THE ECONOMIC
AND SOCIAL
RESEARCH INSTITUTE
MEMORANDUM SERIES
NO.94

BRENDSC: A Computer Program for the Analysis of Survey Data

by Brendan J. Whelan

BRENDSC: A Computer Program for the Analysis of Survey Data

by Brendan J. Whelan *

I . Introduction

Sample surveys are being increasingly used in all the social sciences, and this increase in numbers is matched by the increasing ambitiousness and complexity of the information collected. The analysis of such large quantities of data poses a considerable problem for social scientists. Cross-tabulations alone are no longer sufficient to reflect the richness and depth of the data collected. The sheer number of tables required to cross-tabulate the results of a large survey make it difficult and time-consuming to isolate important relationships, and even when they are found it is difficult to communicate the finding to the reader in a lucid and succinct way. The wealth of data brings with it the temptation to ignore the tiresome ritual of significance testing. It also encourages one to substitute bad tabulations for good theory - much of what masquerades as theory being mere post hoc rationalization.

More powerful methods of analysis are therefore called for. We may divide those which have been developed to date into two groups: (i) methods concerned with analysing the structure of a set of variables, where none of these variables is considered more important than any other. Examples are factor analysis, component analysis and multi-dimensional scaling. (ii) Techniques which investigate the relationship between a specified, important variable, known as the dependent, and a set of independent or "explanatory" variables. Examples are regression, discriminant analysis, AID, MCA [1].

The present program belongs to the second group. Its main advantages are: (i) it forces the researcher to specify an explicit model of human behaviour (ii) it provides a convenient method of compressing and summarizing the net effects of large numbers of variables (iii) it provides tests of significance for both the model as a whole and for the net effects of individual variables.

*The author wishes to thank B.M. Walsh for his constant encouragement and help. Thanks are also due to E. E. Davis, R. C. Geary, P. Neary, N. O'Broin and R. O'Connor for their helpful advice.

The program may be used to calculate either regressions or discriminant functions. The nature of these techniques and the relationship between them are more fully discussed in section II below. For the moment, it is sufficient to note that both techniques involve the calculation of a set of coefficients (weights) for the variables, and that the statistical significance of the coefficient of a certain variable indicates how closely that variable is related to the dependent. The basic difference between the two techniques is the level at which the dependent variable is measured. If the dependent is cardinally measurable, regression analysis is appropriate, while if the dependent is dichotomous (i.e. the dependent is whether or not an individual belongs to a certain group) discriminant analysis should be used.

One of the main factors influencing the design of the program was the persistent problem in the social sciences of the level of measurement which is possible. Frequently, the most important social variables are categorial in nature. For instance, it is easy to divide people into categories on the basis of nationality, but these categories have no necessary ordinal or cardinal relationship with one another. In other words, measurement is only possible on a nominal scale. The program aims to facilitate the inclusion of such variables by assigning a "dummy variable" (see [2] p. 221) to each category of each variable. This gives rise to large numbers of variables and the program is designed to deal economically with them. Although originally developed to handle dichotomons data, the program will also process cardinally measured variables, so that any combination of the two types of variable may be used.

Up to now the program has proved useful in analysing female labour force participation rates [3] and voting patterns [4]. Potential users should consult [3] for a fairly detailed example of the program's use.

This memorandum first of all discusses the nature of regression and discriminant analysis. It then goes on to give technical details of how the program works, a description of the print-out and instructions as to how the data should be set up for the program. The Appendix derives the basic relationship

between regression and discriminant functions, and shows how a priori probabilities may be incorporated in the analysis. Section III and the Appendix are somewhat technical and may be omitted by readers who want only a general idea of what the program does and how to use it.

II. The Nature of Regression Analysis, Discriminant Analysis and Linear Probability Functions

Regressions: (See Johnston [2] Chapters 1 and 4)

The basic objective in regression analysis is to estimate from a sample the relationship which is assumed to exist between a dependent variable, usually designated y, and a set of independent variables, usually designated $x_1 x_2 \ldots\ldots x_k$.

It is presumed that each observed value of y is a weighted sum of the x variables plus a random disturbance i.e. $y = b_o + b_1 x_1 + b_2 x_2 \ldots\ldots b_k x_k + \varepsilon$ where the weights (the b's) are known as the "regression coefficients" and ε is a random disturbance. The regression program provides us with the "best" * possible estimates of these b's, and also tells us how much of the variation in y is "accounted for" or "explained" by variations in the x's.

For instance, say we were interested in explaining consumption of beer and that we have data on the behaviour of a sample of beer drinkers over a period. We might then hypothesize that beer consumption (y) is a function of the price of beer (X_1) the price of spirits (X_2) and the incomes of the individuals (X_3) i.e.:

$$y = b_o + b_1 X_1 + b_2 X_2 + b_3 X_3$$

We could then feed the observed values of the four variables (y, X_1, X_2, X_3) into the regression program and the program will estimate the b's, determine their significance and the significance of the overall relationship. A full description of the information provided by the program is given in section (IV).

