International Journal of Multidisciplinary Research and Development
Online ISSN: 2349-4182, Print ISSN: 2349-5979; Impact Factor: RJIF 5.72
Received: 19-08-2019; Accepted: 21-09-2019
www.allsubjectjournal.com
Volume 6; Issue 11; November 2019; Page No. 122-126

The robustness of binary logistic regression and linear discriminant analysis for the classification and differentiation of BTV cases in goats

Azza B Musa[1], Amal Alsir Alkhidir Abedalraheem[2], H Hamad[3], Siddik Mohamed Ahmed Shaheen[4]
[1] Central laboratory, Ministry of higher education and scientific research, Department of biostatistics, Khartoum, Sudan
[2] Sudan University of science and technology, faculty of science, department of applied statistics, Khartoum, Sudan
[3] Central veterinary research laboratory, Khartoum, Sudan
[4] University of Khartoum, faculty of economics and social studies, Khartoum, Sudan

Abstract
Binary logistic regression (BLR) and linear discriminant analysis (LDA) are both applied in order to predict the probability of a specific categorical outcome based upon several explanatory variables (predictors).A random sample of 642 animals was selected from goats being represented by all predictors. The comparison between BLR and LDA was based on the significance of coefficients, sample size impact to percentage of correct classification rate, sensitivity, specificity and accuracy, and area under receiver operating characteristics (ROC) curve.
Results showed both methods have similar contributors for data classification. The percentages of correct classification for total sample size were 87.3% for both models, and overall sample size the same trend was detected in the percent of correct classification on both analyses as the sample sizes have been changed. The areas under ROC curve (AUCs) were 0.814 and 0.801 for BLR and LDA, respectively. However, BLR showed slight superiority for animals being correctly classified. In conclusion BLR and LDA can be used effectively for classification even with violation of normality assumption. The aim of this work is to evaluate the convergence of these two methods when they are applied in data from the epidemiological studies.

Keywords: discriminant, classification, percentage, LDA, AUCs

1. Introduction
Bluetongue virus (BTV) is an infectious disease transmitted by Culicoides biting midges, affecting mainly domestic and wild ruminants, one of the 22 species or serogroups in the genus Orbivirus in the Reoviridae family. BTV causes severe morbidity and mortality in sheep, while the infection is subclinical in some domestic and wild ruminants [10]. BTV is an arbovirus, and until recently, its transmission was thought to be only mediated in cattle and ruminant through the bite of infected midges. This sole transmission route has been challenged recently with the emergence of reports of direct contact transmission with some serotypes and vertical transmission from mother to fetus [11].
Choosing the exact statistical method for data fitting is a frequent question for researchers. Among the most paramount criteria for the differentiation between statistical methods are, the type of response variable as well as the purpose of the research design. If we have categorical and dichotomous dependent variable, both binary logistic regression (BLR) and linear discriminant analysis (LDA) were suggested as the two multivariate models that have been used for classification of cases into their original groups. To date, there has been an increasing interest in choosing between BLR and LDA for analysis of biological data. Although, the theory behind each method has been extensively published, the comparison between the two methods still represents a problem for researchers who aimed to distinguish between two or more categorical outcomes in practice. Summarizing the findings of previous studies, none of these methods was perfectly superior over the other in term of data classification. Different criteria have been used for evaluating the performance of BLR and LDA in previous investigations. Hair *et al*, [6] revealed that BLR was better than LDA for analyzing categorical binary outcomes, particularly, if the predictor variables were continuous. Moreover, they concluded that the preference of BLR was attributed to its flexibility regarding the assumptions concerning independent variables.

2. Materials and Methods
This study was planned to evaluate the performance of binary logistic regression (BLR) and linear discriminant analysis (LDA) for differentiation between cases of goats having bluetongue virus and others not having a virus on the basis of different predictors. Considering the assumptions behind each method, BLR and LDA were compared according to sample size impact, significance of coefficients, and for each model we plotted the corresponding receiver operating characteristics (ROC) curve. An ROC curve graphically displays sensitivity and 100% minus specificity (false positive rate) at several cutoff points. By plotting the ROC curves for two models on the same axes, one is able to determine which test is better for classification, namely, that test whose curve encloses the larger area beneath it.
There are 642 goats were selected with two types of breeds, from gedarif state in Sudan, the independent variables were Age (continuous variable), Breed, Sex, Locality and Climate (categorical variables).
Groups of different sizes (n = 100, 250, 400, 550, 642) were randomly selected from the total size (n = 642). Data analysis was conducted by BLR and LDA models using

SPSS software (Statistical Package for the Social Sciences, Chicago, Illinois) version 25, and NCSS Data analysis software 2019.

