Information Sciences Letters
An International Journal

http://dx.doi.org/10.18576/isl/120320

Using the Linear Discriminant Analysis Method to Classify Types of Bowels and Esophageal cancer in Jordan

Adeeb Ahmed AL Rahamneh[1,*], *Sami Senyal Jresat*[2], *Faraj Zubaidi*[3] *and Sulieman Ibraheem Shelash Al-Hawary*[4]

[1]Department of Economics, Faculty of Business, Al-Balqa Applied University, Salt, Jordan
[2]Department of Finance and Management, Faculty of Human Sciences, Al-Balqa Applied University, Salt, Jordan
[3]Health Management Department, Batterjee Medical College, Aseer, Saudi Arabia
[4]Department of Business Administration, School of Business, Al al-Bayt University, Mafraq, Jordan

Received: 24 Jun. 2022, Revised: 21 Sep. 2022, Accepted: 5 Nov. 2022
Published online: 1 Mar. 2023

Abstract: The research aims at achieving the best linear model to distinguish between two types of Bowels and Esophageal cancer in Jordan, using the methodofdiscriminant analysis, the SPSS program was used to analyze the data. The study concluded a number of results, the most prominent of which were: the variables sex (x_1), weight (x_3), and Platelets Count P.C (x_8) which have a significant impact in constructing the discriminatory function. The probability of correct classification of a disease belonging to the first group was equal to (62.8%) and to the second group was equal to (77%). The probability of misclassification in the first group, was equal to (37.2%), and for the second group was (23%), the overall correct classification ratio (71.6%) and the false classification ratio (28.4%),the probability of correct classification of a disease belonging to the first group was equal to (66.4%) and the second group was equal to (77.6%). It was noted that the discriminant analysis method was able to identify the most important independent variables in the diagnosis of both types of Bowel and Esophageal cancer.

Keywords: Discriminant analysis, classification, misclassification, Bowelcancer, Esophageal cancer

1 Introduction

Discriminant analysis is one of the most important multivariate statistical analysis methods used in the field of analyzing classified data. The use of discriminant analysis requires that the data of explanatory variables have a multivariate normal distribution. The discriminant analysis aims at categorizing the observations into their correct groups with the lowest possible classification error. The research aims achieving the best linear model, using the method of discriminant analysis to distinguish between two types of cancer, which are:
- Bowel cancer
- Esophageal cancer

2 A linear discriminant function for two populations that follow a multivariate normal distribution

Assuming we have two populations (G_0, G_1) and each has a probability density function $f(\underline{x}/G_0)$, $f(\underline{x}/G_1)$ respectively, and assuming that (π_0, π_1) is Prior probability and that observation $(\underline{x})$ comes from (G_0, G_1) respectively, assuming that the probability density functions follow a multivariate normal distribution under the condition of equal variance and covariance matrices$(\Sigma_0, \Sigma_1)(\underline{x} - \mu_0)$ [1,2].

* Corresponding author e-mail: dr.adeeb@bau.edu.jo

Also, the ratio of the two probability density functions:

$$h(\underline{x}) = \frac{\frac{\pi_0}{(2\pi)^{n/2}[\Sigma]^{1/2}} exp[-1/2(\underline{x}-\underline{\mu}_0)\acute{)}\Sigma^{-1}(\underline{x}-\underline{\mu}_0)]}{\frac{\pi_1}{(2\pi)^{n/2}[\Sigma]^{1/2}} exp[-1/2(\underline{x}-\underline{\mu}_1)\acute{)}\Sigma^{-1}(\underline{x}-\underline{\mu}_1)]} \tag{1}$$

$$h(\underline{x}) = exp\{-\frac{1}{2}[-2\acute{\underline{\mu}}_{0\Sigma^{-1}}\underline{x}+2\acute{\underline{\mu}}_{1\Sigma^{-1}}\underline{x}+\acute{\underline{\mu}}_{0\Sigma^{-1}}\underline{\mu}_0-\acute{\underline{\mu}}_{1\Sigma^{-1}}\underline{\mu}_1]\} \times \frac{\pi_0}{\pi_1} \tag{2}$$

