

On the Application of Linear Discriminant Function to Evaluate Data on Diabetic Patients at the University of Port Harcourt Teaching Hospital, Rivers, Nigeria

Nicholas Pindar Dibal, Christopher Akas Abraham

Department of Mathematical Sciences, University of Maiduguri, Maiduguri, Nigeria

Email address:

pndibal@unimaid.edu.ng (N. P. Dibal), extokas47@gmail.com (C. A. Abraham)

To cite this article:

Nicholas Pindar Dibal, Christopher Akas Abraham. On the Application of Linear Discriminant Function to Evaluate Data on Diabetic Patients at the University of Port Harcourt Teaching Hospital, Rivers, Nigeria. *American Journal of Theoretical and Applied Statistics*. Vol. 9, No. 3, 2020, pp. 53-56. doi: 10.11648/j.ajtas.20200903.14

Received: April 16, 2020; **Accepted:** May 3, 2020; **Published:** May 18, 2020

Abstract: Many real life events involves several interacting variables, hence multivariate statistical tool is necessary for appropriate analysis and interpretation. Discriminant analysis (DA) is one of the commonly used multivariate method in various fields of study including education, finance, environment, medicine etc., where complex data analysis and interpretation is required. This paper demonstrates and illustrate approaches in presenting how the discriminant analysis can be carried out on 335 (40 diabetics and 295 non-diabetic) patients and how the output can be interpreted using the Fisher's linear Discriminant function (FLDF). The performance of FLDF was adjudged based on the percentage of correct reclassification of the original observation to yield the discriminant scores from the functions. Up to 65.4% correct classification was achieved, and similarly 62.7% percent of the cross-validated grouped cases were correctly classified into either being a Diabetic or non-diabetic patient. Patient's age and gender were found to be the two most important contributing variables in classifying a patient between the two groups.

Keywords: Discriminant Analysis, Classification, Diabetes, Fisher's LDF

1. Introduction

Multivariate analysis is a statistical technique for the analysis of two or more variables observed from one or more sample objects [5]. It consists of methods that are suitable for application when several measurements are taken on each unit in one or more samples. The quality and validity of information obtained about the population of study increases with the number of variables measured on each sampled unit, and adequately modeled to explain variations in the response variable. The discriminant function procedure introduced by Fisher [9] is one of the linear models used for the analysis of multivariate data; it gives the best linear function for discriminating the explanatory variables. Discriminant Analysis is a statistical tool widely used primarily for two objectives; finding a predictive equation for classifying new individuals or interpreting the predictive equation to better understand the

relationships that may exist among the variables. It can also be used in *explanatory* or *predictive* frameworks where in both cases, some group assignments must be known before carrying out the Discriminant Analysis. Discriminant function analysis is very similar to logistic regression, and both can be used to answer the same research questions. Logistic regression does not have as many assumptions and restrictions as discriminant analysis. However, when discriminant analysis' assumptions are met, it is more powerful than logistic regression. Discriminant analysis can also be used with small sample sizes. It has been shown that when sample sizes are equal, and homogeneity of variance/covariance holds, discriminant analysis is more accurate [2]. While discriminant analysis and multiple regression analysis are similar, the main difference between these two techniques is that regression analysis deals with a continuous dependent variable, while discriminant analysis must have a discrete dependent variable. The method attempts to express one dependent variable as a linear

combination of other features or measurements.

Discriminant analysis determines which variables discriminate between two or more naturally occurring distinct k groups ($k \geq 2$). It can be used to determine which predictor variables are related to the dependent variable and to predict the value of the dependent variable given certain values of the predictor variables [4]. Discriminant Analysis has found application in numerous fields of study, for example; in health related research, it may be of interest to determine which amongst; religious, demographic, cultural and socio-economic variables discriminate between patients deciding to go for surgery after diagnosis of certain ailment and those that decline. Here, relevant information on many variables are measured from each patient before diagnosis is made. Discriminant Analysis could then be used to determine which variable(s) are the best predictors of the patient's final decision. Other areas where discriminant analysis is commonly used are in; ecology, prediction of financial risk, education, health, market analysis and consumer choice preference, analysis of corporate performance, livestock farming, credit worthiness, urban residential mobility, demographic studies, prediction of degree classification etc., ([2, 8, 12, 13, 1]). The application of classical linear discriminant analysis in medicine for diagnosis as to whether or not an individual patient has the disease he/she presented with some symptoms is of great significance and usually requires time, resources, equipment and experts/specialists. When such requirements are not all available, diagnosis are mostly based on assumptions which may or may not be accurate. This study seeks to apply linear discriminant analysis to classify diabetic patients based on some socio-economic variables. Diabetes is a common illness prevalent among older individuals; correct diagnosis of patients with traits of diabetes depends to a larger extent on the availability and adequacy of facilities and professionals.