* "Best" is used here in the technical series of the linear estimate with minimum variance.

Discriminant Function (See Kendall and Stuart [5])

Say we are given samples from each of two populations. On the basis of the samples, we wish to set up a rule which will enable us to allot a new individual to the correct population when we do not know which population he belongs to. A discriminant function is a set of weights, derived from the samples, which, when applied to the measurements of the individual to be classified, gives his discriminant function score. If this score is above a certain level the individual is allocated to one population and if it is below this level he is allocated to the other population.

For instance, say we have a sample of Americans and a sample of Irishmen and we wish to set up a discriminant function based on height and weight. That is, we wish to derive from the samples two discriminant function coefficients, one for the weight variable and one for height. If we are then given height and weight measurements for a new individual of unknown nationality and asked to allocate him to one population or the other, we would multiply his weight by one coefficient, his height by the other and add the results. If the resulting "discriminant function score" is above a certain "critical" value, the individual is classified as American. If it is below that value he is classified as Irish.

In this example we have assumed that the individual to be classified is equally likely to come from either population. This may not be realistic in practice. Say we know that 20 per cent of the population from which we were sampling were Irish and 80 per cent were American. It would then be reasonable to make it more "difficult" for a new individual to be classified as Irish. That is, the critical value should be adjusted downward. This is an example of the incorporation of "a priori probabilities" in the discriminant. In survey work, we frequently do not have any control over the proportions of our sample which come from each population (i.e. the proportion of women who work; people who smoke; people with high need achievement etc.) and the sample proportions in the two groups may be used as estimates of the a priori probabilities.

The Linear Probability Functions (LPF)(see [6])

This is a particular type of discriminant function which is calculated by regressing a dummy variable (= 1 if the item is in population 1 and zero if it is in population 2) on a set of independent variables. It can be shown that it is a constant multiple of the standard discriminant function, and it may be modified to take account of different costs of misallocation and different a priori probabilities (see the Appendix for a mathematical derivation of these results).

Since it is essentially a regression analysis all the usual statistics and tests of significance may be applied. These include t-tests, analysis of variance and R^2. A special feature of the present program is the inclusion of F-tests for sets of variables (see Section V, item 9 below). These are particularly useful with dummy variable(categorical)regressors, because in this case one is interested both in the overall explanatory power of, say, the age variable, as well as the explanatory power of individual age categories.

There are, however, three difficulties peculiar to LPF's.

(1) Heterovariance: Johnston [2, p. 227] shows that the variance of the random disturbance is not constant, in contradiction of the usual regression assumptions. Goldberger [7] suggests a two step procedure using generalized least squares for dealing with this problem.

(2) $\hat{y} > 1$ or $\hat{y} < 0$: i.e. estimated values of the dependent above 1 or below zero. This is possible, but the actual y value (=the probability of being in population 1) must be between zero and one. Goldberger [7] suggests the use of probit analysis to solve this problem.

(3) Evaluation of Goodness of Fit: The conventional measure of goodness of fit in regression is the adjusted coefficient of determination, $\bar{R}^2$. However, LPF's usually seem to give low values of $\bar{R}^2$; low, that is, by comparison with conventional

regression analysis. This is not surprising when one considers that $\bar{R}^2$ depends essentially on the unexplained sum of squares, i.e. the sum of the squares of the errors one would make when using the LPF to predict the values of y for the individuals in the sample. Thus, even a prediction of y = 0.9 for a certain individual who is in population 1 still makes a contribution of $(1 - 0.9)^2 = .01$ to the unexplained sum of squares. This hardly seems appropriate in the present case when we know that the true value must be either zero or one i.e. we are considering an "either-or" situation where an item either is or is not a member of population 1. The observed values of $\bar{R}^2$ must thus be interpreted with some caution.

An alternative method of assessing goodness of fit is to use the estimated discriminant function to allocate the members of the two samples and to examine the proportion of correct assignments achieved [8, p. 132]. The program provides this information since it gives the total number of correct allocations, the total number of incorrect allocations and the number of correct allocations in the "unit group" (i.e. the category of the dependent which is scored one). These figures allow the following type of tabulation to be drawn up:

		ACTUAL		
		Unit Group	Zero Group	Total
ASSIGNED	Unit Group	179	135	314
	Zero Group	332	1805	2137
	Total	511	1940	2451

These illustrative figures are taken from [3, Table 5]. Here the unit group represented membership of the labour force while the zero group denoted non-membership. This table allows one to compute the overall percentage of correct assignments (= (1984/2451) X 100 = 81% in the above example) and also to compute

the percentage of correct assignments in each group. (= (179/511) x 100 = 35% in the unit group and (1805/1940) x 100 = 93 % in the zero group). In this way, one can get a clear idea of the effectiveness of the estimated discriminant function in allocating to one group or the other.

An even more rigorous test of the quality of an estimated LPF is provided by estimating the function on one sample and testing it on a different one. This ensures that the coefficients are stable over different samples, and so are unlikely to be the result of purely random fluctuations in the original sample. It is advisable, if sample size permits, to divide one's sample into two randomly chosen half samples, and to estimate the LPF on one half sample and test its predictive power on the other half (see [3, Table 6]).