Adding and subtracting an $(\acute{\underline{\mu}}_{1\Sigma^{-1}}\underline{\mu}_1)$ for the right-hand side of the equation 2 gives:

$$h(\underline{x}) = exp[(\underline{\mu}_0-\underline{\mu}_1)\acute{)}\Sigma^{-1}\underline{x}-\frac{1}{2}(\underline{\mu}_0+\underline{\mu}_1)\acute{)}\Sigma^{-1}(\underline{\mu}_0-\underline{\mu}_1)] \times \frac{\pi_0}{\pi_1} \tag{3}$$

$$Lnh(\underline{x}) = [(\underline{\mu}_0-\underline{\mu}_1)\acute{)}\Sigma^{-1}\underline{x}-\frac{1}{2}(\underline{\mu}_0+\underline{\mu}_1)\acute{)}\Sigma^{-1}(\underline{\mu}_0-\underline{\mu}_1)+Ln\frac{\pi_0}{\pi_1}] \tag{4}$$

Where the first term of the equation 4 Represents the Fisher linear discrimination function, the second term is at the point of separation between the two groups (G_0, G_1) as for the unknown parameters $(\underline{\mu}_0, \underline{\mu}_1, \Sigma)$ in equation 4, it is estimated from the sample data using the maximum likelihood method where the best estimate of $\underline{\mu}_0$ is $\underline{\bar{x}_0}$ and $\underline{\mu}_1$ is $\underline{\bar{x}_1}$. Based on that:

$$\underline{\bar{x}_0} = \frac{\sum_{i=1}^{n_0}\underline{x}_{0i}}{n_0} \tag{5}$$

$$\underline{\bar{x}_1} = \frac{\sum_{i=1}^{n_1}\underline{x}_{1i}}{n_1} \tag{6}$$

As for the combined variance and covariance matrixΣ, it is estimated by the combined variance and covariance matrix of the sampleS_p according to the following formula [3,4]

$$S_p = \frac{(n_0-1)S_0+(n_1-1)S_1}{n_0+n_1-2} \tag{7}$$

Where: S_0, S_1: The estimated variance and covariance matrices for the two groups G_0, G_1. π_0, π_1: are estimated in two ways, or they are assumed to be equal $(\pi_0 = \pi_1)$ to the two populations or estimated as follows: [5,6]

$$\hat{\pi_0} = \frac{n_0}{n}, \hat{\pi_1} = \frac{n_1}{n} \tag{8}$$

Where: n_0: the size of the sample drawn from the population G_0, n_1: the size of the sample drawn from the population G_1, $n = n_0 + n_1$. Substituting the unknown parameters into equation 4, we get the statistic:

$$W = (\underline{\bar{x}_0}-\underline{\bar{x}_1})\acute{)}S_p^{-1}\underline{x}-\frac{1}{2}(\underline{\bar{x}_0}+\underline{\bar{x}_1})\acute{)}S_p^{-1}(\underline{\bar{x}_0}-\underline{\bar{x}_1})+Ln\frac{\hat{\pi_0}}{\hat{\pi_1}} \tag{9}$$

And in case $(\pi_0 = \pi_1)$, then the equation 9 becomes:

$$W^* = (\underline{\bar{x}_0}-\underline{\bar{x}_1})\acute{)}S_p^{-1}x-\frac{1}{2}(\underline{\bar{x}_0}+\underline{\bar{x}_1})\acute{)}S_p^{-1}(\underline{\bar{x}_0}-\underline{\bar{x}_1}) \tag{10}$$

The classification rule would be to classify the observation (x) to the group (G_0) if W> zero or $W^* >$ zero and to the group (G_1) if W$\leq$ zero or $W^* \leq$ zero

