IRAQI STATISTICIANS JOURNAL

https://isj.edu.iq/index.php/isj

ISSN: 3007-1658 (Online)

Weighting Data using the Robust Transformation Matrix to Eliminate the Effect of Outliers in a Robust Discriminant Analysis Model

Khalid Hyal Hussain[1], Fahad Hussein Enad[2]

1Ministry of Planning, Commission of Statistics and GIS، Dhi Qar Statistics Directorate, khaled313@outlook.com

[2]University of Dhi Qar / Department of Studies and Planning , fahadh@utq.edu.iq

ARTICLE INFO

Article history:
Received 26/11/2024
Revised 27/11/2024
Accepted 14/1/2025
Available online 15/5/2025

Keywords:
Discriminant analysis
Robustness
Outlier
Breakdown point
Financial distress

ABSTRACT

Robust statistical methods are of great importance in statistical studies because they provide great resistance in the presence of basic violations of statistical analysis models due to failure to achieve one of the basic assumptions such as the normal distribution of data and others. One of the most important problems facing the researcher is the problem of the presence of outliers in the data under study. Therefore, the goal of this study is to reduce the impact of outliers on the accuracy of the results. The discriminant analysis method was applied to a set of data taken from the Iraqi Stock Exchange, where the outliers were weighted with certain weights to eliminate their impact on the results. The banks under study were classified into two groups based on the cut-off point. The classification error of the mentioned methods was measured and the results were good and reliable.

1.1 Introduction

Robust methods are one of the important measures that recent studies have focused on because of their ability to deal with contaminated data in the event that one of the basic assumptions is violated. The discriminant analysis method is considered one of the methods of multivariate statistical analysis that is concerned with classification and discrimination. It is based primarily on the discriminant function, which is a linear combination between the explanatory variables that increases the variance between groups for the purpose of distinguishing them and reduces the variance within the elements of one group. In this study, we relied on using a robust method to get rid of the effect of outliers by weighting the contaminated data with certain weights and then using the robust discriminant function for the purpose of distinguishing and classifying the banks under study into two groups based on the median financial indicators.

1.2. Importance

Addressing the problem of using the discriminant analysis method in the presence of outliers by using a robust weighting matrix that gives relative weights to the outliers to reduce their impact on the discriminant function.

2.1. Robustness

The concept of Robustness, Robust Statistics, or Robustness Methods in the statistical literature means a set of statistical methods and tools that deal with data contaminated with outliers (contaminated data) and give a high rate of resistance in the event

* Corresponding author. E-**mail** address: khaled313@outlook.com
https://doi.org/10.62933/rctb9q73

that one of the basic assumptions is violated. Traditional statistical methods are based on them. For example, when the assumption of normal distribution of data is violated or in the case of the presence of outliers, the traditional methods, in the presence of these problems, produce inaccurate results in terms of estimation and analysis. However, solid methods provide strong and resistant capabilities. For example, the arithmetic mean is a highly efficient estimate, but it is very sensitive to outliers, as it is affected by the presence of a single outlier, so its breaking point is low, approaching zero. Therefore, the arithmetic mean is considered a vulnerable estimator, while the median is not affected by outliers and has a high breaking point, reaching (50%) Therefore, the median is a robust estimator according to (Huper 1981).

This term means robustness against violations of basic assumptions and insensitivity to the presence of outliers that affects the accuracy of the model used.

First, we must look into the robustness of distributions, as violating natural assumptions such as (independence, congruence, randomness...) leads to the basic distribution deviating from the true distribution. (Hyal.k.2023)

The traditional methods in applied statistics depend on the principle of continuity, but unfortunately, the principle of continuity is baseless because in reality, the traditional methods rely mainly on (the statistically meaningful hypothesis). (B. Klaus 1986)

2.2 Measures of Robustness

There are several standards to measure the level of robustness in capabilities, such as:

A - First: Sensitive Curve:

The sensitivity curve (S.C) is considered one of the most important criteria relied upon in determining the level of robustness of the estimator. It measures the effect of outliers on the statistical estimators and parameters, and its mathematical formula is:

$$SC_n(x,Z) = n\left[Z_{(x_1,x_2,\ldots,x_{n-1},x_n)} - Z_{(x_1,x_2,\ldots,x_{n-1})}\right] \quad (1)$$

Where (Z)It is a typical estimator of a statistic $(x_1, x_2, \ldots, x_{n-1})$ Is the data before it is contaminated with outliers.

