© 2015 IJSRST | Volume 1 | Issue 5 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X
Themed Section: Science and Technology

An Investigation of Malaria Predictors Using Logistic Regression Model

Abubakar Boyi Dalatu[1], Mukhtar Garba[1], Nwoji Jude Oguejiofor[2]

[1]Department of Statistics, Waziri Umaru Federal Polytechnic Birnin Kebbi, Nigeria
[2]Department of Computer Science, Waziri Umaru Federal Polytechnic Birnin Kebbi, Nigeria

ABSTRACT

Although malaria is a disease which is considered the most deadly killer especially to children less than 5 years mainly of African countries, there exists no statistical model for the analysis of its predictors for the case of Kebbi State. In this work a logistic regression model using maximum likelihood estimation is proposed. The application of the model using Kebbi State malaria data established that there is significant relationship between malaria status and such predictors as fever, temperature greater than or equal to 37.5 degree, headache, convulsions, cold, cough or sweating, etc. While age, sex, backache and vomiting are not good predictors of malaria. Doctors, medical practitioners and researchers will find this model useful in predicting malaria prevalence.

Keywords : Logit Function, Logistic regression, Maximum Likelihood Estimation

I. INTRODUCTION

Logistic regression or logit regression is a type of probabilistic statistical classification model that is used to predict a binary response from a binary predictor. Logistic regression is used for predicting the outcome of a categorical dependent variable based on one or more predictor variables (features).

An explanation of logistic regression begins with an explanation of the logistic function, which always takes on values between zero and one i.e. (0 and 1):

$$F(t)=\frac{e^t}{e^t+1}=\frac{1}{1+e^{-t}}$$

Hosmer D.W and Lemeshow S. (2000) -----(1.1)

viewing *t* as a linear function of an explanatory variable *x* (or of a linear combination of explanatory variables), the logistic function can be written as:

$$F(z)=\frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$$

----------(1.2)

This will be interpreted as the probability of the dependent variable equaling a "malarial" rather than a non-malarial.

We also write z as the linear sum

$$z=\alpha+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k$$ ----------(1.3)

Where the x's are independent variables of interest, α and the β_i 's are constant terms representing unknown parameters.

We also define the inverse of the logistic function, the logit:

$$g(x)=\ln\frac{F(x)}{1-F(x)}=\beta_0+\beta_1 x$$ ----------(1.4)

And equivalently: $\frac{F(x)}{1-F(x)}=\beta_0+\beta_1 x$ ----------(1.5)

Whose log value gives the logit, describes the odds for a malaria patient with independent variables specified by x.

Logistic regression uses the **Maximum Likelihood Estimation** method to estimate the model coefficients. This method yields values of α and β which maximize the probability of obtaining the observed set of data. Conceptually, it works as follows:

First construct a likelihood function which expresses the probability of the observed data as a function of the unknown parameters α and β.

For univariate case, the contribution to the likelihood function for a given value of the predictor X, is

$P(Y=1|x)^y * P(Y=0|x)^{1-y}$ Kleinbaum et al, (1994) --(1.6)

Thus when Y = 1, the contribution is: $P(Y=1|x)$
when Y = 0, the contribution is: $P(Y=0|x)$
Since the sample observations are assumed to be independent, the likelihood function for the data set is just the product of the individual contributions:

$L=\Pi P(Y=1|x)^y * P(Y=0|x)^{1-y}$ ---------- (1.7)

A more tractable version of this function is obtained by taking the natural logarithm of the likelihood function, called the Log Likelihood function:

$LL=\Sigma yLog[P(Y=1|x)]+(1-y)Log[P(Y=0|x)]$ -------- (1.8)

To find the values of the parameters that maximize the above function, we differentiate this function with respect to α and β and set the two resulting expressions to zero. An iterative method is used to solve the equations and the resulting values of α and β are called the maximum likelihood estimates of those parameters. The same approach is used in the multiple predictor case where we would have (p+1) equations corresponding to the p predictors and the constant α.
We have that;

$E\left[Y_i\right]=\pi_i$ ---------- (1.9)

Also

$\log it(\pi_i)=\beta_0+\beta_1 x_{i1}+...\beta_p x_{ip}=x_i'\beta=(1,x_{i1},...x_{ip})'$ ---------- (1.10)

and $\beta=(\beta_0,...\beta_p)'$ ---------- (1.11)

