Journal of the Nigerian Association of Mathematical Physics
Volume 37, (*November, 2016*), pp 163 – 172

Health Status Analysis of the Erythrocyte Sedimentation Rate (ESR) of Patients Using Logistic Regression Method

[1]Aye P.O, [2]Ologbonyo J.J and [3]Azuaba E.

[1,2]Department of Mathematical Sciences, Adekunle Ajasin University, Akungba-Akoko, Ondo State, Nigeria
[3]Department of Mathematics, Federal University of Technology, Minna, Nigeria.

Abstract

In this study, a logistic regression model was fitted using a secondary data from the Institute of Medical Research, Kuala Lumpur, Malaysia which contained 32 observations and 3 variables(Fibrinogen (X_f), γ-Globulins (X_g) and Health Status). Three logistic regression models were constructed namely: the null model (model with only the constant), model with the two predictors and model with the best predictor.Backward selection procedure was used in this study to select the best variable that contributes to the outcome of the erythrocyte sedimentation rate (ESR). The logistic regression models were constructed for the data with their corresponding accuracy and precision at 95% confidence level.Various tests were carried out to validate the models obtained.The result shows that only Fibrinogen contributed significantly to the outcome of the erythrocyte sedimentation rate (ESR) while γ-Globulins do not really contributed. It is indeed possible to predict whether a patient is healthy or not based on the values of Fibrinogen (X_F) and γ-Globulins (X_g) of an individual. Also it was discovered that the Fibrinogen (X_f) is the best predictor to use in predicting whether a patient is healthy or not.The best model was the model with the two predictors with 87.5% accuracy.

Keywords: Logistic regression, erythrocytes sedimentation rate, fibrinogen ,y-globulin ,health status, maximum likelihood, prediction, akaike information criteria, deviance, modelling.

1.0 Introduction

Logistic regression model is a mathematical model that describe the relationship of several independent variableless to a binary (dichtomous) dependent variable.Logistic regression model was developed primarily by Cox in 1958 and Walker and Duncan in 1967 [1- 4].Logistic regression can be seen as a special case of generalized linear model and thus analogous to linear regression. The model of logistic regression, however, is based on quite different assumptions (about the relationship between dependent and independent variables) from those of linear regression [5].

In particular the key differences of these two models can be seen in the following two features of logistic regression. First, the conditional distribution $y \mid x$ is a Bernoulli distribution rather than a Gaussian distribution, because the dependent variable is binary. Second, the predicted values are probabilities and are therefore restricted to [0,1] through the logistic distribution function because logistic regression predicts the probability of particular outcomes [6-7].

Logistic regression is an alternative to Fisher's 1936 classification method, linear discriminant analysis. If the assumptions of linear discriminant analysis hold, application of Bayes' rule to reverse the conditioning results in the logistic model, so if linear discriminant assumptions are true, logistic regression assumptions must hold. The converse is not true, so the logistic model has fewer assumptions than discriminant analysis and makes no assumption on the distribution of the independent variables [8-9]. All the analysis were carried out using Statistical Package for Social Sciences [10-11].

1.1 The Erythrocytes Sedimentation Rate

The erythrocyte sedimentation rate (ESR), also called a sedimentation rate or Westergren ESR is the rate at which red blood cells sediment in a period of one hour. It is a common hematology test, and is a non-specific measure of inflammation.

The erythrocyte sedimentation rate (ESR) is the rate at which red blood cells (erythrocytes) settle out of suspension in blood

Corresponding author: Aye P.O, E-mail: ayepatricko@gmail.com, Tel.: +2347030596765

plasma, when measured under standard conditions. The ESR increases if the levels of certain proteins in the blood plasma rise, such as in rheumatic diseases, chronic infections and malignant diseases; this makes the determination of the ESR one of the most commonly used screening tests performed on samples of blood. One aspect of a study carried out by the Institute of Medical Research, Kuala Lumpur, Malaysia, was to examine the extent to which the ESR is related to two plasma proteins, fibrinogen and γ-globulin, both measured in gm/l, for a sample of thirty-two individuals. The ESR for a 'healthy' individual should be less than 20 mm/h and since the absolute value of the ESR is relatively unimportant, the response variable used here will denote whether or not this is the case. A response of zero will signify a healthy individual (ESR < 20) while a response of unity will refer to an unhealthy individual (ESR ≥20) [12].

