LUPINE PUBLISHERS

Archives of Diabetes & Obesity

DOI: 10.32474/ADO.2019.02.000132

ISSN: 2638-5910

Review Article

A Note on the Application of Discriminant Analysis in Medical Research

KC Bhuyan*

Department of Statistics, Dhaka, Bangladesh

***Corresponding author:** KC Bhuyan, Department of Statistics, Dhaka, Bangladesh

Received: May 27, 2019 **Published:** June 06, 2019

Introduction

The social aspects of medical research and research in life science deal with data collected from randomly selected investigating units either by direct interview or by doing some controlled experiments. The objective of investigation is to verify the existence of some pre-assumed hypothesis and belief. Whatever be the objective of the research, the study is done using the collected data, where the data are either categorical or numerical or both. There are many advanced statistical methods [1] to handle the data, especially if the data are numerical. Limited number of methods of analysis is there to handle the categorical data. But proper statistical methods are to be applied to study the unknown characteristics in the population. Once the technique of analysis is identified, the analysis can be done using any one of the Statistical Packages available for the analysis. The major source of data in the field of life science and medical science is related to public health and the main aspect of analysis is to suggest ways and means to improve the health status of the public by controlling the diseases which are fatal as well as causes of health hazard in the society. Any type of health hazard is the cause of economic and social problems for the country as well as for the family. Thus, the health planners need the proper analytical findings to formulate the guidelines so that the health hazard is reduced to a great extent. The data are collected from each and every investigating unit.

Some of the data are age, height, weight, level of education, occupation, income, marital status, type of work, food habit, smoking habit, prevalence of obesity, diabetes, hypertension, etc. The prevalence of obesity is the cause of non-communicable diseases like diabetes and hypertension and the sources of these two are occupation, work type, food habit, smoking habit. There may be other sources of diseases. But the sources mentioned here are categorical (qualitative) variable. The quantitative variable income, age, weight, etc. are also the sources of non-communicable diseases [2]. The data mentioned above are the multivariate data. The analytical procedures of these data are different, and the procedures are broadly classified as

a) Dependence Analysis

b) Interdependence Analysis.

One of the techniques of dependence analysis is the Discriminant Analysis. This analysis is used to discriminate the investigated units according to some categorical variable and to identify the most responsible variable(s) for the discrimination. Accordingly, the health planners can suggest the ways and means so that proper action can be taken to control the sources responsible for the health hazard in the society.

Discriminant Analysis

Let x_{ij} be the i-th variable observed from j-th category of units [i = 1, 2, ……..p ; j= 1, 2, …., k], where

$$\bar{Y}_j = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ .. \\ \bar{x}_p \end{bmatrix}$$

S_j = Variance- Covariance matrix of (x_{ij}) , S_u = Combined variance- covariance matrix, where

$S_u = \sum n_j S_j / (n - k)$. Here n_j = number of observation in j- th category, $n = \sum n_j$.

The k category of units can be discriminated by

a) ML Method of Discrimination

b) Bayes Discriminant Rule and

c) Fisher's Discriminant Rule.

The ML method of discrimination needs to calculate the following statistic

$$D_{ji} = (\bar{Y}_j - \bar{Y}_l)' \, ? \, S_u^{(-1)} x - \frac{1}{2} \bar{Y}_j' \, S_u^{(-1)} \, \bar{Y}_j + \frac{1}{2} \bar{Y}_l \, S_u^{(-1)} \, \bar{Y}_j$$

where $j \neq k$. If we consider that k = 3, then there exists linear relationship of the type

$D_{12} + D_{23} = D_{13.}$

In such a situation the value of x will be allocated to a population as per the rule discussed below:

a) If D_{12}>0 and D_{13}>0, allocate x to the population -1

b) If D_{12}<0 and D_{13}>0, allocate x to the population -2

c) If D_{12}<0 and D_{13}<0, allocate x to the population -3.

The discriminant analysis is meaningful if the k mean vectors of variables are heterogeneous. For the analysis a model is considered when there are k = 2 groups, where the model is

$D = B_0 + B_1 x_1 + B_2 x_2 + \ldots\ldots\ldots\ldots + B_p x_p$

Here B_i (i = 1, 2, ….,p) is the discriminant coefficient of the variable x_i . The interpretation of these coefficients are closely follows the logic of multiple regression. The value B_i indicates the importance of the variable x_i in discriminating between the two groups. For k=3 groups problem, one function is considered to discriminate between first group and combined second and third group. Another function is considered to discriminate between second and third group. The number of discriminant functions are (k – 1) if their k groups. In that case the interpretation of coefficients is made for each function separately. The coefficients are available using Statistical Packages. D is the discriminant score for different values of x when k=2. The correlation coefficient of discriminant score and the variable is used to decide an important variable for discrimination. The highest correlation coefficient indicates the most important variable for discrimination. Different functions may identify different important variables for discrimination.