Therefore $\mathrm{E}\left[Y_i\right]=\frac{\exp(x_i'\beta)}{1+\exp(x_i'\beta)}$ ---------- (1.12)

The likelihood for n observations is then

$\mathrm{L}=\prod_{i=1}^{n}\left(\frac{\exp(x_i'\beta)}{1+\exp(x_i')}\right)^{\sum_{i=1}^{n}Y_i}\left(1-\frac{\exp(x_i'\beta)}{1+\exp(x_i'\beta)}\right)^{n-\sum_{i=1}^{n}Y_i}$ Kleinbaum et al, (1994) ---------- (1.13)

The log-likelihood is

$\sum_{i=1}^{n}\left[Y_i\log\left(\frac{\exp(x_i'\beta)}{1+\exp(x_i'\beta)}\right)+(1-Y_i)\log\left(1-\frac{\exp(x_i'\beta)}{1+\exp(x_i'\beta)}\right)\right]$ --(1.14)

Also using the fact that; $p=\frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$ ----------(1.15)

The joint probability of the data (the likelihood) is given by

$L=\prod_{i=1}^{n}p^{Y_i}(1-p)^{1-Y_i}$ ---------- (1.16)

$=p^{\sum_{i=1}^{n}Y_i}(1-p)^{n-\sum_{i=1}^{n}Y_i}$ ---------- (1.17)

For estimation, we will work with the log-likelihood

$l=\log(L)=\sum_{i=1}^{n}Y_i\log(p)+(n-\sum_{i=1}^{n}Y_i)\log(1-p)$ ---------- (1.18)

The maximum likelihood estimate (MLE) of p is the value that maximizes l (equivalent to maximizing L).
The first derivative of l with respect to p is

$U(p)=\frac{\partial l}{\partial p}=\sum_{i=1}^{n}Y_i/p-(n-\sum_{i=1}^{n}Y_i)/(1-p)$ ---------- (1.19)

And is referred to as the score function. To calculate the MLE of p, we set the score function, U (p) equal to 0 and solve for p. In this case, we get an MLE of p that is

$\hat{p}=\sum_{i=1}^{n}Y_i/n$ Agresti, (1996) ----------(1.20)

II. METHODS AND MATERIAL

A. Wald Test

The Wald test statistic is a function of the difference in the MLE and the hypothesized value, normalized by an estimate of the standard deviation of the MLE.

$w=\frac{(\hat{p}-p_0)^2}{\hat{p}(1-\hat{p})/n}$ Menard, (1995) ---------- (1.21)

For large n, W~χ^2 with 1 degree of freedom.
In R, you will see $\sqrt{W}\sim N(0,1)$*reported*

B. Likelihood Ratio Tests

The likelihood ratio test (LRT) statistic is the ratio of the likelihood at the hypothesized parameter values to the likelihood of the data at the MLE(s).
The LRT statistic is given by

$$LR = -2\left(\frac{L \text{ at } H_0}{L \text{ at MLE(s)}}\right) = -2(H_O) + l(MLE) \quad \text{---------- (1.22)}$$

For large n, LR~ with degrees of freedom equal to the number of parameters being estimated.

For the binary outcome discussed above, if the hypothesis is

$H_0 : P = P_0$ $H_A : p \neq p_0$, then

$$l(H_0) = \sum_{i=1}^{n} Y_i \log(p_0) + (n - \sum_{i=1}^{n} Y_i)\log(1 - p_0), \quad \text{------- (1.23)}$$

$$l(MLE) = \sum_{i=1}^{n} Y_i \log(\hat{p}) + (n - \sum_{i=1}^{n} Y_i)\log(1 - \hat{p}) \quad \text{---------- (1.24)}$$

And the LRT statistic is

$$LR = -2\left[\sum_{i=1}^{n} Y_i \log(p_0 / \hat{p}) + (n - \sum_{i=1}^{n} Y_i)\log\left((1 - p_0)(1 - \hat{p})\right)\right]'$$

Menard, (1995) ----------(1.25)

where $LR \sim \chi_1^2$

C. Literature Review

Riedel et al. (2010) used logistic regression models to develop geographical patterns and predictors of malaria risk in Zambia. Their survey was carried out by the Zambian Ministry of Health and partners with the objective of estimating the coverage of interventions and malaria related burden in children less than five years.

De La Cruz et al. (2006) used logistic regression among other tests to identify factors associated with bed net use in Ghana among children less than five years of age and to compare the characteristics of mothers whose children use bed nets (doers) with those whose children do not (non-doers).