1.2 Problem Statement

There has been need for a statistically verifiable model that can adequately predict the Health status of individual through erythrocyte sedimentation rate .

1.3 Purpose of Study

The aim of this research work is to build a statistically verifiable model that can adequately predicts the health status of an individuals .

2.0 Research Methodology

2.1 Hypotheses of the Research

Three hypotheses of the research are summarized as follows:

(1) $H_0: \chi_L^2 > \chi_{\propto,s-p}^2$: The fit is adequate for the model.
$H_1: \chi_L^2 < \chi_{\propto,s-p}^2$: The fit is not adequate for the model.

(2) $H_0: \mathcal{X}_{cal}^2 > \mathcal{X}_{tab}^2$: There is no difference between observed and model-predicted values.
$H_1: \mathcal{X}_{cal}^2 < \mathcal{X}_{tab}^2$: There is difference between observed and model-predicted values.

(3) $H_0: \chi_L^2 > \chi_{\propto,s-p}^2$: Null model is a better fit than fitted model.
$H_1: \chi_L^2 < \chi_{\propto,s-p}^2$: Model is a better fit than null model.

2.2 Significance of the Study

i This research work is useful and significant as it presents an example of the application of statistics in medicine and describes how logistic regression can be used for assessing erythrocytes sedimentation rate.

ii This research work is reproducible, that is, all resources used in the work are available and accessible thereby creating an enabling environment for further study or research.

iii Medical practitioners could make use of this research to predict the health status of their patient.

2.3 Logistics Regression Assumptions

i The data $Y_1, Y_2, \ldots, Y_n$ are independently distributed, i.e., cases are independent.

ii Distribution of Y_i is $\text{Bin}(n_i, \pi_i)$, i.e., binary logistic regression model assumes binomial distribution of the response. The dependent variable does not need to be normally distributed, but it typically assumes a distribution from an exponential family (e.g. binomial, Poisson, multinomial, normal,etc.)

iii It does not assume a linear relationship between the dependent variable and the independent variables (i.e. multicollinearity), but it does assume linear relationship between the logit of the response and the explanatory variables; $\text{logit}(\pi) = \beta_0 + \beta_X$.

iv Independent (explanatory) variables can even be the power terms or some other nonlinear transformations of the original independent variables.

v The homogeneity of variance does not need to be satisfied. In fact, it is not even possible in many cases given the model structure.

vi Errors need to be independent but not normally distributed.

vii It uses maximum likelihood estimation (MLE) rather than ordinary least squares (OLS) to estimate the parameters, and thus relies on large-sample approximations.

viii Goodness-of-fit measures rely on sufficiently large samples, where a heuristic rule is that not more than 20% of the expected cells counts are less than 5.

2.4 Materials and Methods

The secondary data used in this study was obtained from the screening test perform on samples of blood by the Institute of Medical Research, Kuala Lumpur, Malaysia. It consist of 32 observations both the response variable with two categories and two explanatory variables. The response variable was the Erythrocytes Sedimentation Rate (ESR) classified as 0 or 1, and the explanatory variables were the amounts of fibrinogen and γ-globulin measured in gm/l [13].

Table 2.1: The collected data for the analysis

S/N	X_f	X_g	Y_i
1	2.52	38	0
2	2.56	31	0
3	2.19	33	0
4	2.18	31	0
5	3.41	37	0
6	2.46	36	0
7	3.22	38	0
8	2.21	37	0
9	3.15	39	0
10	2.6	41	0
11	2.29	36	0
12	2.35	29	0
13	5.06	37	1
14	3.34	32	1
15	2.38	37	1
16	3.15	36	1
17	3.53	46	1
18	2.68	34	0
19	2.6	38	0
20	2.23	37	0
21	2.88	30	0
22	2.65	46	0
23	2.09	44	1
24	2.28	36	0
25	2.67	39	0
26	2.29	31	0
27	2.15	31	0
28	2.54	28	0
29	3.93	32	1
30	3.34	30	0
31	2.99	36	0
32	3.32	35	0

2.4 Method of Data Analysis

Binary logistic regression estimates the probability that a characteristic is present (e.g. estimate probability of "success") given the values of explanatory variables, in this case a single categorical variable; π = Pr (Y = 1|X = x). Consider the predictor variable X to be any of the risk factor that might contribute to the disease. Probability of success will depend on levels of the risk factor.