III. How the Program Works

(i) As a Regression Program

Let X be an n x p data matrix whose first column consists entirely of units and p be the total number of (different) variables which will be required in all selections. The program first forms the matrix X'X . Say that k independent variables (including the intercept) are to be included in the first selection. A sub-matrix of X'X which we shall call XPX is then formed containing the (k +1) rows and (k + 1) columns of X'X (i.e. k independents + the dependent) which are required for the first selection. The order of the rows and columns of XPX corresponds to the order in which it is desired to enter the variables into the regression. The row containing the cross-products of the dependent and independents is put on the bottom and the column containing these elements is put on the extreme right. This gives the form of XPX shown in Figure 1.

By "pivoting"* successively on the diagonal elements of XPX, the regression of y on all the variables which have been pivoted on up to that point is obtained. For instance, if we pivot successively on the first m diagonal elements, the first m elements of the last column of XPX will give b_o and b_1 to b_{m-1}, and the lower right-hand element of XPX will give the residual sum of

* The pivoting method used is a variant of the Gauss-Doolittle method of matrix inversion as described in [9, p. 192-196].

squares. (See Figure 2) If we pivot on all diagonal elements down to the k-th we will obtain the regresssion of y on all the variables included (See Figure 3).

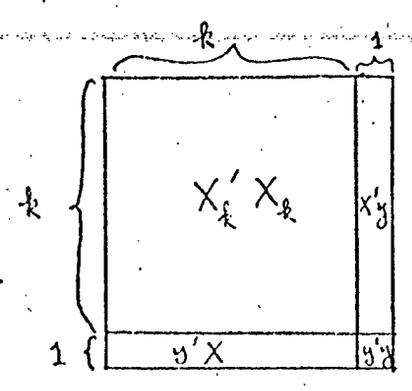

Fig.1: The XPX matrix before pivoting

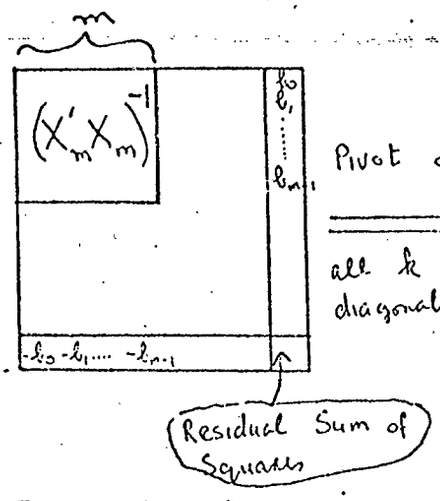

Fig.2: The XPX matrix after pivoting on first m diagonals

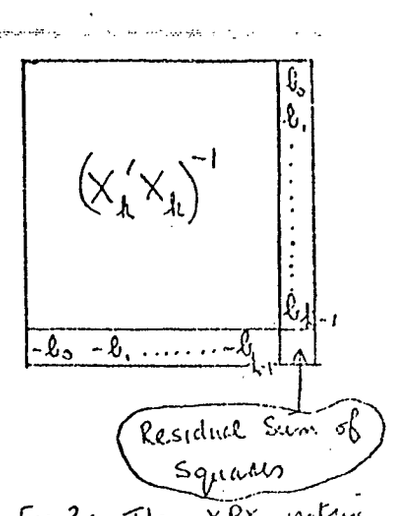

Fig.3: The XPX matrix after pivoting on all k diagonals

This procedure is utilized so that a whole set of equations can be obtained from what is, essentially, a single inversion of the X'X matrix. Thus, for instance, the program will provide the regression of y on X_7, X_{10}, X_{11}; then the regression of y on X_7, X_{10}, X_{11}, X_8, X_9; then the regression of y on X_7, X_8, X_9, X_{10}, X_{11}, X_1, X_2.....X_6. The user has merely to specify the order in which he wants the equation printed out. The order in which the variables appear on the data cards is immaterial, and any number of new variables from 1 to $(k - k_m)$ (where k = the total number of variables and k_m = the number already included) may be added at each stage. Any of the input variables may be specified as the dependent.

(ii) As a Discriminant Program

The procedure used is identical with that described above, except that the dependent variable must be a dummy variable i.e. a variable which has the value zero if an individual is in one group and one if he is in the other. The program

calculates the Discriminant function coefficients and, if desired, it will then go on to calculate the mean of the discriminant function in each group, and the critical value (for details see Appendix below). By means of the critical value, it will then allocate each member of the sample to one or other group. Finally, it will print out the number of correct allocations and the number of incorrect allocations as an indication of the quality of the discriminant function.

The program has two main advantages over the standard IBM [10] and BIOMED [11] discriminant function programs: (i) it can accommodate up to 9999 observations and (ii) it allows "a priori probabilities" to be incorporated in the discriminant function. The program can deal with up to 50 variables, and can be easily modified to include even more. However, users should bear in mind that a considerable amount of computer time is required to handle very large numbers of coefficients.

