Performance of Logistic Regression in Tuberculosis Data

R.E Ogunsakin[*], A. B. Adebayo[**]

[*] Department of Mathematical Sciences, Ekiti State University
[**] Department of Mathematical Sciences, Ekiti State University

Abstract- This paper examined logistic regression for describing the relationship between an indications of suffering from complications pulmonary tuberculosis and its associated risk factors (predictors). Logistic regression was used as a tool to see the performance on tuberculosis data. The data used for this paper was collected from the Records Department of Federal Medical Centre, Ido Ekiti, Ekiti State, Nigeria, between the period of 2010 to 2011. At the end of the analysis, the estimated function

$$\lambda_j = 0.839 - 0.233x_{ij} - 0.311x_{2j} - 0.974x_{3j} + 1.793x_4 - 0.127x_{5j} + 0.58766_j + 0.161x_{7j}$$

revealed that complications of pulmonary tuberculosis were positively associated with social history of the patients , previous exposure to tuberculosis infection but negatively associated with age, nature of occupation. Also, the absence of complications of pulmonary tuberculosis was influenced by the presence of malaria fever.

Index Terms- Logistic regression, tuberculosis, pulmonary, risk factors

I. INTRODUCTION

Tuberculosis is a chronic infection usually of life long duration caused by two species of Mycobateria: Mycobacterium tuberculosis and Mcyobacterium Bovis. It is a serious disease worldwide and it is more common in areas of high incidence of HIV infection (Erhabor, 2002).

Estimates from the World Health Organization shows that each year about 2 million people die worldwide with this condition many of these are never aware they have this disease. Tuberculosis continues to be a major public health problem in many countries, especially in developing third world nations. In the past, tuberculosis was a major health problem in North America and Europe, while the incidence of tuberculosis has declined in the US since the 1990's when it was the leading cause of death in the United States, it is still a major concern and a resurgence of the disease has taken place in the last few years. In the year 2000, 16,377 new cases of tuberculosis were reported in the US. About a century ago, it was the most common cause of death until it was curbed with discovery of effective antibiotics in the 1950s and Rifampicin in 1970, Osuntokun (40). The disease was on the decline between 1853 and 1984 and it became a disease limited to particular risk groups like the elderly, the homeless, alcoholics, refugees, immigrants and people living under poor socio – economic conditions. There were earlier projections that tuberculosis might be eliminated by the year 2010, Tandon (47). The longstanding downward trend suddenly took a reverse turn and began to rise in the mid 1980'a in Europe, the Americans and Africa in part because Mycobacteria Tuberculosis frequently and dramatically infect persons with the AIDS / HIV, Daley (14).

The lethal association between HIV / AIDS and Tuberculosis has directed increasing attention to the problem of Tuberculosis HIV / AIDS destroys a persons immune system, leaving the HIV infected person highly susceptible to tubercle bacilli. This association is responsible for the observed increase in the tuberculosis incidence in areas of high incidence of HIV infection especially in Central and Southern Africa, Onadeko (39). In developed countries, the incidence is usually among the older individuals. Developing countries usually have a high incidence among the younger population, this means an increased likelihood of transmission to infants and young children and in the workplace.

In 1993, the WHO declared tuberculosis as a global emergency and they instituted W.H.O. global tuberculosis control policy as a measure to help combat this epidemic. The logistic model also called growth model, had been used by various statisticians in different fields of specialization. It was used by Pearl and Reed to describe the growth of an albino rat and of a tadpole's tail. Berkson J. employed the logistic model for analyzing bioassay data. Cox (1989) used the logistic model for handling quanta response data. Bishop et al (6) also used the model in the analysis of contingency tables. Besides, Agresti (1), Collett (13), Dobson (1990), Hosmer and Lemeshow (25), Raymond], Draper and Smith (1966), Morgan (1985), have all used this model to classify observations into two or more groups. In this work, the logistic regression model is used to analyze and classify a tuberculosis patient as having complications of pulmonary tuberculosis or otherwise.

II. METHODOLOGY

Logistic Regression Model

This model can give estimated probabilities that lie within the range of zero to one. It is for this important reason that logistic regression model is more suitable to use as a means of modeling probabilities.