$(x_1, x_2, \ldots, x_{n-1,x_n})$ It is data that is polluted by the addition of a single outlier.

The sensitivity curve is considered one of the important tests to measure the robustness of the estimator. When this test is applied, for example, to the arithmetic mean, we find that the arithmetic mean is very sensitive to outliers, as it is greatly affected and may collapse by adding just one outlier. However, in the case of the median, it is not sensitive to outliers and is not affected when... Add one outlier.

B - Influence function:

The influence function was first proposed in 1964 by the scientist Hampel, and he worked on developing it and inferring it with a group of researchers. It is considered an approximate version of the sensitivity curve, through which the estimator (Z) is calculated at a specific distribution such as (F), meaning that the estimator (Z) is a function of the distribution (F).

For example, to express the arithmetic mean, the formula is:

$$Z(F) = E_F(X) \quad (2)$$

In the case of indicating the mediator, the formula is:

$$Z(F) = F^{-1}(0.5) \quad (3)$$

The primary goal of the influence function is to measure the changes that occur in the estimator as a result of contamination of the data under study with a single outlier value such as (X_0).

If we have an estimator (τ) defined on the probability distribution (F) where:

$$F_\varepsilon\left(\frac{X}{\delta_{X_0}}\right) = (1-\varepsilon)F + \varepsilon\delta_{X_0} \quad (4)$$

where:

ε: a certain constant.

δ: standard deviation of the distribution (F).

X_0: outlier.

We assumed that the triple (Ω, γ, P) is the probability space of the independent random variables with identical distribution (x_1, x_2,,x_n) then the influence function for the estimator (τ) is:

$$IF(X_0,T,F) = \lim_{\varepsilon\to 0}\frac{T(F_{\tau,X_0}) - T(F)}{\tau} \quad (5)$$

C – The breakdown point:

The breaking point (BP) is considered one of the important measures to indicate the level of robustness of the estimator. It represents the highest resistance point or percentage of resistance that the estimator can reach and still maintain the properties of the true parameter before it collapses due to data contamination with a number of outliers, according to (Maronna 2006). Hampel (1974) stated that the breaking point represents the highest level at which the estimator can maintain its efficiency before losing it when there is a high percentage of abnormal values in the data, which causes inaccurate results to appear in statistical analysis models. The highest breakdown points that the estimator can reach is (50%) because the amount of contaminated data or outliers cannot exceed (50%) of the amount of real data under study, as mentioned by Giloni (2006). Previous studies, such as (Huper 1983) and (Rousseeuw 1987), have shown that if we have a specific estimator such as (τ) for a set of data contaminated with outliers, the breaking point of this estimator can be calculated from the following formula:

$$BP\left(\frac{\tau}{D}\right) = \min\left\{\frac{m}{n} sup_{DD^*} \left|\left|\tau(D) - \tau(D^*)\right|\right|\right\} \qquad (6)$$

Where:

$BP(\tau/D)$: the breakdown for the estimator (τ).

D: The data set before being contaminated with outliers.

D^*: The data set after being contaminated with outliers.

m: number of outliers.

n: number of observations.

2.3 Outliers

Literature and statistical sources mentioned several definitions and concepts of outliers, such as:

Bross 1961: An outlier is an observation that appears to deviate very significantly from the rest of the sample data.

Freeman 1980: He defined an outlier as an observation that was not generated in the general way that most data in the sample were generated from.

AL Jubouri 1976: An outlier is a value that is inconsistent with other observations in one of the variables or phenomena because it comes from other populations or distributions that differ from the rest of the data.

Keller 2000: An outlier is an observation that falls far from the regression line and has a large error compared to the rest of the data, so it will have a significant impact on the characteristics of the regression model and its estimates.

Barnett and Lewis 1978: They showed that an outlier is an irrational observation when compared to the rest of the data under study.

Hawkins 1980: It is an observation that deviates so much from other observations that it comes to mind that it was generated by different mechanisms or from different other distributions.

Hampel 1986: It is the value that usually comes from different other distributions, and the percentage of abnormal values in real data ranges between (1-10) %.