2. Methodology

Discriminant analysis focuses on the association between multiple independent variables and a categorical dependent variable to create rules for separating distinct groups as much as possible, and for assigning an observation of unknown origin to one of $k \geq 2$ distinct preexisting groups [11]. Fisher's linear discriminant function (FLDF) is frequently used for discrimination, classification and prediction purposes under the usual basic statistical assumptions of; multivariate normality of the independent variables, equality of variance and covariance matrices, and relative equality of groups sample sizes [11]. According to [7], the linear discriminant analysis (LDA) is suitable for supervised classification when the number of observations, n is larger than the number of variables p , i.e. ($n > p$). However, with the development of new technologies, there has been increase in complex problems with high-dimensionality, a situation where the number of variables (the dimension of the data vectors) is much larger than the number of observations

(sample size), that is ($n < p$) in many disciplines such as medicine and epidemiology, genetics, biology, metrology etc. [10]. There are many proposed methods for the analysis high-dimensional data where ($n < p$) such as K-D tree and R-tree [6], however, their performance and efficiency decrease as the dimensionality increases because the methods are designed to operate with small dimensionality.

When discriminating between two groups, the analysis assumes the two samples or populations have the same covariance matrix Σ but distinct mean vectors μ_1 and μ_2 with p variables, where the discriminant function that maximizes the separation of the groups is the linear combination of the p variables [14]. Discriminant analysis focuses on the association between multiple p independent variables and a categorical dependent variable by forming a composite of the independent variables. Fisher's approach transforms the p - variate observations x to univariate y such that the y 's obtained from the two populations were adequately separated. The objective is to select linear combination of x that maximize the ratio of squared distance between sample means of y to its variance. This study uses the method of discriminant analysis on health-related data collected from the University of Port Harcourt Teaching Hospital, Port Harcourt, Rivers State, Nigeria. Data on 335 (40 Diabetic and 295 Non-diabetic) patients from the medical clinic were collected. Measurements on thirteen (13) variables was taken on each patient, the variables are; Age (x_1), Gender (x_2), Marital Status (x_3), Weight (x_4), Body Mass Index (BMI) (x_5), Height (x_6), Systolic (x_7), Diastolic (x_8), Employment Status (x_9), Smoking habit (x_{10}), Alcohol Intake (x_{11}), Snacks Consumption (x_{12}), and Hypertension Status (x_{13}). The linear discriminant function (LDF) model is denoted by;

$$DF = b_0 + b_1x_{i1} + b_2x_{i2} + \dots + b_px_{ip} \quad (1)$$

where: $D_i = DF$ or the predicted score

b_i = the discriminant coefficient or weight for variable i

X_i = the independent score for independent variable i

b_0 = a constant

p = the number of independent variables

The standardized DF coefficients are of immense logical importance; they indicate the importance of each independent variable. The sign on the coefficients (\pm) specifies the type of relationship, that is whether the variable is making a positive or negative influence, and coefficients with large absolute values associated with variables have more excellent differentiating capability [3].