Suppose that we have n Bernoulli observations $Y_1, Y_2 \dots ., Y_n$ in which Y_j = 0 or 1, j = 1, 2, …, n such that $P_r(Y_j = 1) = P_j \; and \; P_r(Y_j = 0) = 1 - P_j$

$E(Y_j) = 0 \; X \; P_r(Y_j = 0)] + (Y_j = 1 \; X \; P_j(Y_j = 1)] = P_r(Y_j = 1) =$

P_j *the probability of success crresponding to teh* j_{th} *response or variable*

.

For each j = 1, 2, …., n, there is a row vector $X_j = (x_{j1}, x_{j2}, \dots ., x_{jk})$ of explanatory variables. The idea here is

to find an equation that related the probability of success of the j_{th} observation i.e. P_j to some factors, i.e. K explanatory variables, $x_{j1}, x_{j2}, \ldots, x_{jk}$ that we think may influence P_j.

The logistic regression model is usually formulated by relating the probability of success of the j_{th} observation i.e. P_j conditional on a vector X_j of explanatory variables, through the logistic distribution functional form. Thus,

$$P_j = P_r\left\{P_j = \frac{1}{X_j}\right\} = \frac{e^{\beta + X_j\beta_i}}{1 + e^{\beta + X_j\beta_i}} 1 + e^{\beta} + X_j\beta_i \ldots \quad \ldots\ldots\ldots\ldots\ldots 1$$

Where

$$X_j = (X_{j1}, X_{j2}, \ldots, X_{jk})_{1xk}$$

$$\beta_i = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_k \end{pmatrix} kx_1$$

And $1 - P_j = P_r\{Y_j = 0 | X_j\} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 2$

$$= \frac{1}{1} - \frac{e^{\beta_o + X_j\beta_i}}{1 + e^{\beta_o + X_j\beta_i}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 3$$

$$= \frac{1}{1 + e^{\beta_o + X_j\beta_i}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 4$$

The $\beta_i i$ = 0, 1, 2, …., k are unknown regression coefficients or parameters that are to be estimated from the data and x_{ji} denotes the set of values of the k explanatory variables x_{ji}, $x_{j2}, \ldots, x_{jk}, i = 1, 2, \ldots, k$ associated with the j_{th} observation.

The linear logistic model for the dependence of p_j on the values of the k explanatory variables x_{ji}, $x_{j2}, \ldots, x_{jk}$ associated with the j_{th} observation is:

$$\frac{e^{\beta + X_j\beta_i}}{1 + e^{\beta + X_j\beta_i}} \cdot \frac{1 + e^{\beta + X_j\beta_i}}{1} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 5$$

$$\frac{P_j}{1 - P_j} = e\beta_o + X_j\beta_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 6$$

$$Log_e\left(\frac{P_j}{1 - P_j}\right) = \beta_o + X_j\beta_i = \beta_o + \sum_{i=1}^{k} \beta_i X_{ji} \ldots\ldots\ldots\ldots\ldots 7$$

When a linear logistic model is fitted to explore the relationship between a binary response variable and one or more predictor variables as in the case of this study, the model is referred to as a logistic regression model.

When the response variable has j mutually exclusive and exhaustive categories denoted by j = 1, 2, …., j and J_{th} category is taken as the reference category for the response variable. The choice of the reference category is arbitrary because the ordering of the categories is also arbitrary.

There are also k explanatory variables $x_1, x_2, \ldots\ldots\ldots, x_k$. Hence, the multinomial logistic regression model is then specified in log odds form as:

$$log_e \frac{P_j}{P_j} = \sum_{k=0}^{k} \beta_{jk} x_k \quad j = 1, 2, \ldots, j - 1 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots 8$$

Where

$$\sum_{j=1}^{j} P_j = 1 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 9$$

And $x_0 = 1$

The Odds and The Logit of P_j

The logit of P_j is derived from the logistic function

$$P_j = \frac{e^{\beta_o + X_j\beta_i}}{1 + e^{\beta_o + X_j\beta_i}} \quad 10$$

From1, it follows that

$$1 - P_j = \frac{1}{e^{\beta_o + X_j\beta_i}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 11$$