Binary logistic regression (BLR) was used to study the association between a dichotomous dependent variable and a given set of one or more explanatory variables. Unlike, ordinary regression analysis, BLR can predict the binary categorical outcome, denoting a probability of success or failure. Hence, the predicted probabilities are ranged from 0 to 1. This feature makes BLR another suitable method for classification of cases into one of the two groups. To derive the BLR model, let p is the probability of success (case classified into group 1), and (1 -p) as the probability of failure (case classified into group 0). Therefore, the BLR model will be:

$$\text{Logit}(P) = \text{Ln}\left(\frac{p}{1-p}\right) = \beta_o + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} \quad (1)$$

The term p / (1 -p) is the odds ratio [1, 5], β_j is the value of the jth coefficient, j = 1, 2, 3…, k and x_{ij} is the value of the i^{th} case of the j^{th} independent variable. The parameters of BLR are β_o, β_1 …, β_k. By taking the exponential function for the previous equation, the probability of occurrence of a condition can be estimated using the following logistic regression model:

$$P(Y_i = 1 \mid X_i) = \frac{\text{odds}}{1+\text{odds}} = \frac{e^{\beta^T X_i}}{1+(e^{\beta^T X_i})} = \frac{1}{1+e^{-\beta^T X_i}} \quad (2)$$

Where Y_i is the binary outcome; X_i is the independent variable, the base e is the exponential function, and $e^{\beta^T X_i}$ is the odds ratio for the independent variable X_i. The choice between LDA and BLR is to greater extent depends on the assumptions beyond each method.

Linear discriminant analysis (LDA) is a statistical method used to examine the association between a categorical outcome and multiple independent variables in the form of discriminant function. This multivariate technique can be used to find out which explanatory variable best discriminate between two or more groups along with classification of cases into their proper group [12, 4]. The number of canonical discriminant functions is mainly determined by the number of categories minus one, or the number of discriminators variables, which is smaller. If we have only two groups or categories, then one discriminant function will be derived, giving the simplest form of LDA. The linear discriminant equation (LDE) is given as follows:

$$LDE = \beta_o + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} \ldots\ldots\ldots (3)$$

Where β_j is the observation of the j^{th} coefficient or weight, j = 1, 2…, k, x_{ij} is the observation of the i^{th} animal, for the j^{th} independent variable. Based on the estimates of the coefficients of LDA, we can identify carefully which explanatory variables would be able to discriminate between the groups of interest. The previous form of LDA is the unstandardized one, in which the equation included the constant term. The standardization process can occur by the same way of z scores. In practice, the coefficients with high magnitude reflect the importance of the corresponding variable in explaining the outcome. Furthermore, this function produces what is called discriminant scores, from which the predicted probabilities will be estimated for each case of the categorical outcome variable. These discriminant scores along with the group means (centroids) contribute in the classification of cases into their groups [13].

The choice between LDA and BLR is to greater extent depends on the assumptions beyond each method. Theoretically, BLR is more flexible regarding the assumptions, particularly those of independent variables. However, both methods require some assumptions in common [9] such as, independency of observations, absence of multicolinearity between predictors, and absence of outliers in datasets.

3. Results

The assumptions required by linear discriminant analysis were carried out, Box's M statistic which has been used to test the homogeneity of covariance matrices revealed the violation of that assumption (Box's M = 137.584, F = 8.991, and P < 0 .05), the dataset denoted non-normal distribution for all variables. The results obtained from the preliminary analysis showed no signs of collinearity between the explanatory variables. The highest correlation (0.231) was observed between locality and climate.

Table1 summarizes the results of descriptive statistics of data, it showed higher frequency for positive result 548 (85.4%), and 94 (14.6%) for negative result for blue tongue virus in gedarif state.

Table2 showed that standardized canonical discriminant function coefficients and the unstandardized function coefficients for discriminant analysis, B coefficient and *Z* statistic (squared Wald statistic) for logistic regression [9], the findings revealed climate has higher coefficients (1.101 and 0.876) for both models, followed by breed (0.846 and 0.790). And age was less successful as predictors (0.049 and 0.029) for both (BLR and LDA) respectively.

