ESTIMATING DISCRIMINATORY PERFORMANCE OF A BINARY LOGISTIC REGRESSION MODEL FOR NUTRITION ANEMIA TEST EVALUATION

P. ASHOK KUMAR[1] M.MUTHUKUMAR[2]

[1]*Research Scholar, Department of Statistics, PSG College of Arts& Science, Coimbatore*
[2]*Assistant Professor, Department of Statistics, PSG College of Arts& Science, Coimbatore*

ABSTRACT

Binary Logistic model is a very important statistical tools often used in analyzing and predicting of biological study. The study proposes to assess the prejudice performance of a binary logistic regression model to rightly classify between the anemic and non-anemic. The prejudice performance of binary logistic regression model is deliberate using two approaches. The first approach is the use of fitted binary logistic regression model to acutely predict the subjects that are anemic and non-anemic among pregnancy women, with the help of the parameters sensitivity and specificity. The alternative approaches is based on Receiver Operating Characteristic (ROC) curve for the fitted binary logistic regression model and then discover the Area Under the curve (AUC) as a measure of discriminatory performance. The value of sensitivity is observed to be greater than the value of 1- specificity, which signifies suitable differentiation for the mentioned cut point. The area under the curve indicates that there is evidence of reasonable differentiation reported by the fitted model.

Key Words: Binary Logistic Regression model, Discriminatory Performance, Sensitivity, Specificity, Receiver Operating Characteristic (ROC) Curve, Area under the Curve (AUC).

1. INTRODUCTION

Many research problem calls for the analysis and prediction of a dichotomous outcome which is used when we want to prediction a categorical (yes/No, success or failure) based on a set of independent variables. In the logistic Regression model, the log of odds of the dependent variable is model as a linear Combination of independent variable. The prediction is useful to the people and organizations to decide their future plans and taking decision-making [1, 2]. The application of binary logistic regression model to predict the classification of subjects as cases and non-cases are extremely significant in health outcome [4]. To fit the binary logistic model using independent data is very important in assessing the appropriateness of a model for specific purpose [5].

The receiver operating characteristic (ROC) curve is one of the statistical tools to predict the correctness of likelihoods of an event. ROC curve provides a comprehensive way to analyze the accuracy of predictions. The procedures for evaluation of accuracy depend on the type of the predictor. The area under the ROC curve has a significant clarification for disease categorization from healthy subjects [6]. which is defined as a plot of test sensitivity as the y coordinate versus its 1-specificity or false positive rate (FPR) as the x coordinate, is an effective method of evaluating the quality or performance of diagnostic tests, and is widely used in biological study.[7]

The objective of the study is to evaluate binary logistic regression model for anemia with respect to the pregnant women. The prejudice performance of binary logistic regression model is measured using two approaches. The first approach for measuring the prejudice performance is the use of fitted binary logistic regression model to correctly predict the subjects that are Anemic (y=1) and non-Anemic (y=0) with the help of the parameters sensitivity and specificity. The alternative approach is based on the Receiver operating characteristic (ROC) curve for the fitted logistic regression model and then determining the area under the curve as a measure of discriminatory performance [8].

2. MATERIALS AND METHODS

The binary logistic model has to be fitted for the dataset containing the information about 160 rural pregnancy women at Udumalpet taluk. Two variables used in the logistic regression equation are dependent variables (Anemic) denoted as Y and independent variable (risk factor) denoted as X. Dichotomous variable is a special case of categorical variable with two outcomes only. Examples of dichotomous variables in Medical fields are nutrition anemia Y = Yes / No, X = Risk factors [Age, Family type and IFA tablet consumed]. In this study relied upon both primary data and secondary data. The rural pregnancy women's are respondents. The Primary data has collected by using interview method with the help of questioner from 160 respondents as a sample. Interviews regarding health, nutrition and socio-economic status, and measurements of weight and height of the women, were conducted. And the secondary data were obtained from various government reports such as ANC reports, private hospital report and lab report.

According to the report of World Health Organization criteria, the cut off level of the hemoglobin concentration in blood for the diagnosis of anemia is less than 11 gm/dl or less in the first trimester, 10.5 gm/dl or less in the second trimester, and 11 gm/dl or less in the third trimester for pregnant women. [10, 11].

2.1 Binary Logistic Regression Model

The logistic model considers the following general epidemiologic study framework

The observed independent variables $X_1, X_2, \ldots\ldots X_K$ on a group of subjects, for determined disease status, as either '1' if "with anemic" or 0 if "without anemic."

The epidemiological information to describe the probability that the disease will develop during a defined study period, say T_0 to T_1, in a disease-free individual with independent variable $X_1, X_2, \ldots\ldots X_K$ values which are measured at T_0 .