When using multivariate data for discriminant analysis, measurements are made on p variables for the n observations. With the classical LDA, it is assumed that $n > p$ and the n observations are divided into $k \geq 2$ predefined distinct groups and that the i^{th} group is denoted by π_i , $i = 1, 2, \dots, k$. The data matrix $X_{(n \times p)}$ represent the measurements of all observations and the values for the p variables are; X_1, X_2, \dots, X_p for each observation. Fisher's linear discriminant analysis looks for the linear function bx which maximized the ratio of the between-groups sum of squares to

the within-groups sum of squares; that is, let

$$y = Xb = \begin{bmatrix} x_1 b \\ \vdots \\ x_g b \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_g \end{bmatrix} \quad (2)$$

be a linear combination of the columns of X . Then y has total sum of squares.

$$y'Hy = b'X'HXb = b'tb \quad (3)$$

which can be partitioned as; sum of the within-groups sum of squares,

$$\sum y_i' H_i y_i = \sum b' x_i' H_i x_i b = b' W b \quad (4)$$

and the between-groups sum of squares,

$$\sum n_i (\bar{y}_1 - \bar{y})^2 = \sum n_i \{b'(\bar{x}_1 - \bar{x})\}^2 = b' B b \quad (5)$$

where \bar{y}_1 is the mean of i^{th} sub-vector y_i of y , H_i is the $(n_i \times n_j)$ centering matrix, $W =$.

The discrimination rule is to

$$\text{allocate } X \text{ to } \begin{cases} \text{group I} & \text{if } d'w^{-1} \left\{ x - \frac{\pi}{2} (\bar{x}_1 + \bar{x}_2) \right\} > 0 \\ \text{group II} & \text{Otherwise} \end{cases} \quad (6)$$

3. Results and Discussion

The Fisher's linear discriminant analysis technique was used to fit a predictive equation based on the measured variables for classifying new individuals, and to re-classify the original data to enable the interpretation of the predictive equation for better understanding of the relationships that may exist among the p variables. The results of the analysis are presented as follows:

Table 1. Fisher's Linear Discriminant Function Classification.

Groups		Predicted Group		
		Membership		Total
		1	2	
Original	1	24	16	40
	% Count	60	40	100
	2	104	191	295
	% Count	35.3	64.7	100
Cross Validated	1	19	21	40
	% Count	47.5	52.5	100
	2	111	184	295
	% Count	37.6	62.4	100

64.2% of Original Grouped Cases Correctly Classified.
60.6% of Cross-Validated Group Cases Correctly Classified.

In Table 1, 16 (40% of diabetic patients) and 102 (35.3% of non-diabetic patients) observation were wrongly classified in group I and II respectively using re-substitution. While 21 (52.5% of diabetic patient) and 110 (37.3% of non-diabetic patients) observation were wrongly classified in group I and

II respectively using cross-validation. In general, the Fisher's linear discriminant function correctly classified 65.4% of the total observation using re-substitution and 62.7% of the total observation using cross-validating method.

Table 2. Test of Equality of Group Means.

Variable	Wilk's				
	Lamda	F	df1	df2	Sig.
Age	0.982	6.062	1	332	0.014
Gender	0.974	8.756	1	332	0.003
Marital Status	1.000	0.041	1	332	0.084
Weight	1.000	0.102	1	332	0.750
BMI	0.996	1.293	1	332	0.256
Height	0.994	1.850	1	332	0.175
Systolic	0.997	0.983	1	332	0.322
Diastolic	0.998	0.507	1	332	0.477
Employment Status	0.994	1.910	1	332	0.168
Smoking Habit	1.000	0.004	1	332	0.945
Alcohol Drinking	0.999	0.458	1	332	0.499
Eat Snacks	0.999	0.174	1	332	0.677
Hypertion	0.994	1.985	1	332	0.160

The Wilks' lambda in Table 2 shows the measure of how well each function separates cases into groups, smaller values of Wilks' lambda indicate greater discriminatory ability of the function. The more important independent variables that contribute significantly to the prediction by the discriminant function as indicated by the smaller values of the Wilks's lambda are, Age(x_1), Gender(x_2).

Table 3. Canonical Discriminat.

Function Coefficients	
Age	-0.043
Gender	1.772
Marital Status	0.161
Weight	-0.102
BMI	0.293
Height	12.459
Systolic	0.005
Diastolic	0.004
Employment Status	0.279
Smoking Habit	0.529
Alcohol Drinking	-1.606
Eat Snacks	0.205
Hypertion	0.509
Constant	-20.950

Table 3 contains the unstandardized discriminant function coefficients, which are similar to unstandardized b_i (regression) coefficients in multiple regressions. The actual prediction equation based on the unstandardized coefficients (Equation 7) can be used as the predictive model for the classification of new patients. These predictor variables provide the best discrimination between groups; the fitted linear discriminant model is;

$$D_i = -20.938 - 0.043x_1 + 1.780x_2 + 0.161x_3 - 0.101x_4 + 0.293x_5 + 12.444x_6 - 0.005x_7 + 0.003x_8 + 0.274x_9 + 0.530x_{10} - 1.606x_{11} + 0.207x_{12} + 0.513x_{13} \quad (7)$$

A discriminant score can be calculated based on the weighted combination of the independent variables in Table 3, where D_i the predicted is or discriminant score.

