International Journal of Research (IJR)

e-ISSN: 2348-6848
p-ISSN: 2348-795X
Vol. 9 Issue 10
October 2022

Divergence and Similarity of Logistic Regression and Linear Discriminant Analyses in Evaluating Factors Associated with the Prevalence of Bronchopneumonia.

CHRIS-CHINEDU, JOY NONSO', Ijomah, Maxwell Azubuike'' and **Onu, Obineke Henry'**
'Mathematics/Statistics, Ignatius Ajuru University of Education, Rumuolumeni, Port Harcourt, Rivers State, Nigeria.
"Mathematics/Statistics department, University of Port Harcourt, Rivers State, Nigeria.
Corresponding email:onuobinekehenry@gmail

Abstract

Logistic regression and Discriminant analyses are both used to forecast the probability of a certain categorical result based on numerous explanatory variables (predictors). The purpose of this study is to assess the convergence and comparability of the two models when applied to data from the health sciences. In this regard, we modeled the association of several factors with the prevalence of bronchopneumonia symptoms with both techniques and compared the result. It was observed that the logistic and discriminant analyses similarly produced the same result.

Keywords: Logistic regression, discriminant analysis, PCA, bronchopneumonia.

1. Introduction

The two multivariate statistical methods; Logistic regression and linear discriminant analyses can be used for the evaluation of the associations between various covariates and a categorical outcome. Both techniques have been widely used in studies, particularly in the fields of health and socio-logical science. According to Onu, et al. (2022), Logistic regression is used when the dependent variable is Dichotomous, that is to say, when the dependent variable is coded (categorical, say male or female) and the explanatory variables are of any kind. In Health sciences, the presence or absence of a stated disease or situation is usually the outcome. Applying the logit transformation, the probability of group membership in relative to several variables which is independent of their distribution is always predicted by logistic regression. This is calculated by dividing the probability of having the outcome by the probability of not having it.

Discriminant analysis, on the other hand, is a classification technique for determining which collection of factors discriminates between two or more naturally occurring groups and then classifying an observation into these known groups. Discriminant analysis accomplishes this by evaluating orthogonal discriminant functions. The linear combination of the standardized independent predictor variables yields the largest mean differences between the existing categories. However, both methods may be used to forecast the probability of a specific result using all or a subset of the information provided.

Many researchers have worked on logistic regression and discriminant function. Maja, et al. (2004) compared the two models using simulated data. The study used simulation to assess the performance of the models under comparison. Ijomah et al. (2018) compared logistic and poisson models using

Received: *19 Sept. 2022*
Revised: *8 Oct. 2022*
Final Accepted for publication: *12 Oct 2022*

e-ISSN: 2348-6848
p-ISSN: 2348-795X
Vol. 9 Issue 10
October 2022

count data of household utilized or not utilized primary health care services. Just recently, Onu et al. (2022) compared Binary Logistic and Poisson models on diabetic Patients in Nigeria using Dichotomous and Non-Dichotomous predictors. All the above quoted literatures failed to compare the divergence and similarities of Logistic and Linear discriminant analyses in evaluating factors associated with the prevalence of Bronchi pneumonia in Nigeria.

Theoretical properties have been extensively examined in the literature; nonetheless, the choosing of the appropriate data analysis technique remains an issue for researchers. After reviewing the attributes of the two discriminating methods, the main goal of this work is to investigate the convergence of the two analytical methods when they are used to analyze categorical health outcomes in pediatric epidemiological research. Using both statistical methodologies, we investigated the relationships between anthropometric and lifestyle factors and the occurrence of Bronchopneumonia in children under the age of five. As a result, the reader will be able to see the differences and similarities between the two models in order to make a better application choice.