IV. Print-Out

This section describes the program's print-out.

1. (Optional) "Transformation": This indicates that the user has specified that the variables listed are to be added. The resulting sum is to be given the number of the first variable in the sum.

   Format: " TRANSFORMATION

   VAR NN REPRESENTS NN + MM + II etc."

2. (Optional) "Interaction" This indicates that the user has specified that the variables listed are to be multiplied and will be stored as indicated.

   Format: "INTERACTION OF VARS NN AND MM STORED AS VAR KK".

3. "Expanded XPX matrix bordered by Dependent". This is the XPX matrix described in III above. The totals of each variable will be given by the first row and the sums of squares by the diagonal elements.

4. "Correlation Matrix". The dependent variable will always appear in the last row.

5. "Explained XPX Matrix after Inversion". The right-most column of this matrix gives the regression coefficients, the bottom row gives minus the regression

coefficients and the element in the lower righthand corner gives the residual sum of squares. The remainder of the matrix is the $(X'X)^{-1}$ matrix for this selection.

6. The regression coefficients, their variances, F-values and t-values.

7. Analysis of Variance for the regression.

8. Coefficient of Determination R^2, unadjusted and adjusted.

9. Addition to explained sum of squares achieved by the most recently introduced set of variables. F-value for this Addition.

10. Determinant of Correlation matrix.

11. Farrar-Glauber test for Multicollinearity (see [14]).

12. Haitovsky test for Multicollinearity (see [15]).

13. (For Discriminant Analyses only) Means of Variables in each group.

14. (For Discriminant Analyses only) A Priori Probabilities (see Appendix)

15. (For Discriminant Analyses only) Value of LPF (Linear Probability Function) in each group.

16. (For Discriminant Analyses only) Critical Value.

17. Optional: Residual analysis: Actual value of dependent
Predicted value of dependent
Residual

(For Discriminant Analyses) Allocation on basis of Discriminant Function (= "Disc. Group")

(For Discriminant Analyses) Whether allocation was right (R) or wrong (W).

Durbin Watson Statistic
No. of Sign Changes

(For Discriminant Analyses) Total Number of Correct Allocations
Total Number of Incorrect "
Number of Correct Allocations of individuals belong to the unit group.

V. How to Set Up Data for the Program

The appropriate JCL cards are available from B. Whelan or J. O'Meara, ESRI. Only the cards which vary from problem to problem are described here. A "field" may be defined as the set of columns on a card allocated to a certain variable.

1.1 Control Card (Format (A4, I4, 5I2))

Col. 1-4: 4 characters to name the problem.

5-8: Number of observations.

9-10: Total Number of (different) variables being used in all selections.

11-12: Number of Variables being read in on cards.

13-14: Number of selections.

15-16: (Optional) Number of addition transformations required. If none leave blank.

17-18: (Optional) Number of interactions (multiplications) desired. If none leave blank.

2.1 Card Giving List of Field Numbers of all Variables (Format (40I2))

The number of these should be the same as that punched in Cols. 9-10 of the control card. If there are k interactions (k punched in cols. 15-16 of card 1) the last k variables listed should be the field numbers which it is desired to give to these interactions.

3.1 Format Card (Format (20A4))

Specifies FORMAT in which data is read in.

4. (Optional) Addition Transformations

4.1 Card giving number of variables in each set to be added (FORMAT (40I2))

There should be as many numbers listed on this card as there are addition transformations required (cols. 13-14 of Card 1).

4.2 Cards listing the Field numbers of the variables which comprise each set (FORMAT (4012))

There should be one card for each transformation. Note that the sum is stored under the field number of the first variable in the transformation, so that the first variable in a transformation cannot be included separately in an equation which also contains the transformation.

5. (Optional) Interaction Transformations

5.1 Card listing the pairs of variables whose interactions are required (FORMAT (40I2))

The number of fields punched on the card should be twice the number which appears in cols 15-16 of the control card.

6. Selection Cards Each selection requires 3 cards

6.1 Parameter Card for Selection (FORMAT (5I2, F4.0)).

Cols. 1-2 Field Number of the dependent for this selection

3-4 Number of independent variables in this selection

5-6 Number of separate equations required in this selection

7-8 Residuals Option:

Blank or zero: Residuals not required

-1 : Residuals required for final equation only

k : Residuals required for first k equations.

If residuals are required for all equations set k equal to the number in cols. 5-6.

Cols. 9-10 Discriminant Option

01 if this is discriminant analysis, zero if regression.

11-14 A priori probability of unit group. Punch the decimal point.

6.2: Independent Variable List Card (FORMAT (40I2))

Containing the field numbers of the independent variables for this selection, in the order in which it is desired to enter them.

6.3: Card listing the number of variables to be included in each equation of this selection. (FORMAT (40I2)).

This gives the number of independent variables to be included in each equation. There should be one number for each equation which it is desired to estimate. Each number punched should be larger than the previous number, and the last number should be equal to that in cols. 3-4 of card 6.1

Cards 6.1, 6.2 and 6.3 should be repeated for each selection.