4. Conclusion

This study applied the Fisher's linear Discriminant Analysis (FLDF) to health data on diabetic patients from the University of Port Harcourt Teaching Hospital, Rivers, Nigeria and created a predictive discriminant model that classifies patients into one of two groups (Diabetic and Non-Diabetic) based on demographic and socio-cultural information from each patient. Using the structure matrix, the results show that age and gender are the most useful variables for segmenting the patient base. The Fisher's linear discriminant function correctly reclassify 65.4% of the total observation using re-substitution and 62.7% of the total observation using cross-validating method. The fitted model also predicts group membership of new patients adequately.

References

- [1] Abdullah, N. A. H., Halim, A., Ahmad, H., and Rus, R. (2008). Predicting corporate failure of Malaysia's listed companies: Comparing multiple discriminant analysis, logistic regression and the hazard Model. *International Research Journal of Finance and Economics*, 15, 201-21
- [2] Alayande, S. A. and Bashir, K. A. (2015). An Overview and Application of Discriminant Analysis in Data Analysis. *IOSR Journal of Mathematics (IOSR-JM)*. 11 (1), 12-15. DOI: 10.9790/5728-11151215.
- [3] AlKubaisi, M., Aziz, W. A., George, S. and Al-Tarawneh, K. (2019). Multivariate Discriminant Analysis Managing Staff Appraisal Case Study. *Academy of Strategic Management Journal*. 18 (5), Online ISSN: 1939-6104
- [4] Antonogeorgos, G., Demosthenes, B. P., Kostas, N. P. and Anastasia, T. (2009). 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. *International Journal of Pediatrics*. doi: 10.1155/2009/952042
- [5] Bhuyan, K. C. (2005). *Multivariate Analysis and its Application*. Department of Statistics, Garyounis University, Libya. New Central Book Agency (P) Ltd.
- [6] Cai, D., He, X. and Han, J. (2008). Srda: An Efficient Algorithm for Large-scale Discriminant Analysis. *Knowledge and Data Engineering*, 20 (1): 1-12.
- [7] Clemmensen, L. K. H. (2013). On Discriminant Analysis Techniques and Correlation Structures in High Dimensions. Kgs. Lyngby: Technical University of Denmark. Technical Report-2013, No. 04
- [8] Erimafa, J. T. (2009), Application of Discriminant Analysis to Predict the Class of Degree for Graduating Students in a University System. *International journal of physical science*, 4 (1), 16 - 21.
- [9] Fisher, R. A. (1938). The Statistical Utilization of Multiple Measurements. *Ann. Eng. Lond.* 7, 179-88.
- [10] Gebru, T. G. (2018). Sparse Linear Discriminant Analysis with more Variables than Observations. Ph. D. Thesis. The Open University
- [11] Hafez, E. I., Abdel-Fatah, E. M., Abdel-Nabi, S. M. and Zeidan, A. S. A. (2015). Discriminant Analysis in View of Statistical and Operations Research Techniques. *Journal of Multidisciplinary Engineering Science and Technology (JMEST)*. 2 (11), 3039-3047
- [12] Mbanasor J. A. and Nto, P. O. O. (2008). Discriminant Analysis of Livestock Farmers' Credit Worthiness Potentials under Rural Banking Scheme in Abia State, Nigeria. *Nigerian Agricultural Journal*, 39 (1), 1-7.
- [13] Micheal, A. B. (2014). Application of Discrimination and Classification on Diabetes Mellitus Data. *International Journal of Applied Science and Technology*, 4 (6). 292-298
- [14] Schlegel, A. (2018). Linear Discriminant Analysis for the Classification of Two Groups <https://aaronshlegel.me/linear-discriminant-analysis-classification-two-groups.html>