Dividing (1) by (2) yields

$$\left(\frac{P_j}{1 - P_j} /\right) = e^{\beta_o + X_j\beta_i} \ldots\ldots\ldots\ldots\ldots\ldots 12$$

Taking the natural logarithm (base e) of both sides, we obtain

$$log_e\left(\frac{P_j}{1 - P_j} /\right) = \beta_o + X_j\beta_i = \beta_o + \sum_{i=1}^{k} \beta_i X_{ji} = \lambda_j \ldots\ldots 13$$

The method is based on the logistic transformation or logit proportion, namely;

$$Logit_{(p)} = \frac{P}{1 - P} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 14$$

Where;

P = $P_r(y = 1)$

$(1 - P) = P_r(y = 0)$

The odds ratio is a measure of association for 2 X 2 contingency table (Agresti, 2007). In 2 X 2 tables, the probability of "success: π_2 in row 2. Within row 1, the odds of success are defined to be:

$$\theta = \frac{P_j}{1 - P_j}$$

The quantity $\frac{P_j}{1 - P_j}$ is called odds denoted as θ and the quantity $log_e\left(\frac{P_j}{1 - P_j}\right)$ is called the log odds or the logit of P_j

And

$$logit\ P_j = log_e\left(\frac{P_j}{1 - P_j}\right) = Log e^{\theta}$$

Fitting the Linear Logistic Regression Model to Binary Data

Let Y_j be a Bernoulli (binary) response variable in which $Y_j = 0 \text{ or } 1$ for all j = 1, 2,…, n depending on k explanation variables $X_{j1}, X_{j2}, \ldots . X_{jk}$. If probability of success is

$P_j = prob\,(Y = \frac{1}{j})$ means probability of Y = 1, given j.

Hence, probability of failure is 1 – probability of success.

$1 - P_j = q_j$

$q_j = P(Y = 0/_j)$

$q_j = 1 - P(Y = 1/_j)$

From linear logistic regression model,

$$P_j = \frac{e^{\beta_o + \sum_{i=1}^{k} \beta_i x_{ji}}}{1 + e^{\beta_o \sum_{i=1}^{k} \beta_i x_{ji}}}$$

The logistic regression function is the logit transformation of P, where;

$$logit\,(P) = lin\frac{P}{1-P} = \beta_o + \beta_i x_i + \ldots + \beta_k x_k$$

Where β_o = the constant of the equation and β_i= the coefficient of the predictor variables i. Using the logistic transformation in this way overcomes problems that might arise if p was modeled directly as a linear function of the explanatory variables; in particular it avoids fitted probabilities outside the range (0, 1). The parameters in the model can be estimated by maximum likelihood estimation.

The Hospital diagnostic index cards and the case notes of these discharged patients were thoroughly studied with particular attention being paid to some of the factors (explanatory variables or predictors) influencing the probability of having complications of pulmonary tuberculosis which formed the main focus of this study.

The dependent variable Y is defined as
Y (Outcome) = 1, Success (π_1)
0, Failure (π_2)

In this project work, a Tuberculosis patient is considered to have attained "Success" if he or she had suffered from complications of pulmonary Tuberculosis after clinical diagnosis; otherwise, he or she is considered to have attained "failure".

The predictors (independent or explanatory) variables available for this work are defined as follows:

Age (X_1) =
1, 15 – 24 years
2, 25 – 34 years
3, 35 – 44 years
4, 45 – 54 years
5, Age ≥ 55years

Nature of occupation before (X_2) = 0, No job
1, Student
2, Unskilled workers e.g. (traders, workers in the cement or tobacco factory or Quarry etc
3, Skilled workers e.g. workers in chest Hospitals or Tuberculosis wards etc.

Previous Contact with a person Having chronic cough or an infected person (X_3)
0, No sign of contact
1, Sign of contact

Social History
(Tobacco Smoking and (X_4) =
0, If the patient had not smoked or drank before
1, If the patient had smoked and drank before
2, If the patient had not drank at all but smoked

Alcohol consumption
3, If the patient had not smoked at all but had drank before.

