Applications of Discriminant Analysis in Medical diagnosis

Hazim M. Gorgess
Anas Kh. Mohammed
Dept. of Mathematics/College of Education for Pure Science (Ibn AL-Haitham) / University of Baghdad

Received in : 19June2013 , Accepted in : 4 December 2013

Abstract

In this paper, the discriminant analysis is used to classify the most wide spread heart diseases known as coronary heart diseases into two groups (patient, not patient) based on the changes of discrimination features of ten predictor variables that we believe they cause the disease . A random sample for each group is employed and the stepwise procedures are performed in order to delete those variables that are not important for separating the groups. Tests of significance of discriminant analysis and estimating the misclassification rates are performed.

Keywords: Discriminant analysis, classification, stepwise procedures, misclassification rates .

Introduction

Discriminant analysis is a technique for the multivariate study of group differences. Discriminant analysis is particularly appropriate when one wishes to describe, summarize, and understand the differences between or among groups.

It is convenient to determine which of a set of variables is best captures or characterizes group differences. The most frequent applications of discriminant analysis are for predictive purpose, that is, for situations in which it is necessary or desirable to classify subjects into groups or categories.**[1]**

Theoretical Part

1. The Discriminant Function for Two Groups

The derived discriminant functions may be used to classify new cases into groups. Prior probabilities of belonging to each group may be entered or derived from the observed data. For the case of two groups, we assume that the two populations to be compared would have the same covariance matrix $\sum_1=\sum_2=\sum$, but distinct mean vectors μ_1 and μ_2. We work with samples $y_{11}, y_{12}, \ldots, y_{1n_1}$ and $y_{21}, y_{22}, \ldots, y_{2n_2}$ from the two populations. As usual ,each vector y_{ij} consists of measurements on p variables. The discriminant function is the linear combination of these p variables that maximizes the distance between the two transformed group mean vectors. A linear combination $z=a'y$ transforms each observation vector to scalar

$z_{1i} = a'y_{1i} = a_1y_{1i1} + a_2y_{1i2} + \ldots + a_py_{1ip}$, i =1, 2, …, n_1

$z_{2i} = a'y_{2i} = a_1y_{2i1} + a_2y_{2i2} + \ldots + a_py_{2ip}$, i =1, 2, …, n_2

Hence, the $n_1 + n_2$ observation vectors in the two samples. $y_{11}, y_{12}, \ldots, y_{1n_1}$ $y_{21}, y_{22}, \ldots, y_{2n_2}$, are transformed to scalars $z_{11}, z_{12}, \ldots, z_{1n_1}, z_{21}, z_{22}, \ldots, z_{2n_2}$

We find the means $\bar{z}_1 = \frac{\sum_{i=1}^{n_1} z_{1i}}{n_1} = a'\bar{y}_1$,and $\bar{z}_2 = \frac{\sum_{i=1}^{n_2} z_{2i}}{n_2} = a'\bar{y}_2$

where $\bar{y}_1 = \frac{\sum_{i=1}^{n_1} y_{1i}}{n_1}$, $\bar{y}_2 = \frac{\sum_{i=1}^{n_2} y_{2i}}{n_2}$

We wish to find the vector a that maximizes the ratio $\frac{(\bar{z}_1 - \bar{z}_2)^2}{s_z^2}$ which can be expressed as [1] :

$$Q = \frac{(\bar{z}_1 - \bar{z}_2)^2}{s_z^2} = \frac{\left[a'(\bar{y}_1 - \bar{y}_2)\right]^2}{a's_pa} \qquad \ldots(2.1)$$

The numerator of this ratio is the square of the difference between the means of z for the two groups and the denominator is the sum of squares within groups. Putting d= $\bar{y}_1 - \bar{y}_2$, D= a'd and w= $a'S_pa$, and substituting in equation (2.1) we get:

$$Q = \frac{D^2}{w} \qquad \ldots(2.2)$$

By differentiating Q with respect to a and putting the derivative equal to zero [2]

we obtain $\frac{\partial Q}{\partial a} = \frac{2wDd - 2D^2S_pa}{w^2} = 0$

This yields $wDd = D^2S_pa$, Dividing by D^2 we obtain $\frac{w}{D}d = S_pa$

and hence $a = S_p^{-1}\frac{w}{D}d$, since $\frac{w}{D}$ is any nonzero constant so let $\frac{w}{D} = 1$ and maximize of (2.1) occur as when :[1]

$$a = S_p^{-1}\,d = S_p^{-1}(\bar{y}_1 - \bar{y}_2) \qquad \ldots(2.3)$$

Or when a is any multiple of $S_p^{-1}(\bar{y}_1 - \bar{y}_2)$. Thus the maximizing vector a is not unique however its direction is unique, that is the relative values or ratios of $a_1 , a_2 , \dots , a_p$ are unique .

