Discriminant Analysis of Bronchitis Cancer Data

Nazera Khalil Dakhil, Yahya Mahdi Al-mayali , and Gahssan Dahair Al-Thabhawee
College of Mathematics and Computer Sciences
University of Kufa
E-mail: n_dakhil@hotmail.com

Abstract
The aim of this research is to predict membership in two mutually exclusive groups of bronchitis cancer patients, and allocating new patients using Discriminant Analysis.
Discrimination is a multivariate technique concerned with separating distinct sets of objects (or observations) and with allocating new objects (observation) to previously defined groups. The results showed 90% , and 98% of dead and alive patients were classified correctly. Only 2% and 10% of dead and alive patients were misclassified

Introduction
Discriminant analysis was first introduced by R.A. Fisher in 1936 [6]. It is rather exploratory in nature. As a separative procedure, it is often used on a one-time basis in order to investigate observed differences when fundamental relationships are not well understood. Classification procedures are less exploratory in the sense that they lead to well-defined rule, which can be used for assigning new objects.

Previous studies showed the used of discriminant functions to classify three types of lesions in three groups: The normal, the benign, and the malignant. It was observed that the correctly classified carcinoma is only 42% and for normal are 100%. [7].

The main contribution of this paper is to propose a simple Fisher-type discriminant method on gene selection in microarray data.[5]
Other study compared the performance of different discrimination methods for the classification of tumors based on gene expression data.[9]

Discriminant Analysis

Classification with Two Multivariate Normal Populations

Classification procedures based on normal populations predominate in statistical practice because of simplicity and reasonably high efficiency across a wide variety of population models. We now assume that $f_1(x)$ and $f_2(x)$are multivariate normal densities, the first with mean vector $\mu_1$and covariance matrix $\sum_1$ and the second with mean vector $\mu_2$and covariance matrix $\sum_2$.
The special case of equal covariance matrices leads to a particularly simple linear classification statistic.

Classification of Normal Populations When $\sum_1 = \sum_2 = \sum$

Suppose that the joint densities of $X'=[x_1,x_2,\ldots\ldots,x_p]$ for populations π_1 and π_2 are given by

$$fi(x)=\frac{1}{(2\Pi)^{p/2}|\sum|^{1/2}}\exp[-\frac{1}{2}(x-\mu_i)'\sum^{-1}(x-\mu_i)] \text{ for } i=1,2 \quad (1)$$

suppose also that the population parameters μ_1, μ_2 and $\sum$ are known, Then, after cancellation of the terms $(2\pi)^{p/2}|\sum|^{1/2}$ the minimum ECM regions is

$$R1:\exp\left[-\frac{1}{2}(x-\mu_1)'\sum^{-1}(x-\mu_1)+\frac{1}{2}(x-\mu_2)'\sum^{-1}(x-\mu_2)\right]\geq\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

$$R2:\exp\left[-\frac{1}{2}(x-\mu_1)'\sum^{-1}(x-\mu_1+\frac{1}{2}(x-\mu_2)'\sum^{-1}(x-\mu_2)\right]<\left(\frac{c(2|1)}{c(1|2)}\right)\left(\frac{p_2}{p_1}\right) \quad (2)$$

Given these regions R_1 and R_2 , we can construct the classification rule given in the following result.
Result 1. Let the populationsπ_1 and π_2 be described by multivariate normal densities of form (1). Then the allocation rule that minimize the ECM is as follows:
Allocate x_0 to π_1 if.

$$(\mu_1-\mu_2)'\sum^{-1}x_0-\frac{1}{2}(\mu_1-\mu_2)'\sum^{-1}(\mu_1+\mu_2)\geq\ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

$$R1:(\mu_1-\mu_2)'\sum^{-1}x-\frac{1}{2}(\mu_1-\mu_2)'\sum^{-1}(\mu_1+\mu_2)\geq\ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] \quad (3)$$

Otherwise. Allocate x_0 to π_2

Proof. Since the quantities in (2-2) are nonnegative for all x, we can take their natural logarithms and preserve the order of inequalities.

$$-\frac{1}{1}(x-\mu)'\Sigma^{-1}(x-\mu_1)+\frac{1}{2}(x-\mu_2)'\Sigma^{-1}(x-\mu_2)$$
$$=(\mu_1-\mu_2)'\Sigma^{-1}x-\frac{1}{2}(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1+\mu_2) \quad (4)$$

and, consequently

$$R2:(\mu_1-\mu_2)'\Sigma^{-1}x-\frac{1}{2}(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1+\mu_2)<\ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p2}{p1}\right)\right] \quad (5)$$

The minimum ECM classification rule follows.

In most practical situations, the population quantities μ_1, and μ_2 and $\sum$ are unknown, so the rule (3) must be modified. Previous researchers have suggested replacing the population parameters by their sample counterparts.([10],[1]).