The probability being modeled can be denoted by the conditional probability statement as

$$P(\text{D}=1/X_1, X_2, \ldots\ldots X_K) = \alpha + \beta_1 X_1 + \beta_2 + \cdots \ldots + \beta_K) = \text{z}$$

The model is defined as logistic if the expression for the probability of developing the disease as follows

$$\text{f(z)} = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{-(\alpha+\Sigma\beta_i X_i)}}$$

The terms α and β_i in this model represent unknown parameters that to be estimate based on data obtained on the X's and on D (disease outcome) for a group of subjects. The above formula to plug in these values and obtain the probability that this individual would develop the disease over some defined follow-up time interval. [12, 13]

2.2 Measuring Discrimination Performance

The discriminatory performance of fitted binary logistic regression model is assessed, by studying the agreement between predicted outcome and the observed outcome, with the help of 2×2 classification table. A subject is predicted to be a anemic or non-anemic based on whether the predicted probability is greater or less than a specified threshold probability value. The 2×2 general classification table is given below:

Table1. General classification table

Predicted Outcome	**Observed Outcome**		
		Y=1(Anemic)	Y=0 (Non-Anemic)
	Y=1(Anemic)	TP (A)	FP (B)	A +B
	Y=0 (Non-Anemic)	FN (C)	TN = (D)	C +D
		n_1 =A +C	n_0= B + D	A+B+C+D

Where, TP is true positives, FP is false positives, TN is true negatives, and FN is false negative [3, 8, 14]. The table can be used to calculate the two major parameters sensitivity and specificity and other indices and they are defined as:

$$Sensitivity\ (Se\ r\ TPR) = \frac{The\ number\ of\ anemic\ correctly\ predicted\ (TP)}{Total\ number\ of\ anemic\ in\ the\ sample\ (n_1)} = \frac{A}{A+C}$$

$$Specificity(Sp\ r\ TNF) = \frac{The\ number\ of\ Non-anemic\ correctly\ predicted\ (TN)}{Total\ number\ of\ Non-anemic\ in\ the\ sample\ (n_0)} = \frac{D}{B+D}$$

$$False\ Postive\ Fraction(FPF) = \frac{The\ number\ of\ fales\ anemic\ predicted\ (FP)}{Total\ number\ of\ anemic\ in\ the\ sample\ (n_1)} = \frac{B}{B+D}$$

$$False\ Negative\ Fraction(FNF) = \frac{The\ number\ of\ fales\ non-anemic\ predicted\ (FP)}{Total\ number\ of\ non-anemic\ in\ the\ sample\ (n_0)} = \frac{c}{A+C}$$

The accuracy of the fitted model is calculated as

$$Accuracy = \frac{A+D}{A+B+C+D}$$

2.3 Receiver operating characteristic (ROC) curve

Receiver operating characteristic curve is a probability curve, it is a performance measurement for categorization issues at different doorstep settings. ROC curve can be is used to determine between or to predict whether cases are more likely to be anemic or Non anemic. The suitable cutoff values for categorization by comparing the sensitivity and specificity of various cutoff values. The ROC curve is the entire set of possible true positive fraction (TPF) and false positive fraction (FPF) attainable by binary outcome variable y with different doorstep. The ROC curve is ROC (.) = {(FPF(c) , (TPF(c)), c $\in$(-∞,∞)}.When the doorstep c increases, both FPF(c) and TPF (c) decrease [12,13,15]

ROC curve is used to predict a binary outcome that tells how well the fitted model. . The area enclosed by ROC curve and horizontal axis, is known as the AUC (area under the curve), it is denoted as $AUC = \int_0^1 ROC(t)dt = P[(y=1) > P(y=0)]$. .If the AUC value is close to one or Se $\geq$ 1-Sp, the model is excellent to classify between the anemic and non anemic. If AUC is 0.5 or Se = 1- Sp, the model is worthless and not capable to distinguish between the Anemic and non-Anemic. Therefore

the positive discrimination is lies between 0.5 and 1.0 and negative discrimination of AUC lies between 0 and 0.5. Hence, the fitted model is expected to provide a proper discrimination if the true anemic have a greater predicted probability than the true non- anemic [15,16,17].

3. RESULTS AND DISCUSSION

The binary logistic model was fitted to the data to test the research hypothesis regarding the relation between different categorical variable with nutrition anemia among respected pregnancy women. Table 1 The output of the model depicts 61(38.13%) subjects were classified in anemic and 99(61.87%) subjects were classified as no- anemic of pregnancy women. The first approach for assessing the discriminatory performance is the use of fitted binary logistic model to correctly predict the subjects that are anemic (y=1) and no anemic (y=0) and then resolve the proportions of parameters sensitivity and specificity.

