University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-6894

Vol.18 No.1 Mar 2023

Selective Overview Discriminant Analysis

Hassan S. Uraibi
Dept. of Statistics University of AL Qadisiya
hassan.uraib@qu.edu.iq

Khalid Hyal Hussain

Dept. of Statistics University of AL- Qadisiya

khaledhyal@outlook.com

https://doi.org/10.32792/utq/utj/vol18/1/1

Abstract

In this paper we will review the concepts of discriminant analysis which is one of the important multivariate statistical analysis methods. It is used in various scientific, particularly the studies that are concerned with classification and prediction. It is having discriminant function which has an ability to distinguish between observations and then is classified the observations into two groups or more than it.

It belongs to it depending on a special function called (the discriminant function), which is a linear combination between the independent variables and the dependent variable, which is of the nominal type. The main objective of this research was to address the subject of discriminant analysis from its early beginnings, present the studies and research that concerned it, how the linear discriminant function developed, and then propose new discriminatory functions such as the quadratic and logistic discriminant functions. It also presented the basic concepts of classification processes, conditions and assumptions that must be available to apply the discriminant analysis method, and then dealt with the Robust Discriminant Analysis method, which dealt with the problems of the presence of outliers among the data under study, which violate the basic assumptions of applying the discriminant analysis method, such as the hypothesis of the normal distribution of data.

Keywords: Discriminant analysis, classification, financial analysis.

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-6894

Vol.18 No.1 Mar 2023

1. Introduction

Discriminant analysis is one of the methods of multivariate statistical analysis and it has wide used in classification, discrimination, prediction and others.it is a statistical technique concerned with studying the relationship between a categorical variable and a continuous data set.

(Maharaj 2014) showed that the purpose of the discriminant analysis is to determine which variables affect the process of distinguishing between two or more groups, and to build a discriminatory base to predict the affiliation of new elements to one of the groups based on the preliminary information available in the data under study.

According The study of (Feinberg 2010), Since discriminant analysis can solve classification problems involving categorical dependent variables, many researchers from various fields such as business, medical, education, environment, sociology, finance, etc. have been drawn into this field. For example, marketing researchers typically like to use discriminant analysis to study market segmentation Marketing researchers want to define linear combinations of predictor variables that help to better distinguish between known combinations. They also like to categorize the anonymous notes into predefined groups. For example, marketing researchers like to predict which customers will renew their contracts in the next year based on specific variables. Therefore, previous studies were interested in the subject of discriminatory analysis from a long time ago, dating back to the beginning of the twentieth century, when the simple beginnings of the idea of classifying and distinguishing between data in a scientific way were.

2. The first beginnings of discriminant analysis

The study of (Tildesley 1921) is one of the first ideas for discriminant processes, as it is the first study used in the field of classification and discrimination of the English scientist (M.L. Tildesley), the main problem of the study was how to distinguish and classify things on the basis of a fixed scientific base. Where it aimed to classify a group of bony skulls dating back to prehistoric human bodies, or the so-called Burmese skulls. The researcher used the modern matching parameters method of the English mathematician (Person), the study concluded that the used method gave good results in classifying the observations under study into several racial groups according to human race. A disadvantage of this study that it is primitive and lacks the modern scientific method.

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-6894

Vol.18 No.1 Mar 2023

After that, the Indian statistician (P.C. Mahalanobis 1936) put the first seed of the idea of discrimination in a statistical way, where he proposed an important statistical scale used to distinguish and classify between two populations called (Mahalanobis Distance), this measure was denoted by (D^2), which is a statistic that measures the square of distance between two centers of two groups, and the basic formula for this scale was as follows:

$$D^2 = (\bar{x}_1 - \bar{x}_2)' \Sigma^{-1} (\bar{x}_1 - \bar{x}_2) \qquad (1)$$

Where:

$\bar{x}_1, \bar{x}_2$:is the means of the first and the second groups respectively

Σ: is the (Var-Cov) matrix.

This statistic was developed for use in classifying a new observation into one of a two groups that follow a normal distribution and have common properties after knowing the means of the two groups and the var-cov matrix. Also, the classification errors should be as minimum as possible.

After deriving the mathematical expectation, variance and moments of this statistic, it was applied to study the ethnic classification of societies based on some characteristics as explanatory variables. It demonstrated its ability to distinguish between the studied totals and with an appropriate classification error.

3. Discriminant Functions

There are several functions of discriminant, each function has certain assumptions and conditions for use, depending on the type of study and data so We will introduce some popular functions of discriminant analysis.

3.1 The linear Discriminant Analysis (LDA):

The distance statistics of (Mahalanobis) did not give accurate results in the classification as it cannot be relied upon in large research and studies, is not a clear discriminant function and is unreliable for complex classification processes so it was necessary to develop a special function that enables researchers to carry out classification operations accurately.

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-6894

Vol.18 No.1 Mar 2023

The British scientist (Ronald Fisher 1936) suggested to the first once the Linear Discriminant Function (LDF), Where he assumed that this function is the best way to classify and discriminant in scientific issues, which is as in the following formula:

$$Z = x'\Sigma^{-1}(\bar{x}_1 - \bar{x}_2) \qquad (2)$$

Where:

$\bar{x}_1, \bar{x}_2$is the mean of the first and the second group respectively.

Σ is the (Var-Cov) matrix.

Assumptions of (LDF):

When applying the linear discriminant function then the variances are homogeneous for the groups also the data must be following a normal distribution, as well as that the relationship is of a linear type for the features. This function was applied to classify a number of plants depending on a set of interrelated variables, and the sample size was (50) observations to be classified into two groups according to the strain. After estimating the parameters of the model and conducting statistical tests and analyzing variance using the (Fisher) criterion and calculating the probability of classification error, the study concluded that the linear discriminatory function has a strong effectiveness in classification between the two groups.