Suppose, Then, that we have n_1 observations of the multivariate random variable $X'=[x_1,x_2,\ldots\ldots,x_p]$ from π_1 and n_2 measurements of this quantity from π_2, with $n_1+n_2-2 \geq p$. Then the respective data matrices are

$$\underset{(n_1\times p)}{X_1}=\begin{bmatrix} x'_{11} \\ x'_{12} \\ \vdots \\ x'_{1n_1} \end{bmatrix} \quad (6)$$

From these data matrices, the sample mean vectors and covariance matrices are determined by

(7)

$$\underset{(p\times 1)}{\bar{x}_1}=\frac{1}{n_1}\sum_{j=1}^{n_1}x_{1j}, \qquad \underset{(p\times p)}{S_1}=\frac{1}{n_1-1}\sum_{j=1}^{n_1}(x_{1j}-\bar{x}_1)(xij-\bar{x}_1)'$$

$$\underset{(p\times 1)}{\bar{x}_2}=\frac{1}{n_2}\sum_{j=2}^{n_2}x_{2j}, \quad \underset{(p\times p)}{S_2}=\frac{1}{n_2-1}\sum_{j=1}^{n_2}(x_{2j}-\bar{x}_2)(x_{2j}-\bar{x}_2)'$$

Since it is assumed that the parent populations have the same covariance matrix $\sum$ the sample covariance matrices S_1 and S_2 are combined (pooled) to derive a single unbiased estimate of $\sum$. In particular, the weighted average

$$Spoold\left[\frac{n_1-1}{(n_1-1)+(n_2-1)}\right]S_1+\left[\frac{n_2-1}{(n_1-1)+(n_2-1)}\right]S_2 \quad (8)$$

Is an unbiased estimate of $\sum$ if the data matrices X_1 and X_2 contain random samples form the populations π_1 and π_2, respectively.

For μ_2 and S_{pooled} for $\sum$ in (3) given the $\overline{X}_2$ for μ_1 , $\overline{x}_1$ Substituting "sample" classification rule:

The Estimated Minimum ECM Rule For Two Normal Populations

Allocate X_0 to π_1 if

$$(\bar{x}_1-\bar{x}_2)'S^{-1}_{pooled}x_0-\frac{1}{2}(\bar{x}_1-\bar{x}_2)'S^{-1}_{pooled}(\bar{x}_1+\bar{x}_2)\geq\ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] \quad (9)$$

Allocate X_0 to π_2 if otherwise

If, in(9)

$$\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)=1$$

Then ln(1) = 0, and the estimated minimum ECM rule for two normal populations amounts to comparing the scalar variable

$$\hat{Y}=(\bar{x}_1-\bar{x}_2)'S^{-1}_{pooled} \quad X=\hat{a}'X \quad (10)$$

Evaluated at X_0, with the number

$$\hat{m}=\frac{1}{2}(\bar{x}_1-\bar{x}_2)'S^{-1}_{pooled}(\bar{x}_1+\bar{x}_2)$$
$$=\frac{1}{2}(\bar{y}_1+\bar{y}_2) \quad (11)$$

Where

$$\bar{y}_1=(\bar{x}_1-\bar{x}_2)'S^{-1}_{pooled}\bar{x}_1=\hat{a}'\bar{x}_1$$

and

$$\bar{y}_2=(\bar{x}_1-\bar{x}_2)'S^{-1}_{pooled}\bar{x}_2=\hat{a}'\bar{x}_2$$

That is, the estimated minimum ECM rule for two normal populations is equivalent to creating two univariate populations for the y values by taking an appropriate linear combination of the observations from populations π_1 and π_2 and then assigning a new observation x_0 to π_1 or π_2, depending upon whether $\hat{y}_0=\hat{a}'x_0$ falls to the right or left of the midpoint $\hat{m}$ Between the two univarite means $\hat{y}_1$and $\hat{y}_2$.

Once parameter estimates are inserted for the corresponding unknown population quantities, there is no assurance that the resulting rule will minimize the expected cost of misclassification in a particular application.

This is because the optimal rule in (3) was derived assuming that the multivariate normal densities $f_1(x)$ and $f_2(x)$ were known completely. Expression (9) is simply an estimate of the optimal rule. However, it seems reasonable to expect that it should perform well it the sample sizes are large.

To summarize, if the data appear to be multivariate normal. the classification statistic to the left of the inequality in (9) can be calculated for each new observation X_0. these observations are classified by comparing the values of the statistic with the value of $\ln[(c(1|2)/c(2|1))(p_2/p_1)]$.

Testing for Equality of Covariance Matrices

One of the assumptions made when comparing two or more multivariate mean vectors is that the covariance matrices of the populations are the same. Before pooling the variation across sample to from a pooled covariance matrix When comparing mean vectors, it can be worthwhile to test the equality of the population covariance matrices. One commonly used test for equal covariance matrices is Box's M-test ([2],[3])

With g population the null hypothesis is

$$H_0 = \sum_1 = \sum_2 = \cdots = \sum_g = \sum \tag{12}$$

Where $\sum_l$ is the covariance matrix for the lth population, l =1,2,…,g, and $\sum$ is the presumed common covariance matrix. The alternative hypothesis is that at least two of the covariance matrices are not equal.