Some Results of Discriminant Analysis

The following Discriminant Analysis was done [3] to discriminate 900 randomly selected adults of Bangladesh classified by levels of obesity. It was observed that levels of obesity were varied differently with the variation of different social factors. Thus, there were in search of identification of most important variables to discriminate the respondents according to various of levels of obesity. This was done by discriminant analysis. The analysis helps to identify the important variables for which the groups of respondents were significantly different [4]. The variables which were included in the analysis were sufficient to discriminate the different groups of respondents according to their level of obesity as Box's M = 287.926 and the corresponding F= 1.403 with p –value = 0.000. The analysis provided 3 discriminant functions for 4 groups of respondents. The first function was significant as values of Wilk's $\wedge$ for first, second and third functions were 0.918, 0.973 and 0.994, respectively and the corresponding x^2 values were 75.920(p-value=0.0000, 24.493 (p-value=0.222) and 5.065 (p-value=0.829). The standardized canonical discriminant function coefficients were presented in Table 1. From the discriminant analysis the correlation coefficients of variables and the discriminant functions scores were calculated. These coefficients were shown in Table 2. The analysis indicated that the respondents of different levels of obesity were significantly different according to socio-demographic variables. The important variable for discrimination was residence followed by age. The other important variables were gender and marital status.

Table 1: Coefficients of functions.

Variable	1	2	3
Residence	0.274	-0.310	0.301
Age	0.385	0.401	-0.409
Gender	0.690	0.151	-0.019
Marital status	-0.102	-0.138	0.514
Religion	0.106	-0.067	0.216
Education	-0.096	0.613	-0.480
Occupation	0.113	-0.333	-0.579
Type of work	-0.033	0.033	0.173
Income	0.526	0.440	0.639
Smoking habit	0.070	-0.378	0.017
Prevalence of diabetes	-0.102	-0.035	0.013

Table 2: Pooled within group correlations between discriminating variables and standardized discriminant function.

Variable	1	2	3
Residence	0.686*	-0.402	=0.122
Age	0.529*	-0.412	-0.357
Gender	0.224*	-0.030	0.191
Marital status	-0.118*	-0.050	0.104
Religion	-0.057	0.560*	-0.121
Education	0.486	0.508*	0.437
Occupation	0.396	-0.492*	-0.102
Type of work	0.249	0.345*	-0.328
Income	0.164	-0.226*	0.048
Smoking habit	-0.065	-0.164	0.408*
Prevalence of diabetes	0.072	-0.087	-179*

As a second example of discriminant analysis, 662 children and adolescents of some randomly selected affluent families were classified by their level of obesity [5]. There were 4 groups of respondents and for these 4 groups the variables age of the children, food habit of children, utilization of time by the children, father's education, mother's education, father's occupation, mother's occupation and family income were different and most of them were associated with the level of obesity. Therefore, these variables were included to discriminate the children. For 4 groups of children 3 Fisher's linear discriminant functions were available. The coefficients of these functions for different variables were shown in Table 3. First function explained 92.7% variation of the children's level of obesity and most important variable to explain this variation is father's occupation followed by mother's education and father's education. This phenomenon was observed from pooled within groups correlations between discriminating variables and standardized canonical discriminant functions. The results of this pooled within groups correlations were given in Table 4. The most important variables identified by functions were shown by given asterix. Since first functions explained 92.7% variation in level of obesity and this function was statistically significant [Wilk's Lamda=0.834, Chi-square =97.811, p-value=0.000], the pooled within groups correlations were shown for this first function. Some variables were also found important by second function to discriminate children by level of obesity. However, the 2nd and 3rd functions were not statistically significant and the pooled within groups correlations of variables and 3rd function were not shown.

Table 3: Discriminant coefficients of different variables: (. Indicates significant correlation).