2. Discriminant Analysis For Several Groups

In discriminant analysis for several groups, we are concerned with finding linear combinations of variables that would best separate the k groups of multivariate observations. For k groups (samples) with n_i observations in the i^{th} group, we transform each observation vector y_{ij} to obtain

$$z_{ij} = a' y_{ij} \qquad i=1,2, \dots ,k \; ; \; j=1,2, \dots , n_i$$

and find the means $\bar{z}_i = a' \bar{y}_i$, where $\bar{y}_i = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i}$. As in the two group case , we seek the vector a that maximally separates $\bar{z}_1, \bar{z}_2, \dots , \bar{z}_k$. To express separation among $\bar{z}_1, \bar{z}_2, \dots , \bar{z}_k$ we extend the separation criterion to the k group case

Since $a'(\bar{y}_1 - \bar{y}_2)= (\bar{y}_1 - \bar{y}_2)' a$ we can write :[1]

$$\frac{(\bar{z}_1 - \bar{z}_2)^2}{S_z^2} = \frac{[a'(\bar{y}_1 - \bar{y}_2)]^2}{a' S_p a} = \frac{a'(\bar{y}_1 - \bar{y}_2)(\bar{y}_1 - \bar{y}_2)' a}{a' S_p a} \qquad \dots(2.4)$$

To extend (2.4) to k groups, we use the H matrix in place of $(\bar{y}_1 - \bar{y}_2)(\bar{y}_1 - \bar{y}_2)'$ and E in place of S_p to obtain:

$$\lambda = \frac{a' H a}{a' E a} \text{ where :} \qquad \dots(2.5)$$

$$\mathbf{H}= \sum_{i=1}^{k} n_i (\bar{y}_{i.} - \bar{y}_{..})(\bar{y}_{i.} - \bar{y}_{..})' \, , \, \mathbf{E}= \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.}) (y_{ij} - \bar{y}_{i.})'$$

$$\bar{y}_{i.} = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i} = \frac{y_{i.}}{n_i} \, , \, \bar{y}_{..} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \frac{y_{ij}}{nk} = \frac{y_{..}}{nk}$$

The p×p matrix H has a between sum of squares on the diagonal for each of the p variables .off diagonal elements are analogous sums of products for each pair of variables. The p×p error matrix E has a within sum of squares for each variable on the diagonal ,with analogous sums of products off diagonal. Thus H has the form:

$$\mathbf{H} = \begin{pmatrix} SSH_{11} & SPH_{12} & . & . & . & SPH_{1p} \\ SPH_{12} & SSH_{22} & . & . & . & SPH_{2p} \\ . & . & & & & . \\ . & . & & & & . \\ . & . & & & & . \\ SPH_{1p} & SPH_{2p} & . & . & . & SSH_{pp} \end{pmatrix} \qquad \dots(2.6)$$

The matrix E can be expressed in a form similar to (2.6)

$$\mathbf{E} = \begin{pmatrix} SSE_{11} & SPE_{12} & . & . & . & SPE_{1p} \\ SPE_{12} & SSE_{22} & . & . & . & SPE_{2p} \\ . & . & & & & . \\ . & . & & & & . \\ . & . & & & & . \\ SPE_{1p} & SPE_{2p} & . & . & . & SSE_{pp} \end{pmatrix} \qquad \dots(2.7)$$

We can write the ratio in (2.5) as $a'\mathrm{H}a = \lambda a'\mathrm{E}a$

$a'\ (\mathrm{H}a - \lambda \mathrm{E}a) = 0$ ...(2.8)

We examine values of λ and a that are solutions of (2.8) in a search for the value of a that results in maximum λ. The solution $a' = 0'$ is not permissible because it gives $\lambda = \frac{0}{0}$ in (2.5).Other solutions are found from

$\mathrm{H}a - \lambda \mathrm{E}a = 0$ ...(2.9)

Which can be written in the form

$(\mathrm{E}^{-1}\mathrm{H} - \lambda \mathrm{I})\ a = 0$ ...(2.10)