Table1. Classification Table

Predicted Outcome	**Observed Outcome**		
		Y=1(Anemic)	Y=0 (No Anenic)
	Y=1(Anemic)	TP = 41	FP = 12
	Y=0 (No Anemic)	FN = 20	TN = 87
		n_1 = 61	n_0= 99

The table 2 reports the output of logistic model with significant predictors. The model delivers good discriminatory performance, if the covariates age, type of family and Iron folic acid tablet consumers are included in the model.

Table2. Fitting logistic regression model

Parameter	Estimate (B)	Std. Err.	Wald Chi Sq.	Df	Sig.	Exp(B)	95% C.I. for Exp (B)	
							Lower	Upper
Age	0.627	0.204	9.436	1	0.002	1.873	1.255	2.795
Family type	-1.348	0.481	7.852	1	0.005	0.26	0.101	0.667
IFA	2.817	0.467	36.395	1	0.000	16.725	6.697	41.764
Constant	-4.186	0.946	19.568	1	0.000	0.015		

The fitted logistic regression model can be expressed as

$$\log\left[\frac{p_i}{1-p_i}\right] = -4.186 + 0.627 * Age - 1.345 * Family\ type + 2.817 * IFA\ tablet\ consumer$$

The fitted model helps to predict which subjects will be anemic and which will not be anemic. If the predicted probability of subjects is greater than or equal to 0.5, it can be predicted that the subject will be anemic and else, not a anemic. The classification table 1 shows the observed and predicted outcomes of the subjects at a cut point Cp= 0.5

Here, the numbers of true or observed anemic subjects that are predicted to be anemic are true positives (TP = 41) and the number of observed no anemic subjects that are predicted to be no anemic are true negatives (TN = 87) with 80.0% if the accuracy of the model. The sensitivity is the proportion of true positives among all the subjects that are anemic and specificity is the proportions of true negatives among all the no anemic subjects. If both sensitivity and specificity values are equal to 1, then the condition of perfect discrimination will be obtained. The sensitivity (Se) = TP/ n_1 = 41/61 = 0.672 or 67.2% and specificity (Sp) = TN/ n_0 = 87/91 = 0.956 or 95.6%. The false positive count is FP = 12. The expected value of 1- Specificity = FP/ n_0 = 12/ 99= 0.121 which is close to zero, and hence Se= 0.672 is expected to be greater than the value of 1-Sp, which signifies the suitable discrimination for this cut point.

The problem of measuring discrimination is that the two values, sensitivity and specificity may vary on the basis of cut-points point which is the drawback to measuring discrimination using the first approach. Hence, an alternative approach to assess the discriminatory performance is applied.

The alternative approach is based on receiver operating characteristic (ROC) curve for the fitted binary logistic model and then determining the area under the curve as a measure of discriminatory performance. The ROC curve for prediction probability is shown in the figure1. The ROC curve in the figure, when applied to a logistic model, is a graph of sensitivity and 1-specificity obtained from the range of cut points for the predicted value.

The value of 1-specificity gives the proportion of observed no anemic subjects that are falsely predicted to be anemic. The highest value of sum of sensitivity and specificity gives optimum cutoff value. The ROC curve begins at the origin (0,0), y-axis takes the value from 0 to 1, and x-axis takes the value from 0 to 1. The diagonal line joining (0,0) and (1,1) is the reference line. The figure 1 shows that the probability prediction is deliver a suitable sign of anemic than the waist to height ratio.

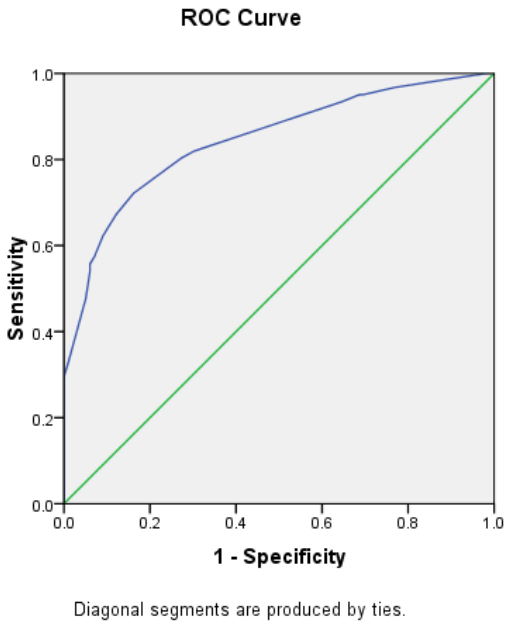

Figure 1. Receiver operating characteristic (ROC) curve

If the area under the curve (AUC) is larger, then the model discriminates better and hence the area under the curve is calculated and presented in the table 3.