3.2 The Quadratic discriminant function (QDF):

Among the disadvantages of the two previous studies, (Mahalanobis 1936) and (Fisher 1936), is the inability to apply them except in the light of certain hypotheses such as the normal distribution of data and the homogeneity of variance for the samples under study, as well as the linearity of the relationship between the dependent variable and the independent variables.

The scientist (Smith 1946) developed a new study to solve the problem of using the method of discriminant analysis in the case of heterogeneity of the variances of the groups under study, where the main problem addressed by this study is the method of solving the problems of discrimination between two societies when the variances are heterogeneous, so a new discriminatory function was proposed It is called the Quadratic Discriminant Function (QDF) and its formula is as follows:

$$Q_{(z)} = (z - \bar{y})'\Sigma^{-1}{}_y(z - \bar{y}) - (z - \bar{x})'\Sigma^{-1}{}_x(z - \bar{x}) + \ln\frac{/\Sigma_y/}{/\Sigma_x/} \qquad (3)$$

Where:

Z: Vector of new observations which we want to classification

$\bar{y}$: Vector means of Groupe y

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-6894

Vol.18 No.1 Mar 2023

$\bar{x}$: Vector means of Groupe x

$\Sigma^{-1}{}_y$: The inverse for covariance matrix of the first group

$\Sigma^{-1}{}_x$: The inverse for covariance matrix of the second group

It is a generalization of the linear discriminant function of (Fisher), which requires a multiple normal distribution of the data. After applying the study to the real data, it was concluded that the new function is highly effective in discrimination compared to the linear discriminatory function in the presence of the problem of heterogeneity of variance, as it gave accurate results with less classification error.

As for (Hussain 1999), he used the Quadratic Discriminant Function in the processes of distinguishing numbers, the main goal was to find a scientific way to distinguish between writing numbers by designing an electronic system to analyze the patterns of Arabic numbers. Observations were used for a sample size of (150) people. After deriving the quadratic function and performing the statistical analysis, the quadratic function proved its ability to distinguish, as good results were obtained in distinguishing writing patterns, and the classification error was close to zero.

3.3 The Logistic Discriminant Function (LODF):

The previous study had addressed the problem of the heterogeneity of variance for the samples under study, but the quadratic discriminant function (QDF) is ineffective in the case of violating the hypothesis of the normal distribution of the data and the linearity of the relationship between the dependent variable and the independent variables, so (Cox and Chambers 1967) suggested a new discriminatory function called the Logistics Discriminant Function (LODF) in which the response variable is of a binary response type that takes the two values (0,1) and the logistic response function is in the following form:

$$P_i = \frac{e^{(\beta_0+\beta_1 X_1+\cdots+\beta_P X_P)}}{1+e^{(\beta_0+\beta_1 X_1+\cdots+\beta_P X_P)}} \qquad (4)$$

Where: β_i is the parameters of model, i=0,…,p and p is the number of independent variables.

The logistic discrimination function depends mainly on the initial distribution of the data of the first and second communities, and its formula is:

$$Z = \text{Ln}\frac{P(X/Y_1)}{P(X/Y_2)} \qquad (5)$$

The function was applied to a group of sick people who were given doses of several drugs and then followed up on their death or life status. The study concluded that the new logistic function has a high ability to distinguish in such circumstances and with an appropriate classification error.

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-6894

Vol.18 No.1 Mar 2023

4. **Discrimination Function for more than two groups:**

The above studies mentioned the methods of discriminant analysis and their application into two groups only, but the scientist (Anderson1968) published a new study at the University of (Oxford), the main problem that it addressed was how to distinguish between more than two populations that follow the same characteristics in terms of multiple normal distribution and homogeneity of variance. A new method was proposed called (the restricted method) that was used to distinguish between more than two groups by placing (k) restrictions in the form of equations with the number of groups included in the study. Accuracy of classification in the case of multiple communities, as below:

$L_0(x) = 0$

$L_1(x) = \pi_1 f_1(x) - \gamma_{21} f_2(x) - \gamma_{31} f_3(x)$

$L_2(x) = \pi_2 f_2(x) - \gamma_{12} f_1(x) - \gamma_{32} f_3(x)$

.

.

.

$L_i(x) = \pi_i f_i(x) - \gamma_{ji} f_{i+1}(x) - \gamma_{(j+1)i} f_{i+2}(x)$ (6)

(Al-katib 2010) dealt with the linear characteristic function in the case of more than two groups, by dealing with a new method for distinguishing between digital images through the (Wilks Lambda) test, which takes the following formula:

$$\Lambda = \prod_{i=1}^{p} \frac{1}{1 + \lambda_i} \qquad (7)$$

Where: p is the number of independent variables andλ_i is the eigen values for x.

The aim was to identify the significance of the discriminatory function, that is, its ability to distinguish between digital images, and then move on to the second stage, which represents finding the discriminatory function that will be used in distinguishing between digital images, where the (Mahalanobis Distance) scale is used to identify the distance between any two societies according to the form (1). The classification errors resulting from the discrimination function were calculated and it was found that the function is able to distinguish between digital images with an appropriate classification error.

5. Applications of discriminant analysis

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-6894

Vol.18 No.1 Mar 2023

The discriminant analysis method is used in various scientific fields and issues due to its high ability to classify and predict, and we will present some important uses.