The solution of (2.10) are the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_s$ and associated eigenvectors $a_1, a_2, \ldots, a_s$ of $\mathrm{E}^{-1}\mathrm{H}$. The eigenvalues are considered to be ranked $\lambda_1 > \lambda_2 > \ldots > \lambda_s$. The number of nonzero eigenvalues s is the rank of H which can be found as the smaller of k−1

and p. Thus the largest eigenvalue λ_1 is the maximum value of $\lambda = \frac{a'\mathrm{H}a}{a'\mathrm{E}a}$ in (2.10) and the coefficient vector that produces the maximum is the corresponding eigenvector a_1. Eq(2.10) can be verified by using calculus as follows: Differentiating $\lambda = \frac{a'\mathrm{H}a}{a'\mathrm{E}a}$ with respect to a then putting the derivative equal to zero, we obtain :[55]

$$\frac{\partial \lambda}{\partial a} = \frac{2(a'\mathrm{E}a)\mathrm{H}a - 2(a'\mathrm{H}a)\mathrm{E}a}{(a'\mathrm{E}a)^2} = 0$$

This yields : $(a'\mathrm{E}a)\mathrm{H}a - (a'\mathrm{H}a)\mathrm{E}a = 0$, dividing by $a'\mathrm{E}a$, we obtain :

$\mathrm{H}a - \lambda\ \mathrm{E}a = 0$, or $(\ \mathrm{H} - \lambda\ \mathrm{E})\ a = 0$.

Which can be written as $(\mathrm{E}^{-1}\mathrm{H} - \lambda \mathrm{I}\ )\ a = 0$ hence , the discriminant function that maximally separates the means is $z_1 = a_1'\ y$ that is, it represents the dimension that maximally separates the means. From the s eigenvectors $a_1, a_2, \ldots, a_s$ of $\mathrm{E}^{-1}\mathrm{H}$ corresponding to $\lambda_1, \lambda_2, \ldots, \lambda_s$, we obtain s discriminant functions.

$z_1 = a_1'\ y$, $z_2 = a_2'\ y$, . . . , $z_s = a_s'\ y$. The relative importance of each discriminant function z_i, i=1,2, . . . ,s. can be assessed by considering its eigenvalue as a proportion of the total [37]

$$\frac{\lambda_i}{\sum_{j=1}^{s} \lambda_j} \qquad \ldots(\ 2.11\ )$$

By this criterion two or three discriminant functions will often suffice to describe the group differences. The discriminant function associated with small eigenvalues can be neglected.

3. Test of Significance of Discriminant Analysis

For the case of two groups, we wish to test $\mathrm{H}_0 : \mu_1 = \mu_2$ Vs $\mathrm{H}_1 : \mu_1 \neq \mu_2$ the discriminant function coefficient vector a is significantly different form 0 if T^2 is significant, where :[5]

$$\mathrm{T}^2 = \frac{n_1\, n_2}{n_1 + n_2} (\bar{\mathrm{y}}_1 - \bar{\mathrm{y}}_2)'\ \mathrm{S}_\mathrm{p}^{-1}\ (\bar{\mathrm{y}}_1 - \bar{\mathrm{y}}_2) \qquad \ldots(2.12)$$

Which is distributed as $\mathrm{T}^2_{\mathrm{p}, n_1 + n_2 - 2}$ when H_0: $\mu_1 = \mu_2$ is true. We reject H_0 if $\mathrm{T}^2 > \mathrm{T}^2_{\alpha, \mathrm{p}, n_1 + n_2 - 2}$. Also, we can use F approximation test where :[5]

$$F = \frac{n_1 + n_2 - \mathrm{p} - 1}{(n_1 + n_2 - 2)\mathrm{p}} \mathrm{T}^2 \qquad \text{... (2.13)}$$

Which is distributed as $\mathrm{F}_{\mathrm{p}, n_1+n_2-\mathrm{p}-1}$ when H_0: $\mu_1=\mu_2$ is true. We reject H_0 if
$\mathrm{F} > \mathrm{F}_{\mathrm{p}, n_1+n_2-\mathrm{p}-1}$

For several groups, to test $\mathrm{H}_0 : \mu_1 = \mu_2 = . . . =\mu_\mathrm{k}$, we use the Wilk's lambda statistic defined as : [6]

$$\Lambda = \frac{|\mathrm{E}|}{|\mathrm{E+H}|} \qquad \text{...(2.14)}$$

We reject H_0 if $\Lambda > \Lambda_{\alpha, \mathrm{p}, \mathrm{v_H}, \mathrm{v_E}}$. The parameters in Wilk's Λ distribution are p=number of variables, $\mathrm{v_H} = \mathrm{k}-1$ degrees of freedom of hypothesis,

$\mathrm{v_E} = \mathrm{N} - \mathrm{k}$ with $\mathrm{N} = \sum_{\mathrm{i}=1}^{\mathrm{k}} n_\mathrm{i}$ degrees of freedom for error.