From the data in Table3, it can be seen that LDA used F-distribution and Wilkes' lambda statistic, while as BLR relied on chi-square distribution and Wald statistic for testing the contribution of explanatory variables in discrimination of animals of the two groups, in term of determining the best set of predictors which significantly differentiate between positive and negative, results of both BLR and LDA revealed that age and climate have significant (p < 0.05) contribution in data classification, using the total sample of this study (n = 642) but locality revealed significant contribution in LDA model only.

The effect of sample size on the classification abilities of LDA and BLR, five different random samples were chosen from the studied real dataset. The percentages of correct classification were recorded for the two analytical methods along with the variation in the sample sizes (100, 250, 400, 550 and 642). Referring to the findings in Table4, the percentage of correct classification increase as sample size have been increased, so the higher sample size (642) recorded higher percentage of correct classification (87.3%) for both models.

Table5 presents sensitivity, specificity, and accuracy of both approaches at various cutoffs of the probability of having disease. Although some differences are observed between the methods, as we can see in the table5, the aforementioned models clearly indicate that the binary logistic model is similar to the linear discriminant analysis model.

In this study, we have plotted the ROC curve for both BLR and LDA at two different sample size, the whole sample size (n = 642) and another smaller one (n = 100). As Table 6 shows, the area under ROC curve for BLR was 0.814 (n= 642, SE = 0.024, 95% C.I = 0.766- 0.862, P < 0.001), on the other hand the area under ROC curve for LDA was 0.801 (n= 642, SE = 0.025, 95% C.I = 0.752- 0.850, P < 0.001). When using samples of smaller size (n=100), it was apparent that the AUC for BLR was 0.766 (SE = 0.048, 95% C.I = 0.673 - 0.859, P < 0.001), and for LDA the AUC was 0.767 (SE = 0.047, 95% C.I = 0.674 - 0.859, P < 0.001). Furthermore, Figure 1 and 2 present the ROC curves for BLR and LDA using a real dataset, with small samples the AUC were slightly different between BLR and LDA but with the total examined sample (n = 642), the curves revealed that the differences in AUC for the two models were quite small and may be ignored.

Table 1: The Class Distribution of the BTV in Goat

Class name	Class size	Class distribution
Positive	548	85.4%
Negative	94	14.6%
Total	642	100%

Source: Prepared by researcher using SPSS version 25

Table 2: Variables and Coefficients for the Discriminant Analysis and the Binary Logistic Regression Models

Variables	Binary Logistic Regression		Linear Discriminant analysis	
	B Coefficients	Z Statistics	Canonical discriminant coefficients	Standardizes coefficients
Breed	0.846	3.049	0.790	0.170
Age	0.049	1052.418	0.029	0.575
Sex	-0.061	0.001	0.064	0.021
Locality	0.121	4.700	0.181	0.355
Climate	1.101	1130.170	0.876	0.747
Constant	-3.178	74.546	-4.799	-

Source: Prepared by researcher using SPSS version 25

Table 3: The role of predictors in explaining the outcome using linear discriminant analysis and binary logistic regression models

Predictors	Binary Logistic Regression		Linear Discriminant analysis		
	Wald statistic	P value	Wilks' lambda	F	P value
Breed	1.746	0.186	0.999	0.651	0.420
Age	32.441	0.000	0.958	28.154	0.000
Sex	0.026	0.872	0.999	0.336	0.562
Locality	2.168	0.141	0.968	21.209	0.000
Climate	33.618	0.000	0.895	74.801	0.000
Constant	8.634	0.003	-	-	-

Source: Prepared by researcher using SPSS version 25

Table 4: Percentages of correct classifications of animals conducted by linear discriminant analysis and logistic regression models, at different sample sizes

Variables	Percentage of correct classification	
	BLR	LDA
100	69.0%	69.0%
250	83.9%	79.9%
400	86.9%	84.4%
550	86.9%	86.9%
642	87.3%	87.3%

Source: Prepared by researcher using SPSS version 25

Table 5: Comparison of Logistic Regression and Linear Discriminant Analysis in terms of Sensitivity, Specificity and Classification Accuracy