Previous exposure
to diseases (X_5) =
0, None
1, Presence of HIV/AIDS as the main disease
2, Presence of at least one from diseases that can depress immunity apart from HIV/AIDS (Diabetes, Leprosy, Cancer, Malnutrition, Measles)
3, Presence of at least one from Hypertension, Pneumonia with or without Malaria fever
4, Presence of only Malaria fever

Previous exposure
to Tuberculosis (X_6) infection
0, No previous Tuberculosis infection
1, If the patient had been infected before but failed to complete his/her treatment
2, If the patient had been infected before but completed his or her treatment.

Length of time of reporting
to the right hospital after (X_7) =
noticing persistent cough or discomfort.
Other discomfort
1, if the Patient had reported after 1-3 weeks of persistent cough or other discomfort.
2, if the patient had reported after 1-5 months of persistent cough or other discomfort.
3, if the patient had reported after 6-10 months of persistent cough or other
4, if the patient had reported after 11-15 months of persistent cough or other discomfort
5, if the patient had reported after 16-20 months of persistent cough

III. RESULTS

The data collected from fifty randomly selected discharged Tuberculosis patients consisting of the dependent variable (outcome) Y and the explanatory variables (predictors) X_1, X_2, X_3, X_4, X_5, X_6 and X_7 were analyzed using the SPSS (17.0). The cross-tabulation of each independent variable $X_¿$ with dependent variable Y was examined and the chi-squared Test was carried out for each independent variable in-order to ascertain whether they are dependent or not.

The table below shows the result of the chi-squared test for the eight independent variables.

Table1

Variable	Calculated Chi-Squared value	Tabulated Chi-Squared Value	Df	Significant P-Value
X_1	4.200	9.488	4	0.380
X_2	38.480	7.815	3	0.000
X_3	28.880	3.841	1	0.000
X_4	44.920	5.991	2	0.000
X_5	12.200	9.488	4	0.016
X_6	54.760	5.991	2	0.000
X_7	15.600	9.488	4	0.004

Test of Independence

It is important to test for the independence of variables which will tell us whether variable $X_¿$ is dependent or not with variable $Y¡$. That is, we wish to test the hypothesis

$H_0 : P_{¿¡} = P_{¡} P_{·j}$

Against

$H_1 : P_{¿¡} \neq P_{¡ \cdot} P_{·j}$

For all ¡ and j where P¡ j is the probability that both X¿and Yj occur

For testing independence in r X c contingency tables, the calculated chi-squared is obtained from

$$X^2 = \Sigma \frac{(0 - E)^2}{E} \quad (3.5.1)$$

Based on (r - 1) (c - 1) degrees of freedom

Where

O are the observed frequencies

$E = \frac{\text{Row Total x Column Total}}{n}$ are the expected frequency values.

Since the chi-square calculated for variables X_2, X_3, X_4, X_5, X_6 and X_7 as shown in the table 3.0 above are greater than their corresponding chi-square values from the table at $\alpha = 0.05$, we therefore reject the null hypothesis of statistical independence of these variables and the dependent variable and conclude that variables X_2, X_3, X_4, X_5, X_6 and X_7 are not independent of the observed outcome Y. this is also in conformity with their significant P – values [prob ($x^2 \geq$ observed] which are less than 0.05.

Correlation Matrix

The correlation analysis of the dependent variable and independent variables with one another were carried out using the SPSS (17.0) computer program and the results is shown below.

Table 2: Correlation Matrix

Variable	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7
Y	1.000	-0.379	-0.416	-0.313	-0.021	-0.529	-0.246	-0.579
X_1	-0.379	1.000	-0.475	0.340	-0.299	0.150	0.218	0.106
X_2	-0.416	-0.475	1.000	-0.035	0.105	0.052	0.219	0.164
X_3	-0.313	0.340	-0.035	1.000	-0.030	0.030	-0.104	0.143
X_4	-0.021	-0.299	0.105	-0.030	1.000	0.210	0.149	-0.076
X_5	-0.529	0.150	0.052	0.030	0.210	1.000	0.161	-0.075
X_6	-0.246	-0.218	0.219	-0.104	0.149	0.161	1.000	0.203
X_7	-0.579	0.106	0.164	0.143	-0.076	-0.075	0.203	1.000

It was discovered that the independent variables are correlated some are highly positive correlated while some are highly negative correlated and also low positive correlated with response variable Y, hence they could all be used for the analysis.