The uses of discriminatory analysis have expanded in many scientific fields, and one of the most important fields is its use in the field of financial analysis to study the problems of financial default based on ratios and financial data, as in the study (Altman 1968), The researcher explained the importance of financial analysis of financial ratios and indicators and its role in economic studies. Predicting financial default was the main problem of the study, as the study aimed to use a statistical analysis method, which is a method of discriminatory analysis, to predict the financial default of a sample of American industrial companies based on ratios and indicators that were taken from the financial statements. Where the sample size was (66) companies, he first classified them into two groups: a defaulting group, numbering 33 company, and a non-performing group, numbering 33 company, as well. Also, the number of financial ratios that were used in the analysis amounted to (22) and they represent the explanatory variables of the study. The study reached to build a discriminatory analysis model capable of predicting the financial failure of the studied companies for a period of up to two years before it occurs. When measuring the accuracy of the forecast, it exceeded (80%), but this percentage decreases when forecasting for a period longer than two years. The influential and important ratios in the forecasting process were also determined. After that, (Altman) developed a fixed model to study the prediction of financial failure based on a number of important financial ratios and according to the following formula:

$$Z = 1.2X_1 + 1.4X_2 + 3.3X_3 + 0.6X_4 + X_5 \qquad (8)$$

Where:

X_1: represents to one of the activity ratios and is equal to (capital / assets)

X_2: Profitability Ratio (Profit / Assets)

X_3: It is one of the profitability ratios and is equal to (profit before taxes / assets)

X_4: It is the financial leverage ratio equal to (stockholders' equity / liabilities)

X_5: Represents the activity ratio and equals (sales / assets)

According to the above model, (Altman) considers that the institution is not a failure if $Z > 2.99$ and it is a failure if $Z < 1.81$ and if

$2.99 > Z > 1.81$ then the institution is considered to be in the (grey zone) and it is not possible to predict its future state with certainty.

One of the recent studies that used the method of discriminatory analysis in predicting the problem of financial default is the study (Khawaldy 2017), where the goal was to predict the

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-6894

Vol.18 No.1 Mar 2023

financial failure of a group of small and medium-sized companies and financial institutions based on financial indicators and identify the most important variables in the forecasting process. The study was applied to a sample of financial institutions the size of (30) institutions in the state in Algeria, where financial ratios and indicators were adopted as independent variables. The results showed the efficiency of the model used and its ability to predict financial failure and to determine the financial ratios affecting the forecasting process, which are (current assets turnover rate and total assets turnover rate).

Most of the previous studies assumed that the initial distribution of the data is the normal distribution, which gives a large bias to the estimates. therefore, (Akeyede and Ailobhio 2021) study aimed to use a statistical analysis method for discrimination and classification that has a property of less bias instead of the traditional methods of discrimination, which is the discriminant analysis method that uses rank data and compares it with the classic method. The practical side took a sample size of (350) observations. From the rank data of the financial loan applications and the goal was to propose a mathematical model that would enable the owners of financial institutions to determine the accepted and rejected loan applications. The hypothesis that will be tested is that the discriminatory analysis model in the case of the monotonous data will give better results than the model in the case of the basic data to determine Accepted applications for financial loans with as little bias as possible the proposed model has proven its ability to rank for loans (83%), the bias rate was very low. The results showed the factors affecting the decision of accepting or rejecting the loan applications of customers.

Another study was published in the discriminant analysis by (Hand 1983) which included a comparison of two methods of discriminant analysis. The first method is classified within the parametric methods, which is the method of linear discriminant analysis, which uses the linear discriminant function of the English scientist (R. Fisher 1936) that was presented in previous studies either the second method is one of the nonparametric methods and uses the Kernal function according to the following formula:

$$\widehat{P}(X/W_i) = \frac{1}{n_i}\sum_{i=1}^{n_i}\prod_{k=1}^{p} \Lambda_i^{1-/X_k - X_{ijk}/}\ (1-\Lambda_i)^{/X_k - X_{ijk}/} \qquad (9)$$

Where X_k is the principle components for x, k=1,…,p

P: number of independent variables.

Λ: is called (smoothing parameter) where $1 \geq \Lambda_i \geq \frac{1}{2}$

The study was applied to real binary response data in a multiple model for a group of patients with (urinary incontinence). after estimating the parameters for both functions and conducting a statistical analysis of the data with the presence of 10 variables, it was concluded that the

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-6894

Vol.18 No.1 Mar 2023

(Kernal) method is more effective and responsive in distinguishing with binary data it gives less classification error.

(Bull and Dunner 1987) published a study in the Journal of the American Association for Statistical Sciences, the study aimed to compare two methods of multivariate analysis, namely, the multiple logistic regression method and the multigroup discriminatory analysis method. The comparison was based on two criteria: efficiency and bias in estimating parameters. Two sets of explanatory variables were used, one with a normal distribution and the other with a different distribution. After deriving the relationships of the two functions and applying them to the real data and using the (Efron) scale, it was found that the logistic regression estimates are more efficient in the case of data that do not follow a normal distribution as well when there is a strong correlation between observations. As for the discriminant analysis, it is better in the case of the natural hypothesis of the data and when the distance is large between aggregates and no common areas between views.

(Al-Suleimani 1998) used the linear discriminant analysis method that depends on the linear discriminant function of (Fisher) in the case of two groups and in the case of more than two groups to show its ability to discriminate. The study was applied to a sample of infants with inflammatory bowel conditions, based on 20 variables. After estimating the parameters of the model and performing the statistical analysis, the study showed the ability of the linear discriminant function to identify the variables of importance that affect the diagnosis of these diseases, and the study recommended focusing on them when collecting data in the future.

One of the important studies of the method of discriminant analysis in the medical aspect is the study (Zhang 2000), this study dealt with two parametric methods of discriminant analysis, namely linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA). Finding an ideal classification of the data into two groups according to the type of DNA of the studied observations (DNA or RNA) and with an appropriate classification error. The classification process relied on some of the inherent characteristics that distinguish the item from the surrounding data. After conducting the statistical analysis, it was found that the two functions gave close results in the classification process.