Wilk's Λ in (2.14) can be expressed in terms of the eigenvalues λ_1 , λ_2 , . . . , λ_s of $\mathrm{E}^{-1}\mathrm{H}$ as follows :

$$\Lambda_1 = \prod_{i=1}^{\mathrm{s}} \frac{1}{1+\lambda_\mathrm{i}} \qquad \text{...(2.15)}$$

The number of nonzero eigenvalues of $\mathrm{E}^{-1}\mathrm{H}$ is s= min (p, $\mathrm{v_H}$) which is the rank of H. The range of Λ is $0 \leq \Lambda \leq 1$ and the test based on wilk's Λ is an inverse test in the sense that we reject H_0 for small value of Λ. Since Λ_1 is small if one or more λ_i's are large, Wilk's Λ tests for significance of the eigenvalues and thereby for the discriminant functions. The s eigenvalues represent s dimensions of separation of the mean vectors $\bar{\mathrm{y}}_1, \bar{\mathrm{y}}_2, \ldots, \bar{\mathrm{y}}_\mathrm{k}$. We are interested in which, if any of these dimensions is significant. In addition to the Wilk's Λ test.

We can use χ^2 approximation for Λ_1 with $\mathrm{v_H}$, $\mathrm{v_E}$ degrees of freedom.

$$\mathrm{V}_1 = -\left[\mathrm{v_E} - \tfrac{1}{2}\,(\mathrm{p} - \mathrm{v_H}+1)\right] \ln \Lambda_1 \qquad \text{...(2.16)}$$

$$= -\left[\mathrm{N} - 1 - \frac{1}{2}\,(\mathrm{p+k})\right] \ln \prod_{\mathrm{i}=1}^{\mathrm{S}} \frac{1}{1+\lambda_\mathrm{i}} = \left[\mathrm{N} - 1 - \frac{1}{2}\,(\mathrm{p} + \mathrm{k})\right] \sum_{\mathrm{i}=1}^{\mathrm{s}} \ln(1+\lambda_\mathrm{i})$$

Which is approximately χ^2 with p(k−1) degrees of freedom. The test statistic Λ_1 and its approximations(2.16) test the significance of all of λ_1 , λ_2 , . . . , λ_s. If the test leads to rejection of H_0, we conclude that at least one of the λ's is significantly different from zero, and therefore there is at least one dimension of separation of mean vectors. Since λ_1 is the largest, we are only sure of its significance along with that of $\mathrm{z}_1 = a_1^{\prime}\,\mathrm{y}$. To test the significance of λ_2 , . . . , λ_s ,we delete λ_1 from Wilk's Λ and the associated χ^2 approximation to obtain

$$\Lambda_2 = \prod_{\mathrm{i}=2}^{\mathrm{s}} \frac{1}{1+\lambda_i}$$

$$\mathrm{V}_2 = -\left[\mathrm{N} - 1 - \frac{1}{2}\,(\mathrm{p} + \mathrm{k})\right] \sum_{\mathrm{i}=2}^{\mathrm{s}} \ln(1+\lambda_\mathrm{i})$$

Which is approximately χ^2 with (p−1)(k−2) degree of freedom.

If this test leads to rejection of H_0, we conclude that at least λ_2 is significant along with the associated discriminant function $\mathrm{z}_2 = a_2^{\prime}\,\mathrm{y}$. We can continue in this fashion, testing each λ_i in turn until a test fails to reject H_0. The test statistic at the mth step is :

$\Lambda_m = \prod_{i=m}^{s} \frac{1}{1+\lambda_i}$ which is distributed as $\Lambda_{p-m+1,\, k-m,\, N-k-m+1}$ ...(2.17)

The statistic $V_m = -\left[N - 1 - \frac{1}{2}\,(p+k)\right] \sum_{i=2}^{s} \ln \Lambda_m$ ...(2.18)

$$= \left[N - 1 - \frac{1}{2}\,(p+k)\right] \sum_{i=m}^{s} \ln(1+\lambda_i)$$

has an approximate χ^2 distribution with (p−m+1)(k−m) degrees of freedom.