2. Methodology

Linear Discriminant Analysis (LDA)

Linear Discriminant analysis focuses on the relationship between multiple independent variables and a categorical dependent variable by forming a composite of the independent variables. The extent by which any of the composite variables discriminates among two or more pre-existing groups of subjects can be determined by this type of multivariable model, and also can generate a classification model for prediction of the group membership of new observations. When the dependent variables have two groups, the discriminant analysis becomes so simple. In other words, a linear discriminant function that runs across the means of the two groups can be used to distinguish subjects between groups. The linear discriminant analysis explanatory variables are assumed to be normally distributed with equal covariance matrices in each of the categories. The evaluated coefficient for the independent variables in each of the cases, is the product of the cases' score on that variable. These multiplications are totalled and added to the constant, which resulted in a composite score, that is, the discriminant score for that case. The linear discriminant function (LDF) is represented by

$$LDF = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_k x_{ik} = bX \quad (1)$$

Where

b_j = the value of the j[th] coefficient, j=1,…,k; and

x_{ij} = the value of the i[th] case of the j[th] predictor.

The LDF can also be expressed in a recognized and permanent form that allows the examination resemblances or differences of variables moderated on different scales. In the standardized LDF, each variable is altered to accommodate to certain requisition by subtraction of its mean value and division by its standard deviation. The greater discriminating ability is shown by the coefficients with large

Received: *19 Sept. 2022*
Revised: *8 Oct. 2022*
Final Accepted for publication: *12 Oct 2022*

absolute values to their variable correspondence. For all the cases on the dependent variables from the LDF, the predicted probabilities and group membership can as well be estimated by the scores. This approach is being founded on the rationale that it is more likely that the independent and dependent variables are connected as the between-groups sum of square is bigger with respect to within-group sum of squares. Likewise, the ratio of between group divided by total sum of squares or of within-group divided by total sum of squares is used in evaluating the relationship. Seeing this, the ratio of between-group divided by within-group sum of squares is parallel to the ratio of variances, which is the F statistic, a test that curbs the possibility that the noticed relationship is due to chance.

The discriminant coefficients are chosen by the principle that they make the most of the distance among the two groups means (centroid) $|\overline{y}_1 - \overline{y}_2|$. Fisher was the first person that intimated to convert the multivariate observation x to univariate observations y in such a way that the $y's$ developed from groups 1 and 2 have the greatest distance among them. So, to this degree the linear combination $y = a'x$ is the one that makes the most of the ratio (squared distance among sample means)/(sample variance y). The vector of a number of a factor is given by the eigenvectors of the matrix $B * S^{-1}$, where B = $(\overline{X}_1 - \overline{X}_2)'$ is the between-group matrix and S is an estimate of Σ. A very crucial features of these composite sums of squares is that they surround the variability and the co-variability of each variable. The discriminant coefficients can be estimated in unstandardized or standardized form but they are unconnected of the form, less communicative than those in regression. Let's consider the mean of two groups, $\overline{X}_1, \overline{X}_2$, and the pooled covariance matrix, the rule of allocation based on Fisher's discriminant functions is as follows:

$$x_i \mathrm{E} \begin{cases} group.1, if\ .y = (\overline{X}_1 - \overline{X}_2)' S^{-1} X_i \geq 1/2 (\overline{X}_1 - \overline{X}_2)' S^{-1} (\overline{X}_1 + \overline{X}_2) \\ \\ group.2, if\ .y = (\overline{X}_1 - \overline{X}_2)' S' X_i < 1/2 (\overline{X}_1 - \overline{X}_2)' S^{-1} (\overline{X}_1 + \overline{X}_2) \end{cases} \quad (2)$$

Logistic Regression Analysis (LRA).

