APPLICATION OF DISCRIMINANT ANALYSIS ON CLASSIFIED GROUP

IWEKA FIDELIS, PhD.
Department of Educational Psychology, Guidance and Counseling
University of Port Harcourt
Port Harcourt

And

TEI-FIRSTMAN, ROSE
Department of Educational Psychology, Guidance and Counseling
University of Port Harcourt
Port Harcourt

ABSTRACT

This paper explored the application of discriminant analysis on classified groups. Discriminant analysis is a statistical procedure that is related to multiple regressions which uses a number of predictor's variables to classify subjects into two or more distinct groups. It highlighted the steps applied for the two-group discriminant analysis, illustrated the steps with practical examples and explained the directions for multiple discriminant analysis. The paper also explained related problems in discriminatory analysis, elements of discriminant analysis, assumptions and uses of discriminant function analysis.

Introduction

Discriminant analysis was initially developed by Fisher (1936) for the purpose of classifying objectives into one or two clearly defined groups (Nunnally 1981). Classification is an important component of virtually all scientific research. Statistical techniques concerned with classification are essentially of two types. The first is cluster and the other is discriminant function analysis (Evertt and Landau 2004).

The most well-known technique is fishers linear (discriminant function analysis. Discriminant analysis is a statistical procedure related to multiple regression, but it differs in that, the criterion is a categorical variable rather than a continuous ne. discriminant analysis uses a number of predictor variables to classify subjects into two or more distinct groups, such as dropouts versus persisters, successful versus unsuccessful students, delinquents versus non-delinquents, etc. the criterion in discriminant analysis is thus a person's group membership. We might predict teacher retention based on measures of self-efficacy and job satisfaction. The procedure results in an equation or discriminant function, where the scores on the predictors are multiplied by weights to predict the classification of subjects into groups. When there are just two groups, the discriminant function is essentially a multiple-regression equation with the group membership criterion coded 0 or 1. But with three or more groups as the criterion, discriminant analysis goes beyond multiple regressions. In most predict studies, the criterion variable is quantitative that is, it involves scores that can fall anywhere along a continuum from low to high. For example, the CGPA of university is quantitative variable for scores on the variable can fall anywhere at or between 0.00 and 1.00. (Panneersalvan 2004).

Sometimes, however the dependent variable may be categorical variable that is, it involves membership in a group (or category) rather than score along a continuum for example a researcher might be interested in predicting whether an individual is more engineering majors or business majors. In this instance, the criterion variable is dichotomous- an individual is either in one group or the other of course, a categorical variable can have more than just two categories (for example engineering majors, business major, education majors, science majors and so on) the technique of multiple regression cannot be used when the criterion variable is categorical, instead, a technique known as discriminant function analysis is used. Often, this design is very appropriate for vocational and careers development.

One of the main goals of education is preparation of people for various occupations or professions that are most suitable for them (Kpolovie, 2010). Here, profession, career, vocation or occupation is of course a categorical variable. Discriminant analysis will actually predict students who will excel in engineering and those who will best excel in law on the basis of several tests (Independent variables) administered to them on the one hand, the design will adequately sort out all student that do well on engineering test or subtests and do poorly on law-related tests or subtest and put them together. On the other hand the design allows for identification of students who will excel in law via their high performance in law related test on subtests and poor performance in test or subtest that are engineering related. Based on this the best possible predictive decision as to which profession or vocation each of the students should study or get engaged can be made.

Discriminant, analysis aims at studying the effect of two or more predictor on certain evaluation criterion. The evaluation criterion may be good or bad, like and dislike. Successful or unsuccessful, etc. hence it is one of the advanced topics of multivariate analysis. In some situation, it may be essential to study the effect of two or more predictor variable on certain evaluation criterion as listed below.

EXAMPLE: Above expected level or below expected level

- While grouping students after a training program in terms of their enhanced skills, the criterion.
- While grouping the performance of educational institutions in terms of their achievements, the criterion will be “above expected level or below expected level.

The researcher will be keen in checking whether the predictor variables discriminant among the groups. More especially, it is necessary to identify the predictor variable (independent variable) which is more important when compared to the other predictor variables such analysis is called discriminant analysis. In this presentation, two group discriminant analyses are presented with an illustration. The different stages of the discriminant analysis as listed by Panneersalvam, (2004) are as follows;

- Designing a discriminant function like the one shown below

  $Y=ax_1\ b\ x_2$

  Where T is the linear composite representing the discriminant function, x_1 and x_2 are the predictor variables (Independent variables) which are having effect on the evaluation criterion of the problem of interest of the problem of interest.

- Finding the discriminant ratio (k) and determining the variables which account for inter-group differences in terms of group means.
  This ratio is maximum possible ratio between the "Variability within groups".
- Finding the critical value which can be used to include a new data set (i.e new combination of instances for the predictor variables) into its appropriate group.
- Testing the null hypothesis, Ho: the group means are equal is importance, against the alternate hypothesis. H.: the group means are not equal in importance, using F test at a given significance level.