2.4 kinds of outliers:

Studies have shown the classification of outliers according to their location in the data. Those anomalies that appear in the regression residuals for single models are called (Outliers), but in the case of using multiple models, the outliers are called Leverage Points, which are the outliers that occur in the independent variables and are Their direction in the values of (x) and their values are very far from most of the values of the matrix Some of these attraction points have a negative impact on the data line, causing the general average to deviate away from the true center of the data, so they are called (Bad leverage points). Others have a positive impact on the data line, so they are called (Good Leverage Point), as shown in the following figures:

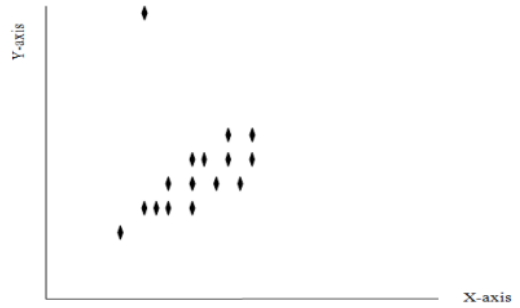


Figure (2-1) an outlier value towards the (y) values

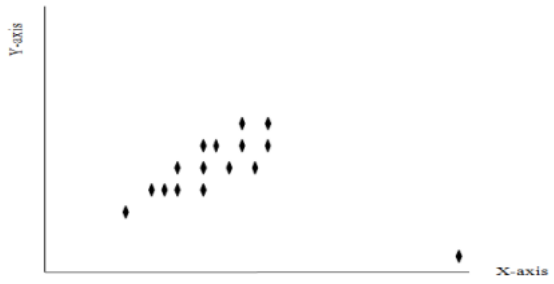


Figure (2-3) A good outlier (located on the data regression line)

2.5 Sources of outliers

The outliers are often a small group compared to the other values in the sample, and there are many reasons that lead to the appearance of outliers in the data, the most important of which are:

1- The data should be taken from asymmetric distributions, meaning that it contains a high skewness towards the right or left.

2- Outliers come from contaminated distributions, unlike good data that comes from the Basic Distribution.

3- Abnormal values are generated as a result of measurement errors that the researcher makes when collecting data, or because of a defect in some equipment, such as laboratory equipment, or because of an error in transmitting information correctly, or what is called a typographical error.

4- The occurrence of sudden changes in the study population due to an emergency event or a compelling reason, such as a sudden loss or sudden profit that occurs in financial institutions or companies.

2.6 The effect of outliers on multivariate statistical analysis models

Statistical work in the event of the presence of outliers is subject to many difficulties because the researcher rarely expects or knows the sources of this contaminated data, and the various types of statistical analysis models are the ones that determine and measure the relationship between the variables, and when estimating the model parameters, we can know the strength and importance of this model.

The least squares method is considered one of the most important statistical methods used in estimation due to the good advantages it enjoys, such as the ease of its practical application and the accuracy of its results provided that its assumptions are met. However, in the event of the presence of outliers, this method becomes inefficient and inaccurate, according to (Huper 1981). The presence of a single outlier will destroy the good advantages of the least squares method. There are several effects caused by the presence of outliers, such as:

-Violating the assumption of normal distribution only (Huper 1981). The presence of a single outlier can lead to the collapse of the assumption of normal distribution of data.

- Causes an increase in error variance (MSE).

- Reduces the value of the coefficient of determination (R2)

- Reduces the value of (F) calculated in the variance analysis table

- It causes the problem of autocorrelation to arise.

The presence of outliers gives an inaccurate and inefficient estimate of the model parameters using the least squares (OLS) method. (kleinbaum1988.

2.7 The effect of outliers on the application of the discriminant analysis method

It has been shown that there is a significant impact of outliers on statistical models. Likewise, the method of discriminant analysis is also affected by these anomalies. From the theoretical side, discriminant analysis is one of the methods of multivariate statistical analysis, and the presence of outliers is in the form of leverage points, which attract the general line of the data. Changing its direction gives

inaccurate results. From the statistical side, the linear discriminant function is stated in the following formula:

$$Z = x^{'}V^{-1}(\bar{x}_1 - \bar{x}_2) \quad (7)$$

It is clear that the linear discriminant function depends mainly on the variance and covariance matrix, which depends mainly on the arithmetic mean. Likewise, the cut point that we use in the classification decision also depends on extracting the arithmetic mean. We mentioned previously that the arithmetic mean is very sensitive to outliers. It collapses in the presence of a single outlier. Therefore, the estimates of this function are inaccurate and do not accurately represent the study data. Also, discriminant analysis depends on the vectors of averages for each sample and the general average and is sensitive to outliers. All of these factors lead to the emergence of inefficient discriminatory coefficients and misclassification of observations. Therefore, these matrices must be fortified using robust methods, such as replacing the arithmetic mean and variance with the secure location and measurement parameters.