We can also use F approximation for each Λ_i. For $\Lambda_1 = \prod_{i=1}^{s} \frac{1}{1+\lambda_i}$ we use

$$F = \frac{1-\Lambda_1^{1/t}}{\Lambda_1^{1/t}} \frac{df_2}{df_1} \qquad \text{...(2.19)}$$

where $t = \sqrt{\frac{P^2(k-1)^2-4}{P^2+(k-1)^2-5}}$

Putting w=$N - 1 - \frac{1}{2}$(p+k) then $df_1 = P(k-1)$, $df_2 = wt - \frac{1}{2}[p(k-1)-2]$

For $\Lambda_m = \prod_{i=m}^{s} \frac{1}{1+\lambda_i}$, m=2,3,…,s

We use $F = \frac{1-\Lambda_m^{1/t}}{\Lambda_m^{1/t}} \frac{df_2}{df_1}$ with p − m+1 and k − m in place of p and k − 1

$$t = \sqrt{\frac{(p-m+1)^2(k-m)^2-4}{(p-m+1)^2+(k-m)^2-5}}$$

$w = N - 1 - \frac{1}{2}$(p+k) , df_1=(p − m+1)(k − m), df_2=wt $- \frac{1}{2}[(p-m+1)(k-m)-2]$

4. Tests of Equality of Covariance Matrices [1]

For k multivariate populations, the hypothesis of equality of covariance matrices is
$H_0 : \sum_1 = \sum_2 = \cdots = \sum_k$

Calculate $C_1 = \left[\sum_{i=1}^{k} \frac{1}{v_i} - \frac{1}{\sum_{i=1}^{k} v_i}\right]\left[\frac{2p^2+3p-1}{6(p+1)(k-1)}\right]$ …(2.20)

Then : U= −2(1−C₁) ln M is approximately $\chi^2 \left[\frac{1}{2}\,(k-1)p(p+1\right]$ …(2.21)

Where M is $M = \frac{|S_1|^{\frac{v_1}{2}}\,|S_2|^{\frac{v_2}{2}} \ldots\, |S_k|^{\frac{v_k}{2}}}{|S_p|^{\frac{\sum_i v_i}{2}}}$ …(2.22)

$$\text{and, } \ln M = \frac{1}{2} \sum_{i=1}^{k} v_i \ln|S_i|$$

$$-\frac{1}{2}\,\left(\sum_{i=1}^{k} v_i\right) \ln|S_p| \qquad \text{... (2.23)}$$

We reject H_0 if $U > \chi^2_\alpha$ …(2.24)

5. Stepwise Selection Of Variables [6]

The stepwise method for selecting variables in discriminant analysis is rather like doing a stepwise regressions and is especially useful in similar circumstances, namely when we have rather a long list of possible classification variables and it is unlikely that all will make a useful contribution to a set of discriminant functions. We would like to find the best subset, or else something close to that. The single variable that gives the significant classification into our groups is chosen first, then we look at the remaining variables and add the one that gives the biggest improvement. We check the two variables now, and make sure that each makes a significant contribution in the presence of the other. At each step we see whether another variables can be added that will make a significant improvement, and whether any previous ones can be removed. The process stops when no more variables can be added or removed at the level of significance we are using.

6. Classification Procedures

6.1. Classification use Discriminant Function

A simple procedure for classification can be based on the discriminant function,

$$z= a'y_0= (\bar{y}_1 - \bar{y}_2)' S_p^{-1} y_0 \qquad \text{...(2.25)}$$

Where y_0 is the vector of measurements on a new sampling unit that we wish to classify into one of the two groups (populations).