Variables:

i. Let Y be a binary response variable
   $Y_i = 1$ if the trait is present in the observation (person, unit, e.t.c..)
   $Y_i = 0$ if the trait is NOT present in the observation.

ii. X = (x_1, x_2, ...,x_k) be a set of explanatory variables which can be discrete, continuous, or a combination. x_i is the observed value of the explanatory variables.

2.4.1 Logit Function

The logit function is a function $F(t) of$ the form:

$$\pi(t) = \frac{e^t}{1+e^t} \quad \text{.....} \quad (2.1)$$

The logit function takes on only values between zero and one.

2.4.2 Model Fit:

Overall goodness-of-fit statistics of the model; we will consider:

1. Likelihood ratio test
2. Hosmer-Lemeshow test and statistic
3. Residual analysis: deviance

2.5 Parameter Estimation

Since the logistic regression is a modified form of the classical linear regression model, the logistic regression model bears some similarities with the classical linear regression model. Given a dependent variable Y and a predictor variableX, then we have that:

$Y_i = \beta_0 + \beta_1 X_i + e_i$……..... (2.2)

Then

$E[Y_i \backslash X_i, \beta_0, \beta_1] = \beta_0 + \beta_1 X_i$……..... (2.3)

Where $E[Y_i \backslash X_i, \beta_0, \beta_1]$ *i*s the expected value of the dependent variable for a given value of the predictor variable.

Then the logistic regression model is given by

$\pi(x) = Pr[Y_i \backslash X_i, \beta_0, \beta_1] = \frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}}$……..... (2.4)

or

$log\left(\frac{Pr[Y_i \backslash X_i, \beta_0, \beta_1]}{1-Pr[Y_i \backslash X_i, \beta_0, \beta_1]}\right) = \beta_0 + \beta_1 X_i$……..... (2.5)

Where Y_i is the outcome or dependent variable with two categories and X_i is the independent variable.

The term

$\frac{Pr[Y_i \backslash X_i, \beta_0, \beta_1]}{1-Pr[Y_i \backslash X_i, \beta_0, \beta_1]}$……..... (2.6)

is usually referred to as the **odds ratio**.

while the term

$log\left(\frac{Pr[Y_i \backslash X_i, \beta_0, \beta_1]}{1-Pr[Y_i \backslash X_i, \beta_0, \beta_1]}\right)$……..... (2.7)

is usually referred to as the **log odds.**

The logistic regression model given above, models probability that a dependent variable Y_i takes on a value for a given independent variable X_i using the logit function.

The general method of estimation that leads to the least squares function under the linear regression model (when the error terms are normally distributed) is called maximum likelihood. This method provides the foundation for the approach to estimation with the logistic regression model. In a general sense, the method of maximum likelihood yields values for the unknown parameters that maximize the probability of obtaining the observed set of data. In order to apply this method one must first construct a function, called the likelihood function. This function expresses the probability of the observed data as a function of the unknown parameters. The maximum likelihood estimators of the parameters are the values that maximize this function. Thus, the resulting estimators are those that agree most closely with the observed data. We now describe how to find these values for the logistic regression model.

If *Y* is coded as 0 or 1 then the expression for $\pi(x)$ given in equation (2.4) provides (for an arbitrary value of $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$, the vector of parameters) the conditional probability that *Y* is equal to 1 given *x*. This is denoted as $\pi(x)$. It follows that the quantity $(1 - \pi(x))$ gives the conditional probability that *Y* is equal to zero given *x*, $\Pr(Y = 0|x)$. Thus, for those pairs (X_i, Y_i), where $Y_i = 1$, the contribution to the likelihood function is $\pi(x_i)$, and for those pairs where$Y_i = 0$, the contribution to the likelihood function is $1 - \pi(x_i)$, where the quantity $\pi(x_i)$ denotes the value of $\pi(x)$ computed at x_i. A convenient way to express the contribution to the likelihood function for the pair (x_i, Y_i) is through the expression

$\boldsymbol{\pi}(x_i)^{Y_i}[1 - \boldsymbol{\pi}(x_i)]^{1-Y_i}$……..... (2.8)