Steps of Two-Group Discriminant Analysis

The steps of discriminant analysis applied to two groups are presented below:

Step 1: Input the data, let the predictor variables representing the two factors be x_1 and X_2

Step 2: Classify the data into mutually exclusive and collectively exhaustive groups G1 and G2. Some sample combinations of G1 and G2 are shown below:

Group 1 (G1)	Group 2 (G2)
Dislike	Like
Bad	Good
Below	Above

Let n1 and n2, be the number of sets of observations in the Group – 1 and Group – 2, respectively.

Step 3: Find the mean of X_1 as well as X_2 in each group

Let $\overline{X}_1$ $(_{G1})$ be the mean of X_1 in group -1, and X_2 $(_{G2})$, be the mean of X_2 in group -2 also find the grand mean of X_1, as well as X_2.

Step 4: In each group, find $\sum X_1^{2,}$ $\sum X_2^2$ and $\sum X_1, X_2$

Step 5: Define the linear composite as $Y = aX, + bX_2$

Step 6: Find the values of a and b by solving the following normal equations: $a \sum (X_1 - \overline{X}_1)^2 + b \sum (X_1 - \overline{X},^2) (X_2 - \overline{X}_2) = X_1 (_{G2}) - X_1(_{G1})$ $a \sum(X_1 - \overline{X}_1) (X_2 - \overline{X}_2) + b \sum(X_2 - \overline{X}_2)^2 = X_2 (_{G2}) - X2 (_{G1})$

The sum of squares in the above normal equation can be substituted with the following simple formulas:

$\sum(X_1 - \overline{X}_1)^2 = \sum(X_1 - \overline{X}_1(_{G1}))^2 + \sum(X_1 - \overline{X}_1(_{G2}))^2$

$\sum(X_2 - \overline{X}_2)^2 = \sum(X_2 - \overline{X}_2(_{G1}))^2 + \sum(X_2 - \overline{X}_2(_{G2}))^2$

$\sum(X_1 - \overline{X}_1) (X_2 - \overline{X}_2) = \sum(X_1 - \overline{X}_1(_{G1})) (X_2 - \overline{X}_2(_{G1})) + \sum(X_1 - \overline{X}_1(_{G2})) (X_2 - \overline{X}_2(_{G2})$

Where $\sum(X_1 - \overline{X}_1(_{GI}))^2 = \sum X_{2,} - n_1 \overline{X}_2(_{G1})$ $\sum(X_1 - \overline{X}_1(_{G2}))^2 = \sum X_{2,} - \overline{2} X^2(_{G1})$

$\sum(X_2 - \overline{X}_2(_{G1}))^2 = \sum X^2_{2,} - n_1 \overline{X}^2_2(_{G2})$ $\sum(X_2 - \overline{X}_2(_{G2}))^2 = \sum X^2_{2,} - n_2 \overline{X}^2_2(_{G2})$

$\sum (X_1 - X_{1\,(G1)}) (X_2 - X_{2\,(G1)})$ $\sum X_1 X_2$ $n_1 X_{1(G1)} \sum(X_1 - X_1(_{G2})) (X_2 - X_2\,(_{G2})) = \sum X_1 X_2 - n_2 X_{1\,(G2)} X_{2\,(G2)}$

Note: In each of the above equations $\sum X_1^{2,}$ $\sum X_2^2$, and $\sum X_1 X_2$ are computed using the data sets of the corresponding group.

Step- 7: In each group find the discriminant score for each combination of the variables X_1 and X_2. Then, find the average of the discriminant scores of each group and the grand mean of the discriminant scores for the entire problem.

Step 8: Find the variability between groups (V_{BG}) using the formula
$V_{BG} = n_1, (\overline{S}_1 - S)^2 + n_2 (\overline{S}_2 - \overline{S})^2$ where S_1 and S_2 are the means of the discriminant scores in the group 1 and group 2, respectively, and S is the grand mean of the discriminant score of the entire problem.

Step 9: Find the variability within groups (V_{BG}) using the following formula:

$$V_{wG} = \sum_{J=I}^{n1} \left(S_{1j} - \overline{S}_1 \right)^2 + \sum_{J=I}^{n2} \left(S_{2j} - \overline{S}_2 \right)^2$$

Where S_{1j} and S_{2j} are the discriminant scores for the jth set of observation in Group – 1 and -2, respectively; $\overline{S}_1$ and $\overline{S}_2$ are already defined in step -8.

Step 10: Find the discriminant ration, (K) Using the Following formula and identify the predictor variable (S) which has/have more importance when compared to the other predictor variable.

$$K = \frac{V_{BG}}{V_{WG}}$$

This is the maximum possible ratio between "the variability between groups and the variability within groups".

Step 11: Validate the discriminant function using the given data sets by forming groups based on the critical discriminant scores (grand mean of discriminant scores). If the discriminant score of the data set is less than the critical discriminant score, then include the member of the entity representing that data set in the "Below" category; otherwise, include it in the "Above" category.