The study (Bassiouni 2021) is one of the most recent studies that used the method of discriminatory analysis. The problem of the study was how to use a multivariate scientific method to distinguish and classify and apply it to reduce the prevalence of cases of diabetes and reduce the number of deaths. The study aimed to determine the factors affecting the incidence of diabetes through the use of a discriminatory function that has the ability to distinguish and separate people into two groups (infected and uninfected), determine the factors affecting the incidence and indicate the importance of each factor or variable as well as predicting the likelihood of developing the disease in the future, through The use of a multivariate statistical analysis method, which is the method of discriminatory analysis.

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-6894

Vol.18 No.1 Mar 2023

Where the discriminant function was built and the number of independent variables was (10), the study was applied to a sample size of (350) observations. The study reached to identify the variables affecting the incidence of diabetes, and the non-significant variables were excluded. The efficiency of the discriminant function was (90.6%) and the classification error was (9.4 %).

6. **Robust Discriminant Analysis**

The studies mentioned above did not take into account the presence of outliers among the data under study, which greatly affects the accuracy of the results, as the previous functions such as the linear, quadratic and logistic discriminant functions, all depend mainly on the (Var-Cov) matrix, which is calculated by the mean, we know that the mean is very sensitive to outliers, it is collapses with the presence of one outlier value because its breakdown point is (1/n), and the presence of outliers leads to a violation of the hypothesis of the normal distribution of the data, which causes major classification errors when applying the mentioned functions. Robust means a method that resists the presence of outliers in the data under study and gives accurate results with a violation of the normal distribution hypothesis.

(Randles et all, 1978) published a study in the Journal of the American Society for Statistical Sciences, where the main problem of the study was how to increase the accuracy of estimation and classification in the event of outliers? the aim of the study was to build a discriminant model. It is immune to the linear and quadratic functions that are resistant and give accurate estimators in the event that there are outliers in the data under study. Two methods have been proposed to construct robust discriminant functions, the first is linear and the second is quadratic. The first method is a generalization of the linear discriminant function of (Fisher), where a little weight was given to the observations that are located far from the common area between the two communities, where the weights are calculated from the following formula:

$$W_i = \frac{2}{D_i} \qquad \text{if } D_i > 2$$

$$= 1 \qquad \text{if } D_i \leq 2$$

Where (D_i) represents the distance of the observation is from the data center and it is called (Mahalanobis distance) and its formula as (1):

$$D_i = (x_i - \bar{x})' s^{-1} (x_i - \bar{x})$$

After that, the mean and variance are calculated by weight method through the following formula:

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-6894

Vol.18 No.1 Mar 2023

$$\bar{X}_w = \frac{\sum w_i x_i}{\sum w_i} \qquad (10)$$

$$S_w = \frac{\sum w_i{}^2 (x_i - \bar{x}_w)(x_i - \bar{x}_w)'}{\sum w_i{}^2} \qquad (11)$$

We then recalculate (D_i) using the new weighted estimators, estimate the cut-off point and classify the new data. As for the second method, the robust (M-estimates) method was used to estimate the averages, the var-cov matrix then the immune estimators were applied to the linear and quadratic discriminant functions and the required classification was performed. The new functions were applied to a set of data polluted with anomalous values that were generated in a random way and it was found that these methods give accurate estimators with little classification error compared to the traditional linear and quadratic discriminant functions.

The researcher (Titterington 1981) presented a comparative study between the traditional discriminant analysis methods (linear, quadratic and logistic) with the robust discriminant analysis method. These methods were applied to a set of data for a sample of patients who had severe head injuries. The data were multidimensionality and varied in nature (monotonous, binary, continuous, discontinuous) with missing values and extreme and outliers' values. The aim was to predict the patient's future condition using the mentioned discriminant functions and to identify the variables affecting the prediction process and to estimate the propensity parameters. After conducting the final analysis, the study concluded that the robust method is the best method in estimation because it addresses the problem of outliers, abnormal and missing data.

(AL Rawashdeh, ate all 2018) presented a comparative study between the linear discriminant analysis (LDA) method with some strong discriminatory methods such as the method of least specific covariance (DMCD) and method of least specific covariance (fast) (FMCD), the problem was The basic principle of the study is that the traditional linear discriminant function gives inaccurate estimators in the event that there are outliers among the data under study, so the goal was to build a discriminatory function in a strong manner that resists the presence of outliers and gives real, efficient estimates with a high percentage. Initially, the three methods were applied to the data in a simulation process to estimate the parameters of the model for all functions and calculate the capabilities of the model under study after polluting the data by adding anomalous values and calculating the model capabilities, new algorithms were applied to immunize the functions on the same polluted data, while the real data are the financial indicators of two groups of banks (Islamic and non-Islamic) in Malaysia. Where the number of observations was (271) collected for the time period 2003-2011, the sample size was (96) observations from Islamic banks and (175) observations from other banks, the number of financial ratios is (23) representing the explanatory variables. The

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-6894

Vol.18 No.1 Mar 2023

robust discriminant functions were applied and classification errors were calculated for both groups of banks, and it was found that the results are very close to the simulation results and that the robust discriminant functions gave highly efficient estimates compared to other methods.

From a disadvantage of this studies is the increase in the misclassification error, where the efficiency of the discriminant analysis depends mainly on the classification error, misclassification must be as little as possible to getting an efficient model and gives accurate results, so the researchers took great care in reducing this error.