Assuming multivariate normal populations, a likelihood ratio statistic for testing (12) is given by [1]

$$\Lambda = \prod_l \left(\frac{|S_l|}{|S_{pooled}|} \right)^{(n_l-1)/2} \tag{13}$$

Here n_l is the covariance matrix for lth group sample covariance matrix and S_{pooled} is the pooled sample covariance matrix given by

$$s_{pooled} = \frac{1}{\sum_l (n_l-1)} \{(n_1-1)S_1 + (n_2-1)S_2 + \cdots + (n_g-1)S_g \tag{14}$$

Box's test is based on has X^2 approximation to the sampling distribution of $-2\ln\Lambda$. Setting $-2\ln\Lambda =$ M (Box's M statistic) gives

$$M = \left[\sum_l (n_l - 1) \right] \ln|S_{pooled}| - \sum_l (n_l - 1) \ln|S_l| \tag{15}$$

If the null hypothesis is true, the individual sample covariance are not expected to differ too much and, consequently, do not differ too much from the pooled covariance matrix. In this case, the ratio of the determinants in (13) will all be close to 1, Λ will be near 1 and Box's M statistic will be small. If the null hypothesis is false, the sample covariance matrices can differ more and the differences in their determinants will be more pronounced.

Box's Test for Equality of Covariance matrices

Set

$$u = \left[\sum_l \frac{1}{(n_l - 1)} - \frac{1}{\sum_l (n_l - 1)} \right] \left[\frac{2p^2 + 3p - 1}{6(p+1)(g-1)} \right] \tag{16}$$

Where p is number of variable and g is the number of groups. Then

$$C = (1-u)M = (1-u)\left\{ \left[\sum_l (n_l - 1) \right] \ln|S_{pooled}| - \sum_l [(n_l - 1) \ln|S_l|] \right\} \tag{17}$$

Has an approximate x^2 distribution with

$$v = g\frac{1}{2}p(p+1) - \frac{1}{2}p(p+1) = \frac{1}{2}p(p-1)(g-1) \tag{18}$$

Degrees of freedom. At significance level α, reject H_0 if $C > \chi^2 p(p+1)(g-1)/2(\alpha)$.

Box's x^2 Approximation works well if each n_l exceeds 20 and if p and g do not exceed 5. In situations where these conditions do not hold , Box ([2],[4]) has provided a more precise F approximation to sampling distribution of M.

Comparing Several Multivariate Population Means (One-Wey MANOVA)

Often, more then two populations need to be compared. Random sample, collected from each of g population, are arranged as

$$\begin{aligned} &\text{Population:} && X_{11}, X_{12}, \cdots, X_{1n_1} \\ &\text{Population:} && X_{21}, X_{22}, \cdots, X_{2n_2} \\ &\vdots && \vdots \\ &\text{Population:} && X_{g1}, X_{g2}, \cdots, X_{gn_g} \end{aligned} \qquad (19)$$

MANOVA is used first to investigate whether the population mean vectors are the same and, if not, which mean components differ significantly.

Assumptions about the Structure of the Data for One-Way MANOPVA

1. $X_{l1}, X_{l2}, \cdots, X_{ln_l}$, is random sample of sizes n_l from a population with mean μ_l, $l = 1,2,\cdots, g$. The random sample from different populations are independent.

2. All populations have a common covariance matrix ∑.

3. Each population is multivariate normal.

Condition 3 can be relaxed by appealing to the central limit theorem when the sample sizes n_l arelarge.

A review of the univariate analysis of variance (ANOVA) will facilitate our discussion of the multivariate assumptions and solution methods.

A Summary of Univariate ANOVA

In the univariate situation, the assumptions are that $X_{l1}, X_{l2}, \cdots, X_{ln_l}$ is random sample from an $N(\mu_L, \sigma^2)$ population , $l = 1,2,\cdots, g$, and that the random samples are independent. Although null hypothesis of equality of means could be formulated as $\mu_1 = \mu_2 = \cdots = \mu_g$, it is customary to regard μ_l as the sum of an overall mean component, such as μ, and a component due to the specific population. For instance, we can write $\mu_l = \mu + (\mu_l - \mu)$ or $\mu_l = \mu + T_l$ where $T_l = \mu_l - \mu$.

Populations usually correspond to different sets of experimental conditions, and therefore, it is convenient to investigate the deviations T_l associated with the lth population (treatment).

The reparameterization

$$\underset{\left(l^{th}\text{ population mean}\right)}{\mu_l} = \underset{\left(overall\ mean\right)}{\mu} + \underset{\left(l^{th}\text{ population (treatment) effect}\right)}{T_l} \qquad (20)$$

Leads to a restatement of the hypothesis of equality means. The null hypothesis becomes

$$H_0 = T_1 = T_2 = \cdots T_g = 0$$

The response X_{lj}, distributed as $N(\mu + T_l, \sigma^2)$, can be expressed in the suggestive from

$$\underset{}{X_{lj}} = \underset{\left(overall\text{ mean}\right)}{\mu} + \underset{\left(\begin{matrix}treatment\\effect\end{matrix}\right)}{T_l} + \underset{\left(\begin{matrix}random\\error\end{matrix}\right)}{e_{lj}} \qquad (21)$$

Where the e_{lj} are independent N(0,σ^2) random variables. To define uniquely the model parameters and their least squares estimates, it is customary to impose the constraint $\sum_{l=1}^{g} n_l T_l = 0$.