Fitting the logistic regression model, as the observations are assumed to be independent, the likelihood function is obtained as the product of the terms given in equation (2.8) as follows:

$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(x_i)^{Y_i}[1 - \pi(x_i)]^{1-Y_i}$..... (2.9)

The principle of maximum likelihood states that we use as our estimate of $\boldsymbol{\beta}$ the value that maximizes the expression in equation (2.9). However, it is easiermathematically to work with the log of equation (2.9). This expression, the *loglikelihood*,is defined as

$L(\boldsymbol{\beta}) = \ln L(\boldsymbol{\beta}) = \sum_{i=1}^{n} Y_i \ln[\boldsymbol{\pi}(x_i)] + (1 - Y_i)\ln[1 - \boldsymbol{\pi}(x_i)]$……..... (2.10)

To find the value of $\boldsymbol{\beta}$ that maximizes $L(\boldsymbol{\beta})$ we differentiate $L(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ and set the resulting expressions equal to zero. These equations, known as the *likelihood equations*, are:

$\sum[Y_i - \boldsymbol{\pi}(x_i)] = 0$……..... (2.11)

$\sum x_i[Y_i - \boldsymbol{\pi}(x_i)] = 0$……..... (2.12)

In equations (2.11) and (2.12) it is understood that the summation is over *i* varying from 1 to *n*.

In linear regression, the likelihood equations, obtained by differentiating the sum-of-squared deviations function with respect to $\boldsymbol{\beta}$ are linear in the unknown parameters and thus are easily solved. For logistic regression the expressions in equations (2.11) and (3.12) are nonlinear in $\boldsymbol{\beta_0}$and $\boldsymbol{\beta_1}$, and thus require special methods for their solution. These methods are iterative in nature and have been programmed into logistic regression software such as SPSS used in this research work [10].

2.6 Evaluating Model Fit

Listed below are the measures used for evaluating the adequacy and fitness of our model.

1. The Akaike Information Criteria (AIC): The Akiake Information Criteria (AIC) is a measure of how good a statistical model is for a given data set. AIC helps provide a means for model selection because the preferred model is always the one with the minimum AIC value. Mathematically, AIC is given by

$\mathrm{AIC} = 2\mathrm{k} - 2\ln(\mathrm{L})$……..... (2.13)

Where k is the number of parameters in the model and L is the likelihood function for the estimated model. Generally, the lower the AIC value of a model, the better the fit.

2. Deviance: The deviance is a measure of the lack of fit of the logistic model to the dataset. It is analogous to sum of squares in ordinary least squares. It is a quality of fit statistic for assessing fit of a model through hypothesis testing. Mathematically, Deviance of a given model is given by

$D = -2ln\left(\frac{likelihoodofthefittedmodel}{likelihoodofthesaturatedmodel}\right)$……..... (2.14)

Where the saturated model is a model with a theoretically perfect fit.

The Deviance statistic is used in carrying out the likelihood ratio test which is used in assessing the contribution of a predictor or set of predictors to the model.

2.7 Hypothesis Testing

In this project, Likelihood ratio test using deviance to assess the model fit was used.

H_0:The fit is adequate for the model

H_1:The fit is not adequate for the model

Test statistics $\chi_L^2 = Nulldeviance - residualdeviance$

$\chi_L^2 \sim \chi_{s-p}^2$where$s - p$ is the difference in the number of parameters of the null model and the prediction model.

Decision rule:

Reject H_0 if $\chi_L^2 > \chi_{\propto,s-p}^2$, otherwise do not reject.

2.8 Forecasting with Linear Logistic Regression

Logistic regression is used to predict a categorical (usually dichotomous) variable from a set of predictor variables. logistic regression is often chosen if the predictor variables are a mix of continuous and categorical variables and/or if they are not nicely distributed (logistic regression makes no assumptions about the distributions of the predictor variables). Logistic regression has been especially popular with medical research in which the dependent variable is whether or not a patient has a disease, of which in the case of this study, the dependent variable is whether or not a patient's health status is normal.