Direction: To classify entity member of future data set.

In future, if the values of the predictor variables are known, then the discriminant score of that data set can be obtained using discriminant function. Then, as per the guidelines stated earlier, the corresponding member of the entity representing that data set can be included in the appropriate group.

Step 12: Find F ratio using the following formula:

$$F = \frac{n_1 n_2 (n_1 + n_2 - m - 1)}{m (n_1 + n_2)(n_1 + n_2 - 2)} D^2$$

and $D^2 = (n_1 + n_2 - 2) (a (\overline{X}_{1 (G2)} - \overline{X}_{1 (G1)}) + b (\overline{(X}_{1(G2)} n - \overline{X}_{1 (G1)})$

Where m is the number of predictor variables (in the case, it is equal to 2); D^2 is known as Mahalanobi's squared distance and degree of freedom = (m, $(n_1 + n_2 - m - 1)$.

Step 13: Find the table F value for (m, $(n_1 + n_2 - m - 1)$ degree of freedom at a significant level of

Step 14: If the calculated F value is more than the table F value, reject the null hypothesis, Ho; otherwise, accept the hypothesis (Ho) where
Ho: The group mean are equal in importance
H_1: The group mean are not equal in importance.
(This means that one variable is more important than other variables).

Example: The Dean of faculty of education wants to do discriminant analysis concerning the effect of two factors, namely, the yearly spending on infrastructures of the facility (X_1) and the yearly spending on the interface events of the faculty (X_2) on the grading of the faculty by an inspection team. He has collected for the past 12 years and submitted to the inspection team as shown in the table. 1. 1. Base on the data, the committee had awarded on the following grades for each year, as shown in the same table. 1. 1.

a. Design the discriminant function,
   $Y = aX_1 + bX_2$
b. Computer the discriminant ratio, K and identify the variable which is more important in relation to the other variable.
c. Validate the discriminant function using the given data by forming groups based on the critical discriminant score.
d. Test whether the group mean are equal in importance at a significance level of 0.05.

Solution

The combination of the hypothesis of this example is;
Ho: The group mean are equal in importance.
H_1: The group mean are not equal in importance
a) Design of the discriminant function, $Y = aX_1 + bX_2$

Step 1 & 2: Input data and award of grades by the inspection team are as already shown in table 1.1.

Table 1.1: Yearly spending on infrastructures and interface events

Year	Grades	X_1 infrastructure	X_2 Interface Event
1	Below	3	4
2	Below	4	5
3	Above	10	7
4	Below	5	4
5	Below	6	6
6	Above	11	4
7	Below	7	4
8	Above	12	5
9	Below	8	7
10	Below	9	5
11	Above	13	6
12	Above	14	6

Step 3: The mean of X_1 as well as X_2 in each group is shown in Table 1.2. The grand means of X_1 and X_2 are also shown in the same table.

Step 4: The calculations of necessary results to solve normal equations are summarized in Table 1.3. Also, they are summarized in refined form, as in table 1.4.

Table 1.2. Summary of means of X_1 and X_2 and for each group.

Group	Year	Standard	X_1 infrastructure	Interface Event X_2
G1	1	Below	3	4
	2	Below	4	5
	4	Below	5	4
	5	Below	6	6
	7	Below	7	4
	9	Below	8	7
	10	Below	9	5
		Total Below	**42**	**35**
		Mean Below	**6**	**5**
G2	3	Above	10	7
	6	Above	11	4
	8	Above	12	5
	11	Above	13	6
	12	Above	14	8
		Total Above	**60**	**30**
		Mean Above	**12**	**6**
		GRAND MEAN	**8.5**	**5.41668**

Table 1.3 Calculations of Necessary Results to Solve Normal Equations

GROUP	YEAR	STANDARD	X_1	X_2	X_1^2	X_2^2	$X_1 X_2$
	1	Below	3	4	9	16	12
G1	2	Below	4	5	16	25	20
	4	Below	5	4	25	16	20
	5	Below	6	6	36	36	36
	7	Below	7	4	49	16	28
	9	Below	8	7	64	49	56
	10	Below	9	5	25	25	45
Total Below			**42**	**35**	**280**	**183**	**217**
Mean Below			**6**	**5**			
G2	3	Above	10	7	100	49	70
	6	Above	11	4	121	16	44
	8	Above	12	5	144	25	60
	11	Above	13	6	169	36	78
	12	Above	14	8	196	64	112
Total Above			**60**	**30**	**730**	**190**	**364**
Mean Above			**12**	**6**			
Grand Mean			**8.5**	**5.41668**			

Table 1.4 Refined Forms of Sum of Squares

SUM OF SQUARES	BELOW	ABOVE	TOTAL
$\sum (X_1 - X_1)^{2} = \sum X^2_1 - n X^2_1$	28	10	38
$\sum (X_2 - X_2)^{2} = \sum X^2_2 - n X^2_2$	8	10	18
$\sum (X_1 - X_1) (X_2 - X_2) = \sum X_1 - X_2 - n X_1 X_2$	7	4	11