Motivated by the decomposition in (21), the analysis of variance is based upon an analogous decomposition of the observation,

$$\underset{\left(observation\right)}{X_{lj}} = \underset{\left(\begin{matrix}overall\\mean\end{matrix}\right)}{\bar{X}} + \underset{\left(\begin{matrix}estimated\\treatment\ effect\end{matrix}\right)}{\left(\bar{X}_l - \bar{X}\right)} + \underset{\left(residual\right)}{\left(X_{lj} - \bar{X}_l\right)} \qquad (22)$$

Where $\bar{X}$ is an estimate of μ, $\hat{T}_l = (\bar{X}_l - \bar{X})$ is an estimate of T_l, and $(X_{lj} - \bar{X}_l)$ is an estimate of the error e_{lj} .

Subtracting $\bar{X}$ from both sides of (21) and squaring gives

$$\left(X_{lj} - \bar{X}\right)^2 = \left(\bar{X}_l - \bar{X}\right)^2 + \left(X_{lj} - \bar{X}_l\right)^2 + 2\left(\bar{X}_l - \bar{X}\right)\left(X_{lj} - \bar{X}_l\right)$$

We can sum both sides over j, note that $\sum_{j=1}^{n_l} (X_{lj} - \bar{X}_l) = 0$, and obtain

$$\sum_{j=1}^{n_l}\left(X_{lj}-\overline{X}\right)^2 = n_l\left(\overline{X}_l-\overline{X}\right)^2+\sum_{j=1}^{n_l}\left(X_{lj}-\overline{X}_l\right)^2$$

Next, summing both sides over l we get

$$\sum_{l=1}^{g}\sum_{j=1}^{n_l}\left(X_{lj}-\overline{X}\right)^2 = \sum_{l=1}^{g} n_l\left(\overline{X}_l-\overline{X}\right)^2 + \sum_{l=1}^{g}\sum_{j=1}^{n_l}\left(X_{lj}-\overline{X}_l\right)^2$$

$$\begin{pmatrix} SS_{cor} \\ total\ (corrected)\ SS \end{pmatrix} = \begin{pmatrix} SS_{tr} \\ between\ (samples)\ SS \end{pmatrix} + \begin{pmatrix} SS_{res} \\ within\ (samples)\ SS \end{pmatrix} \quad (23)$$

Or

$$\sum_{l=1}^{g}\sum_{j=1}^{n_l} X_{lj}^2 = (n_1+n_2,\cdots,n_g)\overline{X}^2 + \sum_{l=1}^{g} n_l\left(\overline{X}_l-\overline{X}\right)^2 + \sum_{l=1}^{g}\sum_{j=1}^{n_l}\left(X_{lj}-\overline{X}_l\right)^2$$

$$(SS_{obs}) = (SS_{mean}) + (SS_{tr}) + SS_{res} \quad (24)$$

In the course of establishing (24) ,we have verified that the arrays representing the mean, treatment effects, and residuals are orthogonal. That is, these arrays, considered as vectors, are perpendicular whatever the observation vector

$$Y' = \left[X_{11},\cdots,X_{1n_1},X_{21},\cdots,X_{2n_2},\cdots,X_{gn_g}\right]$$

. Consequently, we could obtain SS_{res} by subtraction, without having to calculate the individual residuals, because $SS_{res} = SS_{obs} - SS_{mean} - SS_{tr}$. However, this is false economy because plots of the residuals provide checks on the assumptions of the model.

The vector representations of the arrays involved in the decomposition (22) also have geometric interpretations that provide the degrees of freedom. For an arbitrary set of observations, let. The observation vector y can lie anywhere in n $=n_1 + n_2 + \ldots + n_g$ dimensions; the mean vector $\overline{X}1 = \left[\overline{X},\cdots,\overline{X}\right]'$ must lie along the equiangular line of 1, and the treatment effect vector

$$\left(\overline{X}_1-\overline{X}\right)\begin{bmatrix}1\\ \vdots\\ 1\\ 0\\ \vdots\\ 0\\ 0\\ \vdots\\ 0\end{bmatrix} + \left(\overline{X}_2-\overline{X}\right)\begin{bmatrix}0\\ \vdots\\ 0\\ 1\\ \vdots\\ 1\\ 0\\ \vdots\\ 0\end{bmatrix} + \cdots + \left(\overline{X}_g-\overline{X}\right)\begin{bmatrix}0\\ \vdots\\ 0\\ 0\\ \vdots\\ 0\\ 1\\ \vdots\\ 1\end{bmatrix}$$

$$= \left(\overline{X}_1-\overline{X}\right)u_1 + \left(\overline{X}_2-\overline{X}\right)u_2 + \cdots + \left(\overline{X}_g-\overline{X}\right)u_g$$

Lies in the hyperplane of linear combinations of the g vectors $u_1,u_2, \ldots,u_g$. Since $1 = u_1+ u_2 + \ldots +$ u_g , the mean vector also lies in hyper plane, and it is always perpendicular to the treatment vector. Thus, the mean vector has the freedom to anywhere along the one-dimensional equiangular line, and the treatment vector has the freedom to lie anywhere in the other g-1 dimensions. The residual vector, $\hat{e} = y-(\overline{X}1)-\left[\left(\overline{X}_1-\overline{X}\right)u_1+\left(\overline{X}_2-\overline{X}\right)u_2+\cdots+\left(\overline{X}_g-\overline{X}\right)u_g\right]$ is perpendicular to both the mean vector and the treatment effect vector and has the freedom to lie anywhere in the subspace of dimension n-(g-1)-1=n-g that is perpendicular to their hyperplane.