Estimation of the Linear Logistic Regression Parameters

Data used for the analysis comprised of fifty randomly selected discharged Tuberculosis patients consisting of the outcome variable Y (dichotomous values) and the predictors X_1 X_2 X_3 X_4 X_5 X_6 and X_7 were analyzed. SPSS software package was used for the analysis, the maximum likelihood method is used to estimate the coefficients and its standard error in addition the Newton – Raphson method solve the non linear equations for the logistic model maximum likelihood estimations.

Table3

Variable	Beta Estimate	Standard Error of Beta	Wald statistic value	Degree of freedom	Significant P-value
X_1	-0.233	0.345	0.457	1	0.499
X_2	-0.311	0.537	0.335	1	0.563
X_3	-0.974	1.031	0.893	1	0.345
X_4	1.793	1.156	2.407	1	0.121
X_5	-0.127	0.269	0.225	1	0.635
X_6	0.587	0.532	1.221	1	0.269
X_7	0.161	0.264	0.372	1	0.542
Constant	0.839	1.674	0.251	1	0.616

From the result shown in the table 3.3 above the estimated function is:

$\lambda j = 0.839 - 0.233x_1j - 0.311x_{2j} - 0.974x_3j + 1.793x_4j - 0.127x_5j + 0.587x_6j + 0.161x_7j$ …….. (3.7.1)

At 0.05 level of significant, Table 3.4 shows that variable X_4, X_6, and X_7 increases the probability of complications of pulmonary Tuberculosis while X_1, X_2, X_3, and X_5 decreases the probability of complications of tuberculosis. X_4 is highly significant in order words strongly contributes to the complications of pulmonary Tuberculosis next is X_3 while X_7 make negligible contributions to it. This then implies the larger the value of coefficient for a variable, the bigger is the impact of such a variable to the outcome variable.

Therefore, the logistic regression model is

$$Pj = \frac{e(\lambda_j)}{1 + e(\lambda_j)}$$

The Odds Ratio Results

The following odds ratios were calculated using the formula;

$\theta = \frac{P}{1-P_j}$ and 95% confidence intervals.

While the formula for the upper and lower limit of the odd ratio is given by

$\exp(\beta \pm Z\alpha/_2 S_\beta$

Where

β is the maximum likelihood estimate of β

α is the level of significance which is 0.05

$Z\alpha/_2$ is the upper (one sided) $\alpha/2$ point of the standard normal distribution which is 1.96

And

$S\beta$ is the standard error of β

The table below gives the odds ratio for each predictor variable and their corresponding 95% confidence interval.

TABLE4: Odd Ratio Results

Variable	Odds Ratio	95% C.I.	
		Lower	Upper
X_1	0.792	0.403	1.558
X_2	0.733	0.256	2.099
X_3	0.377	0.050	2.848
X_4	6.008	0.623	57.903
X_5	0.880	0.520	1.492
X_6	1.799	0.634	5.102
X_7	1.175	0.700	1.971

From table4 it is evident that variables X_4, X_6, and X_7 are susceptible for complications of pulmonary tuberculosis.

The Hypothesis Testing

The interest here is to find out which among the logistic regression coefficients or beta estimates contributes to the significance, by testing for the individual beta estimates using our Wald statistic. Hence, the hypothesis becomes:

$H_0 : \beta_¡ = 0$

against

$H_¡ : \beta_¡ \neq 0$

Hence from the test statistic, we conclude that since the Wald statistic value calculated for variables X_1, X_2, X_3, X_4, X_5, X_6 and X_7 as shown in table 3.1 are less than the chi-squared value from the table at $\alpha = 0.05$, we therefore accept H_0 and conclude that the variables are not significant. Hence, X_1, X_2, X_3, X_4, X_5, X_6, and X_7 does not significantly help to predict the complications of pulmonary tuberculosis. We can only base our assumptions on the chi-square test which is more powerful and reliable as an alternative to the Wald Test.

The Test Of Goodness-Of-Fit of The Model

It is always desirable to test for the goodness of fit for the logistic model which will tell us whether a model of this form provides a good fit to the data or not.