Table 3. Area Under the Curve

Area under the curve	Std.Error	*p*-value	95% Confidence Interval	
			Lower	Upper
0.845	0.330	0.000	0.780	0.910

The area under the curve is 0.845 with95% confidence interval (0.780, 0.910). Also, the area under the curve is significantly different from 0.5since *p-value is 0.000* meaning that logistic regression classifies the group significantly better then by chance.

4.CONCLUSION

A binary logistic regression analysis was conducted to predict the persons sick with nutritional anemia. It has been used to find whether independent variables are important or not. It was concluded that the odds of securing choice of age, type of family and iron folic acid consumed pregnancy women and the receiver operating characteristic curve suggests a suitable process to estimating the discriminatory performance of a fitted logistic model. The value of sensitivity is noted to be greater than the value of 1-Specificity, which signifies is appropriate discrimination for the mentioned cut point and the area under the curve ≥0.9. The observed area under the curve for the variables age, type of family, IFA tablet consumed pregnancy women and prediction probability values indicates that there is evidence of reasonable discrimination subjects to anemic and non anemic group using the fitted logistic model.

REFERENCES

[1] J.E. Thornes, and D.B. Stephenson, "How to judge the quality and value of weather forecast products," *Meteorological Applications*, vol.8 (3), pp. 307–314, September 2001.

[2] M. S. Pepe, *The Statistical Evaluation of Medical Tests for Classification and Prediction,* Oxford University Press, New York, 2003.

[3] J. Pearce, and S. Ferrier, "Evaluating the predictive performance of habitat models developed using logistic regression," *Ecological Modeling*, vol.133, pp. 225-245, 2000, 10.1016/S0304-3800(00)00322-7.

[4] S. Noora, "Application of binary logistic regression model to assess the likelihood of overweight," *American Journal of Theoretical and Applied Statistics*, vol. 8, No. 1, pp. 18-25, 2019, doi: 10.11648/j.ajtas.20190801.13

[5] D.P. Allison, *Logistic Regression Using SAS®: Theory and Application, Second Edition*, SAS Institute Inc., Cary, NC,2012.

[6] K. Hajian-Tilaki, "Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation," *Caspian Journal of Internal Medicine,* vol. 4(2), pp. 627-35, 2013.

[7] V. Bewick, L. Cheek, and J. Ball, "Statistics review 13: receiver operating characteristic curves," *Critical care (London, England),* vol. 8 (6), pp. 508-12, 2004, doi:10.1186/cc3000.

[8] D.G. Kleinbaum, and M. Klein, *Logistic regression, statistics for biology and health*, Springer Science – Business Media, 2010, doi: 10.1007/978-1-4419-1742-3_10.

[9] Kadry S, Sleem C, Samad RA. Hemoglobin levels in pregnant women and its outcomes. *Biom Biostat Int J*. 2018;7(4):326-336. DOI: 10.15406/bbij.2018.07.00226.

[10] Elise M Laflamme. Maternal Hemoglobin Concentration and Pregnancy Outcome: A Study of the Effects of Elevation in El Alto, Bolivia. *Mcgill J Med*. 2011;13(1):47.

[11] S. Noora, "Neck circumference as an indicator of overweight and obesity in young adults," *American Journal of Applied Mathematics and Statistics*, vol.6(5), pp. 176-180, 2018, doi: 10.12691/ajams-6-5-1.

[12] D. W. Hosmer, and S. Lemeshow, *Applied logistic regression, Second edition*, John Wiley & Sons, New York, 2000.

[13] D. S. Young, *Handbook of regression methods*, CRC Press Taylor & Francis Group, Broken Sound Parkway, NW, Suite 300 Boca Raton, FL, 2017.

[14] H.J. Motulsky, and A. Christopoulos, *Fitting models to biological data using linear and nonlinear regression. A practical guide to curve fitting*, GraphPad Software Inc., San Diego CA, 2003.

[15] S.M.Pepe, *The statistical evaluation of medical tests for classification and prediction*, Oxford: Oxford University Press,2003.

[16] W. J. Krzanowski, and H. J. David , *Roc curves for continuous data*, CRC Press, Boca Raton, 2009.

[17] M. Gonen, *Analyzing receiver operating characteristic curves with SAS*. SAS Institute Inc., Cary, NC, USA, 2007.