To anticipate the probabilities of the presence or absence of a particular epidemic, feature, or an outcome in general being founded on a set of independent of explanatory variables of any kind (discrete, continuous, or categorical), we consider the analysis. Since the anticipated probability must lie between 0 and 1, simple linear regression models are not enough to derive that, because they permit the dependent variable to exceed these limits and to produce unstable results. We defined p_1 as the probability of an object that belongs to group 1, and p_0, as the probability of an object that belongs to group 0. The logistic regression model has the form

Received: *19 Sept. 2022*
Revised: *8 Oct. 2022*
Final Accepted for publication: *12 Oct 2022*

$$Z_i = \log(\frac{p_{i1}}{p_{i0}}) \quad (3)$$

Where,

p_{i1}/p_{i0} = odds ratio,

The constant in the equation of a curve which can be varied (parameters) form (b_0 to b_k) of the logistic model are evaluated with the use of maximum likelihood method. Using the regression method, the probability of an event to happen was estimated as:

$$P(Y_i = 1 \setminus X_i) = \frac{\ell^{b^T X_i}}{1 + \ell^{b^T X_i}}$$

$$= \frac{1}{1 + \ell^{-b^T X_i}} \quad (4)$$

Where;

$\ell^{b^T X_i}$ = the linear predictor of the logistic regression function, and

Y_i = the study event (dependent variable).

When we use 0.5 as the probability cutoff, then an object can be classified to group 1 if the calculated $p_1 > 0.5$ and also to group 0 if $p_1 < 0.5$. In estimating the parameters of the logistic regression model, the coefficients of the log-likelihood function are being maximized using the method of maximum likelihood, a statistic which gives a summary of the knowledge acquired through study of the predictor variables.

Both multivariable techniques have the same functional structure; a complexity of the independent variables and a formula for classification.

Assumptions

Sequel to the discriminant analysis, the assumptions for ordinary regression are greatly similar in the following ways:

i. They are normally distributed

Received: *19 Sept. 2022*
Revised: *8 Oct. 2022*
Final Accepted for publication: *12 Oct 2022*

ii. They are homogeneous
iii. They are independently distributed in each case.

For reliable estimation of the discriminant function parameters, a sample size of at least 20 cases for each predictor variable and at least 20 cases for each of the dependent variable groups is required; in other words, estimating the coefficients is unstable and may result in misleading findings. The Categorical nature, dichotomous or more than dichotomous natures must be the features of the discriminant analysis. The dependent variable must be mutually exclusive and exhaustible to the population groups. The independent variable of the discriminant assumes to be continuous, and when the categorical variables are included, the reliability of discrimination decreases.

Limitations: Initially, logistic regression takes the assumption that there is an s-shaped dependency among the probabilities of group memberships and a linear function of the predictor variables. It also has effect on the assumption of independency among the observations.

Residual analysis may display some patterns that point out the presence of multi-colinearity or can discover outliers, which can deform the valid evaluation of the logistic coefficients. Furthermore, for logistic regression to give a trusted and trusty estimates, it needs a large number of cases. The more unequal groups are formed from the dependent variables; the more cases are needed. Also, for the variable prediction, logistic regression does not need multivariate normality or homoscedasticity, however, the predictive power increases if these conditions are fulfilled. In OLS regression, the outliers can significantly affect results. The researcher analyzed the standardized difference between the mean of several observations for outliers and as well consider deleting them or separately model them. Likewise, not like OLS regression, logistic regression uses maximum likelihood estimation (MLE) to generate parameters instead of ordinary least squares (OLS). The reliability of the estimates deteriorates when the cases are few for each observation of independent variables, that is, the MLE depends on large-sample asymptotic normality.

Evaluating the two techniques, sensitivity, specificity, and accuracy are the measures used in the same data.

Sensitivity: Sensitivity of a dual classification test as regards to some class is a measurement scale used to check the condition of a test and expresses the proportion of true positives of all the populations with positive cases.

Specificity: This is an expression of the proportion of the true negatives dual classification test, which is, used in testing the proportion of true negatives of all the populations with negative cases.

Accuracy: The degree of correctness of a calculated quantity to the exact value is referred to accuracy. It is estimated as the proportion of the true results of a dual classification test (true positive and true negative) among other possible outcomes.