Variables	Discriminant Functions		
	1	2	3
Constant	-41.241	-38.455	-39.591
Age of children	13.689	13.938	13.868
Food habit of children	2.008	2.021	2.031
Utilization of time	4.568	4.137	4.596
Father's education	1.446	.772	.534
Mother's education	-.550	.676	.959
Father's occupation	8.673	7.403	7.690
Mother's occupation	-.141	-.221	-.346
Family income	.172	-.144	-.279

Table 4: Correlation coefficients of variables with discriminant score.

Variables	Functions	
	1	2
Father's occupation	-.518*	.192
Mother's education	-.240*	.024
Father's education	-.118*	-.080
Utilization of time	.131	.766*
Family income	.450	-.489*
Age of children	-.069	-.270*
Mother's occupation	.210	-.215*
Food habit of children	0.068	-.116*

As a third example, let us present the results of discriminant analysis when discrimination was done according to the prevalence of non-communicable diseases [NCDs]. A group of adults of Bangladesh was investigated [2] to identify the responsible variable for the prevalence of NCDs. The data were recorded from randomly selected 785 adult people of Bangladesh. Among them, 49.4 percent were affected by at least one of the NCDs. The two groups of respondents were discriminated to identify the factors responsible for discrimination. The analysis indicated that the variables age, followed by marital status and weight were the most important variables in discriminating the two groups of respondents. The analytical results were presented in Table 5. As a fourth example, let us discuss the discrimination of students of public and private universities in respect of some social characters. The number of investigated students were 893 from private universities and 119 from public universities [6].

Table 5: Coefficient of discriminant function and pooled within groups correlation between variables and discriminant function score.

Variable	Discriminant coefficient	Correlation coefficient of variables and discriminant function score
Age	0.480	0.701
Marital status	-0.340	-0.570
Weight	0.392	0.307
Education	-0.279	-0.307
Habituated in process food	0.148	0.281
Height	-0.230	-0.270
Residence	-0.221	-0.240
Income	0.475	0.205
Gender	0.010	0.194
Food from restaurant	-0.010	0.191
Doing physical work	0.147	0.138
Change of food habit	0.008	0.047
Smoking habit	-0.027	0.036

As there were two groups of students, viz. students of public university and students of private university, one discriminant function was derived. The function was

$$D =-0.041+0.551x_1 -0.784x_2+0.455x_3-0.633x_4+0.080x_5+0.057x_6+0.121x_7-0.004x_8+0.631x_9$$

Table 6: Correlation coefficient between variables and discriminant score in descending order of magnitude.

Variables	Correlation Coefficients
Father's education	-0.705
Mother's education	-0.703
Residential origin	0.470
Age	0.457
Father's occupation	0.397
Mother's occupation	0.175
Income	-0.104
Awareness of health hazard	0.033
Smoking habit	0.002

This function was significant as Wilks Lambda is 0.773 [χ^2= 258.758, p=0.000, Bartlett (1947)] and it indicated that the students of private and public universities were significantly different in respect of some of the socioeconomic characteristics. The important socioeconomic characteristics were identified by the canonical correlation coefficients of the variables and the discriminant score. The correlation coefficients are shown in Table 6 in descending order of magnitude. It is seen that father's education is very important social factor to discriminate between student of private and public universities followed by mother's education, residential origin and age of students. As a fifth example, let us discuss the discrimination of 900 randomly selected adults of Bangladesh [7] in respect of the prevalence of diabetes. There were two groups of respondents, one group of 635 diabetic patients and another group of 235 normal respondents. In doing the discriminant analysis, there was an attempt to decide the inclusion of variables in the discriminant analysis. For this the value 1- r^2 was calculated and was shown in Table 7. Here r is the multiple correlation coefficient when one variable was considered as dependent variable and others as independent variable. None of these calculated values was low and hence all the nine variables were included in the analysis.

Table 7: Results showing the importance of inclusion of variable in the discriminant analysis.