3 A linear discriminant function according to Mahalanobis distance

There is a possibility to find a linear discriminant function if we have (K) of groups by expanding Mahalanobis distance (D_i^2), where: [7,8]

$$\mathrm{D}_i^2(\underline{x}) = (\underline{x}-\underline{\bar{x}_i})\acute{)}S_p^{-1}(\underline{x}-\underline{\bar{x}_i}) \tag{11}$$

where: $i = 1, 2, 3, ..., K$

Equation 11 can be rewritten as follows:

$$\underline{\acute{x}}S_p^{-1}\underline{x} - 2\underline{\bar{x}}_i S_p^{-1}\underline{x} + \underline{\bar{x}}_i S_p^{-1}\underline{\bar{x}}_i \tag{12}$$

The first term on the right side of the equation 12, it is constant for all groups,the second term is a linear function in the vector $(\underline{x})$ as for the last term it does not depend on the vector $(\underline{x})$. Therefore, the first term of equation 11 can be neglected and a linear classification function can be obtained and expressed as $d_i^2(\underline{x})$ in order for this formula to be compatible with the linear discrimination function, this function is multiplied by the expression $(-\frac{1}{2})$so, the linear discriminant function is as follows [9]

$$d_i(\underline{x}) = \underline{\bar{x}}_i S_P^{-1}\underline{x} - \frac{1}{2}\underline{\bar{x}}_i S_p^{-1}\underline{\bar{x}}_i \tag{13}$$

Assuming that the probability density functions for each group follow a multivariate normal distribution under the condition of equal variance and covariance matrices and with initial probabilities $(\pi_1, \pi_2, ..., \pi_k)$.

$$d_i^*(\underline{x}) = Ln\hat{\pi}_i + \underline{\bar{x}}_i S_p^{-1}\underline{x} - \frac{1}{2}\underline{\bar{x}}_i S_p^{-1}\underline{\bar{x}}_i \tag{14}$$

The rule of classification is to classify the observation $(\underline{x})$ to set (i) if the value of the function $d_i(\underline{x})$ as big as possible from the rest of the other groups and using the same method in the case of the probability density functions, they follow a multivariate normal distribution then the observation $(\underline{x})$. It is classified as belonging to group (i) if the value of the function $d_i^*(\underline{x})$ as big as possible from the rest of the other groups.

4 Probabilistic formula for a linear discriminant function

An alternative formula for the linear discrimination function can be derived, known as the probabilistic formula, according to the following formula [8,10]:

Assuming we have two groups $(p.d.f)$ has a probability density function (G_0, G_1) $p(\underline{x}/G_0)$, $p(\underline{x}/G_1)$ and assuming that (π_0, pi_1) these are initial probabilities and that is the vector of the observation $(\underline{x})$ come from the population (G_0, G_1) respectively, then the subsequent distribution that vector $(\underline{x})$ comes from the group (G_1) be: [11,12]

$$P(G_1/\underline{x}) = \frac{\pi_1 f(\underline{x}/G_1)}{\pi_0 f(\underline{x}/G_0) + \pi_1 f(\underline{x}/G_1)} \tag{15}$$

Assuming that the probability density functions $f(\underline{x}/G_0), f(\underline{x}/G_1)$ they are normally distributed multivariate, assuming the covariance and covariance matrices are equal (Σ_0, Σ_1)

$$P(G_1/\underline{x}) = \frac{1}{1 + \frac{\pi_0}{\pi_1} \times \frac{exp[-\frac{1}{2}(\underline{x}-\underline{\mu}_0)\acute{)}\Sigma^{-1}(\underline{x}-\underline{\mu}_0)]}{exp[-\frac{1}{2}(\underline{x}-\underline{\mu}_1)\acute{)}\Sigma^{-1}(\underline{x}-\underline{\mu}_1)]}}$$

$$= \frac{1}{1 + \frac{\pi_0}{\pi_1} \times exp[-\frac{1}{2}(\underline{x}-\underline{\mu}_0)\acute{)}\Sigma^{-1}(\underline{x}-\underline{\mu}_0) + \frac{1}{2}(\underline{x}-\underline{\mu}_1)\acute{)}\Sigma^{-1}(\underline{x}-\underline{\mu}_1)]}$$