APPENDIX

A Note on the Relationship between Discriminant Functions and Linear Probability Functions

This note presents the following: (i) a derivation of the standard discriminant function (ii) a derivation of the linear probability function (i.e. a discriminant function derived from a dummy variable regression) (iii) a demonstration that these functions are proportional (iv) a method of including a priori probabilities in the L P F formulation. The following discussion draws heavily on Anderson [12] and Kendall and Stuart [5].

(i) The Standard (Two-population) Discriminant Function

Say that there are two populations containing individuals each of which is measured on p variables. An example would be two types of flower, each flower being measured on four variables (p = 4), sepal length, sepal width, petal length and petal width. We wish to derive a set of weights ("coefficients") for these variables such that for any new individual a score can be calculated and the new individual allocated to one population on the other according as this score is above or below a certain critical level. To continue the flower example, we would measure the sepal length, sepal width, petal length and petal width of a new flower of unknown origin, then multiply each of these measurements by the appropriate discriminant function coefficient, sum the results and allocate the new flower to one population or the other depending on the magnitude of the resulting number.

Let the population be represented by two (partially overlapping) clusters of points in a p-dimensional space, (a simple case, p = 2, is shown in Figure 4). Each point represents an individual. Our objective is to set up a boundary in the space such that as many as possible of population 1 lie on one side and as many as possible of population

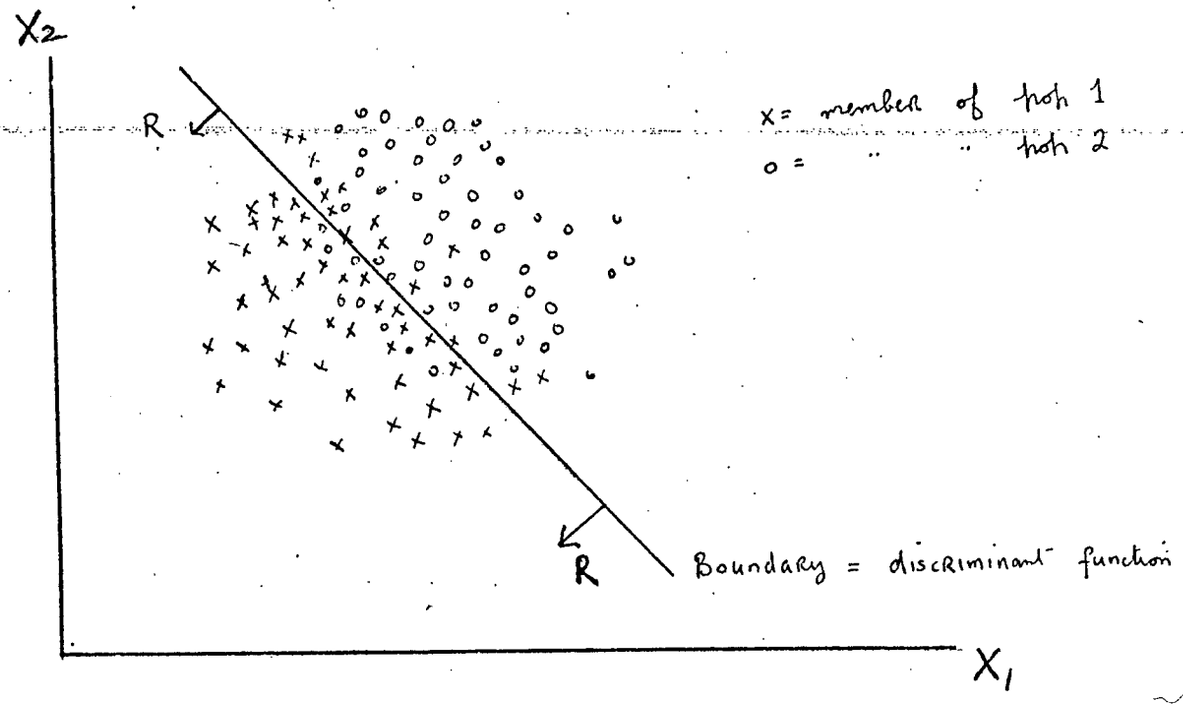

Figure 4: A simple discriminant function (p = 2)

2 on the other. In the case shown in Figure 4 the boundary is a line. In general, it is a (hyper-) plane. The new individual to be classified is represented by a p-dimensional vector, x.

Two further complications may now be introduced:

(1) We may know that members of population 1 have a different chance of occurrence from those of population 2. Such "a priori probabilities" are designated π_1 and π_2. In our flower example, if it is known that type 1 is four times more common than type 2, then $\pi_1 = .8$ and $\pi_2 = .2$.

(ii) The consequences (cost) of misallocation may be different. In a medical application, it is much less dangerous to diagnose a healthy person as unhealthy (because the error is likely to be discovered by subsequent tests) than an unhealthy person as healthy. The cost of misallocating a member of population 2 to population 1 is denoted by C(1/2) and of misallocating

a member of population 1 to population 2 by C(2/1).