Note: n is the number of sets of observations in each category (above/below)

Step 5: Let the linear composite discriminant function be as follows:

$Y = aX_1 + bX_2$

Step 6: The normal equations are as shown below

$a\sum(X_1 - X_1)^2$ 1 $b\sum(X_1 - X_1)(X_2 - X_2)$ $X_{1\ (G2)} - X_{1\ (G1)}$

$a\sum(X_1 - X_1)^2$ 1 $b\sum(X_1 - X_1)(X_2 - X_2)$ $X_{2\ (G2)} - X_{1\ (G1)}$

Subtraction of the result from the column of table 1.4, the normal equations are as presented below:

38a + 11b = 12-6 = 6 --- (1)

11a + 18b = 6-5 = 1 --- (2)

The solution to the above two simultaneous equations are:

a = 0.17229 and

b = 0.04973

Hence, the discriminant function is as shown below:

$Y = 0.17229x_1 - 0.04973x_2$

b. Computation of discriminant ratio, K and identification of the variable which is more important in relation to the other variable.

Step 7: Discriminant score for each combination of X1 and X2 in each group. The average of discriminant scores of each group and the grand mean of discriminant scores are summarized in table 1.5.

Alternately, the mean discriminant score of each group as well the grand mean discriminant score can be obtained using the discriminant function, as shown below.

$\overline{Y}$ (Below) = $0.17229\ \overline{X}_1 - 0.04973\overline{X}_2$

= 0.17229x6 – 0.04973 x 5 = 0.78509

Table 1.5: Summary of discriminant scores and their group averages

Discriminant Function: $Y = 0.17229X_1 - 0.4973X_2$							
Below Group –G_1				**Above Group –G_2**			
Data set (ij)	Year	Discriminant Score (sij)	$(sij-S_1)^2$	Data set (ij)	Year	Discriminant score (sij)	$(sij-S_2)^2$
1	1	0.31795	0.218220	1	3	1.37479	0.155480
2	2	0.44051	0.118735	2	6	1.69627	0.005304
3	4	0.66253	0.015021	3	8	1.81883	0.002473
4	5	0.73536	0.002473	4	11	1.94139	0.029684
5	7	1.00711	0.049293	5	12	2.01422	0.060084
6	9	1.03021	0.06084				
7	10	1.30196	0.267155				
Total		5.49563	0.730981	**Total**		**8.84550**	0.253025

Mean Mean (S_1) = 0.78509 Mean (S_2) = 1.7691

Grand total of discriminant scores = 14.34113
Grand mean of discriminant scores (s) = 1.195094

$\overline{Y}$ (above) = $0.17229\overline{X}_1 - 0.04973\overline{X}_2$
= 0.17229 x 12 – 0.04973 x 6
= 1.7691

$\overline{Y}$ (Grand Mean) = $0.17229\overline{X}_1 - 0.04973\overline{X}_2$
= 0.17229 x 8.5 – 0.04973 x 5.41666
= 1.195094

Step 8: The variability between groups (V_{BG}) is computed using the formula. Sum of squares between groups:
$V_{BG} = n_1 (\overline{S}_1 - \overline{S}_2)^2 + n_2(\overline{S}_2 - \overline{S})^2$
= $7(0.78509 - 1.1950942) + 5 (1.7691 - 1.195094)^2$
= 2.824137

Step 9: The variability within groups (VWG) is computed using the formula:
Sum of square within groups;
$V_{WG} = \sum (\overline{S}_{ij} - S_1)^2 + \sum (S_{2j} - \overline{S}_2)^2$
= 0.730981 + 0.253035
= 0.984006

Step 10: The discriminant ratio, K is determined as shown below:

$$K = \frac{V_{BG}}{V_{WG}} = \frac{2.824137}{0,984006} = 2.87$$

This is the maximum possible ratio between the variability between groups and the "Variability within groups". This means that if the variables X_2 is suppressed in the discriminant function, than the revise value of the discriminant ratio will be less than the previous value got.

In the Discriminant Function,
$Y = 0.17229X_1 - 0.04973X_2$, the coefficient of X_2 has negative sign which indicate that the variable X_1 (spending on infrastructure) is more important than the variable X_2 (spending on interface event).

C. Validate the discriminant function using the given data by forming groups based on the critical discriminant score.

Step 11: the discriminant function can be validated based on the grand mean of the discriminant scores which is known as the critical discriminant score.

If the discriminant score of a data set is less than the critical discriminant score (1.195094), include that data set in to group corresponding to "Below" – category; otherwise, include that data set into "Above" category.

As per these guidelines, the classification of the twelve data sets of the given problems is shown in Table 1.6. one can notice that the discriminant function has grouped all the years into their original category as given in the problem expect the year.