Using algebraic methods then:

$$P(G_1/\underline{x}) = \frac{1}{1 + \{e^{log\frac{\pi_0}{\pi_1}\frac{1}{2}(\underline{\mu}_1+\underline{\mu}_0)\acute{)}\Sigma^{-1}(\underline{\mu}_1-\underline{\mu}_0)+(\underline{\mu}_1-\underline{\mu}0)\Sigma^{-1}\underline{x}}\}^{-1}} \tag{16}$$

Note that the equation 16 can be rewritten as follows:

$$P(G_1/\underline{x}) = \frac{1}{1 + \{e^{\alpha+\underline{\alpha}\underline{x}}\}^{-1}} \tag{17}$$

Where:

$$\alpha = log\frac{\pi_0}{\pi_1} - \frac{1}{2}(\underline{\mu}_1 + \underline{\mu}_0)\acute{)}\Sigma^{-1}(\underline{\mu}_1 - \underline{\mu}_0) \tag{18}$$

$$\underline{\beta} = (\underline{\mu}_1 - \underline{\mu}_0)\acute{)}\Sigma^{-1}\underline{x} \tag{19}$$

$(\underline{\mu}_0, \underline{\mu}_1, \Sigma, \pi_0, \pi_1)$, are unknown parameters, are estimated from the sample data through formulas 5, 6, 7, 8 respectively.

Based on the above the estimate of the probability of observation (x) belong to group (G_1) is:

$$\hat{p}(G_1/\underline{x}) = \frac{1}{1+\{e^{\hat{\alpha}+\underline{\hat{\beta}}\underline{x}}\}^{-1}} \tag{20}$$

The estimation of the probability of observations belong to group (G_0) is:

$$\hat{p}(G_0/\underline{x}) = 1 - \hat{p}(G_1/\underline{x}) \tag{21}$$

The classification rule based on the formula of the probability function in the event that we have two groups is as follows:[13,14]

The observation $\underline{x}$ belongs to group G_0 if $\hat{p}(G_0/\underline{x}) > \hat{p}(G_1/\underline{x})$

The observation *underlinex* belongs to group G_0 if $\hat{p}(G_0/\underline{x}) < \hat{p}(G_1/\underline{x})$

5 The practical side

This study was based on a sample of (200) cancer patients, including (100) patients with bowel cancer and (100) patients with esophageal cancer from a hospital in Jordan, the incidence of this disease was considered a dependent variable, and coding was adopted (1) for a patient with bowel cancer and (2) for a patient with esophageal cancer as for the related independent variables, the following variables were adopted:

Sex (x_1): (1) for male and (2) for female. Age (x_2): (1) for $<$49 and (2) for $\geq$50. Weigh (x_3): (1) for $<$39 and (2) for$\geq$39. Packed Cell Volume (x_4) : (1) for $\leq$35, and (2) for $>$35. Hemoglobin (x_5): (1) for $\leq$11.5 and (2) for $>$11.5. White Blood Cell Count (x_6): (1) for $\leq$7.5 and (2) for $>$7.5. Erythrocyte Sedimentation Rate (x_7): (1) for $\leq$22.5 and (2) for $>$22.5. Platelets Count (x_8): (1) for $\leq$150 and (2) for $>$150.

Table 1: The results of the Kolmogorov-Smirnov test for a normal distribution of data

	Kolmogorov-Smirnov		
	Statistic	d.f	Sig.
Sex (x_1)	2.311	200	0.000
Age (x_2)	0.161	200	1.010
Weigh (x_3)	1.601	200	0.015
Packed Cell Volume P.C.V (x_4)	0.152	200	1.013
Hemoglobin H.B (x_5)	0.001	200	1.012
White Blood Cell Count W.B.C (x_6)	0.481	200	0.861
Erythrocyte Sedimentation Rate E.S.R (x_7)	0.327	200	1.014
Platelets Count P.C (x_8)	1.735	200	0.096

It is noted from the results in Table 1 that most of the variables are distributed normally, except for the variables of sex (x_1) and weight (x_3), because the sample size is large $(n > 30)$ the sample data can be considered to follow a normal distribution based on the central limit theorem. In order to form a discriminatory function with statistical significance, the significance of the differences between the averages of the two groups related to the study was tested, as follows:

$H_0 : \mu_0 = \mu_1$

$H_1 : \mu_0 \neq \mu_1$

According to the Wilk scale and chi-square statistic (χ^2),the results highlighted, as shown in Table 2, there are significant differences between the two means, which means the emergence of two groups of patients, and this means that the discriminatory function has the ability to separate the two groups and thus classify any new observation into one of these two groups.