3.0 The Model Analysis

The Model: $\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \cdots + \beta_k x_k$ ..... (3.1)

Table 3.1 : Number and percentage of healthy and unhealthy

OUTCOME	FREQUENCY	PERCENTAGE
HEALTHY	25	78.1%
UNHEALTHY	7	21.9%
TOTAL	32	100.0%

Table 3.2: The initial model table

		B	S.E.	Wald	Df	Sig.	Exp(B)
Step 0	Constant	-1.273	0.428	8.862	1	0.003	0.280

Constant only model,

Only constant is in the model and our predictors are not in the equation yet.

Table 3.3: Variables not in the equation 3.2

			Score	Df	Sig.
Step 0	Variables	X_f	7.289	1	0.007
		X_g	1.865	1	0.172
	Overall Statistics		8.614	2	0.013

Table 3.4: Omnibus tests of model coefficients

		Chi-square	Df	Sig.
Step 1	Step	8.621	2	0.013
	Block	8.621	2	0.013
	Model	8.621	2	0.013

The omnibus test of the model's coefficient gives us chi-square of 8.621 on 2 degree of freedom with p-value of 0.013

Table 3.5: Model summary

Step	-2 Log likelihood	Cox & Snell R square	Nagelkerke R Square
1	25.000	0.236	0.363

Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.-2 log likelihood statistic is 25.000, this statistic measures how good the model predicts the decision (the small the statistic, the better the model). Cox and Snell R square gives the proportion of the variable in outcome of the patient's ESR by 23.6% and Nagelkerke R Square by 36.3% which is also good for the model.

3.1 Hosmer and Lemeshow's Goodness of Fit Test

Table 3.6: Contingency table for Hosmer and Lemeshow test

		$Y_i = 0$		$Y_i = 1$		Total
		Observed	Expected	Observed	Expected	
Step 1	1	3	2.918	0	0.082	3
	2	3	2.892	0	0.108	3
	3	3	2.809	0	0.191	3
	4	2	2.771	1	0.229	3
	5	3	2.707	0	0.293	3
	6	2	2.576	1	0.424	3
	7	3	2.377	0	0.623	3
	8	1	2.214	2	0.786	3
	9	3	1.874	0	1.126	3
	10	2	1.864	3	3.136	5
Step 2	1	2	2.826	1	0.174	3
	2	3	2.803	0	0.197	3
	3	3	2.773	0	0.227	3
	4	2	2.723	1	0.277	3
	5	3	2.646	0	0.354	3
	6	3	2.596	0	0.404	3
	7	3	2.497	0	0.503	3
	8	2	2.146	1	0.854	3
	9	3	2.507	1	1.493	4
	10	1	1.482	3	2.518	4

Table 3.7: Hosmer and Lemeshow test

Step	Chi-square	Df	Sig.
1	9.584	8	0.295
2	8.726	8	0.366

Table 3.8: Classification table

	Observed		Predicted		
			Y_i		Percentage Correct
			0	1	
Step 1	Y_i	0	25	0	100.0
		1	4	3	42.9
	Overall Percentage				87.5
Step 2	Y_i	0	25	0	100.0
		1	5	2	28.6
	Overall Percentage				84.4

H_0: There is no difference between observed and model-predicted values.
H_1: There is difference between observed and model-predicted values
$\mathcal{X}^2_{8,0.05} = 15.51$
Critical region: $\mathcal{X}^2_{cal} > \mathcal{X}^2_{tab}$
Decision rule:H_0 is not rejected
Decision: H_0 is not rejected since P-value is greater than 0.05
Conclusion: there is no significant difference between observed and the model predicted values at α=0.05 level of significance which implies that the model fits the data and can be used to predict the ESR status of a patient.
Moreover from Table 3.8, it shows that the model is accurate enough to be used for prediction since it is of 87.5% accuracy.

Table 3.9: Variables in the equation 3.2

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	X_f	2.036	0.981	4.308	1	0.038	7.662
	X_g	0.145	0.115	1.589	1	0.207	1.156
	Constant	-12.506	5.648	4.903	1	0.027	0.000

Variables entered on step 1: Fibrinogen (X_f) and γ-Globulins (X_g).

Table 3.9 shows that$\beta_0 = -12.506, \beta_1 = 2.036, \beta_2 = 0.145$

Thus the binary logistic regression model is:

$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -12.506 + 2.036X_f + 0.145X_g$ ..... (3.2)

Thus, this suggests that model selection procedure actually select the variable that contributes to the outcome of the ESR test of a patient.