To summarize, we attribute 1 d.f. to SS_{res}. g-1 d.f. to SS_{tr}, and $n-g = (n_1+n_2+\cdots+n_g)-g$ d.f. to SS_{res}. The total number of freedom is $n = n_1+n_2+\cdots+n_g$. Alternatively , by appealing to the univariate distribution theory, we find that these are the degree of freedom for the chi-square distributions associated with the corresponding sums of squares.

The calculations of the sums of squares and the associated degree of freedom are conveniently summarized by an ANOVA table.

ANOVA Table for Comparing Univariate Population Means

Source Of variation	Sum of squares (SS)	Degree of freedom(d.f.)
Treatments	$SS_{tr} = \sum_{l=1}^{g} n_l\left(\overline{X}_l-\overline{X}\right)^2$	g – 1
Residual (error)	$SS_{res} = \sum_{l=1}^{g}\sum_{j}^{n_l}\left(X_{lj}-\overline{X}_l\right)^2$	$\sum_{l=1}^{g} n_l - g$
Total (corrected for the mean)	$SS_{cor} = \sum_{l=1}^{g}\sum_{j=1}^{n_l}\left(X_{lj}-\overline{X}\right)^2$	$\sum_{l=1}^{g} n_l - 1$

The usual F-test rejects $H_0: T_1 = T_2 = \cdots = T_g = 0$ at level α if

$$F = \frac{SS_{tr}/(g-1)}{SS_{res}\Big/\left(\sum_{l=1}^{g} n_l - g\right)} > F_{g-1,\sum n_l - g}(\alpha)$$

Where $F_{g-1,\sum n_l - g(\alpha)}$ is the upper (100α)th percentile of the F-distribution with g-1 and $\sum n_l - g(\alpha)$ degrees of freedom. This is equivalent to rejection H_0 for large values of SS_{tr}/SS_{res} or for large values of $1+SS_{tr}/SS_{res}$. The statistic appropriate for a multivariate generalization rejects H_0 for small values of the reciprocal

$$\frac{1}{1+SS_{tr}/SS_{res}} = \frac{SS_{res}}{SS_{res}+SS_{tr}} \qquad (25)$$

Multivariate Analysis of variance (MANOVA)

Paralleling the univariate reparameterization, we specify the MANOVA model:

MANOVA Model For Comparing Population Mean Vectors

$$X_{lj} = \mu + T_l + e_{lj},\ \ j = 1,2,\cdots,n_l \ \text{ and } l = 1,2,\cdots,g \qquad (26)$$

Where the e_{lj} are independent $N_p(0,\sum)$ variable. Here the parameter vector μ is an overall mean (level), and T_l represents the lth treatment effect with $\sum_{l=1}^{g} n_l T_l = 0$.

According to the model in (26), each component of the observation vector X_{lj} satisfies the univariate model (25). The errors for the components of X_{lj} are correlated, but the covariance matrix $\sum$ is the same for all populations.

A vector of observations may be decomposed as suggested by the model. Thus,

$$\begin{pmatrix} X_{lj} \\ observation \end{pmatrix} = \begin{pmatrix} \overline{X} \\ overall\ mean\ \mu \end{pmatrix} + \begin{pmatrix} (\overline{X}_l - \overline{X}) \\ estimated\ treatment\ effect\ T_l \end{pmatrix} + \begin{pmatrix} (X_{lj} - \overline{X}_l) \\ residual \\ e_{ij} \end{pmatrix} \qquad (27)$$

The decomposition in (27) leads to the multivariate analog of the univariate sum of squares breakup in (22). First we note that the product

$$(X_{lj} - \overline{X})(X_{lj} - \overline{X})'$$

Can be written as

$$\begin{aligned}(X_{lj} - \overline{X})(X_{lj} - \overline{X})' &= [(X_{lj} - \overline{X}_l) + (\overline{X}_l - \overline{X})][(X_{lj} - \overline{X}_l) + (\overline{X}_l - \overline{X})]' \\ &= (X_{lj} - \overline{X}_l)(X_{lj} - \overline{X}_l)' + (X_{lj} - \overline{X}_l)(\overline{X}_l - \overline{X})' \\ &\quad + (\overline{X}_l - \overline{X})(X_{lj} - \overline{X}_l)' + (\overline{X}_l - \overline{X})(\overline{X}_l - \overline{X})'\end{aligned}$$