2.8. Discriminant analysis:

It is a multivariate statistical analysis method used to distinguish and classify the data under study into two or more groups and then predict the classification of new elements into their appropriate groups according to a special rule called the discriminant function.

The discriminant function can be written in the following form for two groups:

$$Z = x^{'}V^{-1}(\bar{x}_1 - \bar{x}_2) \quad (8)$$

where:

Z: The predictive value of the new view to be classified.

$x^{'}$: The trance of independent variables (x_1, x_2, …, x_n)

$(\bar{x}_1 - \bar{x}_2)$: the vector of difference between the means of groups.

V^{-1}: the invers of var-cov matrix for two groups.

The discriminant function increases the variance between groups and reduces the variance within the elements of the group for the purposes of discrimination.

To use discriminant analysis, the data must follow a normal distribution, the number of groups must be greater than or equal to two groups, the explanatory variables must be independent, and the sample must be drawn randomly.

Using the discriminant analysis method requires that the data be free of outliers to ensure accurate results.

The final step is to apply the cut of point, which determines the observation and classifies it into one of the two groups, which is in the following formula:

$$L = \frac{\bar{y}_1 + \bar{y}_2}{2} \quad (9)$$

Where:

$$\bar{y}_1 = \bar{x}_1^{'} s^{-1}(\bar{x}_1 - \bar{x}_2) \quad (10)$$

$$\bar{y}_2 = \bar{x}_2^{'} s^{-1}(\bar{x}_1 - \bar{x}_2) \quad (11)$$

when $\bar{y}_1 < \bar{y}_2$The decision of discriminant will be made by this rule:

- the new observation (x_0) is classified into first group if $L \leq Z_{x_0}$
- the new observation (x_0) is classified into second group if $L > Z_{x_0}$

2.9. Reweighting matrix

In order to protect Fisher's linear discriminant analysis method from the influence of outliers, we decided to adopt a robust method to reduce the influence of those anomalies in the data under study. It was necessary to resort to transforming the data using a specific location and scale matrix, which is called (RMVN). Then use it to fortify the Mahalanobis Distance method according to (Uraibi 2017).

(Olive and Hawkins 2010) presented a new matrix used to weight multiple normal distribution estimators using an efficient algorithm and has a high breaking point, which can be summarized in the following steps:

1- The construction of this algorithm begins with finding estimates of the mean, variance and covariance matrix and can be denoted by $(L_{0,1}, S_{0,1})$.

2- Finding the Mahalanobis Distance by substituting the above estimators, which we denote (MD). Then we arrange the values in ascending order to obtain the median (MD) values and consider it as a standard limit that can be relied upon to obtain a new row of observations that are less than the (MD) values. We repeat the previous steps five times and each time we calculate the (MD) values for all

the data to obtain $((L_{5,1}, S_{5,1})$ The latter considers the location and measurement estimates of the (MD) matrix.

3- We repeat the same previous steps, but using the median and not the arithmetic mean in the variance and covariance matrix to find the value of (MD) and thus we obtain $(L_{5,2}, S_{5,2})$. Finding the FCH estimators by calculating the traditional distance between the previous estimators $)L_{5,1}S_{5,2}($. If the distance is less than the last value of the threshold, which is the median (MD), in the fifth step, then we choose one of these two estimators, so the location estimator is:

$$L_{FCH} = \begin{cases} L_{5,1} & if \ \sqrt{|S_{5,1}|} < \sqrt{|S_{5,2}|} \\ L_{5,2} & Otherwise \end{cases} \quad (12)$$

As for the scale estimator, it is also calculated based on the same condition, but the estimator here is multiplied by a constant or correction factor as follows:

$$L_{FCH} = \begin{cases} \frac{MED\left(MD_i^2((L_{5,1}, S_{5,1}))\right)}{\chi^2_{(p,0.5)}} \times S_{5,1}, & if \ \sqrt{|S_{5,1}|} < \sqrt{|S_{5,2}|} \\ \frac{MED\left(MD_i^2((L_{5,2}, S_{5,2}))\right)}{\chi^2_{(p,0.5)}} \times S_{5,2}, & \text{Otherwise} \end{cases} \quad (13)$$