3.2 Model Selection

Since this is an exploratory analysis, where the analysis begins with a full and saturated model, backward selection procedure was used in this study to actually select the best variable that contributes to the outcome of the ESR.

3.2.1 Backward Stepwise (Likelihood Ratio)

Table 3.10: Omnibus tests of model coefficients

		Chi-square	Df	Sig.
Step 1	Step	8.621	2	.013
	Block	8.621	2	.013
	Model	8.621	2	.013
Step 2	Step	-1.750	1	.186
	Block	6.871	1	.009
	Model	6.871	1	.009

A negative Chi-squares value indicates that the Chi squares value has decreased from the previous step.

Table 3.11: Model summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	25.000	0.236	0.363
2	26.750	0.193	0.297

At a step 1, -2log likelihood statistic is 25.000, at step 2,-2loglikelihood statistic has increased to 26.750. Cox & Snell R square gives the proportion of variation in outcome of patient's ESR as 23.6% and Nagelkerke R Square by 36.3% at step 1. For step 2, Cox & Snell R square gives 19.3% and Nagelkerke R Square reduces to 29.7%.

Table 3.12: Variables in the equation 3.3

		B	S.E.	Wald	Df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	X_f	2.036	.981	4.308	1	0.038	7.662	1.120	52.415
	X_g	0.145	.115	1.589	1	0.207	1.156	0.923	1.449
	Constant	-12.506	5.648	4.903	1	0.027	0.000		
Step 2[a]	X_f	1.950	.917	4.523	1	0.033	7.028	1.165	42.390
	Constant	-6.964	2.777	6.288	1	0.012	0.001		

Variable(S) entered on step 1: Fibrinogen (X_f) and γ-Globulins (X_g).

Variable entered on step 2: Fibrinogen (X_f)

Thus the binary logistic regression equation of the model becomes:

$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -6.964 + 1.950X_f$ ..... (3.3)

Constant = -6.964, this is the expected value of the log odds of the outcome when all independent variables are held constant.

Table 3.13: Model if term odds ratio removed

Variable		Model Log Likelihood	Change in -2 Log Likelihood	Df	Sig. of the Change
Step 1	X_f	-15.896	6.793	1	0.009
	X_g	-13.375	1.750	1	0.186
Step 2	X_f	-16.810	6.871	1	0.009

Table 3.14: Variables not in the equation 3.3

			Score	df	Sig.
Step 2	Variables	X_g	1.727	1	0.189
	Overall Statistics		1.727	1	0.189

Variable(s) removed on step 2: X_g

3.3 The Akaike Information Criteria (AIC)

$\mathrm{AIC} = 2\mathrm{k} - 2\ln(\mathrm{L})$

The three formulated models, their summary and AIC are calculated as follows:

Table 3.15: Model developed using Fibrinogen (X_f) as the only predictor

Model Summary			
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	26.750[a]	.193	.297

Table 3.16: Model developed using γ-Globulins (X_g) as the only predictor

Model Summary			
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	31.793[a]	.056	.085

Table 3.17: Model developed using Fibrinogen (X_f) and γ-Globulins (X_g) as predictors

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	25.000	0.236	0.363

Hence the best model created is when Fibrinogen (X_f) and γ-Globulins (X_g) are used as predictors as shown in Table 3.17 since the model has the lowest -2loglikelihood value, also its shows that the predictor that predicts best is Fibrinogen (X_f) as also shown by the backward stepwise (likelihood ratio) in Table 3:10.

3.4 Hypothesis Tested

Likelihood ratio test using deviance to access the model fit.

H_0:Null model is a better fit than fitted model

H_1: Model is a better fit than null model

Test statistic $\chi_L^2 = Nulldeviance - residualdeviance$

$\chi_L^2 \sim \chi_{s-p}^2$where$s-p$ is the difference in the number of parameters of the null model and the prediction model, where the null model is the one with constant alone.

at $\alpha = 0.05$

Decision rule:

Reject H_0 if $\chi_L^2 > \chi_{\propto,s-p}^2$

Table 3.18: Model fitting information

Model	-2 Log Likelihood	Chi-Square	Df	Sig.
Intercept Only	33.621			
Final	25.000	8.621	2	.013