Lindsay et al. (2010) assessed the future threat from vivax malaria in the United Kingdom using two markedly different modeling approaches. They used logistic-regression model to predict historical malaria incidence between 1917 and 1918 and simple temperature-dependent, process-based model of malaria growth transmitted in the UK, based on environmental and demographic data.

D. Methodology

i. Data Collection

The data used in this study were obtained as secondary data from Kebbi State epidemic control department, a regional World Health Organization (WHO) Health Service directorate Kebbi State.

ii. Organization of Data

From the data (see appendix A), the following variables were derived that is coded as: Malaria status (1= patient has malaria and 0 = patient has no malaria).

Any of either backache or vomiting (1 = YES 0 = NO).

Ages (between 1-20years=1, 21-40years=2, 40years and above=3)

Sex (1 = male, 0 = female).

Fever less than 7days i.e. 2 to 3 days (1 = YES, 0= NO).
Temperature ≥ 37.5°C (1 = YES, 0 = NO)
Others (any of headache, convulsions, cold, cough or sweating, etc.) (1 = YES, 0 = NO).

iii. Descriptive Statistics

It was observed that out of five hundred (500) number patients on the data collected about four hundred and sixty six (466) representing 93.2% were malarial patients and about thirty four (34) representing 6.8% were not malarial patients but rather other related diseases. Moreover, out of 500 number of patients on the data collected about two hundred and twenty three (233) representing 46.6% were female and about two hundred and sixty seven (267) representing 53.4% were males. Data also showed that about two hundred and eighty two (282) representing 56.4% were between the age of 1-20 years, one hundred and one (101) representing 20.2% were between the age of 20-40 years, while, one hundred and seven (107) representing 23.4% were between the age of 40 and above. About 37 (7.4%) of

the patients have no fever while 463 (92.6%) patients have fever. Also out of 500 total number of the data collected, 42 (8.4%) patients have either backache, vomiting or all while 458 (91.6%) doesn't have either any, Also 54 (10.8%) patients do not have temperature greater than 37.5 degree while 446 (89.2%) patients have temperature greater than 37.5 degree, Also about 276 (55.2%) of the patients experience others, (i.e. either headache, convulsions, cold, cough or sweating, etc.) While 224 (44.8%) of the patients do not experience others (i.e. either headache, convulsions, cold, cough or sweating, etc.).

III. RESULT AND DISCUSSION

The following are the output for the analysis of data

Table 1.0 : Logistic Regression Predicting Likelihood of Malaria.

	B	S.E.	Wald	Df	Sig.	Odd Ratio	95% C.I.for Odd Ratio	
							Lower	Upper
Sex	-0.077	1.154	0.004	1	0.947	0.926	0.096	8.893
Ages	-0.477	0.758	0.395	1	0.529	0.621	0.141	2.742
Fever	7.309	1.429	26.150	1	0.000	1493.212	90.687	24586.53
Backache or vomiting	-0.528	3.123	0.029	1	0.866	0.590	0.001	268.787
Temperature ≥ 37.5°	3.036	1.282	5.608	1	0.018	20.814	1.687	256.749
Others(headache, e.t)	3.814	1.762	4.689	1	0.030	45.344	1.436	1431.954
Constant	-2.617	3.532	0.549	1	0.459	0.073		

It can be noted from Table 1.0 that the predictors such as fever, temperature greater than or equal to 37.5 degree and others (either any of headache, convulsions, cold, cough or sweating, etc.) with the significance values 0.000, 0.018, and 0.030 respectively are each less than $\alpha = 0.05$. Therefore we reject the null hypothesis (H_0) and conclude that there is enough evidence to show that these variables (predictors) are each not equal to zero at 95% confidence interval. This means that these predictors are each important to be included in the final model. Therefore, we can conclude that these predictors are relevant in predicting malaria in Kebbi State. From the same table, it is revealing to note that, the predictor's age, sex, backache and vomiting were dropped from the model. Since the p-values 0.947, 0.529 and 0.866 were each greater than $\alpha = 0.05$, we fail to reject the null hypothesis and conclude that there is sufficient evidence to indicate that each of the predictor's age, sex, backache and vomiting are each equal to zero. This shows that these predictors were not important to be included in the model. Hence the predictor's age, sex and backache and vomiting were not relevant in predicting malaria.