Thus, both multivariable techniques can be applied to evaluate the same research problems. Their functional structure is the same though they vary in the method of their coefficients estimation. Discriminant analysis generates a score in synonymous to the generation of logit of the logistic regression. With the suitable mathematical calculations, both techniques produce the anticipated

Received: *19 Sept. 2022*
Revised: *8 Oct. 2022*
Final Accepted for publication: *12 Oct 2022*

probability of the classification of a case into a group of the dependent variable, and we can as well generate the categories of each predicted observation with the use of the suitable cutoff value. Discriminant analysis robustly makes the estimations since the normality assumption is fulfilled.

Contrarily, logistic regression invariably yields robust estimations as it did not assume the distribution of the explanatory variables with the dependent variable and equal variance among the group. Therefore, at the violation of the assumptions, we should not use the discriminant analysis rather we analyze the data with logistic regression, which gives a robust outcome since both continuous and categorical variables can be handled with the technique.

Application

Evaluating the prevalence of Bronchopneumonia using the epidemiological data.

In this study, we compared the outcomes of discriminant and logistic regression analyses in predicting the presence of any bronchopneumonia symptoms among fewer than 5 years' children in Rivers State University Teaching Hospital, Port Harcourt. In 2019, 700 under 5 children (313 males and 387 females), who were presented with the Bronchopneumonia symptoms.

We used the Principal Component Analysis(PCA) to enter the associated independent variables of "presence of any bronchopneumonia symptoms" at $\alpha = 0.05$ significance level. The Eighteen variables were drawn out from the PCA and fulfilled the above criterion. We withheld these eight factors which were the predictors and mutually independent after applying Kaiser's criterion (eigenvalue >1); Anthropometric characteristics, Athletic refreshment frequency consumption, Parental Body mass index, Shortness of breath, Birth-weight and breastfeeding, Cheese pie eating, and listening to music frequency.

The two models fulfilled their assumptions and variance covariance matrices of the groups were equal to Box's M test of equivalence P-value >0.05 for the discriminant analysis; predictors independency, lack of multi-colinearity after checking the residual, and for logistic regression; the large number of observations. For discriminant analysis, we used the standardized canonical discriminant function coefficients and the unstandardized function coefficients; for logistic regression, we used the Z-statistic (squared Wald statistic) to assess how much each variable contributes to discrimination in both models. The respective variables contributions to the discrimination rely on how big the coefficients are. Also, we compared the sign and extent of coefficients. The equality of the covariance matrices was tested using Box's M test, and it was shown that they were equal (P>0.05), thus this met the assumption of discriminant analysis.

The Response operating characteristics (ROC) curve for each of the model were plotted. An ROC curve displays graphically the sensitivity and minus 100% specificity (false positive rate) at distinct cutoff points. We can determine the test that is better for classification by plotting the ROC curves for two methods on the same axes, namely, that test whose curve encompasses the larger area under it. All analyses were carried out using the SPSS version 13.0 software.

3. Results

Received: *19 Sept. 2022*
Revised: *8 Oct. 2022*
Final Accepted for publication: *12 Oct 2022*

International Journal of Research (IJR)
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Vol. 9 Issue 10
October 2022

Applying the principal component Analysis (PCA) and Kaiser's criterion, Eight (8) patterns of our main data were determined, showing the anthropometric indexes of the children, consumption of breakfast, athletic refreshments consumption frequency, parental Body mass indexes, shortness of breath, weight at birth and breastfeeding, eating of cheese pies, and listening to music frequency. These variables were used in both models, and both techniques indicate that anthropometric characteristics, athletic refreshment consumption frequency, and eating cheese pies were the most important contributors as seen in Table 1.

Table 1: Predictors, Standardized and Unstandardized Coefficients for the Discriminant Analysis Model and Logistic Regression.