Table 2: Discriminant function significance test

Test Function	Wilks' (Lambda)	Chi-Square	df	Sig.
1	0.71	60.413	8	0.000

The homogeneity test between the two groups.
$H_0 : \Sigma_0 = \Sigma_1$
$H_1 : \Sigma_0 \neq \Sigma_1$

Table 3: Homogeneity Test between the two groups

Box's M	28.211
F Approx	1.401
Df1	36
Sig	0.158

It can be seen from the results of table 3 that Sig.>0.05. Therefore, the decision is rejected H_1 and accepted H_0 That is, the condition for using the linear discriminant function is fulfilled. In order to obtain two distinguishing functions for the two groups, the significance of the independent variables was tested to know the importance of each variable and its impact on the dependent variable. This is evident from the results in Table 4

Table 4: The significance of the independent variables test

	Wilks' Lambda	F	df1	df1	Sig.
Sex (x_1)	0.738	38.604	1	198	0.013
Age (x_2)	0.893	0.113	1	198	0.732
Weigh (x_3)	0.916	11.992	1	198	0.004
Packed Cell Volume P.C.V (x_4)	0.935	0.119	1	198	1.644
Hemoglobin H.B (x_5)	1.003	0.000	1	198	1.053
White Blood Cell Count W.B.C (x_6)	0.976	1.227	1	198	0.287
Erythrocyte Sedimentation Rate E.S.R (x_7)	0.989	4.001	1	198	0.492
Platelets Count P.C (x_8)	0.93.6	11.996	1	198	0.003

It is noted from the results presented in table 4 that the variables sex (x_1), weight (x_3), and Platelets Count P.C (x_8) have a significant impact in constructing the discriminatory function, while the rest of the variables did not show any significant impact.

The two samples mean vectors and the combined variance and covariance matrix were estimated using Maximum Likelihood estimators according to formulas 5, 6, and 7. The two linear discrimination functions for the two groups were estimated according to the formula 14 and the following results are in Table 5

Table 5: linear discriminant functions

	First discriminant function	Second discriminant function
Constant	-56.847	-67.204
Sex (x_1)	17.993	22.003
Age (x_2)	9.701	10.119
Weigh (x_3)	7.788	8.993
Packed Cell Volume P.C.V (x_4)	11.533	14.345
Hemoglobin H.B (x_5)	-1.092	-1.873
White Blood Cell Count W.B.C (x_6)	11.567	12.002
Erythrocyte Sedimentation Rate E.S.R (x_7)	9.001	8.997
Platelets Count P.C (x_8)	6.488	5.008

It is noted from the results in Table 5 that the following variables are: Age, Hemoglobin, and White Blood Cell Count have the same coefficient values in both functions, which means that the three variables mentioned will not have a role in the process of classifying the variables. The results of data classification are shown below by the sensory functions of the linear discrimination functions, as shown in Table 6

Table 6: The results of classifying the observations by the classification of the linear classification functions

Case	classification		
	The patient belongs to the first group	The patient belongs to the second group	Correct classification ratio
The patient belongs to the first group	55	25	62.80%
The patient belongs to the second group	18	60	77%
Overall classification ratio	44.70%	56.30%	71.60%

It is evident from Table 6 that the probability of correct classification of a disease belonging to the first group was equal (62.8%) and belonging to the second group was equal (77%).As for the probability of misclassification into the first group, it is equal (37.2%),for the second group (23%). It was observed that the overall correct classification ratio (71.6%) and the false classification ratio (28.4%). The data was also classified using the probabilistic formulas with numbers 20, and 21 as an alternative method for classification, and the results were as shown in Table 7

Table 7: Shows the classification of data according to the probabilistic formula of the linear discriminant function

Case	classification		
	The patient belongs to the first group	The patient belongs to the second group	Correct classification ratio
The patient belongs to the first group	56	23	66.40%
The patient belongs to the second group	21	59	77.60%
Overall classification ratio	45.20%	53.80%	72.10%

In Table 7 the results have shown that the probability of correct classification of a disease belonging to the first group was equal to (66.4%) and to the second group was equal to (77.6%).As for the probability of misclassification into the first group, it is equal (33.6%),andto the second group (27.9%).