The sum over j of middle two expressions is the zero matrix, because $\sum_{n=0}^{n_l}(X_{lj} - \overline{X}_l) = 0.$ Hence, summing the cross product over l and j yields

$$\underbrace{\sum_{l=1}^{g}\sum_{j=1}^{n_l}(X_{lj}-\overline{X})(X_{lj}-\overline{X})'}_{\begin{pmatrix} total\ (corrected)\ sum \\ of\ squares\ and\ across \\ product \end{pmatrix}} = \underbrace{\sum_{l=1}^{g} n_l(\overline{X}_l-\overline{X})(\overline{X}_l-\overline{X})'}_{\begin{pmatrix} treatment\ (Between) \\ sum\ of\ squares\ and \\ cross\ product \end{pmatrix}} \underbrace{\sum_{l=1}^{g}\sum_{j=1}^{n_l}(X_{lj}-\overline{X}_l)(\overline{X}_l-\overline{X})'}_{\begin{pmatrix} redidual\ (Within)\ sum \\ of\ squares\ and\ cross \\ product \end{pmatrix}} \qquad (28)$$

The Within sum of squares and cross products matrix can be expressed as

$$\begin{aligned} W &= \sum_{l=1}^{g}\sum_{j=1}^{n_l}(X_{lj} - \overline{X}_l)(X_{lj} - \overline{X}_l)' \\ &= (n_1 - 1)S_1 + (n_2 - 1)S_2 + \cdots + (n_g - 1)S_g \end{aligned} \qquad (29)$$

Where S_l is the sample covariance matrix for the l th sample. This matrix is generalization of the $(n_1 + n_2 - 2)S_{pooled}$ matrix encountered in the two-sample case. It plays a dominant role in testing for the presence of no treatment effects.

Analogous to the univariate result, the hypothesis of no treatment effects,

$$H_0 : T_1 = T_2 = \cdots = T_g = 0$$

is tested by considering the relative sizes of the treatment and residual sums of squares and cross products. Equivalently, we may consider the relative sizes of the residual and total (corrected) sum of squares and cross products. Formally, we summarize the calculations leading to test statistic in a MANOVA table.

ANOVA Table for Comparing Population Mean Vectors

Source of variation	Degrees of freedom (d.f.)	Matrix of sum of squares and cross products (SSP)
Treatment	g-1	$B=\sum_{l=1}^{g} n_l(\bar{X}_l-\bar{X})(\bar{X}_l-\bar{X})'$
Residual (Error)	$\sum_{l=1}^{n} n_l - g$	$W=\sum_{l=1}^{g}\sum_{j=1}^{n_l}(X_{lj}-\bar{X}_l)(X_{lj}-\bar{X}_l)'$
Total (corrected for the mean)	$\sum_{l=1}^{n} n_l - 1$	$B+W=\sum_{l=1}^{g}\sum_{j=1}^{n_l}(X_{lj}-\bar{X})(X_{lj}-\bar{X})'$

This table is exactly the same from, component, as the ANOVA table, except that squares of scalars are replaced by their vector counterparts. For example, $(\bar{X}_l-\bar{X})^2$ becomes $(\bar{X}_l-\bar{X})(\bar{X}_l-\bar{X})'$. The degree of freedom correspond to the univariate geometry and also to some multivariate distribution theory involving wishart densities. [6].

One test of $H_0: T_1 = T_2 = \cdots = T_g = 0$ involves generalized variances. We reject H_0 if the ratio of generalized variances

$$\Lambda^* = \frac{|W|}{|B+W|} = \frac{\left|\sum_{l=1}^{g}\sum_{j=1}^{n_l}(X_{lj}-\bar{X}_l)(X_{lj}-\bar{X}_l)'\right|}{\left|\sum_{l=1}^{g}\sum_{j=1}^{n_l}(X_{lj}-\bar{X})(X_{lj}-\bar{X})'\right|} \quad (30)$$

is too small. The quantity $\Lambda^* = |W|/|B+W|$, proposed originally by Wills (see [11]), corresponds to the equivalent from (25) of the F-test of H_0 : no treatment of effects in the univariate case. Wilks' lambds has the virtue of being convernient and related to the likelihood ratio criterion. The exact distribution of Λ^* can be derived for the special cases listed in table distribution of Wilks' Lambds. For other cases and large sample sizes, a modification of Λ^* due to Bartlet [9] can be used to test H_0.

Table Distribution of Wilks' Lambds. $\Lambda^* = \|W\|/\|B+W\|$		
No. of Variables	No. of groups	Sampling distribution
P = 1	g ≥ 2	$\left(\frac{\sum n_l - g}{g-1}\right)\left(\frac{1-\Lambda^*}{\Lambda^*}\right) \sim F_{g-1,\sum n_l - g}$
P = 2	g ≥ 2	$\left(\frac{\sum n_l - g - 1}{g-1}\right)\left(\frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right) \sim F_{2(g-1),2(\sum n_l - g - 1)}$
P ≥ 1	g = 2	$\left(\frac{\sum n_l - p - 1}{p}\right)\left(\frac{1-\Lambda^*}{\Lambda^*}\right) \sim F_{p,\sum n_l - p - 1}$
P ≥ 1	g = 3	$\left(\frac{\sum n_l - p - 2}{p}\right)\left(\frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right) \sim F_{2p,2(\sum n_l - p - 2)}$

