Mansoura Veterinary Medical Journal 23:3 (2022) 54-58

DOI: 10.21608/mvmj.2022.162138.1121

Original Article *Statistics*

Differentiation and classification of productive efficiency of chicken farms using logistic regression and linear discriminant analysis

Eman A. Abo Elfadl, Hend A. Radwan, Usama A. Abou-Ismail

Department of Husbandry & Development of Animal Wealth, Faculty of Veterinary Medicine, Mansoura University, Gomhoria St., Mansoura, P.O. box 35516, Mansoura, Egypt.

ARTICLE HISTORY

Received: September 23, 2022

Revised: September 27, 2022

Accepted: September 27, 2022

Corresponding author: Eman A. Abo Elfadl;
Email: emmy_f1984@yahoo.com; Tel.
01222570629 ; Fax. 002 (0)5023799 ;
Orcid. 0000-0001-9184-2761

ABSTRACT

Objective: This study was carried out to compare the accuracy of linear discriminant analysis (LDA) and binary logistic regression (BLR) in classification of level of production in layers (high and low) as a dependent variable using breed, total ration, number of mortality, marketing weight and marketing age as independent variables. Regarding the assumptions of each method, LDA and BLR were also compared with respect to the effect of sample size with consecration to lack of multivariate normality of predictors. Procedures: Record data of 12500 layers were collected from private farms in Dakahlia Governorate during the period from 2018 to 2020). The comparison between LDA and BLR based on the significance of coefficients, classification rate, and areas under ROC curve (AUC). Results: showed that both methods selected breed, total ration consumed and marketing age as significant predictors ($P < 0.01$) for classification process. The percentages of correct classification for LDA and BLR were 67.7% and 88.9%, respectively. The AUCs were 0.682 and 0.734, for LDA and BLR, respectively. In addition, the sample size effect had the same impact on both analyses, whereas the accuracy of correctly classified cases was higher in BLR than LDA. Conclusion: It could therefore be concluded that LDA and BLR can be used effectively for classification and prediction of level of production in layers even with the lack of normality assumption.

Keywords: Layer farms, Binary logistic regression, ROC Curve, Linear discriminant analysis.

1. INTRODUCTION

Choosing the suitable statistical model for data fitting is an important approach for all researchers. Among the most paramount criteria for differentiation between statistical methods are the purpose of the research design and the type of variables [1]. In case of categorical dependent variable, both logistic regression (LR) and linear discriminant analysis (LDA) are suggested as two multivariate models that are used for classification of cases into their original groups [2].

Till now, researchers have not shown a consensus in choosing between LR and LDA for analysis of biological data, although the theory behind the use of each method has been extensively studied in research. Therefore, the comparison between the two methods still, to some extent, problematic for researchers who aim to distinguish between two or more categorical outcomes in practice.

On one hand, linear discriminant analysis (LDA) has been shown as a good choice for classification compared to other predictive methods, such as logistic regression, multinomial logistic regression, random forests, support-vector machines, and the K-nearest neighbor algorithm [3]. On the other hand, LR has been recommended, in other instances, better than LDA for analyzing categorical data, particularly, if the predictor variables are continuous [4]. The preference for the use of LR over LDA has been referred to its flexibility regarding the assumptions concerning independent variables. However, a moderate stance between the two previous approaches was adopted by [5] who suggested that LR is similar to LDA when the assumptions of discriminant analysis have been met.

Several attempts have been made to address the advantages of LR and LDA and divergences between them, however, debate about how to choose between the two analytical methods still present. The majority of previous studies showed that when the assumptions of discriminant analysis were verified, particularly, the multivariate normality of explanatory variables and homogeneity of covariance matrices, LDA can perform better than LR. In contrast, other studies still recommend the use of LR for data classification because they fail to practically verify the assumptions of LDA.

What is not clear is the effect of sample size on the performance of the two methods particularly because most of research compared between LR and LDA has considered the assumptions of each method, type of predictors, presence or absence of multicollinearity, and the number of categories of dependent variable, while the effect of sample size remains insufficiently approached [2]. Moreover, most of those studies that examined the impact of sample size on both methods [6-8] were performed using simulation rather than real datasets.