Also, one of the fortified studies of the method of discriminatory analysis is the study (Badr 2013). In this research, an attempt was made to find robust and efficient capabilities of the discriminative function such as the parameters of location and scatter and using them in constructing the linear and quadratic discriminatory functions. The problem of the study was to treat with outliers' values found in the medical data, so the study aimed to apply the method of discriminant analysis in the presence of anomalies by finding robust estimators where a number of strong statistical methods were used such as (the least common variance estimator method, the least determinant covariance estimator method, the least determinant variance estimator method). Combined reweights and estimator method (smallest volume ellipsoid). These methods were compared with the traditional methods and the study was applied to the data of two types of blood diseases in Basra. The study concluded that the best robust method is the method of the smallest co-variance re-weighted method, which gave efficient results with an appropriate classification error.

(Beckman and Johnson 1981) was concerned with reducing the classification error, as the main problem of the study was how to address the increase in the size of classification errors resulting from the discrimination processes? because the classification error may cause serious losses in the case of sensitive research. The classification error is a probability value that can be understood from the following formula:

$$\varepsilon = \mathrm{p}(\mathrm{z\ is\ classifeid\ to}\ \mathrm{G}_2/\mathrm{z}\ \in\ \mathrm{G}_1) \qquad (12)$$

Where (Z) is the new observation to be categorized, $(\mathrm{G}_1, \mathrm{G}_2)$ the first and second groups, respectively. A partial discriminant analysis method based on the rank method was developed for the purposes of discriminating in very complex problems and constraining classification errors to control for the size of the error. It aimed to compare between the Monte Carlo method and the rank method for partial discrimination, and rank data was used instead of the real values for the purpose of distinguishing between two groups. After applying the two methods, it was concluded that the rank method is the best in distinguishing as it gives a very small classification error compared to the other method.

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-6894

Vol.18 No.1 Mar 2023

(Roush and Kelly 2009) presented a comparative study between four types of discriminant analysis (linear discriminant analysis, logistic discriminant analysis, rank discriminant analysis and mixed discriminant analysis). After applying the Monte Carlo simulation method, the study concluded that the best method that gave accurate results in discrimination is the rank method, and the linear and logistic discriminant analysis had the same accuracy in classification, while the mixed discriminant analysis is very useful when the data do not follow a normal distribution. And the ordinal method is the best of the previous methods in terms of classification accuracy, as the classification error was the least possible.

After reviewing the previous studies of the discriminatory analysis method, the main problem was the deviation of the data from the normal distribution, which occurs due to the presence of outliers. This greatly affects the quality of estimators such as mean and variance because these estimators are known to be highly sensitive to outliers as (Sajobi 2012) showed. Therefore, many researches and studies have been developed in the field of classification using discriminant analysis, and many efforts have been made to develop a robust discriminatory rule that are resistant to violations of some assumptions. A number of robust discrimination analysis methods have been proposed by replacing the classic estimators such as (mean and variance) with a number of robust estimators such as (M-Estimators), (S-estimators), Minimum covariance determinant (MCD), minimum volume ellipsoid (MVE) estimator, estimators based on trimmed Mahalanobis distance (M-distance) and feasible solution algorithm (FSA), These methods are mentioned in many studies like (Campbell 1982),(Randles 1978a), (Croux 2001), (Alrawashdeh 2012) ,(Chork1992) and (Lachenbruch 1977).

7.Simulation and application side

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-6894

Vol.18 No.1 Mar 2023

A simulation study was conducted to test each of the linear discriminant analysis (LDA) model, the Sherrod model, and Robust discriminant analysis (RLDA) model, and then make a comparison between them by generating random variables with a number (P = 28) variable distributed according to the multiple normal distribution Variables (Multivariate Normal Distribution) with mean (μ=0) and ($\sigma^2 = \rho^{|i-j|}$):

$x \sim N(0, \rho^{|i-j|})$

Where $\rho = \{0.10, 0.50, 0.90\}$

It represents the degree of correlation, which was taken with three degrees of correlation to generate five different sizes of samples for each degree of correlation, namely:

$n = \{100, 125, 150, 200, 500\}$

For the purpose of testing the accuracy of the three methods, we polluted each group of the generated data set three times with different polluting percentages (0.15, 0.10, 0.05) out of the outliers generated from the chi-square distribution with a degree of freedom of 25.

To classify the observations in each generated data set into two groups, one faltering and the other non-faltering, we used the following multiple linear regression model:

$$Z = x\beta + \varepsilon$$

where β is the unit vector of the parameters of this model and my agency:

$$\beta = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}_{28x1}$$

As for (ε), it is the vector of random errors of the model.

After that, the Inverts of standard logistic function was used, which was formulated in the following form:

$$Pr = \frac{1}{1 + \exp(-Z)}$$

For the purpose of obtaining the conditional probabilities that will be used to generate the random variable (Y) from the binomial distribution, if the value is equal to (0), it is considered a default condition, and if it is equal to (1), it is considered a non-faltering condition.

The above procedures were repeated (1000) times to calculate the probability of miss classification and the probability of its correctness in each of the simulation samples using the traditional discriminant analysis (LDA), the Sherrod model (Sh.mod.) and the high immunity

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-6894

Vol.18 No.1 Mar 2023

discriminant analysis (RLDA) model. Classification error rates have been calculated for each of these criteria and the method that obtains the least error among these rates will be the best method.

P	N	Mis.LDA	Hit.LDA	Mis.Sherrod	Hit.Sherrod	Mis.RLDA	Hit.RLDA
0.10	100	0.125	0.875	0.495	0.505	0.133	0.867
	125	0.139	0.861	0.504	0.496	0.14	0.86
	150	0.148	0.852	0.502	0.498	0.149	0.851
	200	0.153	0.847	0.495	0.505	0.156	0.844
	500	0.17	0.83	0.5	0.5	0.169	0.831
	100	0.105	0.895	0.503	0.497	0.108	0.892

The first case: measuring the classification error of the mentioned methods with correlation degrees (0.1, 0.5, 0.9) with changing the size of the samples for each degree of correlation for clean data and observing the results.