Denote the two groups by G_1 and G_2. Fisher's (1936) linear classification procedure assigns y_0 to G_1 if $z_0=a'y_0$ is closer to $\bar{z}_1$ than to $\bar{z}_2$ and assigns y_0 to G_2if $z_0$is closer to $\bar{z}_2$than to $\bar{z}_1$, $z_0$is closer to $\bar{z}_1$ if $z_0 > \frac{1}{2}(\bar{z}_1 + \bar{z}_2)$

where $\bar{z}_1 = \sum_{i=1}^{n_1} \frac{z_{1i}}{n_1} = a'\bar{y}_1 = (\bar{y}_1 - \bar{y}_2)' S_p^{-1} \bar{y}_1$

To express the classification rule in terms of y, we first write $\frac{1}{2}(\bar{z}_1 + \bar{z}_2)$ in the form:

$$\frac{1}{2}(\bar{z}_1 + \bar{z}_2) = \frac{1}{2} a' (\bar{y}_1 + \bar{y}_2) = \frac{1}{2}(\bar{y}_1 - \bar{y}_2)' S_p^{-1} (\bar{y}_1 + \bar{y}_2) \qquad \text{...(2.26)}$$

Then the classification rule becomes, assign y_0 to G_1 if [1]

$$a'y_0 = (\bar{y}_1 - \bar{y}_2)' S_p^{-1} y_0 > \frac{1}{2}(\bar{y}_1 - \bar{y}_2)' S_p^{-1} (\bar{y}_1 + \bar{y}_2) \qquad \text{...(2.27)}$$

and assign y_0 to G_2 if

$$a'y_0 = (\bar{y}_1 - \bar{y}_2)' S_p^{-1} y_0 < \frac{1}{2}(\bar{y}_1 - \bar{y}_2)' S_p^{-1} (\bar{y}_1 + \bar{y}_2) \qquad \text{...(2.28)}$$

6.2. Classification Use Simple Classification Function [6]

Fisher (1936) proposed a simple classification function for each group based on a linear combination of the discriminating variables. For the case of k groups and p discriminating variables, the simple classification function has the form

$$z_g = b_{g_0} + b_{g_1} X_1 + b_{g_2} X_2 + \ldots + b_{g_p} X_p \text{ , } g= 1,2, \ldots ,k \qquad \text{...(2.29)}$$

The coefficient b_{g_i} associated with variable i in group g is given as :

$$b_{g_i} = (N - g) \sum_{j=1}^{p} w_{ij}\, y_{jk}$$

Where w_{ij} represents the ij^{th} element from the inverse matrix of within groups sums of squares and cross products, N represents the whole number of observations.

The constant b_{k_0} is given as :

$$b_{k_0} = -0.5 \sum_{j=1}^{p} b_{kj}\, y_{jk} \qquad \ldots (2.30)$$

The rule of classification is simply to classify the new observation to the group that yields a maximum value of the classification function h after substituting all discriminating variables into classification functions.

7. Estimating Misclassification Rates

A simple estimate of the error rate can be obtained by trying out the classification procedure on the same data set that has been used to compute the classification function.

This method is referred to as resubstitution. Each observation vector y_{ij} is substituted to the classification functions and assigned to a group .[1]

We then count the number of correct classifications and the number of misclassifications.

The proportion of misclassifications resulting from resubtitution is called the apparent error rate (APER). The results can be conveniently displayed in a classification table as shown below:

Table ,classification table for two groups

Actual group	Number of observations	Predicted group	
		1	2
1	n_1	n_{11}	n_{12}
2	n_2	n_{21}	n_{22}

Let us denote the first and second groups by G_1 and G_2 respectively. Among n_1 observations in G_1,n_{11} are correctly classified into G_1 and n_{12} are misclassified into G_2 ,where $n_1=n_{11}+n_{12}$.

Similarly of the n_2 observations in G_2, n_{21} are misclassified into G_1, and n_{22} are correctly classified into G_2 where $n_2=n_{21}+n_{22}$ thus:[6]

$$APER = \frac{n_{12}+n_{21}}{n_1+n_2} \qquad \ldots(2.31)$$

Similarly, we can define apparent correct classification rate (APCR) as:

$$APCR= \frac{n_{11}+n_{22}}{n_1+n_2} \qquad \ldots(2.32)$$

The method of resubstitution can be readily extended to the case of several groups.