Cut-off points	Binary Logistic regression			Linear Discriminant Analysis		
	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
0.1	100	0	85.3	100	0	85.3
0.2	100	0	85.3	100	0	85.3
0.3	99.6	6.4	85.9	99.6	8.5	86.3
0.4	99.3	16.0	87.0	98.9	20.2	87.3
0.5	97.8	26.6	87.3	97.8	26.6	87.3
0.6	97.6	30.9	87.8	95.6	39.4	87.3
0.7	89.2	51.1	83.6	91.4	50.0	85.3
0.8	82.2	58.5	78.8	82.2	58.5	78.8
0.9	60.8	87.2	64.7	65.4	76.6	67.0

Source: Prepared by researcher using NCSS analysis software 2019

Table 6: Area under the ROC curve (AUC), standard error (SE), 95% confidence interval (CI), and significance tests for logistic regression and discriminant analysis

Sample size	Model	AUC	SE	P-value	Asymptotic 95% Confidence Interval	
					Lower Bound	Upper Bound
Total sample size (n = 642)	BLR	0.814	0.024	0.000	0.766	0.862
	LDA	0.801	0.025	0.000	0.752	0.850
n = 100	BLR	0.766	0.048	0.000	0.673	0.859
	LDA	0.767	0.047	0.000	0.674	0.859

Source: Prepared by researcher using SPSS version 25

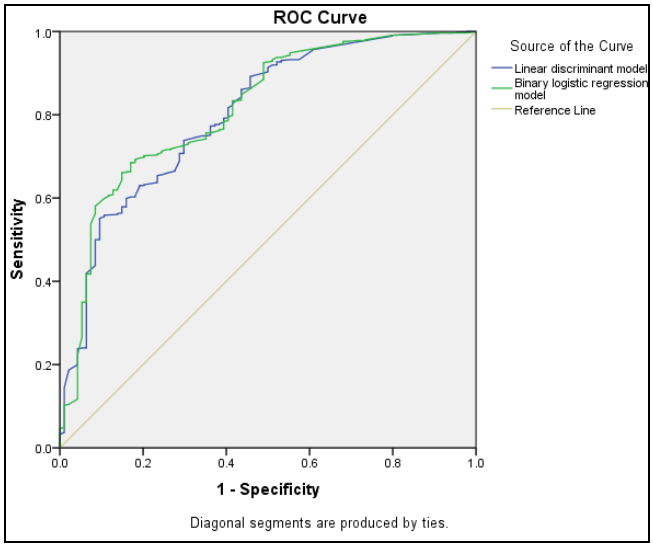

Fig 1: ROC curve for Linear discriminant analysis and Binary logistic regression (n = 642)

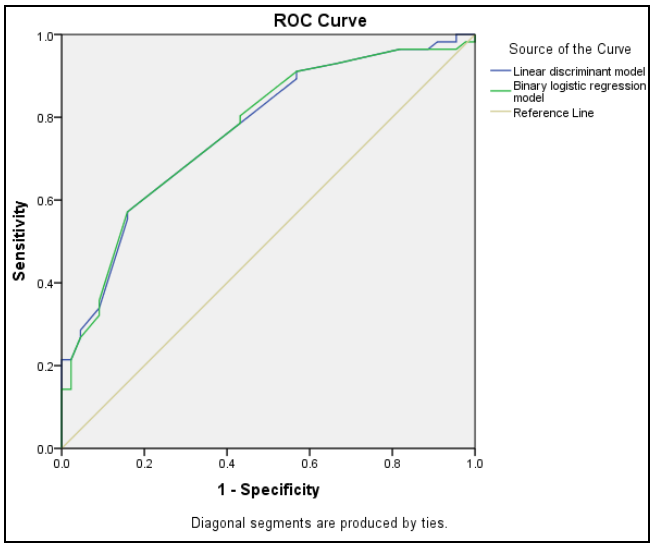

Fig 2: ROC curve for Linear discriminant analysis and Binary logistic regression (n = 100)

4. Discussion

In order to compare the two methods, we applied them in a real dataset, and we did not use simulation methods, as the number of the observations in the dataset, although not very large, but was sufficient to provide reliable results.

Both methods estimated similar effect size and direction of coefficients, with the same statistical significant effect of coefficients, except locality predictor, it showed significant result in Wilks' lambda of LDA but not in wald statistics for testing the overall performance of BLR. The overall classification rate for both was good and same trend was detected in the percent of correct classification for BLR and LDA as the sample sizes have been changed, and either can be helpful in predicting the possibility of animals having BTV symptoms in the general population.

Sensitivity, Specificity and classification accuracy showed similar values for different cutpoints except 0.9 cutpoint showed some differences between two models.