The second case: measuring the classification error of the mentioned methods with different degrees of correlation with changing the size of the samples, but this time by polluting the data with a percentage of outliers and then discussing the results.

The results will be posted and commented on as shown in the tables below.

University of Thi-Qar Journal
ISSN (print): 2706-6908, ISSN (online): 2706-6894
Vol.18 No.1 Mar 2023

ρ	N	Mis.LDA	Hit.LDA	Mis.Sherrod	Hit.Sherrod	Mis.RLDA	Hit.RLDA
0.10	100	0.269	0.735	0.482	0.518	0.148	0.852
	125	0.279	0.721	0.481	0.519	0.158	0.842
	150	0.293	0.707	0.467	0.533	0.159	0.841
	200	0.297	0.703	0.487	0.513	0.163	0.837
	500	0.325	0.675	0.481	0.519	0.177	0.823
0.50	125	0.11	0.89	0.501	0.499	0.111	0.889
	150	0.213	0.787	0.489	0.511	0.121	0.879
	200	0.115	0.885	0.5	0.5	0.114	0.885
0.50	125	0.256	0.744	0.492	0.508	0.141	0.859
	150	0.236	0.744	0.475	0.525	0.132	0.868
	500	0.133	0.867	0.495	0.505	0.133	0.867
0.90	100	0.082	0.918	0.504	0.496	0.081	0.919
	200	0.263	0.737	0.488	0.512	0.139	0.861
	125	0.076	0.924	0.493	0.507	0.08	0.92
	150	0.084	0.916	0.499	0.501	0.085	0.915
	200	0.087	0.913	0.498	0.502	0.086	0.914
	500	0.099	0.901	0.503	0.497	0.099	0.901

Table 1: Classification error rates for the three methods for 1000 clean simulated samples ρ={0.10,0.50,0.90}

The classification error was measured for the three methods (traditional discriminant analysis, Sherrod's model, and high immunity discriminant analysis) taking into account the change in sample size and the degree of correlation, noting that the data is not contaminated with outliers. We note that the classification error for (LDA) and (RLDA) is very close and is less than the classification error for (she.mod) at (ρ = 0.1), and the two methods are superior to (she.mod.) as the sample size increases, and this condition continues even when increasing the degree of correlation to (0.9, 0.5) and the reason is natural because the data is clean and not contaminated with abnormal values, so there is no clear difference for the immune method (RLDA) from the traditional (LDA).

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-6894

Vol.18 No.1 Mar 2023

ρ	N	Mis.LDA	Hit.LDA	Mis.Sherrod	Hit.Sherrod	Mis.RLDA	Hit.RLDA
	500	0.279	0.721	0.486	0.514	0.148	0.852
0.90	100	0.116	0.884	0.484	0.516	0.071	0.929
	125	0.131	0.869	0.474	0.526	0.084	0.916
	150	0.139	0.861	0.482	0.518	0.088	0.912
	200	0.133	0.867	0.469	0.531	0.083	0.917
	500	0.148	0.852	0.484	0.516	0.098	0.902

Table 2: Classification error rates for the three methods for 1000 samples contaminated with 5% outliers, ρ={0.10,0.50,0.90}

In this table, the data was contaminated by abnormal values of (0.05) from the total data, and the results were as follows:

when (ρ=0.1):

We note that the traditional method (LDA) was superior to the (she.mod) method because it has a lower classification error, while the robust method (RLDA) was superior to the two methods, as its classification error was less, reaching (0.14). When the sample size is gradually increased, we notice that the impregnable method maintains a lower classification error than the other two methods.

when (ρ=0.5):

When increasing the degree of correlation, we notice that the classification error of the (LDA) method is much less than the classification error of the (she.mod) model, and this case increases as the sample size increases. As for the (RLDA) method, it gives greater accuracy than the two classification methods, because its classification error is less much more than the rest of the methods.

when (ρ=0.9):

When observing the classification error for (LDA) when (n = 100), we find that it is equal to (0.116), while the error for (she.mod) is equal to (0.482), it is clear that there is a big difference between the two ratios, but when comparing these two ratios with the error for (RLDA) which is equal to (0.071) shows the significant difference of the fortified method from the rest of the methods because it has a very small classification error. Also, the larger the sample size, the more the fortified method outperforms the rest. We note that when (n = 500), the classification error for both (LDA) and (she.mod) methods is (0.148) and (0.484), respectively. As for the classification error of (RLDA) It is (0.098), which is a very small error compared to other methods, which makes the robust method highly accurate in classification.

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-6894

Vol.18 No.1 Mar 2023

0. 10	100	0.251	0.749	0.476	0.524	0.138	0.862
	125	0.271	0.729	0.469	0.531	0.159	0.841
	150	0.284	0.716	0.452	0.548	0.161	0.839
	200	0.281	0.719	0.462	0.538	0.173	0.827
	500	0.314	0.686	0.465	0.535	0.185	0.815
0.50	100	0.216	0.784	0.464	0.536	0.118	0.882
	125	0.236	0.764	0.465	0.535	0.122	0.878
	150	0.237	0.763	0.46	0.54	0.126	0.874
	200	0.236	0.764	0.458	0.542	0.142	0.858
	500	0.265	0.735	0.461	0.539	0.154	0.846
0.90	100	0.117	0.883	0.468	0.532	0.072	0.928
	125	0.125	0.875	0.454	0.546	0.072	0.928
	150	0.129	0.871	0.448	0.552	0.076	0.924
	200	0.131	0.869	0.462	0.538	0.084	0.916
	500	0.141	0.859	0.456	0.544	0.092	0.908