4- RFCH estimator: After obtaining the two $(L_{1,RFCH}S_{1,RFCH})$ from the previous step, (MD) is calculated again for the entire observation and then we repeat the previous step to find $(L_{2,RFCH})$ Taking into account the correction factor and my agencies:

$$L_{2,RFCH} = \frac{MED\left(MD_i^2((L_{1,RFCH}, S_{1,RFCH}))\right)}{\chi^2_{(p,0.975)}} \times S_{1,RFCH} \quad (14)$$

5- RMVN estimators: The algorithm of this method first seeks to find a new matrix of observations for the variables based on the previous location and measurement estimator, as follows:

Let's $U^0 = \sum_{j=1}^{n_1} X_K$

Where:

$$X_K = \left\{X_K : MD_i(L_{2,RFCH}, S_{2,RFCH}) \leq MED\left(MD_i(L_{2,RFCH}, S_{2,RFCH})\right)\right\}$$

K=1,…,n_1

Let's $Q^{(1)} = min\{0.5 \times 0.975 \times n/U^0, 0.995\}$

To get the first estimator for scale:

$$S^{(1)}_{RMVN} = \frac{MED\left(D_i^2(L_{RFCH}, S_{RFCH}))\right)}{\chi^2_{(p,Q^{(1)})}}$$

6- Strengthening Mahalanobis distance law by substituting the fortified location and measurement estimators instead of the arithmetic mean, covariance and covariance matrix, as follows:

$$RMD_i = (\boldsymbol{x_i} - L_{RMVN})' S^{(2)}_{RMVN}{}^{-1} (\boldsymbol{x_i} - T_{RMVN})$$

- After that, a critical value is found according to the following method:
- The following chi-square value $x^2_{(p,0.975)}$is adopted, where (p) is the number of variables.
- take the sum of the RMD values whose values are less than or equal to the critical value, and we symbolize this sum with the symbol (V1).
- Compute (ϑ):$\vartheta = (0.5 * 0.975 * n)/V_1$

Where n is the size of sample.

Now compering (ϑ) with (0.975)

If $\vartheta < 0.975$ then it will be a new critical value as $(x^2_{(p,\vartheta)})$

- weights are calculated through the ratio:

$$w_i = \sqrt[2]{\frac{x^2_{(p,\vartheta)}}{RMD_i}}$$

Weights will be obtained by the number of rows, and each row will be multiplied by its corresponding weight, and the new weighted matrix will be named (w_x)

- e- After that, the traditional method is applied to the weighted data after reducing the effect of the outliers by multiplying the outlier value by a certain weight, so the robust linear discriminant function (RLDAF) becomes as follows:
- After that, the traditional method is applied to the weighted data after reducing the effect of the outliers by multiplying the outlier value by a certain weight, so the robust linear discriminant function (RLDAF) becomes as follows:

$$Z_R = x' s^{-1}(\bar{x}_1 - \bar{x}_2)$$

3-The applied side

3.1 Real data

A random sample was taken from a group of publicly traded financial banks.

The data was taken for a period of 8 years, and the number of explanatory variables was (28), which are the financial ratios mentioned in the reports and financial statements.

First, the median value of each financial ratio was calculated as the table:

Variables	Median	variables	median	variables	Median
X_1	0.039	X_{13}	0.071	X_{25}	0.629
X_2	0.067	X_{14}	0.067	X_{26}	0.043
X_3	0.629	X_{15}	0.088	X_{27}	1.400
X_4	8.350	X_{16}	0.094	X_{28}	8.199
X_5	2.526	X_{17}	0.297		
X_6	1.090	X_{18}	0.470		
X_7	0.081	X_{19}	0.106		
X_8	0.825	X_{20}	0.423		
X_9	0.975	X_{21}	0.629		
X_{10}	0.715	X_{22}	0.962		
X_{11}	0.790	X_{23}	0.534		
X_{12}	0.131	X_{24}	0.933		

Stepwise forward selection:

In this procedure use some tests to get the best discrimination equation this test is (wilks lambda, f test, P.value).