Variable	Wilk's ^	F	d.f	p-value	1- r2
Residence	0.984	14.603	1, 898	0.00	0.769
Age	0.926	71.634	1, 898	0.00	0.924
Sex	1.000	0.206	1, 898	0.65	0.531
Weight	0.999	1.063	1, 898	0.30	0.802
Height	0.985	13.308	1, 898	0.00	0.718
Education	0.971	26.634	1, 898	0.00	0.619
Occupation	1.000	0.007	1, 898	0.93	0.628
Work type	0.989	9.908	1, 898	0.00	0.731
Income	0.997	2.343	1, 898	0.13	0.821

The discriminant coefficients were shown in Table 8 below. The results indicated that the variable residence had the highest discriminating power followed by work type, income and age. The importance of the variables was also observed from the study of the correlation coefficients of the variables with discriminant score. The correlation coefficients in descending order were shown below in Table 9. The function was found highly significant by Bartlett's test (p<0.001). The test indicates that diabetic and non-diabetic respondents were significantly different. The important variable for discrimination was age followed by education and residence. This result was observed from the study of correlation coefficient of the variables and discriminant score. The same of respondents were also discriminated by the type of disease. The total diabetic patients were classified in to four classes, viz. patients of type I, type II, type III diabetes and another group of 269 patients who were ignorant about their type of diabetes. In the first three groups the number of patients were 136, 215 and 19 respectively.

Table 8: Discriminant coefficients of different variables.

Variable	Constant	Age	Edu-Cation	Residence	Height	Work type	Income	Weight	Sex	Occu-pation
Coefficient	-1.012	0.206	-1.009	0.616	0.018	0.559	0.372	-0.043	-0.042	0.041

Table 9: Correlation coefficients with discriminant score.

Variable	Age	Education	Residence	Height	Work type	Income	Weight	Sex	Occupation
Correlation coefficient	-0.788	0.481	0.356	0.340	-0.293	0.143	0.093	0.042	-0.008

Thus, the patients were classified into 4 groups and identified the groups by 1, 2, 3 and 4 respectively. The multivariate analysis of variance showed that the mean vectors of four groups of patients by type were significantly different (Wilk's ^ = 0.891, F= 2.715, p$\leq$ 0.01The discriminant analysis also showed that the 3 discriminant functions were significantly different (p$\leq$0.01). The results were shown in Table 10. The pooled within- groups correlations between discriminating variables and the standardized canonical discriminant functions were shown in Table 11. The first function discriminated well among groups of patients and the variables age and education were important to discriminate among patients of different types of diabetes. The second function discriminated well and the important variables for discrimination were occupation and work type. The third function discriminated well among different groups of patients of different types and the variables income, residence and sex were very important to discriminate well.

Table 10: Discriminant coefficients of different variables.

Variables	Coefficients of functions		
	1	2	3
Constant	-4.947	-1.252	-1.283
Age	0.797	0.764	-0.201
Education	-0.786	0.131	0.077
Residence	0.458	1.162	0.769
Height	0.529	-1.042	-0.447
Work type	-0.253	0.040	0.086
Income	0.194	-0.229	0.442
Weight	0.294	0.157	-0.015
Sex	0.786	-0.698	0.993
Occupation	0.038	0.106	-0.336

Table 11: Correlation coefficients with discriminant score.

Variables	Functions		
	1	2	3
Age	0.533*	0.508	-0.227
Education	-0.442*	0.119	0.225
Weight	0.295*	0.120	-0.006
Height	0.195	-0.503*	-0.377
Occupation	0.148	0.161*	0.012
Work type	0.026	0.088*	-0.063
Income	0.242	-0.126	0.728*
Residence	-0.016	0.365	0.408*
Sex	0.184	-0.109	0.377*
Occupation	0.038	0.106	-0.336

References

1. Bhuyan KC (2018) A note on the application of advanced statistical methods in medical Research. Biomed J Sci & Tech Res 11(3): 1-4.
2. Bhuyan KC, Mortuza A Md, Fardus J (2018) Discriminating patients suffering from non-communicable diseases: A case study among Bangladeshi adults. Biomed J Sci & Tech Res 10(1): 7571-7577.
3. Bhuyan KC, Fardus J (2019) Discriminating Bangladeshi adults by level of obesity. LOJ Med Sci 3(1): 184-190.
4. Bhuyan KC (2004) Multivariate Analysis and its Application, New Central Book Agency (P) Ltd, India.
5. Urmi AF, Bhuyan KC (2019) Discriminating Bangladeshi children and adolescents of affluent families by level of obesity. AJSE.
6. Bhuyan KC, Urmi AF, Fardus J (2017) Discriminating students of public and private universities in respect of some social characters. Jour Stats Studies 34: 13-23.
7. Fardus J, Bhuyan KC (2016) Discriminating diabetic patients of some urban and rural areas of Bangladesh: A Discriminant Analysis Approach. Euro Biomed Jour for Young Doctors 11(19): 134-140.

DOI: 10.32474/ADO.2019.02.000132