Table 3: Classification error rates for the three methods for 1000 samples contaminated with 10% outliers ρ={0.10,0.50,0.90}

In this table, the data was contaminated by an abnormal value greater than the previous one, amounting to (0.10) from the total data, so the results were as follows:

when (ρ=0.1):

We note that the traditional method (LDA) was superior to the (she.mod) method because it has a lower classification error, while the robust method (RLDA) was superior to the two methods, as its classification error was less, reaching (0.14). When the sample size is gradually increased, we notice that the impregnable method maintains a lower classification error than the other two methods.

when (ρ=0.5):

When increasing the degree of correlation, we notice that the classification error of the (LDA) method is much less than the classification error of the (she.mod) model, and this case increases as the sample size increases. As for the (RLDA) method, it gives greater accuracy than the two classification methods, because its classification error is less much more than the rest of the methods.

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-6894

Vol.18 No.1 Mar 2023

when (ρ=0.9):

When observing the classification error for (LDA) when (n = 100), we find that it is equal to (0.116), while the error for (she.mod) is equal to (0.482), it is clear that there is a big difference between the two ratios, but when comparing these two ratios with the error for (RLDA) which is equal to (0.071) shows the significant difference of the fortified method from the rest of the methods because it has a very small classification error. Also, the larger the sample size, the more the fortified method outperforms the rest. We note that when (n = 500), the classification error for both (LDA) and (she.mod) methods is (0.148) and (0.484), respectively. As for the classification error of (RLDA) It is (0.098), which is a very small

P	N	Mis.LDA	Hit.LDA	Mis.Sherrod	Hit.Sherrod	Mis.RLDA	Hit.RLDA
	100	0.243	0.757	0.448	0.552	0.14	0.86

error compared to other methods, which makes the robust method highly accurate in classification.

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-6894

Vol.18 No.1 Mar 2023

0. 10	125	0.256	0.744	0.46	0.54	0.151	0.849
	150	0.262	0.738	0.438	0.562	0.158	0.842
	200	0.258	0.742	0.454	0.546	0.161	0.839
	500	0.292	0.708	0.441	0.559	0.182	0.818
0.50	100	0.199	0.801	0.445	0.555	0.099	0.901
	125	0.228	0.772	0.446	0.554	0.124	0.876
	150	0.226	0.773	0.451	0.549	0.121	0.879
	200	0.232	0.768	0.44	0.56	0.142	0.858
	500	0.248	0.752	0.446	0.554	0.152	0.848
0.90	100	0.122	0.878	0.439	0.561	0.067	0.933
	125	0.126	0.874	0.434	0.566	0.071	0.929
	150	0.114	0.886	0.438	0.562	0.067	0.933
	200	0.121	0.879	0.44	0.56	0.073	0.927
	500	0.134	0.866	0.437	0.563	0.085	0.915

Table 4: Classification error rates for the three methods for 1000 samples contaminated with 15% outliers ρ={0.10,0.50,0.90}

This time, the data was contaminated by an abnormal value greater than the previous one, amounting to (0.15) from the total data, so the results were as follows:

when (ρ=0.1):

We note that the traditional method (LDA) was superior to the (she.mod) method because it has a lower classification error, while the robust method (RLDA) was superior to the two methods, as its classification error was less as it reached (0.14) when (n = 100). When the sample size is gradually increased, we notice that the robust method maintains a lower classification error than the rest of the methods, while the (she.mod) method continues to produce the largest classification error, which means a lack of accuracy in classification and prediction.

when (ρ=0.5):

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-6894

Vol.18 No.1 Mar 2023

When increasing the degree of correlation, we notice that the classification error of the (LDA) method has reached (0.199), which is much less than the classification error of the (she.mod) model, which amounted to (0.445) at a sample size of (n = 100), and this situation continues as the sample size increases. As for the (RLDA) method, it gives greater accuracy than the two methods in classification, because its classification error is much less than the rest of the methods, as the error rate reached (0.099) when (n = 100), which is a very small percentage compared to the previous values.

when (ρ=0.9):

When observing the classification error for (LDA) with a sample size (n = 100), we find that it is equal to (0.122), while the error for (she.mod) is equal to (0.439), and it is clear that there is a big difference between the two percentages, but when comparing these two percentages with the error for (RLDA), which is equal to (0.067), shows the significant difference of the fortified method from the rest of the methods because it has a very small classification error. Also, the larger the sample size, the more the fortified method outweighs the rest. We note that when (n = 500), the classification error for both (LDA) and (she.mod) methods is (0.134) and (0.437), respectively. As for the classification error of (RLDA) It is (0.085), which is a very small error compared to other methods, which makes the robust method superior and has a high accuracy in classification.

7. Summary

The simulation study that we conducted aimed to demonstrate the importance of each of the methods under study, which are (LDA), (she.mod), and (RLDA), and then to find out which methods are more accurate in classification and prediction. The main criterion was misclassification, as the method used It gives the lowest rate of classification error, which is the most efficient in terms of the accuracy of the results. Data were generated for random samples and classification error rates were measured in two cases before and after contamination with abnormal values, taking into account changing sample sizes and degrees of correlation. And as we noticed in the results above, the (she.mod) method gave a very high classification error in all cases, so it has low efficiency in terms of the accuracy of the results in classification and prediction. As for the (LDA) and (RLDA) methods, the performance was balanced in the case of clean data, where the classification error was very close with changing sample sizes and degrees of correlation, but when the data was contaminated with abnormal values, very high differences appeared in the classification errors for the two mentioned methods, especially when increasing the sample size and increasing the degree of Correlation, which gives a clear superiority to (RLDA), which represents the immune method, so it is the most efficient and dependable on its results in classification and prediction.