	Logistic Regression		Discriminate Analysis	
Predictors	b-coefficient	z-statistics	**Un-standardized coefficient**	**Standardized coefficient**
Anthropometric characteristics	0.5235	2.6705	0.3195	0.3135
Breakfast eating frequency	0.0005	0.0045	-0.0165	-0.0165
Athletic reshment frequency consumption	-0.6205	2.7785	-0.4645	-0.4545
Parental body mass index	0.2625	1.3915	0.0975	0.0975
Shortness of breath during acting	0.2315	1.1565	0.1425	0.1425
Birth weight and breastfeeding	-0.2945	1.3645	-0.1875	-0.1875
Chest pain	0.3495	0.6895	0.2205	0.2195
Listening to music frequency	-0.2995	1.3875	-0.1315	0.1215

More so, we observe that the direction of the relationships was the same and the difference were not much in the magnitude of the coefficients. The grand total correct classification rate was 77.5% for discriminate analysis while that of logistic regression was 78.7%.

Table 2 presents sensitivity, specificity, and accuracy of both methods at various cutoffs of the probability of having any bronchopneumonia symptoms.

Table 2: Sensivity of Logistic Regression and Discriminant Analysis models at various Cutoff Points for the Probability of having any Bronchopnuemonia Symptoms.

	Logistic Regression			**Discriminate analysis**		
Cut off value (%)	**Sensivity (%)**	**Specificity (%)**	**Accuracy**	**Sensivity (%)**	**Specificity (%)**	**Accuracy (%)**
0.05	94.35	7.79	29.01	99.49	0.29	24.59
0.1	91.75	22.79	39.69	99.49	1.19	23.29
0.25	68.45	68.69	68.69	91.75	18.69	36.59
0.5	27.65	95.29	78.69	71.29	69.49	69.89
0.75	4.55	99.49	76.29	25.09	94.49	77.49
0.9	0	99.49	74.99	4.59	99.49	76.29

Received: *19 Sept. 2022*
Revised: *8 Oct. 2022*
Final Accepted for publication: *12 Oct 2022*

e-ISSN: 2348-6848
p-ISSN: 2348-795X
Vol. 9 Issue 10
October 2022

NOTE: P (Bronchopnuemonia Symptoms): Values less than or equal to the cutoff value indicates that the child is not having Bronchopnuemonia Symptoms; those greater than the cutoff value indicates that a child is having one of the Bronchopnuemonia Symptoms.

Despite some variations which were observed between the models, as can be seen in Figure 1 below, the ROC curves of the previously mentioned models clearly shows that the logistic model is similar to the discriminant analysis model (i.e., no difference in the area under the curve (AUC), 75% versus 76.3%, P = 0.9).

