefo, E. C. (2021). Multilevel modelling of the risk of malaria among children aged under five years in Nigeria

15.Mutambayi, R., James, N., Adeboye, A., Akinwumi, O., & Song, Q. Y. (2017). Assessment of Risk Determinants in the Regularity of Malaria Using the Binary Logistic Approach.

16. Nkiruka, O., Prasad, R., & Clement, O. (2021). Prediction of malaria incidence using climate variability and machine learning. .

17.Yang, D., He, Y., Wu, B., Deng, Y., Li, M., Yang, Q., ... & Liu, Y. (2020). Drinking water and sanitation conditions are associated with the risk of malaria among children under five years old in sub-Saharan Africa: a logistic regression model analysis of national survey data. *Journal of advanced research*, *21*, 1-13.

18.Workineh, L., Lakew, M., Dires, S., Kiros, T., Damtie, S., Hailemichael, W., ... & Eyayu, T. (2021). Prevalence of malaria and associated factors among children attending health institutions at South Gondar Zone, Northwest Ethiopia: a cross-sectional study.

19.Workineh, L., Lakew, M., Dires, S., Kiros, T., Damtie, S., Hailemichael, W., ... & Eyayu, T. (2021). Prevalence of malaria and associated factors among children attending health institutions at South Gondar Zone, Northwest Ethiopia: a cross-sectional study.

20. Goshu, E. M., Zerefa, M. D., & Tola, H. H. (2022). Occurrence of asymptomatic malaria infection and living conditions in the lowlands of Ethiopia: a community-based cross-sectional study.

21.Boadu, I., Nsemani, W., Ubachukwu, P., & Okafor, F. (2020). Knowledge and prevalence of malaria among rural households in Ghana.

22. Stuck, L., Fakih, B. S., Abdul-Wahid, H., Hofmann, N. E., Holzschuh, A., Grossenbacher, B., ... & Yukich, J. (2020). Malaria infection prevalence and sensitivity of reactive case detection in Zanzibar