Bartlett (see [4]) has shown that if H_0 is true and $\sum n_l = n$ is large,

$$-\left(n-1-\frac{(p+g)}{2}\right)\ln\Lambda^* = -\left(n-1-\frac{(p+g)}{2}\right)\ln\left(\frac{|W|}{|B+W|}\right) \quad (31)$$

has approximately a chi-square distribution with p((g-1) d.f. Consequently, for $\sum n_l = n$ large, we reject H_0 at significance level α if

$$-\left(n-1-\frac{(p+g)}{2}\right)\ln\left(\frac{|W|}{|B+W|}\right) > X^2_{p(g-1)}(\alpha) \quad (32)$$

Where $X^2_{p(g-1)}(\alpha)$ is the upper (100α)th percentile of a chi-square distribution with p(g-1) d.f..

Results and Conclusion

The data was collected from AL-Margin hospital in Al-Hula province. They used Iraqi Board Cancer Registry Form. This Form that was used internationally by WHO to Register cancer patients. It therefore the Form will be not be accepted if the information that was interned were unreliable. The data was mainly concerned with bronchitis cancer in Al-Hula province. Data were then analyzed using SPSS software. The following variables were recorded in the Form and were included in our analysis (see table 1).

Table (1) The Independent Variables

Variables	Description
Sex	Patient's sex
Age	Age of patient in years
Occupation	Occupation of patients
Grade	Grade of tumor
Degree of Tumor	The level of tumor
Lymph Extent	Whether the tumor spread to lymph nodes or not
Metastasis	Spread of the tumor to other organs
Cancer Extent	The extent that cancer spread to
Duration	The time from first diagnosis to the last visit

Data were consisted of 484 patients who suffered from bronchitis cancer. The dependent variable is the status (1= Dead, 2= Alive). The independents variable shown in table 1: sex were categorized to female and male. Age in years were between 4 and 95 years; occupation were categorized to employed , unemployed and unknown; grade were categorized to grade 1 to 4 and unknown; degree of tumor were categorized to advanced, and localized; lymph extent were categorized to regional lymph nodes, and no regional lymph nodes; metastasis were categorized to distance metastasis, and no distance metastasis; cancer extent were categorized to in situ, localized, regional direct extent, regional lymph nodes extent, regional direct extent plus lymph nodes, distance Metastasis, not applicable, and unknown. Mean and standard deviation for all variables are shown in table 2.

Table 3 is a matrix composed of the means of each corresponding value within the two 9X 9 matrices of the two levels of category of outcome variable.

Table (2) Descriptive Statistics for independent variables

Variables	Mean	Std. Deviation
Sex	1.31	.465
Age	65.97	12.214
Occupation	2.86	.457
Grade	8.97	.456
Degree of tumor	1.00	.000
Lymph Extent	.95	.211
Metastasis	.61	.489
Cancer Extent	6.51	1.342
Duration	8.82	11.554
Sex	1.29	.457
Age	62.42	11.572
Occupation	1.76	.463
Grade	7.78	2.513
Degree of tumor	1.19	.393
Lymph Extent	.57	.496
Metastasis	.23	.424
Cancer Extent	4.44	1.826
Duration	54.71	68.594

Table (3). Tests of Equality of Group Means

Variables	Wilks' Lambda	F	df1	df2	Sig.
Sex	1.000	.210	1	482	.647
	.978	10.812	1	482	.001
Occupation	.414	683.268	1	482	.000
Grade	.904	51.132	1	482	.000
Degree of tumor	.898	54.956	1	482	.000
Lymph Extent	.799	121.034	1	482	.000
Metastasis	.855	82.065	1	482	.000
Cancer Extent	.707	199.429	1	482	.000
Duration	.824	102.850	1	482	.000

Table 4 showed Wilk's lambda, F, and significance values contribute information about difference means for each variable. The F and significance

values identify for which variables the two groups differ significantly. This type of information that we will consider before running a discriminant analysis. Wilks' Lambda is the ratio of the within groups sum of square to the total sum of squares. This is the proportion of the total variance in the discriminant scores not explained by differences among groups. A lambda of 1 for sex variable indicated that observed group means are equal (all the variance is explained by factors other than difference between theses means), while a small lambda for occupation variable indicated that group means appear to differ. The table showed that all variable were statistically significant at 1% level except sex that was not significant ($p = 0.647 > 0.05$).

Box's test of equality of covariate matrices showed that ($P= 0.00 < 0.05$) might arouse concern. However, it had been found that even when multivariate normality was violated, the discrininant function can still often performed surprisingly well. If a significance value is low, it would be well to look at the univariate normality of some of the included variables. For instance, we know that the sex variable is not normally distributed but inclusion of sex improved the discriminating ability of the equation.