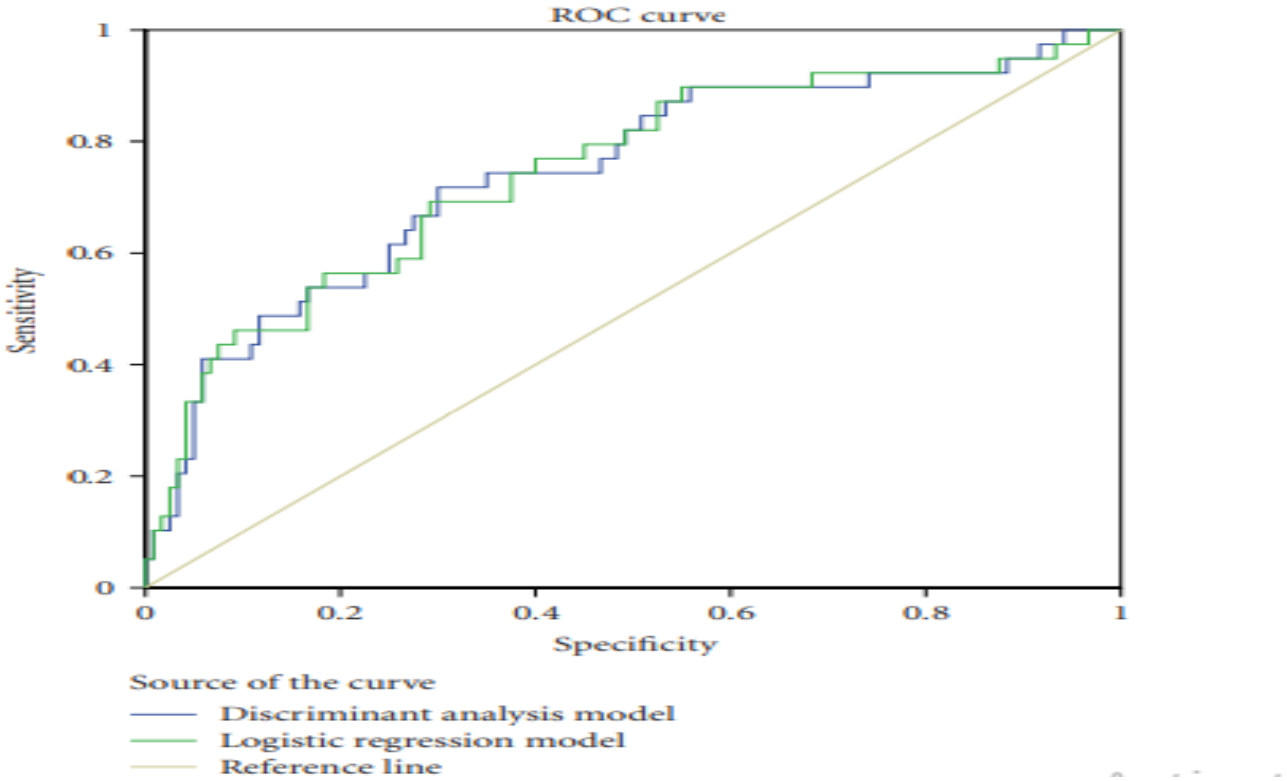

Figure 1: Receiver operating characteristics (ROC) curves for the discriminant analysis and logistic regression models.

4. Discussion

Generally, both logistic regression and discriminant analyses converged in similar results. They both evaluated the same statistical significant coefficients, with their effects similar in size and direction, despite the lager coefficients estimates of logistic regression. Their categorization rate was generally good, and either could be useful in predicting the likelihood of a kid developing bronchopneumonia symptoms in the general population. In terms of correct classification rate, logistic regression nearly outperforms discriminant function, although the variations in AUC were minor, indicating that there was no distinguishing difference between the models.

Received: *19 Sept. 2022*
Revised: *8 Oct. 2022*
Final Accepted for publication: *12 Oct 2022*

e-ISSN: 2348-6848
p-ISSN: 2348-795X
Vol. 9 Issue 10
October 2022

The number of the anticipated discriminant functions equals with the number of the categorical variables minus 1. They all have different sets of coefficients and generates scores for each of the cases, with different classification ability. Then for a four level categorical independent variable entering discriminant analysis, we derive three discriminant functions with their respective scores, and one or two only have the necessary power to gain the optimum classification rates. The risen question here is about the number of functions that is needed to retain from the available functional set.

Brenn and Arnesen (1985); compared the ability of discriminant analysis, logistic regression, and Cox model when applied in a dataset of 6595 men aged 20– 49, who were followed for 9 years for total and coronary deaths, in order to select possible risk factors. The population was split into two groups, one with a morbidity rate of 5 per 1000 and the other with a rate of 93 per 1000. The set of factors generated by logistic regression and the Cox model were identical, whereas the set of variables derived by discriminant analysis only differed slightly. The researchers also discovered that a time-saving option offered for both the logistic and Cox selections had no advantage over discriminant analysis, because the logistic and Cox methods consumed 80 and 10 times more computer time than discriminant analysis when analysing over 3800 subjects, respectively. As a result, the researchers concluded that discriminant analysis is preferable for preliminary or stepwise analysis, while the Cox technique should be employed otherwise.