Therefore, the aim of this study was to determine the robustness of LR and LDA for classification of productive efficiency of chicken farms with three breeds namely Fayomi, Lohman and Bovans, using a set of predictor variables (season, locality, farm size and mortality). Another aim of the study was to explore the

effect of sample size variation. The performance of each method was examined using non normal explanatory variables. The comparison between the two approaches depends on the coefficients of each model, the area under ROC curve (AUC), and the percentage of correct classification of animals.

2. MATERIAL AND METHODS

2.1. Data source

Records data of 12500 layer were collected from Dakahlia farms at period from 2018 to 2020. The comparison between LDA and BLR was based on the significance of coefficients, classification rate, and area under ROC curve (AUC).

2.2. Linear Discriminant Analysis (LDA)

Linear discriminant analysis (LDA) is a statistical method that can be 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 detect which predictor best discriminate between two or more groups.

LDA may be preferable to logistic regression and multinomial logistic regression for group classification. More specifically, LDA can be used for classification of three or more groups (unlike logistic regression) and does not require specification of a reference group (unlike multinomial logistic regression). LDA also has the advantage of use to estimate model parameters under conditions of separability [9].

The number of canonical discriminant functions is determined by the number of categories of the dependent variable minus one, so, in case only two groups or categories are present, then one discriminant function will be derived, giving the simplest form of LDA [2].

The linear discriminant equation (LDE) is given as follows:

LDE = βo + β1xi1 + β2xi2 + … + βkxik

Where βj is the observation of the jth coefficient or weight,

j = 1, 2…, k; xij is the observation of the ith animal, for the jth independent variable [6].

Estimating coefficient of LDA could help in identifying which explanatory variable would be best predictor to discriminate between the groups of interest.

The used form of LDA here is the unstandardized form, in which the equation included the constant term. The standardization form can occur by the same way of z scores. In practice, the coefficients with high magnitude indicate the importance of the corresponding variable in explaining the dependent variable.

2.3. Logistic Regression Analysis (LR)

Binary logistic regression (BLR) is used to study the association between a categorical dependent variable and a given set of one or more explanatory variables. BLR can predict the binary categorical outcome, denoting a probability of success or failure. Hence, the predicted probabilities are ranged from 0 to 1.

LR is more appropriate when researcher is interested in the underlying structure of the prediction e.g. what are the most important predictors? or what is the role that different variables play in the prediction, rather than in the specific prediction of which group people belong to which is the emphasis of LDA [1].

This feature makes BLR another suitable method for classification of cases into one of 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 LR model will be as follows:

Logit (P) = Ln (P1 − P) = βo + β1xi1 + β2xi2 + … + βkxik (2)

The term p / (1-p) is the odds ratio; βj is the value of the jth coefficient, j = 1, 2, 3…, k and xij is the value of the ith Case of the jth independent variable. The parameters of BLR are βo, β1 …, βk.

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 (Yi = 1 | Xi) =TiT Ti iβ Xβ X β Xodds e 11 odds 1 (e) 1 e−= = + + + (3)

Where Yi is the binary outcome; Xi is the independent Variable; the base e is the exponential function [10].

2.4. Data Analysis

Data were collected, organized, summarized and analysed using SPSS statistical program (Version 23.0 for windows). LDA and BLR models were used to check the significance of different determinants to classify level of egg production as dependent variables using breed, total ration consumed, mortality number, marketing weight and marketing age as independent variables. The level of production (dependent variable) was coded before the analysis as following (high producing= 1, low producing= 2). The classification power of linear logistics model and discriminant analysis was compared to determine best fit model and the important predictor for classification process.

3. RESULTS

Results of the preliminary analysis showed no signs of collinearity between the explanatory variables. For determining the best set of predictors which significantly differentiate between different levels of egg production, results of LDA and BLR revealed that breed, total ration consumed, and marketing age showed a significant ($p < 0.05$) contribution in data classification (Table 1), using the total sample of this study (n = 12500).

Table 1. The predictors for classification process using discriminant analysis and binary logistic regression.

Predictor	Linear discriminant analysis LDA)(			Binary logistic regression (BLR)	
	Wilks' lambda	F	p-value	Wald test	p-value
breed	0.962	7.68	0.006	6.07	0.014
Total ration consumed	0.991	1.77	0.185	22.78	<0.0001
Mortality	0.990	1.93	0.166	0.00	0.99
Marketing weight	1.000	0.029	0.864	0.002	0.97
Marketing age	0.974	5.21	0.024	6.88	0.009

Table 3. Testing the significant of linear discriminant analysis and binary logistic model

Linear discriminant analysis			Binary logistic regression	
Wilks' lambda	Chi square	p-value	Wald test	p-value
0.831	35.79	<0.0001	50.16	<0.0001

Table 4. The area under curve and standard error for linear discriminant analysis and binary logistic regression.