A. Interpretation of The ODDS Ratios

As in Table 1.0 the strongest predictor of the outcome of malaria patient was fever, recording an odds ratio of 1493.212 (95 % C.I. = 90.687 - 24586.528). This indicated that patients who had been checked and referred as having fever is likely to estimate the success of malaria as to those who were not referred, controlling for all other factors in the model. The odds ratio 45.344 (95% C.I. = 1.436 - 1431.954) for others (either any of headache, convulsions, cold, cough or sweating, etc.) indicating that for every treatment per patient, there were more malaria due to a pain in the head lasting for some time caused by changes in pressure in the blood vessels leading to and from the brain, controlling for other factors in the model. Again, the odds ratio with respect to temperature greater than or equal to 37.5 degree was 20.814 (95% C.I. = 1.687 - 256.749) meaning that more of the malaria was estimated by temperature greater than or equal to 37.5 degree compared to temperature less than 37.5 degree, holding other factors constant.

B. Findings

This research indicates that there is a linear relationship between malaria and predictors such as fever, temperature greater than or equal to 37.5 degree and others (either any of headache, convulsions, cold, cough or sweating, etc.). The overall logistic model obtained was: Logit (P(y=1)) = -2.617 + 7.309fever + 3.036temperature + 3.814others

Again, for test of significance of the coefficients of the predictors, the study found that the predictors; age, sex, backache and vomiting were not good predictors of malaria.

However, the covariates; fever, temperature greater than or equal to 37.5 degree and others (either headache, convulsions, cold, cough or sweating, etc.) malaria.

IV. CONCLUSION

The study provides evidence of the predictors which influence malaria among children's and adults both males and females in Kebbi State metropolis. Based on the data obtained, about 99.6% were correctly predicted in the model when the relevant factors were added to predict malaria. The model indicates that fever contributes more among other factors, in terms of influencing malaria due to pain in the head lasting for some time caused by changes due to pressure in the blood vessels leading to and from the brain, controlling other factors in the model. It was also observed that temperature at the excess of 37.5o were at higher chances of malaria in patient, even after adjusting for fever.

Therefore based on the analysis from data collected, we therefore, conclude that predictors which actually influence malaria are fever, temperature greater than or equal to 37.5 degree and others (i.e. either headache, convulsions, cold, cough or sweating, etc.).

V. RECOMMENDATIONS

Malaria is a very dangerous disease and using mosquito treated nets are the best solution to avoid it especially now that mosquitoes are developing resistance to drugs. Efforts should be made to eliminate favorable habitats where anopheles can produce eggs. Cleaning of our environment can also help in reducing cases of malaria. In the light of the above it is recommended that Doctors and Clinics should adopt the use of models designed by this research to detect prevalence of malaria among people, so that adequate measures for prevention and control of malaria can be taken early enough to avert danger of the full manifestation of the disease.

VI. REFERENCES

[1] Agresti, A. (1996). An Introduction to Categorical Data Analysis. John Wiley and Sons, Inc. http://lib.stat.cmu.edu/datasets/agresti

[2] De La Cruz N., Benjamin C., Kirk D., Bobbi G., Natasha I., Stephen A., and Robb D. (2006). Who sleeps under bed nets in Ghana; A doer/non-doer analysis of malaria prevention behaviors. Malaria Journal. Vol. 5, issue 1. pp 61-65.

[3] Hosmer D.W and Lemeshow S. (2000). Applied Logistic Regression. 2nd ed. New York, USA: John Wiley and Sons.

[4] Kleinbaum D. G. and Klein M. (1994). 2nd ed. Logistic Regression: A Self-Learning Text. John Wiley and Sons Publishers, New York. pp 15-26.

[5] Lindsay S.W., David G. H., Robert A. H., Shane A. R., and Stephen G. W. (2010). Assessing the future threat from vivax malaria in the United Kingdom using two markedly different modeling approaches. Malaria Journal. Vol. 9, Issue 1. pp 70-79.

[6] Menard, S. (1995). Applied Logistic Regression Analysis. Sage Publications. Series: Quantitative Applications in the Social Sciences, pp. 106.

[7] Riedel N., Penelope V. J. M. M., Laura G., Elizabeth C. K., Victor M. and Rick W. S. (2010). Geographical patterns and predictors of malaria risk in Zambia: Bayesian geo-statistical modeling of the 2006 Zambia national malaria indicator survey. Malaria Journal. Vol. 9. pp 37.