When the linear discriminant assumptions for normality of the distribution of explanatory variables are met, when they are violated, and when they are categorised for various parameters of the explanatory variables such as sample size, covariance matrix, Mahalanobis distance, and the direction of the distance between the group means, Pohar *et al.* (2004) used several simulated datasets and discrimination indexes, the convergence of the two methods is examined. For regular distributed explanatory variables, the authors determined that linear discriminant analysis is the better method. Linear discriminant analysis is still preferred for classified predictor variables, and logistic regression only outperforms discriminant analysis when the number of categories is limited (2 or 3). When the assumptions of linear discriminant analysis are not met, it is not reasonable to use it, but logistic regression produces satisfactory results regardless of the predictor distribution.

Montgomery *et al.* (1987) compared the two methods in veterinary data, using stepwise linear discriminant analysis and logistic regression in a first dataset and comparing the selected variables, the order of selection, and the sign and magnitude of the estimated coefficients of the discriminating models in a second dataset, and found that, while both methods converged, logistic regression is preferable to discriminant analysis, particularly when the assumptions of normality and equal variance are not met.

We compared the two models using a real dataset rather than simulation approaches since the amount of observations in the dataset, while not large, was adequate to produce trustworthy results. When the normality requirements are met, the linear discriminant function is a superior approach than logistic regression; however, when the sample size is high enough (>50 observations), the differences between them become insignificant.

5. Conclusion

Received: *19 Sept. 2022*
Revised: *8 Oct. 2022*
Final Accepted for publication: *12 Oct 2022*

Conclusively, logistic regression and discriminant analyses were similar in the model analysis. To determine which method should be used, the assumptions for the each of the applications must be taken into consideration.

References

Antonogeorgos, *et al.*, (2007): "*Factors associated with asthma symptoms in schoolchildren from Greece: the Physical Activity, Nutrition and Allergies in Children Examined in Athens (PANACEA) study,"* Journal of Asthma, 44 (7), 521–52.

Brenn T. and Arnesen .E. (1985): *"Selecting risk factors: a comparison of discriminant analysis, logistic regression and Cox's regression model using data from the Tromsø heart study,"* Statistics in Medicine, 4 (4), 413–423.

Hair J. F. et. al. (1998): *Multivariate Data Analysis with Readings, Prentice-Hall, Englewood Cliffs, NJ, USA, 5th edition.*

Hosmer D. W. and Lemeshow S., Applied Logistic Regression, *John Wiley & Sons, New York, NY, USA, 1989.*

Ijomah, M. A., Biu, E. O. & Mgbeahurike, C. (2018). Assessing Logistic and Poisson regression model in analyzing count data. International Journal of Applied Science and Mathematical Theory, 4(1); 42-68.

Maja, P., Matya, B. & Sandra, T.(2004). Comparison of Logistic regression and Linear Discriminant Analyses: A simulation study . Metodoloski zvezki, 1(1); 143-161.

Montgomery M. E., White M. E., and Martin S. W., "A comparison of discriminant analysis and logistic regression for the prediction of coliform mastitis in dairy cows," *Canadian Journal of Veterinary Research, vol. 51, no. 4, pp. 495–498, 1987.*

Onu, O. H., Amakuro, O. V. & Alabge, S. A. (2022).Study of Binary Logistic and Poisson regression models of Diabetic Patients in Nigeria using Dichotomous and Non-Dichotomous predictors. Asian Journal of Probability and Statistics. 17(3); 37-48

Pampel, F. C. (2000): *Logistic Regression: A Primer*, Sage, Thousand Oaks, Calif, USA.

Pohar, M. et. al. (2004): *"Comparison of logistic regression and linear discriminant analysis: a simulation study*," MetodoloˇskiZvezki, 1(1), 143–161.

Stevens, J. P. (2002): *Applied Multivariate Statistics for the Social Sciences, Lawrence Erlbaum, Hillsdale, NJ, USA, 4th edition.*

Tabachnink B. G. & Fidell, L. S. (1996): *Using Multivariate Statistics, HarperCollins, New York, NY, USA.*

***Received**: 19 Sept. 2022*
***Revised**: 8 Oct. 2022*
***Final Accepted for publication**: 12 Oct 2022*