Logistic regression slightly exceeds discriminant function at AUC for total size (n = 642) but the differences in the AUC for size (n = 100) were negligibly, thus indicating no discriminating difference between the models.

In a study by Montgomery *et al.* [7], who 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 the magnitude of the estimated coefficients of the discriminating models in a second dataset, resulted that although both methods converged logistic regression is preferable to discriminant analysis particularly when the assumptions of normality and equal variance are not met. Moreover, George Antonogeorgos *et al.*, [3] concluded that the differences between BLR and LDA may be neglected if we have large sample sizes. They expected that small samples may lead to unstable and invalid estimates.

Also Marcos *et al.* [8], presented an automatic obstructive sleep apnea syndrome detection algorithm based on classification of nocturnal oxygen saturation using LR and LDA. They showed that the overall accuracy and AUC were similar.

Demler *et al.* [2] assumed multivariate normality and equal covariance matrices to estimate coefficients using LDA and LR that were identical. LDA and LR had the same true AUC, but the results of real data suggest that the finding is sensitive to the assumption of normality.

In general, in this study both binary logistic regression and linear discriminant analysis converged in similar results and the findings were in agreement with most other previous studies.

5. Conclusion

In summary, the first finding that can be drawn from this study was that both methods have selected the same predictors for significant differentiation between studied groups.

The second major outcome was that the sample size, the percentages of animals being correctly classified was similar in higher sample sizes (n = 550, 642), and regarding the area under the roc curve (AUC), BLR is robust when using large sample size (more than 100).

Taken together, these findings suggest that both LDA and BLR are helpful statistical methods in classifying animals. In conclusion, this investigation provides additional evidence that LDA is robust technique for violation of normality assumption. Besides, researchers can ignore the differences between the two methods, if they have used large samples.

6. Reference

1. Bernard Rosner. Fundamentals of Biostatistics. Seven edition. Harvard University, 2010.
2. Demler OV, Pencina MJ, D'Agostino RB. Equivalence of improvement in area under roc curve and linear discriminant analysis coefficient under assumption of normality. Statistic in Medicine. 2011; 30:1410-1418.
3. George Antonogeorgos, Demosthenes B. panagiotakos, Kostas N. Priftis and Anastasia Tzonou. Logistic Regression and Linear Discriminant Analyses in Evaluating Factors Associated with Asthma Prevalence among 10-12- Years-Old Children: Divergence and Similarity of the Two Statistical Methods. International Journal of Pediatrics, 2009.
4. Hamid H. A new approach for classifying large number of mixed Variables. International Journal of

Mathematical, Computational, Physical, Electrical and Computer Engineering. 2010; 4(10):1355-1360.
5. Hosmer DW, Lemeshow S. Applied Logistic Regression. 2nd Edition. John Wiley and Sons. Canada, 2000.
6. Jose M Rojas, *et al*. Diagnosing bluetongue virus in domestic ruminants: current perspectives. veterinary medicine research and report, 2019.
7. ME Montgomery, ME White, SW Martin. A comparison of discriminant analysis and logistic regression for the prediction of coliform mastitis in dairy cows. Canadian Journal of Veterinary Research. 1987; 51(4):495–498.
8. Marcos JV, Hornero R, lvarez D, Del Campo F, Aboy M. Automated detection of obstructive sleep apnoea syndrome from oxygen saturation recordings using linear discriminant analysis. Medical and Biological Engineering and Computing. 2010; 48:895-902.
9. Pampel FC. Logistic Regression: A Primer, Sage. Thousand Oaks. Calif. USA, 2000.
10. Park Hyeoun Ae. An Introduction to Logistic Regression: From Basic Concepts to Interpretation with Particular Attention to Nursing Domain. J Korean Acad Nurs. 2013; 43(2):154-164.
11. Sherif A. Moawed, Mohamed M. Osman. The Robustness of Binary Logistic Regression and Linear Discriminant Analysis for the Classification and Differentiation between Dairy Cows and Buffaloes. International Journal of Statistics and Applications. 2017; 7(6): 304-310.
12. Timm NH. Applied multivariate analysis. 2nd ed. Springer Texts in statistics, 2002.
13. Worth AP, Cronin MTD. The use of discriminant analysis, logistic regression and classification tree analysis in the development of classification models for human health effects. Journal of Molecular Structure. 2003; 622:97-111.