6 Results

1.The variables sex (x_1), weigh (x_3), Platelets Count P.C (x_8) have a significant impact in constructing the discriminatory function.
2.The following variables: Age, Hemoglobin, and White Blood Cell Count have the same coefficient values in both functions, which means that the three variables mentioned will not have a role in the process of classifying the variables.
3.The probability of correct classification of a disease belonging to the first group was equal to (62.8%) and belonging to the second group was equal to (77%).
4.The probability of misclassification into the first group, is equal to (37.2%), to the second group (23%).
5.The overall correct classification ratio was (71.6%) and the false classification ratio was (28.4%).
6.The probability of correct classification of a disease belonging to the first group was equal to (66.4%) and belonging to the second group was equal to (77.6%).
7.The probability of misclassification into the first group, was equal to (33.6%), and to the second group was equal to (27.9%).
8.It was noted that the discriminant analysis method was able to identify the most important independent variables in the diagnosis of both types of Bowel and Esophageal cancer in Jordan.

Recommendations

1.Expanding the use of the discriminatory analysis method to classify data due to the effectiveness of this method.
2.Conducting extensive studies on cancer due to the high death rates caused by this disease of all kinds.

References

[1] T.W. Anderson. *An Introduction to Multivariate Statistical Analysis*. John Wiley and Sons, New York (2003).
[2] W. Hardle and Z. Hlavka. (Multivariate Statistics). Springer Science and business media LLC, Germany (2007).
[3] J.F. Hair, W.C. Black, R.E. Anderson and R.L. Tatham. *Multivariate Data Analysis*. Prentice Hall Inc., New Jersey (1998).

[4] A.C. Rencher. *Methods of Multivariate Analysis*. Wiley, New York (2012).
[5] C.A. Rencher. *Methods of Multivariate Analysis*. John Wily and Sons publication, New York (2002).
[6] B.G. Tabachnick and L.S. Fidell. *Using multivariate statistics*. Pearson, Boston (2007).
[7] A.M. Al-Bakri and H.L. Hussein. Reducing False Notification in Identifying Malicious Application Programming Interface (API) to Detect Malwares Using Artificial Neural Network with Discriminant Analysis. *Ibn Al-Haitham Journal for Pure and Applied Science*,**27**, 556-565 (2014).
[8] M.S.A NazikEhsan. Use of Discriminant function method for forecasting students result, *Iraq journal of market research and consumer protection*,**3**, 131-146 (2011).
[9] N.H. Timm. *Applied Multivariate Analysis*. Springer-Verlag Inc., New York (2002).
[10] A.F. Ahmed. A Comparative Study of Human Faces Recognition Using Principle Components Analysis and Linear Discriminant Analysis Techniques. *Journal of Engineering and Sustainable Development*,**20**, 1-12 (2016).
[11] H.M. Gorgess and A. Mohammed. Applications of Discriminant Analysis in Medical diagnosis, *Ibn Al-Haitham Journal For Pure And Applied Science*,**27**, 331-342 (2014).
[12] W.R. Klecka. *Discriminant Analysis*. Beverly Hills, London (1984).
[13] I.Abdul Rahmannoaman. Discriminant analysis for evaluation evidence Employment human in Iraq. *Diyala Journal for Pure Science*,**11**, 97-115 (2015).
[14] S. ManfiRida, R.A. Saleh and A. Abdul Latif. A comparison between the logistic regression model and Linear Discriminant analysis using Principal Component unemployment data for the province of Baghdad. *Journal of Economics And Administrative Sciences*,**23**, 367-386 (2017).