From the result of this tests, we can determine the important variables in the model:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{28} X_{28} \qquad (15)$$

Statistical analysis was conducted using the R programming language and the results of the aforementioned tests were obtained. The results contained 6 variables, which are the best and most important among the financial ratios, as in the following table (3-1):

table (3-1): selective variables

Important variables	Wilks. Test	F. test	P.value
X_6	0.5723	14.64777	1.04E-10
X_7	0.548405	13.31274	5.94E-11
X_{15}	0.727093	12.51133	5.17E-07
X_{20}	0.793054	13.1779	8.22E-06
X_{27}	0.643386	13.71833	6.17E-09
X_{28}	0.854149	17.41712	6.33E-05

After observing the results of the analysis, there are (6) financial ratios that are the most important explanatory variables in the discriminant equation. The highest value of the Wilks' Lambda statistic was for the variable (X28), which is the leverage index, as it reached (0.854149), and thus it is considered one of the most powerful variables in the discriminant model. Then comes Followed by the rest of the variables in order of importance.

Thus, the main model becomes according to the important variables according to the following formula:

$$Y = \beta_6 X_6 + \beta_7 X_7 + \beta_{15} X_{15} + \beta_{20} X_{20} + \beta_{27} X_{27} + \beta_{28} X_{28} \qquad (16)$$

Where $(\beta_6, \beta_7, \beta_{15}, \beta_{20}, \beta_{27}, \beta_{28})$ is the discriminant coefficient.

3.2 Estimating the parameters and medians:

After selecting the important variables, the medians of each indicator were calculated for **distressed** and non- **distressed** banks, as well as the discriminatory coefficients for the discriminant model, as shown in the table (3-2):

table (3-2):value of parameters

Best variables	The estimators	The med. For Distressed banks	The med. For non-distressed banks
X_6	-15.512	0.217	0.205
X_7	8.669	0.224	0.215
X_{15}	133.536	0.014	0.017
X_{20}	-1.452	0.415	0.244
X_{27}	-2.944	0.479	0.457
X_{28}	0.158	1.996	7.91

The model after estimating parameters is:

$$Y = -15.512X_6 + 8.669X_7 + 133.536X_{15} - 1.452X_{20} - 2.944X_{27} + 0.158X_{28}$$

It is clear that the variable (X_{15}) has the largest discriminatory coefficient with a value of (133.536), and this indicates that it is the most influential variable in the equation as it is directly proportional. As for the variable (X_6), it has the largest discriminatory coefficient with a negative value (-15.512), so its effect on the predictive value Inversely proportional.

3.3 Cut-off point:

The cut-off point is considered the last forecasting step in classifying new observations into one of the two groups (distressed and non-distressed) and can be calculated from the formula:

$$L = \frac{\bar{y}_1 + \bar{y}_2}{2} \qquad (17)$$

Where $(\bar{y}_1, \bar{y}_2)$ is the means of the variables to first and second groups respectively.

The results of statistical analysis give the prediction values for the robust discriminant function as the table:

Table (3-3): Predictive values of the discriminant model and classification of banks into two groups based on the medians of the financial indicators of the Bank of