The hypothesis then becomes
$H_0 : y_i = \mu_i$
against
$H_i : y_i \neq \mu_i$

The Hosmer Lemeshow goodness-of-fit test divides the subjects (i.e. cases used in the analysis which are fifty discharged tuberculosis patients) into deciles based on predicted probabilities, then computes a chi-square from observed and expected frequencies.

The Hosmer Lemeshow statistic is employed to test for the goodness-of-fit of the model. The calculated Hosmer Lemeshow goodness-of-fit test statistic is obtain,

TABLE5: Contingency Table for Hosmer and Lemeshow Test

Group	Y = 0 Observed	Expected	Y = 1 Observed	Expected	Total
1.	4.000	4.463	2.000	1.537	6.000
2.	4.000	3.425	1.000	1.575	5.000
3.	3.000	3.201	2.000	1.799	5.000
4.	4.000	2.920	1.000	2.080	5.000
5.	2.000	2.596	3.000	2.404	5.000
6.	2.000	2.005	3.000	2.995	5.000
7.	0.000	1.706	5.000	3.294	5.000
8.	2.000	1.222	3.000	3.778	5.000
9.	1.000	0.448	4.000	4.552	5.000
10.	0.000	0.014	4.000	3.986	4.000

Chi-square	df	Sig. P-value
Goodness-of-fit-test	5.776	8
	0.672	

Since the calculated Hosmer Lemeshow Goodness-of-fit test statistic is less than X^2 (8, 0.05) value obtained from the table (i.e. 15.507), we then accept H_0 and conclude that there is no difference between the observed and the model-predicted or fitted values of the dependent. This then implies that the model's estimates fit the data at $\alpha = 0.05$. Also the value of Hosmer Lemeshow goodness-of-fit statistic computed for the full model is C = 5.776 at the corresponding p-value computed from the chi-square distribution with 8 degree of freedom is 0.672 this indicates that the model seems to fit quite well.

Classification of Tuberculosis Patients

Table 6 gives the classification table. Using the obtained λ_j function observations are classified as follows using a prior probability of 0.56.

TABLE6: Classification Results

Observed	Predicted Y		
	Failure π_0	Success π_1	Percentage correct
Failure π_0	17	5	77.3
Success π_1	9	19	67.9
Overall percentage			72.0

From the table above, we conclude that 77.3% of all Tuberculosis patients not having complications of pulmonary tuberculosis are correctly classified, and 32.1% are incorrectly classified. 67.9% of all Tuberculosis patients having complications of pulmonary Tuberculosis are correctly classified, and 22.7% are incorrectly classified.

Therefore, the overall percent correctly classified by this model is 72% ($\frac{17 + 19}{}$ x 100%) while the overall percent incorrectly classified is 28% ($\frac{9+5}{}$) x 100%.

IV. Discussion of Results

The estimated logistic regression function classified 17 of the 22 tuberculosis patients in the observed group (failure) correctly for 77.3% and also classified correctly 19 of 28 tuberculosis patients in the observed group (success) for 67.9%. The model incorrectly classified 5 of the 22 tuberculosis patients in the failure group as having complications of pulmonary tuberculosis (success group) when did not, for 22.7%. And also classified incorrectly 9 of 28 tuberculosis patients in the success group as not having complications of pulmonary tuberculosis (failure group) when they did for 32.1%.

In order to establish the association that exists between risk factors (predictor variables) and the complications of pulmonary tuberculosis (outcome variable) the estimator of the predictor variables for the logistic regression function were obtained and presented in table 3.4.

The estimated function is:

$$\lambda_j = 0.839 - 0.233x_{ij} - 0.311x_{2j} - 0.974x_{3j} + 1.793x_4 - 0.127x_{5j} + 0.58766_j + 0.161x_{7j}$$

The function obtained from (4.1.1) show that complications of pulmonary tuberculosis were positively associated with social history of the patient, previous exposure to tuberculosis infection and length of time or reporting to the right hospital but negatively associated with age, nature of occupation, previous contact with an infected person. It is also observed that absence of complications of pulmonary tuberculosis was influenced mainly by the presence of malaria fever than presence of complications of pulmonary tuberculosis. Presence of HIV / AIDS as the main cause disease associated most strongly to the occurrence of complications of pulmonary tuberculosis. Absence of complications of pulmonary tuberculosis was also associated with previous contact with an infected person. The presence of complications of pulmonary tuberculosis was strongly associated with previous exposure to tuberculosis infection, social history of the patients and length of time of reporting to the hospital.