Table 1.6: Classification of data sets based on critical discriminant score (each entry of the table represents year)

Below	Above
1	10
2	3
4	6
5	8
7	11
9	12

Direction for including future data set: In future, say for the 11^{th} year, if the values of the predictor variables X_1 and X_2 are known, then its discriminant score can be obtained using discriminant function. Then, as per the guidelines stated earlier, that year can be included in the appropriate group.

d. Testing' whether the group means are equal in importance at a significance level 0.05.

Step 11: The formula to computer f is as shown below:

$$F = \frac{n_1 n_2 (n_1 n_2 - m - 1)}{m(n_1 + n_2)(n_1 + n_2 - 2)} \quad D^2$$

Where m is the number of predictor variables, (in this case, it is equal to 2)

$D^2 = (n_1 + n_2 - 2)\,[a(\bar{X}_{1(G2)} - \bar{X}_{1(G1)} + b(\bar{X}_{2(G2)} - \bar{X}_{2(G2)}))$

$= (7 + 5 - 2)\,(0.17229 \times 6 - 0.04973 \times 1)$

$= 9.8401$

And

$F = \frac{7x5x\,(7x5-2-1)}{2x\,(7+5)\,x\,7+5-2} \times 9.8401 = 12.915$

Step 12: The degrees of freedom for the F ratio is given by the formula:
[m1 (n1+n2-m-1)], where m is the number of factors (2). The table F value for (2,9) degrees of freedom and at the given significance level of 0.05 is 4.26.

Step 13: The calculated F value (12.915) is more than the table F value (4.26). Hence, reject the null hypothesis, Ho and accept H1. Thus we get the alternative hypothesis,
H_1: The group mean are not equal in importance.
Based on H_1 and the discriminant function, it is clear that variable X_1 (annual spending on infrastructures) is more important than the other variable X_2 (annual spending in interface events) of the faculty.

Directions for Multiple Discriminant Analysis

In a discriminant analysis, let the number of groups be N and the number of predictor variables be M (which should be at least 2). If the number of groups in a discriminant analysis is more than 2, then the number of non-redundant discriminant functions to be proposed can be the minimum of (N – 1) and M, where N is the number of groups for classification of entity members and M is the number of predictor variables. If we plan to propose two non-redundant discriminant function, than sample combination of the discriminant function is

$Y_1 = aX_1\ bX_2 \quad Y_2 = eX_1 + dX_1$.

The discriminant module of SPSS package or Discriminant Module of SAS package can be used to do multiple discriminant analysis. To use the SPSS the following steps are follows:

- Enter data for each variable in a column. The dependent variable must be categorical.
- Select; analysis-classify-discriminant.
- Select the dependent variable and transfer it to grouping variable box clock define range and enter the lowest and highest codes for the groups.
- Click continue ---------- (etc.) see (UK Wujie & Orluwene 2012)

There are three related problems in Discriminatory Analysis

- Determining whether differences in score profiles for two or more group are statistically significant.
- Maximizing the discrimination among groups by combing the variables in some manner.
- Establishing rules for the placement of new individuals into one of the groups.

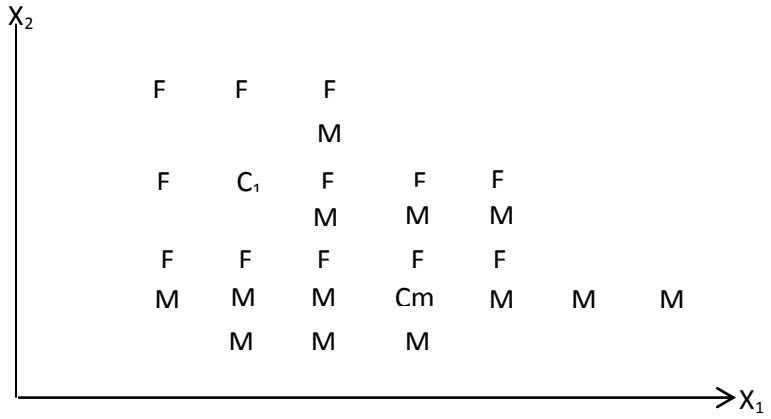
The first of these is the least important for most research in psychology; however, appropriate statistical test are available. Hoteling's T test can be uses to test the statistical significance of the difference between the average profiles of two groups. It could be used, for example in testing the significance of difference in profiles of physiological variables for males and females. If the null hypothesis is rejected in the case, it is inferred that the total profiles for the sexes are different.

It was said that statistical test like those described above are not highly important in most research problems. First, the results of such test frequently are difficult to interpret. It is impossible for two groups to have non-significant different on each of the variables but for the overall difference between profiles to be significant. Such tests combines all the information from the different variables in one overall test of significance on some of the variables, preferably on a majority of them, it is difficult to interpret the significance of difference provided rather meager information about significance differences. Secondly, merely finding that the average profile for two or more groups are significantly different does not solve the major problems.