Letting f_1 and f_2 be the frequency functions of populations 1 and 2 respectively, the expected cost of misallocation is

$$M = \int_R \pi_2 \, c(1/2) \, f_2 \, dx + \int_{1-R} \pi_1 \, c(2/1) f_1 \, dx$$

where R is the region in the p-space in which individuals are allocated to population 1.

$$M = c(2|1)\,\pi_1 + \int_R \left(c(1/2)\,\pi_2 f_2 - c(2/1)\pi_1 f_1 \right) dx$$

In order to determine the boundary, we must minimize M. This will be achieved by taking into R all those points, and only those points where

$C(1/2)\ \pi_2 f_2 - C(2/1)\,\pi_1 f_1 < 0$

$$\therefore \quad \frac{C(2/1)\,\pi_1 f_1}{C(1/2)\,\pi_2 f_2} \geq 1 \qquad \text{will determine the boundary.}$$

Let us assume that f_1 and f_2 are multivariate normal distributions with common variance-covariance matrix Σ.

$$\text{i.e.} \quad f_i(x) = \frac{1}{(2\pi)^{\frac{1}{2}n} |\Sigma|^{\frac{1}{2}}} \left\{ \exp\left[-\tfrac{1}{2} \left(x - \mu^{(i)}\right)' \Sigma^{-1} \left(x - \mu^{(i)}\right) \right] \right\}$$

where $\mu^{(i)}$ is the vector of means from population i and Σ is the matrix of variances and covariance of each population.

Let $$\frac{C(2/1)\,\pi_1}{C(1/2)\,\pi_2} = K$$

$$\therefore \frac{c(2|1)\,\pi_1 f_1}{c(1|2)\,\pi_2 f_2} \geq 1 \Rightarrow K \left[\frac{\exp\left[-\frac{1}{2} \left(x - \mu^{(1)}\right)' \Sigma^{-1} \left(x - \mu^{(1)}\right) \right]}{\exp\left[-\frac{1}{2} \left(x - \mu^{(2)}\right)' \Sigma^{-1} \left(x - \mu^{(2)}\right) \right]} \geq 1 \right.$$

Since log is a monotonically increasing function, this inequality can be written in terms of logs :

$$-\frac{1}{2}\left[(x-\mu^{(1)})'\Sigma^{-1}(x-\mu^{(1)}) - (x-\mu^{(2)})'\Sigma^{-1}(x-\mu^{(2)})\right]$$

must be greater than or equal to log (1/K).

Expanding,

$$-\frac{1}{2}\left[x'\Sigma^{-1}x - x'\Sigma^{-1}\mu^{(1)} - \mu^{(1)}\Sigma^{-1}x + \mu^{(1)}\Sigma^{-1}\mu^{(1)} - x'\Sigma^{-1}x + x'\Sigma^{-1}\mu^{(2)} + \mu^{(2)\prime}\Sigma^{-1}x - \mu^{(2)}\Sigma^{-1}\mu^{(2)}\right] \geq \log(1/K)$$

This gives, by rearrangement of terms

$$x'\Sigma^{-1}(\mu^{(1)}-\mu^{(2)}) - \frac{1}{2}(\mu^{(1)}+\mu^{(2)})'\Sigma^{-1}(\mu^{(1)}-\mu^{(2)}) \geq \log(1/K)$$

If this inequality holds then the new item is allocated to population 1 and if it does not hold it is allocated to population 2. The "discriminant function" coefficients are given by the vector $\Sigma^{-1}(\mu^{(1)}-\mu^{(2)})$. The first term on the left is the discriminant function, evaluated for the individual to be classified. The second term is the discriminant function evaluated at a point mid-way between the population means.

If C(1/2) = C(2/1) (i.e. the costs of misallocation are equal) and $\pi_1 = \pi_2$ (i.e. the a priori probabilities are equal, or unknown) then K = 1 and log (1/K) =O. Hence, in this case, the second term on the left is the critical value of the discriminant function i.e. the value above which items are allocated to population 1. Thus, the effect of introducing (differential) costs of misallocation and (differential) a priori probabilities is to displace the critical value by a constant, equal to $\log\left(\frac{C(1/2)\,\pi_2}{C(2/1)\,\pi_1}\right)$

The above is the parental form of the discriminant. In practice, we must replace each term by its sample counterpart i.e. $\mu^{(i)}$ by $\bar{x}^{(i)}$ and Σ by the matrix of pooled variances and covariances of the

variables. In a survey where the sample of N was chosen without reference to which population an item belonged, then π_i may be estimated by $\frac{N_i}{N}$ where N_i is the number of the items found to be in the i-th population. (Note that $N_1 + N_2 = N$ the total sample).