In the context of this work, it was observed that is interesting however to note that the areas with high predicted probability of 'success' coincide with areas of presence of either HIV or other immune – suppressive disease with a longer duration of persistent cough before reporting to the hospital. The longer the patient stays at home before reporting after noticing persistent cough or other discomfort, the higher is the chance or probability of suffering from complications of pulmonary tuberculosis. It was observed that patients that smoke tobacco with poor socio – economic status are also prone to complications of pulmonary tuberculosis. It was also observed that some of this patient who had complications of pulmonary tuberculosis had RIP on their case notes, which supports the fact that tuberculosis is indeed a chronic disease.

V. CONCLUSION

The age, nature of occupation before infection, previous contact with person having chronic cough, social history, previous exposure to diseases, previous exposure to tuberculosis infection which are the predictor variables, and the complications of pulmonary tuberculosis (outcome variables) has been used to establish the logistic regression function for the complications of pulmonary tuberculosis.

The project work has successfully found logistic regression function for the patients owing to the fact that social history, previous exposure to tuberculosis and length of time of reporting to the hospital contributed significantly to the complications of pulmonary tuberculosis.

We conclude that the most powerful variables in determining complications of pulmonary tuberculosis are social history of the patients, followed by previous exposure to tuberculosis infection and length of time reporting to the right hospital.

REFERENCES

[1] Afifi, A., Clark, V. A. and May, S. (2004). Computer Aided Multivariate Analysis. Fourth Edition. Chapman and Hall, London.

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

[3] Anderson, D. A. (1988). Some Models for Overdispersed Binomial Data. Australian Journal of Statistics, 30, 125 – 148.

[4] Cayla, J. A., M. T. Brugal (1987 – 1997). Factors Predicting Non – Completion of Tuberculosis Treatment among HIV infected Patients in Barcelona. INT J. TUBERCLUNG DIS 2000; 4(1): 55 – 60.

[5] Cox, D. R. (1972). Regression Models and Life Tables (with discussion). Journal of Royal Statistical Society, B. Volume 34, pages 55 – 71.

[6] D. Antoine, J. Jones, M. Watson (2001). Tuberculosis Treatment Outcome Monitoring in England, Wales and Northern Ireland for Cases Reported in Journal of Epidemiol Community Health, 61: 302 – 307.

[7] Draper, N. R. and Smith, H. (1966). Applied Regression Analysis, John Wiley and Sons Inc. New York.

[8] Everitt, B. S. (1998). The Cambridge Dictionary of Statistics. Cambridge University Press.

[9] Floyd, K., L. Blane, J. Lee (2002). Resource required for Global Tuberculosis Control". Science, Vol. 295, pp. 2040 – 2041.

[10] Hauck, W. W. and Donner, A. (1977). Wald's Test as Applied to Hypothesis in Logit Analysis. Journal of the American Statistical Association, 72, pp. 851 – 853.

[11] Hosmer, D. W., Lemeshow, S. (2000). Applied Logistic Regression, Second Edition, Wiley, Inc. New York.

[12] Johnson, W. (1985). Influence Measures for Logistic Regression: Another Point of View. Biometrika, 72, 59 – 65.

[13] WHO/HTM/TB/2006.35. Geneva: Tuberculosis Research and Development. Report of a WHO Working Group Meeting, Geneva, 9 – 11 September, Geneva.

AUTHORS

First Author – R.E. O gunsakin, M.Sc, B.Sc, Ekiti State University, Nigeria. re.korede@yahoo.com
Second Author – A.B. Adebayo. B.Sc, Ekiti State University, Nigeria. oreropo@gmail.com.

Correspondence Author – R.E. Ogunsakin, M.Sc, B.Sc, Ekiti State University, Nigeria. re.korede@yahoo.com , +234(0)8062512714