Particular Application

The real data were collected from records of (51) real patients suffering from coronary heart disease (CHD) from Ibn-Al-Nafees Hospital, moreover, the same informations were obtained about (54) healthy persons. The discriminant analysis were then performed with two groups (patient, not patient) and ten predictor variables that we belive they cause the disease.

the variables for each group are :

A. The dependent variable which represents (0 for Not patient) and (1 for patient)

B. Ten independent variables are ascribed below :

1. Age (X_1)
2. S.cholestrol (X_2)

3. Triglyceride (X_3)
4. LDL (high low density cholesterol) (X_4)
5. HDL(low high density cholesterol) (X_5)
6. Diabetes mellitus (Sugar) (X_6)
7. Hypertension (systolic Blood pressure) (X_7)
8. Sex (0 for Male) (1 for Female) (X_8)
9. Smoking (0 for not smoker) (1 for X-smoker) (2 for smoker) (X_9)
10. Family History (Heredity factor) (X_{10})
(0 for no heritage factor) (1 for heritage factor)

The mean for each group and the total mean are presented in (Table 1).

Applying the rules of stepwise method for discriminant analysis stated earlier we found that only four predictor variables, namely X_4, X_5, X_{10}, X_2 give the significant classification into our groups.

1. The Discriminant Function

We find the linear discriminant function coefficients by using equation (2.3) the linear discriminant function is :

$Z= (0.1300X_4) + (0.3385\ X_5) – (3.6060\ X_{10}) – (0.1958\ X_2)$

2. Test of Significance of Discriminant Analysis

Now to test the significance of discriminant function we calculate the statistic T^2, it found the statistic to be $T^2 = 381.1326171$

Since $T^2 > T^2_{\alpha,\, p, n_1+n_2-2} = T^2_{0.01,\, 4,103} = 14.511$, Then the discriminant function is significant.

Another test of significant can be performed by using eq(2.16) where the value of ∨ was found to be 156.212 comparing this value with $\chi^2_{(0.01,4)}$ = 13.2767, we conclude that the discriminant function is significant.

Also, we can use F approximation for $\Lambda = 0.213$ to test the significance by using equation (2.19), where F = 92.4 and compare with $F_{0.01,4,100}$= 3.51, we conclude that the discriminant function is significant.

3. Tests of Equality of Covariance Matrices

We use the χ^2 approximation test, the value of u was calculated by using equation (2.21), it was found to be 28.01 while the critical value of $\chi^2_{0.001,10}$ =29.59

Thus we accept H_0 since $u < \chi^2_{0.001,10}$.

4. Classification Procedures

4.1. Classification Use Discriminant Function

We must find the mean of discriminant function for the two groups

$\bar{z} = [0.1300*\bar{X}_4] + [0.3385*\bar{X}_5] – [3.6060*\bar{X}_{10}] – [0.1958*\bar{X}_2]$

The mean discriminant function of group 1 is

$\bar{z}^{(1)} = [0.1300*111.69] + [0.3385*64.43] – [3.6060*0.13] – [0.1958*199.37]$
$= -3.176171$

And the mean discriminant function of group 2 is

$\bar{z}^{(2)} = [0.1300*184.94] + [0.3385*34.10] – [3.6060*0.63] – [0.1958*260.57] = -17.706336$

The cut point is $\frac{-3.176171+ -17.706336}{2}$= -10.4412535 , we use equation (2.27), (2.28) to classify the new observations, for example, we want to classify new observation from group 1 (not patients) with the informations LDL=80, HDL=71, S.cholestrol=171 and has no hereditary factor, we found

$z_{(6)} = [0.1300*80] + [0.3385*71] - [3.6060*0] - [0.1958*171] = 0.9517$, By equation (2.80),

$0.9517 > -10.4412535$. Then the observation is correctly classified in Group 1

4.2. Classification use Simple Classification Function

After we find the inverse of within – sum square and cross products marix, we find the functions of group 1 and group 2 by equation (2.29), (2.30) where :

$Z^{(1)} = (-60.037) + (0.337 * X_4) + (0.902 * X_5) + (-2.598 * X_{10}) + (0.117 * X_2)$

$Z^{(2)} = (-70.480) + (0.207 * X_4) + (0.564 * X_5) + (0.997 * X_{10}) + (0.313 * X_2)$

And we use the two functions to classify the new observations. For example, we want to classify a new observation from group 2 (patients) with the informations LDL=183, HDL=28, S.cholestrol=251 and has no hereditary factor, we found

$z^{(1)} = (-60.037) + (0.117 * 251) + (0.337 * 183) + (0.902 * 28) + (-2.598 * 0) = 56.257$

And $z^{(2)} = (-70.480) + (0.313 * 251) + (0.207 * 183) + (0.564 * 28) + (0.997 * 0) = 61.756$

$\because\ z^{(1)} < z^{(2)}$, Then the observation is in group 2

5. Estimating Misclassification Rates

We calculate apparent error rate by equation (2.31)

$\text{APER} = \frac{2+3}{105} = 4.8\%$

And we calculate the apparent correct rate by equation (2.32)

$\text{APCR} = \frac{49+51}{105} = 95.2\%$

the correctly classified into group (1) $\frac{42}{54} = 96.3\%$

The misclassified into group (1)is $\frac{2}{54} = 3.7\%$

The correctly classified into group (2)is $\frac{49}{51} = 96.1\%$

The misclassified into group (2) $\frac{2}{51} = 3.9\%$

The classification is represented in (Table 2).