This gives

$$x' S^{-1} \left(\bar{x}^{(1)} - \bar{x}^{(2)}\right) - \tfrac{1}{2} \left(\bar{x}^{(1)} + \bar{x}^{(2)}\right)' S^{-1} \left(\bar{x}^{(1)} - \bar{x}^{(2)}\right) \geq \log (1/K)$$

where

$$S = \frac{1}{(N_1+N_2-2)} \left\{ \sum_{\alpha=1}^{N_1} \left(x_\alpha^{(1)} - \bar{x}^{(1)}\right)\left(x_\alpha^{(1)} - \bar{x}^{(1)}\right)' + \sum_{\alpha=1}^{N_2} \left(x_\alpha^{(2)} - \bar{x}^{(2)}\right)\left(x_\alpha^{(2)} - \bar{x}^{(2)}\right)' \right.$$

The Linear Probability Function

The following is an alternative derivation of the discriminant function (see Fisher [13]).

Let y be a dummy variable = 1 if the item is in population 1
= 0 if the item is in population 2

Then find the regression of y on the $x_\alpha^{(i)}$ variates by choosing b to minimize

$$\sum_{i=1}^{2} \sum_{\alpha=1}^{N_i} \left\{ y_\alpha^{(i)} - b' \left(x_\alpha^{(i)} - \bar{x}\right) \right\}^2$$

where the first element of x is equal to 1 for i = 1, 2 and

$\bar{x} = (N_1 \bar{x}^{(1)} + N_2 (\bar{x}^{(2)})) / (N_1 + N_2)$ i.e. the overall mean of x

The "normal equations" are

$$\sum_{i=1}^{2} \sum_{\alpha=1}^{N_i} \left(x_\alpha^{(i)} - \bar{x}\right)\left(x_\alpha^{(i)} - \bar{x}\right)' b = \sum_{i=1}^{2} \sum_{\alpha=1}^{N_i} \left\{ y_\alpha^{(i)} \left(x_\alpha^{(i)} - \bar{x}\right) \right\} \quad \text{..... (A)}$$

Since y = 0 for all i = 2, the right hand side of (A) reduces to

$$\sum_{\alpha=1}^{N_1} \left(x_\alpha^{(1)} - \bar{x}\right) = \sum_{\alpha=1}^{N_1} x_\alpha^{(1)} - N_1 \bar{x}$$

$$= N_1 \bar{x}^{(1)} - N_1 \left(N_1 \bar{x}^{(1)} + N_2 \bar{x}^{(2)}\right) / (N_1+N_2)$$

$$= \frac{N_1^2 \bar{x}_1^{(1)} + N_1 N_2 \bar{x}^{(1)} - N_1^2 \bar{x}^{(1)} - N_1 N \bar{x}^{(2)}}{N_1 + N_2}$$

$$= \frac{N_1 N_2}{N_1+N_2} \left\{ \bar{x}^{(1)} - \bar{x}^{(2)} \right\}$$

the

The matrix multiplying b on left-hand side of (A) simplifies as follows

$$\sum_{i=1}^{2} \sum_{\alpha=1}^{N_i} \left(x_\alpha^{(i)} - \bar{x}\right)\left(x_\alpha^{(i)} - \bar{x}\right)' =$$

$$= \sum_{i=1}^{2} \sum_{\alpha=1}^{N_i} \left(x_\alpha^{(i)} - \bar{x}^{(i)}\right)\left(x_\alpha^{(i)} - \bar{x}^{(i)}\right)' + N_1 \left(\bar{x}^{(1)} - \bar{x}\right)\left(\bar{x}^{(1)} - \bar{x}\right)' + N_2 \left(\bar{x}^{(2)} - \bar{x}\right)\left(\bar{x}^{(2)} - \bar{x}\right)'$$

$$= \sum_{i=1}^{2} \sum_{\alpha=1}^{N_i} \left(x_\alpha^{(i)} - \bar{x}^{(i)}\right)\left(x_\alpha^{(i)} - \bar{x}^{(i)}\right)' + \frac{N_1 N_2}{N_1+N_2} \left\{ \left(\bar{x}^{(1)} - \bar{x}^{(2)}\right)\left(\bar{x}^{(1)} - \bar{x}^{(2)}\right)' \right\}$$

(A) can therefore be re-written as

$$C\, \ell = \left(\bar{x}^{(1)} - \bar{x}^{(2)}\right) \left[\frac{N_1 N_2}{N_1+N_2} - \frac{N_1 N_2}{N_1+N_2} \left(\bar{x}^{(1)} - \bar{x}^{(2)}\right)' \ell \right]$$

where

$$C = \sum_{i=1}^{2} \sum_{\alpha=1}^{N_i} \left(x_\alpha^{(i)} - \bar{x}^{(i)}\right)\left(x_\alpha^{(i)} - \bar{x}^{(i)}\right)' = \frac{1}{(N_1+N_2-2)} S$$

But $(\bar{x}^{(1)} - \bar{x}^{(2)})'$ b is a scalar. Hence, b is proportional to $S^{-1} (\bar{x}^{(1)} - \bar{x}^{(2)})$

and the constant of proportionality is

$$P = \left[\frac{N_1 N_2}{N_1+N_2} \left\{ 1 - \left(\bar{x}^{(1)} - \bar{x}^{(2)}\right)' \ell \right\} \right] / (N_1+N_2-2)$$

$$= \left[\frac{N_1 N_2}{N_1+N_2} \left\{ 1 - \hat{\bar{y}}^{(1)} + \hat{\bar{y}}^{(2)} \right\} \right] / (N_1+N_2-2)$$

where $\hat{\bar{y}}^{(i)}$ is the predicted value of y at the mean of population i.