Geometric interpretation of discriminatory analysis: The geometric interpretation given for profile analysis will help one understand some of the issues in discriminatory analysis. If there are N persons and K variables the profile for any person can be represented by a point a K-dimensional space. Each axis of the space consists of one of the variable and the variables are depicted as orthogonal to one another. In discriminatory analysis, it is useful to think of the region of that space occupied by a particular group. If discriminatory analysis is to provide

useful information, it is necessary for the members of different groups to occupy somewhat different parts of the space variables. To the extent that individuals in each group are tightly clustered in a particular region of the K- space and to the extent that there is little overlap between the regions occupied by different groups, discriminatory analysis can provide useful information. A simplied example of a space of two groups on two variables is shown in fig. 1:1. The groups are male's analysis and X_1 and X_2 are raw scores to stress. The profile point for each male is represented by M and the female is represented by F. it can be seen that two groups tend to occupy different regions of the space, males tend to be highly on X_1 and low X_2 and vice versa for females.

However, there is a moderate amount of overlap.



Scores of male and female of physiological reaction to stress

Sum of Squares and Cross Products (SSCP)

After presenting the formal approach to Discriminant analysis, it is necessary to discuss the concept of sum of squares and cross product (SSPC) matrices. Recall that in the Univariate analysis of variance the total sum of squares of the dependent variable is partitioned into two components.

1. Pooled within – Groups sum of squares
2. Between – groups sum of squares
   With multiple dependent variables, it is of course, possible to calculate the within and between sum of squares for each of them. In addition, the total sum of cross products between any two variables can be partitioned into:
1. Pooled within groups sum of products and
2. Between – groups sum of products. With multiple depend variables, it is convenient to assemble the sums of squares and cross products in the following three matrices:
   W = Polled within – groups SSPC;
   B = Between – groups SSCP
   T = Total SSCP

To clarify these notions, assume that there are only two dependent variables. Accordingly, the elements of the above matrices are:

$$W = \begin{pmatrix} SS_{10} & sep_{10} \\ SCP_{10} & SS_{102} \end{pmatrix}$$

Where SS_{10} = polled sum of square within groups for variable 1: SS_{102} = polled sum of squares within groups for variable 2.
Sep_{10} = polled within – group sum of products of variable 1 and 2.

$$B = \begin{pmatrix} SS_{b1} & SCP_b \\ SCP_b & SS_{b2} \end{pmatrix}$$

Where SS_{b1} and SS_{b2} the between –groups sums of squares for variable 1 and 2 respectively: Scp_b is the between – groups. Sum of cross product of variables 1 and 2

$$B = \begin{pmatrix} SS_1 & SCP_{12} \\ SCP_{12} & SS_2 \end{pmatrix}$$

Where SS, and SS2 are the toal sums of squares for variables 1 and 2 respectivelty SCP_{12} is the total sum of cross products of variables 1 and 2. Note that the elements of T are calculated as if all the subejcts belong to a subjects belong to a single group.

Example (1)

Use the data in the table and calculate the elments of the polled within group SSCP matrix (W); the between group SSCP matrix (B); and elements of the total SSCP matrix (T).

Illustration data on two dependnet variables for two groups.

	A_1		A_1	
	X_1	X_2	X_1	X_2
	8	3	4	2
	7	4	3	1
	5	5	3	2
	3	4	2	2
	3	2	2	5
$\sum X$:	26	18	14	12
$\sum X^2$:	156	70	42	38
$\overline{X}$:	5.2	3.6	2.8	2.4
Cp	95		31	

Where:
$\sum X_{t1} = 40$
$\sum X^2_{t1} = 198$
$\sum X_{t2} = 30$
$\sum X_{t2} = 128$
$CP_t = 126$

Solution;

a. Elements of polled within group SSCP (Matrix) W

$$SS_{10} = \left(156 - \frac{(26)^2}{5}\right) + \left(42 - \frac{(14)^2}{5}\right) = 23.6$$

$$SS_{10} = \left(70 - \frac{(18)^2}{5}\right) + \left(38 - \frac{(12)^2}{5}\right) = 14.4$$

$$SCP_{10} = \left(95 - \frac{26 \times 18}{5}\right) + \left(35 - \frac{14 \times 12}{5}\right) = -1.2$$

$$W = \begin{pmatrix} 23.6 & -1.2 \\ -1.2 & 14.4 \end{pmatrix}$$

a. Elements of the between – groups SSCP. Matric (B).

$$SS_{b1} = \left(\frac{(26)^2}{5} + \frac{(14)^2}{5}\right) - \frac{(40)^2}{10} = 14.4$$

$$SS_{b2} = \left(\frac{(18)^2}{5} + \frac{(12)^2}{5}\right) - \frac{(30)^2}{10} = 3.6$$

$$SCP_b = \left(\frac{26 \times 18}{5} + \frac{14 \times 12}{5}\right) - \frac{40 \times 30}{10} = 7.2$$

$$B = \begin{pmatrix} 14.4 & 7.2 \\ 7.2 & 3.6 \end{pmatrix}$$

c.