Table 1: the prediction case

Classification	Discriminant Values	Year	banks	Class.	Discrim. Values	Year	banks	Class.	Discrim. Values	Year	banks	Class.	Discrim. Values	Year	banks
F	-1.084	2016	Estithmar	S	0.195	2014		F	-0.82	2012		F	-0.87	2010	
S	-0.252	2017		S	0.634	2015	Ahly	F	-0.929	2013		F	-1.029	2011	
S	-0.009	2010		S	0.686	2016		F	-0.573	2014	AL Iraqi	F	-1.069	2012	
S	0.082	2011		F	-0.533	2017		F	-0.79	2015		S	-0.232	2013	
S	0.05	2012	Khaleej	F	-0.887	2010		F	-2.729	2016		S	0.051	2014	Babel
S	0.464	2013		S	0.08	2011		F	-1.374	2017		S	0.144	2015	
S	-0.07	2014		S	0.546	2012		F	-1.033	2010		F	-1.864	2016	
S	0.01	2015		S	0.342	2013	Ashure	F	-1.719	2011		F	-0.922	2017	
F	-0.487	2016		S	2.189	2014		S	-0.138	2012		F	-2.283	2010	
S	0.054	2017		S	2.676	2015		F	-1.826	2013	Mansour	F	-1.454	2011	
S	1.199	2010		S	0.31	2016		F	-1.186	2014		F	-1.545	2012	
S	2.392	2011		S	1.361	2017		F	-1.706	2015		F	-2.02	2013	Baghdad
S	1.553	2012		F	-0.881	2010		F	-2.09	2016		F	-1.842	2014	
S	3.325	2013	Tijary	F	-0.709	2011		F	-2.348	2017		F	-0.752	2015	
S	2.753	2014		S	-0.046	2012		S	0.571	2010		F	-2.058	2016	
S	1.519	2015		S	1.238	2013	Etiman	S	0.727	2011		F	-1.185	2017	
S	1.078	2016		S	1.155	2014		S	1.977	2012	Sumer	S	0.183	2010	
F	-1.531	2017		S	0.735	2015		S	0.204	2013		S	-0.326	2011	
F	-1.651	2010		S	-0.362	2016		S	0.037	2014		F	-1.826	2012	
S	2.446	2011		S	0.264	2017		S	0.713	2015		F	-1.444	2013	Elaf
S	2.053	2012	Watany	F	-0.829	2010		S	0.959	2016		F	-2.191	2014	
S	0.818	2013		F	-0.839	2011	Estithmar	S	0.562	2017		S	0.05	2015	
S	0.932	2014		S	0.977	2012		S	2.496	2010		F	-0.82	2016	
S	2.176	2015		S	0.605	2013		S	2.738	2011		F	-0.909	2017	
S	1.757	2016		S	0.857	2014		S	1.497	2012	Ahly	S	0.016	2010	
S	0.717	2017		S	1.526	2015		S	1.408	2013		S	-0.012	2011	AL Iraqi

The table above shows the prediction values and cases of non-distressed and failure for the banks under study and according to the years of study, as there are eight cases for each bank in the number of years, and these values show the status of the bank in that year regarding the case of failure and non-distressed (F, S).

We note that the average of the discriminating values for the distressed banks is (-1.137), while for the second group, which is the non-distressed banks, the average is (0.711), and thus the cut-off point is as follows:

$$\frac{\bar{y}_1 + \bar{y}_2}{2} = \frac{0.711 + (-1.137)}{2} = -0.213$$

Any estimated value that passes this point is considered non distressful, otherwise it is failed.

3.4 Classification errors

Classification error is an important indicator to measure the efficiency of the discriminative model and aims to obtain the lowest classification error.

Classification error is the probability that a view is classified into the first group but belongs to the second group

The previous table shows the predictive values and non-distressed cases for the banks under study, according to the years of study, where there are eight cases for each bank, giving a detailed picture of the failure and non-distressed cases (F, S) for each year.

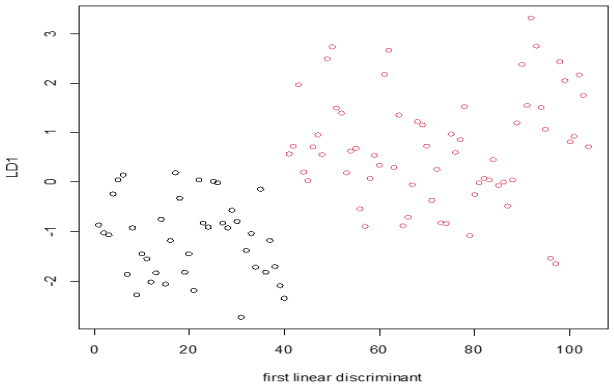
For example, the Iraqi Bank, which was initially classified as faltering depending on the broker, has non distressed cases in some years, such as (2013-2014-2015), while the rest of the years resulted in failure (F). Another example: Let us take the commercial bank that was classified as non-faltering, but it had failures in the years (2010-2011). This is also the case in the rest of the banks. There are cases of matching classification and cases of mismatch, so we generate what is called a classification error. The table below shows the number of success and failing cases that occurred. The results of the discriminant analysis of the model appeared in the previous table (3-4):

Table (3-4) represents the matching cases of default and non-default for the total banks of the study sample

		Hypothesis
Not failure	**failure**	**Prediction**
9	**31**	**Failure**
54	**10**	**Not failure**

The classification error rate is (0.18), which is a certified percentage and gives accurate results in classification and prediction, while the accuracy of the classification is (0.82), and to obtain a match in the classification, the cases that were classified as distressed and are actually distressed are collected, which amount to (31) cases, with the non-distressed cases, which are It was also classified as non-faltering, which is (54), then dividing the result by the total number of views, and the result is

31+54=85/104= 0.82



Shape (3-1): distribute the data into two groups

The figure above shows the spread of new data after classification into two groups. It is clear that there is a slight overlap between the data, which led to the emergence of an intersection area between the two groups that expresses the presence of a classification error, which is very low. This means that the validity of the classification is more accurate in this model.