Conclusions

From the theoretical and practical study, we believe that the following points are considerable :

1. By using the stepwise method, we conclude that the first predictor variable that has the largest significant effect for discriminant between the two groups is high low density cholesterol X_4, followed by the low high density cholesterol X_5 then the Heredity factor X_{10} and finally by s.cholestrol X_2, thus our discriminant function was constructed on the basis of these variables.
2. According to the test of significance we made, namely, the wilk's Λ test and the χ^2 approximation test with level of significance $\alpha = 0.01$ we found that the discriminant function constructed significantly separates the groups.
3. By using the resubstitution method, the resulting classification table revealed that about 5% of the cases are wrongly classified, while about 95% of the cases are correctly classified to the groups.

References

1. Rencher, A. C. (2012), Methods of Multivariate Analysis, Third Edition, Wiley, New York.

2. Kandall, M.G. (1955), The Advanced Theory of Statistics, Vol.II, Third Edition, Charles Griffin, London

3. Samnles, Wilks (1963), Mathematical Statistics, John wiley, New York, London.

4. Klecka, W.R. (1984), "Discriminant Analysis" Beverly Hills/ London

5. صالح، عائدة هادي (2008) " استخدام التحليل المميز لتشخيص بعض امراض العيون"، بحث منشور في مجلة الادارة والاقتصاد، العدد السابع والستون، ص 264-286

6. السليماني، مؤيد سلمان عباس (1998)، استخدام الدالة المميزة لتشخيص حالات التهاب الأمعاء عند الأطفال الرضع، رسالة ماجستير في الإحصاء، مقدمة إلى مجلس كلية الإدارة والاقتصاد، جامعة بغداد.

Table (1) : The Mean for Each Group and the Total Mean

	$\bar{X}_1$	$\bar{X}_2$	$\bar{X}_3$	$\bar{X}_4$	$\bar{X}_5$	$\bar{X}_6$	$\bar{X}_7$	$\bar{X}_8$	$\bar{X}_9$	$\bar{X}_{10}$
Not patients	45.85	199.37	116.80	111.69	64.43	137.61	125.74	0.43	0.63	0.13
patients	63.02	260.57	202.75	184.94	34.10	201.31	155.20	0.31	1.14	0.63
Total	54.19	229.10	158.54	147.27	49.70	168.55	140.05	0.37	0.88	0.37

Table (2) : Classification Result for Discriminant Function

	Actual group	Number of observations	Predicted group	
			Healthy(1)	Disease(2)
	Healthy(1)	54	51	3
	Disease(2)	51	2	49
%	Healthy(1)	100	94.4	5.6
	Disease(2)	100	3.9	96.1

تطبيقات التحليل التمييزي في التشخيص الطبي

حازم منصور كوركيس
أنس خليل محمد
قسم الرياضيات / كلية التربية للعلوم الصرفة (ابن الهيثم) / جامعة بغداد

أستلم البحث في : 19 حزيران 2013 ، قبل البحث في : 4 كانون الاول 2013

الخلاصة

في هذا البحث، استخدم التحليل التمييزي لتصنيف أمراض القلب الأوسع انتشارا والمعروفة باسم أمراض القلب التاجية (انسداد الشرايين) إلى مجموعتين (مريض، غير مريض) على أساس التغيرات التمييزية لعشرة من المتغيرات التنبؤية التي تعتقد انها تسبب المرض.

استخدمت عينة عشوائية لكل مجموعة ونُفّذ اسلوب المراحل لحذف المتغيرات غير المهمة في فصل المجموعات ونُفّذ اختبار معنوية الدالة المميزة ومعدل خطأ التصنيف.

كلمات مفتاحية: التحليل التمييزي ، التصنيف، اسلوب المراحل ، معدل خطأ التصنيف .