T = W + B, the elements of the SSCP matrix (T) can be obtained be adding W and B.

$$T = \begin{pmatrix} 23.6 & -1.2 \\ -1.2 & 14.4 \end{pmatrix} + \begin{pmatrix} 14.4 & 7.2 \\ 7.2 & 36 \end{pmatrix} = \begin{pmatrix} 38.0 & 6.0 \\ 6.0 & 18.0 \end{pmatrix}$$

W B

However, the elements of T can be calculated directly

$SS_t = 198 - \frac{40^2}{10} = 38.0$

$SS_{t2} = 108 - \frac{30^2}{10} = 18.0$

$$SCP = (95 + 31) - \frac{(40) \times (30)}{10} = 6.0$$

Elements of Discriminant Analysis

Although the presentation of discriminant analysis for two groups may be simplied, as the example, it was felt that it will be more instructive to present the general case – that is for two or more, groups. Therefore, in the presentation that follows, the same equations are applicable to Discriminant analyses will any number of groups. DA is generally calculated by the use of a computer. The basic idea of DA is to find a set of weight, V by which to weight the scores of each individual so that the ratio of B to W is maximized. Thereby leading to maximum discrimination among the groups. This can be expressed as follows:

$$\lambda = \frac{V^1_{BV}}{V_{1\,Wv}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Where V1 and V are a row and column vectors of weights, respectively: λ is referred to as the discriminant criterion.

A solution of λ is obtained by solving the following determinant equation:

$$|W^{-1} B - \lambda I| = 0 \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

Where W^{-1}, is the inverse of W, and I is an identity matrix. λ Is referred to as the largest eigenvalue, or characteristics root of the matrix whose determinant set equal to zero-that is equation (2) with two group, only one eigenvalue may be obtained.

$$\left| \underbrace{\begin{pmatrix} 23.6 & -1.2 \\ -1.2 & 14.4 \end{pmatrix}}_{W} + \underbrace{\begin{pmatrix} 14.4 & 7.2 \\ 7.2 & 3.2 \end{pmatrix}}_{B} - \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right| = 0$$

The determinant of W is

$$\begin{vmatrix} 23.6 & -1.2 \\ -1.2 & 144 \end{vmatrix} = (23.6)\,(14.4) - (-1.2)\,(1.2) = 338.4$$

$$\begin{pmatrix} 23 & -1.2 \\ -1.2 & 14.4 \end{pmatrix} = \begin{pmatrix} \frac{14.4}{338.4} & \frac{7.2}{338.4} \\ \frac{7.2}{338.4} & \frac{3.6}{338.4} \end{pmatrix} = \begin{pmatrix} .04255 & .00355 \\ .00355 & 06979 \end{pmatrix}$$

Multiplying W^{-1} by B

$$\begin{pmatrix} .04255 & .00355 \\ .0035 & 06974 \end{pmatrix} \begin{pmatrix} 14.4 & -1.2 \\ 7.2 & 36 \end{pmatrix} = \begin{pmatrix} .63828 & .31914 \\ .55325 & .27662 \end{pmatrix}$$

$$\begin{vmatrix} .63828 - \lambda & .31914 \\ .553235 & .2762 - \lambda \end{vmatrix} = 0$$

$= (.63828 - \lambda)(.27662 - \lambda) - (.31914)(.55325) = 0$

$= .17656 - 63828\lambda - 27662\lambda^2 + - 17656 + 0$

$\lambda^2 - .91490\lambda = 0$

Using $\lambda = b + \sqrt{\frac{b2 - 4ac}{2a}}$

a= 1, b = -91490 c = 0

$\lambda = .91490 + \sqrt{\frac{(-91490)2 - 4x(1)x(0)}{2 x 1}}$

$\lambda = 91490$

Having calculated λ the weights, v, are calculated by solving the following:

$(W^{-1} B - \lambda 1)\, v = 0$ ..... (3)

V = eigen vector

$$\begin{pmatrix} .63828 - .91490 & .31914 \\ :55325 & .27662 - .91490 \end{pmatrix} = \begin{pmatrix} v_2 \\ v_1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\begin{pmatrix} .27662 & .31914 \\ .55325 & .63828 \end{pmatrix} = \begin{pmatrix} v_2 \\ v_1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Using adjoint of a 2x2 matric, we have

$$\begin{pmatrix} -63828 & - & 31914 \\ -55325 & - & 27662 \end{pmatrix}$$

$= \frac{-63828}{-55325} = \frac{-31914}{-27662} = 1.5$

Assumptions of Discriminant Function analysis

There are some assumptions in the use of discriminant function analysis and these are listed (UKWUJIE & ORLUWERE 2012).

Sample Size:

Unequal sample sizes are acceptable. The sample size of the smallest group needs to exceed the numbers of predictor variables. As a rule of thumb, the smallest size should be at least 20 for a few (4 or 5) predictors. The maximum number of independent variables in n- 2, where n is the same size. While this low sample size may work, it is not encouraged, and generally it is best to have 4 or 5 times as many observations and independent variables.