3.5 **Testing the power of the discriminatory function**

To measure the strength of the discriminant function to distinguish between the observations under study, the Wilks' Lambda test is used.

A hypothesis is developed to test the equality of the means of the groups under study. The null hypothesis indicates that there are no statistically significant differences between the means, i.e. the inability of the function to discriminate. As for the alternative hypothesis, it indicates the presence of significant differences between the means, i.e. the ability of the function to discriminate, as follows:

$H_0: \mu_1 = \mu_2$

$H_1: \mu_1 \neq \mu_2$

The results shows that Wilks' Lambda value is (0.5331) and P.value is (0.000000000611) so we reject the null hypothesis and that's mean the discriminant function has a hi ability to discriminant and classification.

3.6. Conclusions

This study was based on how to use the robust discriminant analysis method for the purposes of classification and prediction of financial default for a sample of leading banks in the Iraqi Stock Exchange, through the formation of a robust matrix consisting of the robust location and measurement parameters to resist the influence of outliers.

Contaminated data were weighted with appropriate weights to reduce the impact of outliers on the final results. The discriminant analysis method depends on dividing the study sample into two or more groups of observations, and the Iraqi Stock Exchange has not officially announced its decision to classify banks in terms of faltering or not. Therefore, a new method has been proposed to classify the banks of the study sample, starting with two groups (faltering and non-distressed). Distressed) by relying on a robust statistical measure, which is the median of the values of financial indicators, to eliminate the effect of outliers using an innovative algorithm. The proposed methods and techniques were applied to a random sample of data taken from the Iraqi Stock Exchange for a group of the most widely traded banks.

The results of the analysis showed that this method was highly successful in classifying the study sample into two groups and predicting the classification of new observations into the group that fits them according to the approved cut-off point, with a significant classification error of (0.188), which corresponds to a high classification accuracy of (0.822), which suggests the last method to be used. In classification and forecasting of studies concerned with studying the Iraqi financial market.

3.7. Recommendations

Based on the results of the study, we recommend the following:

1- This study achieved accurate results in classifying and predicting financial distress, so we recommend that researchers take advantage of the methods mentioned in this study.

2- We recommend that researchers in statistical studies use robust standards and methods because they give accurate and efficient results in the event of abnormal values.

3- We recommend that banks and financial institutions focus on the important financial indicators and ratios that were classified as influential variables in this study.

4- We recommend that the competent authorities in the Iraq Stock Exchange provide the financial data and ratios that were used in this study until the year 2023 so that researchers can analyze studies related to financial analysis models.

References

[1] Beavert, W., Journal of Accounting Research ,1966, Financial ratios as predictors of failure. Empirical Research in Accounting.
[2] D. J. Hand, Biometrics, Vol. 39, No. 3, ,1983, A Comparison of Two Methods of Discriminant Analysis Applied to Binary Data.
[3] Edward I. Altman, the Journal of Finance, Vol. 23, N,1968," Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy".
[4] Hampel, F. R. ,1974" The influence curve and its role in robust estimation", Journal of the American Statistical Association.
[5] Huber, P. J. ,1981 "Robust Statistics", New York: John Wiley and Sons.
[6] Huber. J Beter, Der University of Bayreuth, 1977, Germany " Robust Statistical Procedures".
[7] Hyal.k.2023," The high robust statistical analysis with an application".

[8] R.A. Fisher Sc.D., F.R.S.1936."The use of multiple measurements in taxonomic problems".
[9] Richard J. Beckman and Mark E. Johnson, ", Journal of the American Statistical Association, Vol. 76, No. 375,1981," A Ranking Procedure for Partial Discriminant Analysis".
[10] Uraibi, H. S., Midi, H., & Rana, S. (2017). J. Science Asia, 43(1), 56-60," Robust multivariate least angle regression ".
[11] B. Klaus and P. Horn, Robot Vision. Cambridge, MA: MIT Press, 1986.