Linear discriminant analysis		Binary logistic regression	
Area under curve	Standard Error	Area under curve	Standard Error
0.682	0.0189	0.734	0.0197

Table 2. The percentages of overall correctly classified cases using discriminant analysis and binary logistic regression young different sample sizes

Sample size	Percent of correct classified cases	
	Linear discriminant analysis (LDA)	Binary logistic regression BLR)(
1000	84.2%	87.6%
1500	92.6%	94.4%
2000	90.9%	91.8%
2500	100%	100%
3000	100%	100%
3500	100%	100%
Total	67.7%	88.9%

Data presented in Table (2) indicates that LDA used F- distribution and Wilkes' lambda statistic, while BLR relied on the use of chi-square distribution and Wald statistic for testing the contribution of explanatory variables in discrimination of animals regarding its production level. Thus, both LDA and BLR showed no significant differences ($p > 0.05$) between the two breeds on the basis of breed and marketing age of the birds and additional predictor (total ration consumed) were recorded with LDA.

The percentages of correct classification were determined for the two statistical methods with different sample sizes (1000, 1500, 2000, 2500, 3000, 3500). Referring to the data in Table 1, results were recorded in the percent of correct classification for LDA and BLR with different sample sizes. The interesting noticeable finding is that the percent of correct classification of animals was higher when using higher sample sizes (3000 and 3500), for both LDA and BLR, compared to smaller sizes (1000 and 1500). Besides, the ability of BLR to correctly classify animals into their proper level of production was higher than LDA throughout all sample sizes.

Results also showed that with the increase of sample size, the differences in the discrimination and correct classification of cases for both methods became higher compared to those reported for smaller samples. Considering the total sample size (n =12500) of the present study, LDA was able to correctly classify animals by about 67.7%, while 88.9% of cases were classified correctly by BLR.

Results of the present study evaluated the overall fitting of the data made by LDA and BLR. The results of canonical discriminant function were highly significant (Wilks' lambda = 0.831, chi-square = 35.79, $P < 0.0001$). Also, the likelihood ratio test (LRT) for testing the overall performance of BLR was highly significant (chi-square = 50.16, $P < 0.0001$) (Table 3). Therefore, these results provide clear evidence that the two methods can perform well under non-normal data in modelling and predicting layers as high producer or low producer for egg production. Another method to compare LDA with BLR was the Receiver Operating Characteristic (ROC) curve. The area under ROC curve (AUC), 95% confidence interval (C.I.) for the area under ROC curve, and the significance test for AUC are presented for each model (Table 4).

In this study, we plotted the ROC curve for both LDA and BLR, at two different sample sizes, the whole sample (n = 12500). As Table 4 shows, the area under ROC curve for LDA was 0.682 (n = 12500, SE = 0.01890, whereas the area under ROC curve for BLR was 0.734 (n =12500, SE = 0.0197).

4. DISCUSSION

The present study was designed to evaluate the difference between LDA and BLR when dealing with non-normal data, with special consideration for the outcomes potentially attributed to sample size variation. Records of layer farms were used to compare between the two statistical methods. Data of 12500 layers records were collected from Dakahlia layers farms at the period from 2018 to 2020. The comparison between LDA and BLR was based on the significance of coefficients, classification rate, and area under ROC curve (AUC).

Classification of layers level of production (high versus low) was carried out on different sample sizes lacking for the multivariate normality of the independent variables. The results showed that both LDA and BLR selected the same variables to discriminate between the two breeds. Among the significant predictors, as denoted by LDA and BLR, total ration consumed was the most important predictor in differentiation between level of layer production (high and low), followed by marketing age, and the breed came as a last predictor. On the other hand, mortality number and marketing weight appeared non-significant discriminators for layer level of production as showed by the two models. It could therefore be concluded that substantial mean differences between level of layer production is a result of the effect of the independent variable of the model. This finding suggests that the least square estimators of LDA are consistent with the maximum likelihood estimators of BLR [2].