Normal Distribution

It is assumed that the data (for the variables) represent a sample from a multivariate normal distribution you can examine whether or not variables are normally distributed with histograms of frequency distributions however, note that violations of the normality assumption are not "fatal" and the resultant significance test are still reliable as long as non-normality is caused by skewness and not outliers. (Ukwujie and Orluwene 2012).

Homogeneity of Variance/Covariance's:

Discriminant function analysis is very sensitive to heterogeneity of variance – co-variance matrices. Before accepting final conclusions for an important study, it is a good idea to review the within group variance and correlation matrices. Homoscedasticity is evaluated through scatter plots and corrected by transformation of variances.

Outliers:

Discriminant functional analysis is highly sensitive to the inclusion of outliers. Run a test for univariate and multivariate outliers for each group, and transform or eliminate them. If one group in the study contains extreme outliers that impact the mean, they will also increase variability. Over all significance tests are based on polled variance, that is, the average variance across all groups. Thus, the significance tests of the relatively larger means (with the large variances) would be based on the relatively smaller polled variances, resulting erroneously in statistical significance.

Neo- Multicollinearity:

If one of the independent variables is very highly correlated which another or one is a function i.e., the sum of other independent, then the tolerance value for that variable will approach o and the matric will not have a unique discriminant solution. There must be also a multicollinearity of the independents. To the extent that independent are correlated, the standardized discriminant function coefficients will not reliably assess the relative importance of the predictors varies.

Logistic regression may offer an alternative to discriminant function analysis as is usually involves fewer violations of assumptions.

A linear discriminant equation with raw scores is as shown in the equation.

$D = a+b_1X_1 + b_2 X_2 + b_3X_3……..+b_1X_1$

Where

D = Discriminant function,
b = discriminant coefficient or weight for that variable
X = respondent's score for that variable
a = a constant
I = the number of predictor variables

This is similar to regression equation. The b's maximize the distance between the mean of the criterion (dependent) variable. A linear discriminant equation with standard score is as shown below.

$D = B_1Z_1 + B_2Z_2 + B_3Z_3 \ldots\ldots\ldots + B_iZ_i$

Where

D = discriminant function
B = beta weight for that variable
Z = respondents standard score for that variable
i = the number of predictor variables

Standardized discriminant coefficients can also be like beta weight in regression. Good predictors have large weights while poor predictors have small weights. These coefficients reflect the contribution of one variable in the context of all other variables in the model. The weights are chosen so that when your computer a discriminant score (1) for each subject and then do analysis of variance on D, there will be a significant difference between/among the group means or the ratio between groups sum of square to the within group sum of square is as large as possible. The value of this ratio is called Eigen value.

For instance, if we want to predict student's success/failure in their degree programme using variables like age, sex, and self-concept, attitude towards school, previous CGPA and anxiety. In this situation it is simple discriminant analysis function with only categories success/failure.

Uses of Discriminant Function Analysis as stated by Ukwujie & Orluwene (2012)

1. To investigate differences between groups on the basis of the attribute of the cases, indicating which attribute contribute most to group separation. The descriptive technique successively identifies the lower combination of attributes known as canonical discriminant function which contribute maximally to group regression.
2. Predictive discriminant function analysis addresses the equation of how to assign new cases to groups. The discriminant analysis function uses a person's scores on the predictor variables to predict the category to which the individual belongs.
3. To determine the most parsimonious way to distinguish between groups.
4. To classify cases into group. Statistical significance tests using chi square enable you to see how well the function separate the groups.
5. To test whether cases are classified as predicted.

References

Bunday, B.D (1983). Pure Mathematics for Advanced Level (2nd ed), great Britain: Robert Hartnoll LTD.

Jack, R.F & Wallen, E (2000). How to Design and Evaluate Research in Education. U.S.A: MC Grawhill Companies Inc.

Kpolovie P.J (2010). Advanced Research Method. Imo State: Spring field publishers LTD.

Mac'Odod, D.S (1997). Quantitative and Statistical Analysis for business Decisions, Port Harcourt: Linnet Paul Publications.

Mertens, D.M (2005). Research and Evaluation in Education and Psychology: Integrating Diversity with Quantitative, Qualitative and mixed Methods 2nd Edition. California: Saga Publications, Inc.

Nunnally, J.C (1981). Psychometric Theory (2nd Ed). New Delhi: Tata MC Grawhill Publishing Company LTD.

Pannerselvan R (2010). Research Methodology, New Delhi: Asoke K Ghosh, Phi Learning Private LTD.

Sabine landau and Brian S, Everit (2004). A Handbook of Statistical Analysis using SPSS. London: Chapman and Hall Press LLC.

Ukwujie R.P.I and Orluwene G.W (2012). Peanuts Educational Statistics (4th ed). Port Harcourt: Chadik Printing Press.