Table (4) Variables in the Analysis

Step		Tolerance	F to Remove	Wilks' Lambda
1	Occupation 1	1.000	683.268	
2	Occupation 1	.741	935.839	1.000
	Sex	.741	105.315	.414
3	Occupation 1	.736	614.438	.705
	Sex	.741	97.166	.372
	Cancer Extent	.989	46.672	.339
4	Occupation 1	.735	555.072	.619
	Sex	.741	87.014	.339
	Cancer Extent	.988	40.130	.311
	Duration	.996	37.402	.309
5	Occupation 1	.735	521.077	.586
	Sex	.740	81.802	.329
	Cancer Extent	.986	37.539	.303
	Duration	.996	36.223	.302
	Grade	.998	10.730	.287
6	Occupation 1	.734	518.367	.580
	Sex	.740	80.371	.325
	Cancer Extent	.578	10.411	.284
	Duration	.996	35.292	.299
	Grade	.990	11.761	.285
	Lymph Extent	.583	4.123	.281

A stepwise variable selection was used, which entered the variables into the discriminant equation, one at a time, based on designed criterion for inclusion ($F \geq 1.00$); but will drop variables from the equation if the inclusion requirement dropped below the designated level when other variables have been entered. Basically the selection rule here was to minimize Wilks' Lambda at each step.

Table 5 demonstrated which variables indicated the number of variables in the discriminant equation at each step. Notice that all the variables in the analysis have F-values greater than 1.00. Notice that in a total of 6 steps 9 variables were entered. The tolerance level was a measure of linear dependency between one variable and the others. If tolerance was less than 0.001, this indicated a high level of linear

dependency, and program will not enter that variable into the equation.

Canonical correlation is the correlation between the discrininant scores. A high correlation (0.850) indicated a function that discriminated well.

The discriminant equation is:

D = -8.101 + (1.128) sex + 2.154 occupation + 0.100 grade + 0.371 lymph extent + 0.141 cancer extent + (-0.006) duration

Table (5) Number and Percent of subjects classified

Figure (1) Predicted Group Membership

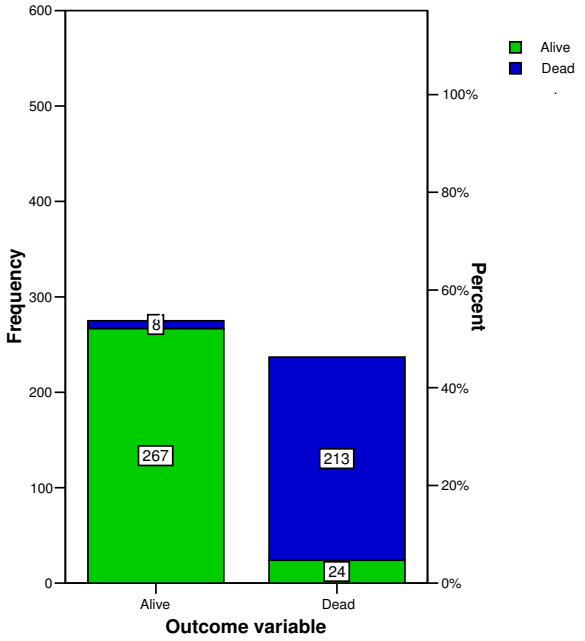

Figure (1) and Table (5) illustrated a simple summary of number and percent of patients classified correctly and incorrectly. In summary 90% , and 98% of dead and alive patients were classified correctly. Only 2% and 10% of dead and alive patients were misclassified.

REFERENCES

1. Anderson, T.W. "An Introduction to Multivariate Statistical Analysis (3rd ed.). New York: John Wiley, 2003.
2. Box, G.E. P., "Problems in the Analysis of Growth and Wear Curves" Biometrics, 6(1950), 362-389.
3. Box, G. E. P., and N. R. Draper. Evolutionary Operation: A Statistical Method for Improvement. New York: John Wiley, 1969.
4. Box, G. E. P.," A General Distribution Theory for a Class of Likelihood Criteria." Biometrika, 36 (1949), 317-346
5. Eric S. Fung and Michael K. Ng (2007), "On sparse Fisher discriminant method formicroarray data", Bioinformation 2(5): 230-234 (2007).
6. Fisher,R.A."The Statistical Utilization of Multiple Measurements." Annals of Eugenics,8(1938),376-386.

Status	**Predicted Group Membership**		**Total**
	Dead	Alive	
Original Dead	212 (89.8%)	24 (10.2%)	236 (100%)
Alive	5 (2.0%)	243 (98.0%)	248 (100%)

7. Mokhtari-Dizadji1,M, Vahed M and Gity M (2003), "The application of discriminant analysis in differentiation of fibroadenoma and ductal carcinoma of breast tissue using ultrasound velocity measurement". ***Iran. J. Radiat. Res., 2003; 1(3): 163 – 169.***
8. Richard A. Johnson and Dean W. Wichern "Applied Multivariate Statistical Analysis" Discrimination and Classifiction,11(2007),575-670.
9. Sandrine Dudoit ,Jane Fridlyand and Terence P. Speed (2000), "Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data". Technical report 576, June 2000.
10. Wald, A. "On a statistical problem Arising in the Classification of an Individual into One of Two Groups." Annals of Mathematical statistics,15(1944), 145-162.
11. Wills, S. S. "Certain Generalizations in the Analysis of Variance." Biomtrika, 24 (1932), 471-494.