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-6894

Vol.18 No.1 Mar 2023

References:

(1) Maharaj, E. A., & Alonso, A. M. (2014). Discriminant analysis of multivariate time series: application to diagnosis based on ECG signals. Computational Statistics & Data Analysis, 70, 67-87. doi:10.1016/j.csda.2013.09.006.

(2) Feinberg, F. M. (2010). Discriminant analysis for marketing research applications. In N. S. Jagdiesh, & K. M. Naresh (Eds.) Wiley International Encyclopedia of Marketing (Vol. 2). New York, NY:John Wiley & Sons, Ltd doi:10.1002/9781444316568.

(3) M.L. Tildesley,1921, A first study of the Burmese skull, University College, London.

(4) P.C. Mahalanobis,1936, On the generalized distance in statistics, India.

(5) R.A. Fisher,1936, The use of multiple measurements in taxonomic problem, the Annals of Eugenics, v. 7, p. 179-188 (1936) with permission of Cambridge University, London.

(6) Cedric A.B. Smith,1946, Some examples for discrimination.

(7) Elizabeth A. Chambers and D.R. Cox,1967, Discrimination between alternative binary response models, Imperial college, British.

(8) J.A. Anderson,1968, Constrained Discrimination between k Populations, Oxford university, British.

(9) Edward I. Altman,1968, Financial ratios, discriminant analysis and the prediction of carport bankruptcy, The Journal of Finance, Vol. 23, No. 4 (Sep., 1968), pp. 589-609, America.

(10) Ronald H. Randles, James D. Broffitt, John S. Ramberg & Robert V. Hogg,1978, Generalized Linear and Quadratic Discriminant Functions Using Robust Estimates, Journal of the American Statistical Association, 73:363, 564-568.

(11) Richard J. Beckman and Mark E. Johnson,1981, A Ranking Procedure for Partial Discriminant Analysis, Journal of the American Statistical Association, Vol. 76, No. 375 (Sep., 1981), pp. 671- 675.

(12) D. M. Titterington, G. D. Murray, L. S. Murray, D. J. Spiegelhalter, A. M. Skene, J. D. F. Habbema, G. J. Gelpke,1981, Comparison of Discrimination Techniques Applied to a Complex Data Set of Head Injured Patients, Journal of the Royal Statistical Society. Series A (General), Vol. 144, No. 2 (1981), pp. 145-175.

(13) D. J. Hand,1983, A Comparison of Two Methods of Discriminant Analysis Applied to Binary Data, Biometrics, Vol. 39, No. 3 (Sep., 1983), pp. 683-694.

(14) Shelley B. Bull and Allan Donner,1987, The Efficiency of Multinomial Logistic Regression Compared with Multiple Group, Journal of the American Statistical Association, Vol. 82, No. 400 (Dec., 1987), pp. 1118- 1122.

(15) Michael Q. Zhang,2000, Discriminant analysis and its application in DNA sequence motif recognition, *Briefings in Bioinformatics*, Volume 1, Issue 4, 1 November 2000, Pages 331–342.

(16) Mufda J. Alrawashdeh, Taha Radwan and Khalid Abunawas,2018, Performance of linear discriminant analysis using different robust methods, European Journal of pure and applied mathematics, Vol. 11, No. 1, 2018, 284-298.

(17) Sajobi, T. T., Lix, L. M., Dansu, B. M., Laverty, W., & Li, L. (2012). Robust descriptive discriminant analysis for repeated measures data. Computational Statistic and Data Analysis, 56(9).

(18) Campbell, N. A. (1982). Robust procedures in multivariate analysis II. Robust canonical variate analysis. Journal of the Royal Statistical Society. Series C (Applied Statistics), 31(1), 1–8. doi: 10.2307/2347068.

(19) Randles, R. H., Broffitt, J. D., Ramberg, J. S., & Hogg, R. V. (1978a). Generalized linear and quadratic discriminant functions using robust estimates. Journal of the American Statistical Association, 73(363), 564-568. doi: 10.2307/2286601.

(20) Croux, C., & Dehon, C. (2001). Robust linear discriminant analysis using S-estimators. Canad. J. Statist., 29(3), 473–493. doi:10.2307/3316042.

(21) Alrawashdeh, M. J., Muhammad Sabri, S. R., & Ismail, M. T. (2012). Robust linear discriminant analysis with financial ratios in special interval. Applied Mathematical Sciences, 6(121), 6021-6034.

(22) Chork, C. Y., & Rousseeuw, P. J. (1992). Integrating a high-breakdown option into discriminant analysis in exploration geochemistry. Journal of Geochemical Exploration, 43(3), 191-203. doi: 10.1016/0375-6742(92)90105-H.

(23) Ahmed, S. W., & Lachenbruch, P. A. (1977). Discriminant analysis when scale contamination is present in the initial sample. In J. Van Ryzin (Ed.), Classification and Clustering (pp. 331-354). New York, NY: Academic Press.

(24) Wina, Herwindiati, D. E., & Isa, S. M. (2014). Robust discriminant analysis for classification of remote sensing data. In A. Wibisono (Ed.), Proceedings of 2014 International Conference on Advanced Computer Science and

Information System (pp. 454-458), IEEE. doi: 10.1109/ICACSIS.2014.7065892.

(25) I.Akeyede, F.G. Ibi, D. T. Ailobhio,2021, A Discriminant Analysis Procedure for Loan Application using Ranked Data, Asian journal for mathematic sciences, ISSN 2581-3463.