Both discriminant analysis and logistic regression can be used to predict the probability of a specified outcome using all or a subset of available variables [11]. This study aims to evaluate the convergence and choosing between two methods when they are applied in epidemiological data and set some guidelines for proper choice; this is the problem that motivated this research. The comparison between the methods is based on several measures of predictive accuracy using blue tongue virus data.

In case of lack of normality, there is a similarity between the findings of the current study and those reported by previous research work [2, 3, 6]. Previous work reported that LDA and BLR have the same accuracy in estimating the practical differences between groups. Regarding our hypothesis, the highest percentages of correct classifications of animals were observed for large sample sizes (3000-3500), using both LDA and BLR. However, in general, the present results showed that BLR was slightly superior and able to classify animals correctly than did LDA, particularly for smaller samples. The results of the current study also demonstrated that with the increase of sample size, the classification rate of the two methods became similar and the differences between the two models might become neglectable, and that the percentage of correct classification was higher than in small sizes group. Inconsistent findings about the performance of LDA and BLR with regard to sample size have been published. For example, [12] reported that LDA was better than BLR when analysing small size datasets.

The differences between LDA and BLR may be small when big sample sizes are considered, and small samples may lead to unstable estimates [7]. The present findings agree with those of [13] who reported that the percent of correctly classified cases was higher in LR than LDA. They also, indicated that the difference in sample size has the same effect on both models. The results of the current study also agree with those of [14] who used both LDA and BLR for differentiation of normal and diabetic patients. They demonstrated that the classification power was higher for BLR than for LDA. Moreover, [8] used real data to compare LDA and BLR on the basis of normality assumption, number of predictors, and sample effect and found that the use of BLR was associated with better results than LDA in classification process. They also showed that the two models perform equally with larger samples.

On contrast, [7] and [15] concluded that both LDA and BLR showed the same classification accuracy in the studies that were conducted on outcomes from health problems. When Veterinary data is considered, another study performed by [16] evaluated the two methods and recommended the use of BLR over LDA especially when the normality assumption and homogeneity of covariance matrices were not verified.

The results of ROC curve and the area under the curve (AUC) can also be considered as another evidence for evaluating the performance and quality of the LDA and BLR. Taking sample size into account, it has been recommended that the clinical conclusions from ROC curves can be regarded if the sample size was 100 and more [17].

The results of ROC curves of this study revealed that the AUC was larger for BLR than LDA. The significant statistics for testing the AUC for both methods indicate that all AUC were significantly different from half. It can be therefore concluded that both LDA and BLR were strongly able to differentiate among Fayomi, Lohman and Bovans, with regard to the non-normal explanatory variables. Moreover, the results of Wilks' lambda and LRT for testing the overall performance of LDA and BLR confirm the conclusion that both methods are robust when using nonnormal data. Comparing the results of two methods according to AUC, the present findings agree with those reported by previous studies [e.g. 13, 16, 18].

A recent study used real datasets to evaluate the differences between LDA and BLR in predicting diabetes by [11] and [19]. Their findings revealed that the AUC for LDA and BLR were similar.

CONCLUSION

Regarding the percentages of correctly classified cases and the finding of this study it can be concluded that accuracy of BLR in data classification and prediction is higher than that of LDA, and that both models selected nearly the same predictors for classification process, using non-normally distributed data. The sample size has the same impact on LDA and BLR, although, the area under the roc curve (AUC) showed that BLR might be slightly superiority than LDA, and classification accuracy of higher cut off points also showed small difference between two models. Therefore, in order to decide which method should be used, the assumptions for the application of each method should be considered.

Conflict of interest statement

The authors declare that there is no any conflict of interest in the current research work.

Animal ethics committee permission

The current research work was executed according to standards of Research Ethics Committee, Faculty of Veterinary Medicine, Mansoura University (30).

Authors' contributions

Eman A. Abo Elfadl conducted the experiment and statistical procedures; Hend A.Radwan wrote, revised and contributed to writing the manuscript; Usama A. Abou-Ismail revised and edited the manuscript.