Table 3.19:Goodness-of-fit values

	Chi-Square	Df	Sig.
Pearson	28.339	29	.500
Deviance	25.000	29	.678

From Table 3.19, the deviance for the fitted model is 25.000 and from Table 3.18 ,the calculated test statistic are as follows :

Test statistics $\chi_L^2 = 33.621 - 25.000 = 8.621$

$\chi_{tab}^2 = \chi_{0.05,2}^2 = 5.99$

Conlusion: we reject H_0 since $\chi_{cal}^2 > \chi_{tab}^2$, i.e. Model is a better fit than null model

4.0 Discussion of Results and Findings

A logistic regression model was fitted using a secondary data set from the Institute of Medical Research, Kuala Lumpur, Malaysia which contains 32 observations and 3 variables(Fibrinogen (X_f), γ-Globulins (X_g) and the health status). Three logistic regression models were constructed namely: the null model (model with only the constant), model with the two predictors and model with the best predictor.

Omnibus test, Hosmer and Lemeshow's test, AIC and the Likelihood ratio's test were carried out to assessed and validated the models.

From the analysis carried out, it was observed that only Fibrinogen contributed significantly to the outcome of the erythrocyte sedimentation rate (ESR) while γ-Globulins do not really contribute to the model. The best model was the model with the two predictors with 87.5% accuracy.

This study has shown that it is indeed possible to predict whether a patient is healthy or not based on the values of Fibrinogen (X_F) and γ-Globulins (X_g) of the individual. Also it was discovered that the Fibrinogen (X_f) is the best predictor to use in predicting whether a patient is healthy or not.

5.0 Acknowledgement

The authors would like to express their sincere gratitude to **Dr Olushoga Fashoranbaku**of the Department of Statistics,Federal University of Technology Akure**,** for stimulating lectures on statistics,constant guidance,helpful remark and useful sugggestion.

6.0 References

[1] Achilleas, K., Panagiota S., Euthimios M., and Anastasios S. (2008): Implementing Logistic Regression Analysis to Identify Incentives for Agricultural Cooperative Unions to adopt Quality Assurance Systems; International Conference on Applied Economics.

[2] Bender R and Grouven U (1997). Ordinal Logistic Regression in Medical Research. Journal of the Royal College of Physicians of London Vol. 31, No 5.

[3] Hosmer D. and Lemeshow S. (2009): Applied Logistic Regression (Second Edition); New York: John Wiley & Sons, Inc.

[4] Stephen and Lucia (2002). Logistic Regression and Artificial Neural Network Classification Models: A Methodology Review. Journal of Biomedical Informatics 35(5-6):352 - 359 .

[5] Pyke S.W and Sheridant P.M (1993). Logistic Regression Analysis of Graduate Student Retension. The Canadian Journal of Higher Education Vol. XXIII -2 .

[6] Anders Skrondal, Sophia Rabe-Hesketh (2004): Generalized latent variable modeling : multilevel, longitudinal, and structural equation Models; Boca Raton London New York Washington, D.C.

[7] Chhatwal J., Alagoz O., Lindstrom M.J., Kahn C.E., Shaffer K.A. and Burnside E.S. (2009): A logistic regression model based on the national mammography database format to aid breast cancer diagnosis; American Journal of Roentgenology.

[8] Adeogun O.A ,Ajana A.M ,Ayinla O.A ,Yarhere M.T , Adeogun M.O (2008). Application of Logit Model in Adoption Decision: A study of hybrid clarias in Lagos State Nigeria. American-Eurasian Journal of Agric & Environ. Sci.,4(4): 468 - 472 .

[9] Brian E .O, Handy R.J ,Cutter G.R (1980). A two compartment regression model applied to compliance in a hypertension treatment program. Journal of Chronic Diseases Vol. 33, issue 10 , Page 645 - 651.

[10] Andy Field (2007): Discovering Statistics Using SPSS 3rd edition; Boca Raton London New York Washington, D.C.

[11] Julie Pallant (2011): spss-survival-manual-4th-edition; Allen &Unwin, 83 Alexander Street, Crows Nest NSW 2065, Australia.

[12] Ganong W.F (2003). Review of Medical Physiology,Mc Graw Hill,New york.

[13] Collett D (1991). Modelling Binary Data . Springer US.