Since the functions are proportional, it is clear that their discriminatory power will be the same. Thus, as Ladd [6] points out "Discriminant analysis and linear probability analysis start from quite different places but end up at nearly the same place".

The Inclusion of Costs of Misallocation and A Priori Probabilities in the Linear Probability Function (LPF)

Morrison [8] shows that the inclusion of a priori probabilities often substantially improves the discriminatory power of estimated discriminant functions. The author of the present note has been unable to find a case where such a priori probabilities were included in an LPF analysis. The LPF formulation has several advantages over the standard discriminant, notably its similarity to regression and the possibility of interpreting the coefficients as conditional probabilities. In order to include a priori probabilities in the estimated LPF, a modified version of the critical value was derived and this is presented below. To achieve greater generality, both costs of misallocation and a priori probabilities are assumed unequal.

Let b_1 be the vector of coefficients derived from a standard discriminant function, and b_2 be the vector of coefficients derived from an LPF analysis. We have shown above that

$$b_2 = b_1 \left[\frac{N_1 N_2}{N_1+N_2} \left\{ 1 - \hat{\bar{y}}^{(1)} + \hat{\bar{y}}^{(2)} \right\} \right] / (N_1+N_2-2)$$

$$= b_1 \; P$$

We have also shown that the optimum classificatory rule when using b_1, allowing for $\pi_1 \neq \pi_2$ and $C(1/2) \neq C(2/1)$ is:

if $x' b_1 \geq \frac{1}{2} \left(\bar{x}^{(1)} b_1 + \bar{x}^{(2)} b_1 \right) + \log (1/K)$

allocate to population 1 and otherwise allocate to population 2.

If we express this rule in terms of b_2 we obtain

$$x' b_2 / P > \tfrac{1}{2} (\bar{x}^{(1)} b_2 / P + \bar{x}^{(2)} b_2 / P) + \log (1/K)$$

$$\text{or} \quad x' b_2 > \tfrac{1}{2} (\bar{x}^{(1)} b_2 + \bar{x}^{(2)} b_2) + P \log (1/K)$$

$$\text{i.e.} \quad \hat{y} > \tfrac{1}{2} (\hat{\bar{y}}^{(1)} + \hat{\bar{y}}^{(2)}) + (\log (1/K)) \left[\frac{N_1 N_2 (1 - \hat{\bar{y}}^{(1)} + \hat{\bar{y}}^{(2)})}{N(N-2)} \right]$$

where $\hat{y}$ is the value of the L P F for the individual to be allocated.

The latter is the rule which the program uses to allocate individuals.

REFERENCES

[1] Sonquist, J. A. "Multivariate Model Building - The Validation of a Search Strategy", Institute for Social Research. Ann Arbor Michigan, 1970.

[2] Johnston, J., Econometric Methods (First Ed.) McGraw-Hill, New York, 1963.

[3] Walsh, B.M. and Whelan, B.J. "The Determinants of Female Labour Force Participation" Journal of the Statistical & Social Inquiry Society of Ireland (forthcoming).

[4] Walsh, B.M. and Robson, C. "Alphabetical Voting: A Study of the February 1973 General Election," ESRI, Paper No. 71, Dublin, 1973.

[5] Kendall, M. G., and Stuart, A. The Advanced Theory of Statistics, Vol. 3, Chapter 24, Griffin, London, 1968.

[6] Ladd, G. W. "Linear Probability Functions and Discriminant Functions" Econometrica Vol. 34, No. 4, October 1966.

[7] Goldberger, A.S. Econometric Theory: John Wiley & Sons, New York, 1964.

[8] Morrison, D. F. Multivariate Statistical Methods: McGraw-Hill New York, 1967.

[9] Anderson, R.L. and Bancroft T. A. Statistical Theory in Research; McGraw-Hill, New York, 1952.

[10] Dixon, W.J., BIOMED Manual, University of California Press, Berkeley, 1968.

[11] IBM SSP Manual; IBM Ltd., New York, 1970.

[12] Anderson, T.W., An Introduction to Multivariate Statistical Analysis, John Wiley & Sons, New York, 1958.

[13] Fisher, R. A. "The use of multiple measurements in taxonomic problems". Ann. Eugen., 7, (pp. 179-188), 1936.

[14] Farrar, D., and Glauber, R., "Multicollinearity in Regression Analysis; The Problem Revisited" Review of Economics and Statistics, Vol. 49, pp. 92-107, 1967.

REFERENCES

[15] Haitovsky, Y. "Multicollinearity in Regression Analysis: Comment" Review of Economics and Statistics, Vol. 51, pp. 486-489, 1969.