5. REFERENCES

[1] Abdulhafedh A. Comparison between Common Statistical Modeling Techniques Used in Research, Including: Discriminant Analysis vs Logistic Regression, Ridge Regression vs LASSO, and Decision Tree vs Random Forest. Open Access Libr 2022; 9, 1-19. http://doi.org/10.4236/oalib.1108414

[2] Moawed SA, Osman MM. The Robustness of Binary Logistic Regression and Linear Discriminant Analysis for the Classification and Differentiation between Dairy Cows and Buffaloes. Int J Stat Appl 2017; 7 (6): 304-310. http://doi.org/10.5923/j.statistics.20170706.05

[3] Boedeker P, Kearns NT. Linear Discriminant Analysis for Prediction of Group Membership: A User-Friendly Primer. AMPPS 2019; 3 (2): 250-263. https://doi.org/10.1177/2515245919849378

[4] Hair JF, Black WC, Babin BJ, Anderson RE, Tatham RL. Multivariate Data Analysis, Upper Saddle River, N.J. Pearson Prentice Hal, 2006.

[5] Kolari J, Glennon D, Shin H, Caputo M. Predicting large US commercial bank failures. J Bus Econ 2002; 54 (4): 361-387.https://doi.org/10.1016/S0148195(02)00089-9

[6] Pohar M, Blas M, Turk S. Comparison of logistic regression and linear discriminant analysis: a simulation study. Metodoloski Zvezki 2004; 1 (1): 143–161. http://doi.org/10.51936/ayrt6204

[7] Antonogeorgos G, Panagiotakos DB, Priftis KN, Tzonou A. Logistic Regression and Linear Discriminant Analyses in Evaluating Factors Associated with Asthma Prevalence among 10- to 12-Years-Old Children: Divergence and Similarity of the Two Statistical Methods. Int J Pediatr 2009; 952042. http://doi.org/10.1155/2009/952042

[8] Liong CY, Foo SF. Comparison of Linear Discriminant Analysis and Logistic Regression for Data Classification. In: Proceedings of the 20th National. Symposium on Mathematical Sciences. Putrajaya, Malaysia: AIP Conf Proc; 2013, p. 1159-1165.

[9] Hastie T, Tibshirani R, Friedman J. Elements of statistical learning. New York, NY: Springer; 2009.

[10] 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. J Mol Struct 2003; 622: 97-111. https://doi.org/10.1016/S0166-1280(02)00622-X

[11] Musa AB, Abedalraheem AAA, Ibrahim MT, Hamad H, Shaheen SMA. Divergence and Similarity of the Binary Logistic Regression and Linear Discriminant Analysis Models in Evaluating Factors Associated with Bluetongue Virus in Cattle. Int J Stat Appl 2019; 9 (6): 180-185. https://doi.org/10.5923/j.statistics.20190906.02

[12] Wilson RL, Hargrave BC. Predicting graduate student success in an MBA program: Regression versus classification. Educ Psychol Meas 1995; 55: 186-195. https://doi.org/10.1177/0013164495055002003

[13] El-habil A, El-Jazzar MA. Comparative study between linear discriminant analysis and multinomial logistic regression. An-Najah University Journal for Research, (Humanities) 2014; 28 (6): 1525-1548. https://www.researchgate.net/publication/274393542

[14] Zandkarimi E, Safavi AA, Rezaei M, Rajabi G. Comparison between logistic regression and discriminant analysis in identifying the determinants of type 2 diabetes among prediabetes of Kermanshah rural areas. J Kermanshah Univ Med Sci 2013;17(5):e77059. https://doi.org/10.22110/JKUMS.V17I5.951

[15] Panagiotakos DB. A comparison between Logistic Regression and Linear Discriminant Analysis for the Prediction of Categorical Health Outcomes. IJSS 2006; 5: 73-84.

[16] Montgomery ME, White ME, Martin SW. A comparison of discriminant analysis and logistic regression for the prediction of Coliform mastitis in dairy cows. Can J Vet Res 1987; 51 (4): 495-498.

[17] Metz CE. Basic principles of ROC analysis. Seminars in Nuclear Medicine 1978; 8: 283-298.

[18] Abledu GK, Buckman A, Adade T, Kwofie S. Comparison of Logistic Regression and Linear Discriminant Analyses of the Determinants of Financial Sustainability of Rural Banks in Ghana. JTAS 2016; 5 (2): 49-57. https://doi.org/10.11648/j.ajtas.20160502.12

[19] Ahmadi MA, Bahrampour A. Comparison of logistic regression and discriminant analysis in predicting type 2 Diabetes. Iran J Epidemiology 2015; 11 (3): 62-69. URL: http://irje.tums.ac.ir/article-1-5442-en.html