ISSN: 2574-1241

DOI: 10.26717/BJSTR.2018.11.002090

KC Bhuyan. Biomed J Sci & Tech Res

BIOMEDICAL
Journal of Scientific & Technical Research

Mini Review Open Access

A Note on the Application of Advanced Statistical Methods in Medical Research

KC Bhuyan*

Professor of Statistics, Bangladesh

Received: : November 10, 2018; **Published:** : November 26, 2018

***Corresponding author:** KC Bhuyan, Professor of Statistics, Savar, Dhaka, Bangladesh

Introduction

The data on many aspect of life science are both categorical and numerically measured values. Some of these data are observed from controlled and/or random experiments. Whatever be the source of data, the analysis of these helps the researchers to infer about the characteristics of the population from which the data are collected. The researchers in the field of medical science often face the problem in doing analysis of the collected data due to lack of knowledge of proper statistical techniques suitable for a particular data. For proper conclusion about the parameter of the population, it needs the application of the proper analytical techniques. Once a technique is identified, the analysis can be performed using any of the Statistical Packages. The main aspect of the data in bioscience is related to public health and main aspect of analysis is to suggest ways and means to control the disease and to suggest the methods so that premature death can be reduced, specially the child and infant death. For some countries, the failure of birth control is also a health hazard. Thus, the health planners need the proper analytical findings in the field of medical science and in the field of other aspects of bioscience.

The empirical analysis in the field of health science needs data related to health hazard collected from several units suffering from any health problem or expect to suffer from some communicable or non-communicable diseases. Any unit under study may provide different types of information (information of values of variable). As the values of the variables are collected from each member of the investigated units, the variables are expected to be correlated. We usually recognize these collected data as Multivariate Data. There are different methods to handle these Multivariate Data. In this note, application of some of the multivariate techniques are discussed. The multivariate analysis has two main aspects, viz.

a) Dependence analysis, and

b) Interdependence analysis

The multivariate regression analysis including logistic regression analysis, discriminant analysis, and multivariate analysis of variance are the topics of dependence analysis. The interdependence analysis, also known as data reduction technique, deals with principle component analysis, factor analysis, cluster analysis and canonical correlation analysis. The data, multivariate or uni-variate, are collected according to some pre-determined objectives. The analytical plan is also pre-determined so that proper conclusion can be drawn according to the objectives. The whole activities along with statistical interpretation are presented concisely. This presentation of analytical results is known as reporting writing. However, the presentation of analytical results along with different activities of the research work varies from work to work and it also varies with the variation of objective of the research. Let us now discuss some of the application of the multivariate analysis using a part of the real data collected by Urmi and Bhuyan [1]. They have already done some analysis and presented in some of the research papers published in home and abroad [2].

Application of the Multivariate Analysis

Multiple Regression Analysis

For regression analysis the general consideration is that when n sample units are investigated to collect information on several variables, it may happen that some of the variables are interrelated. For example, prevalence of diabetes [yes= 1, no = 0] and level of BMI (kg /m^2) are interrelated along with other variables, viz. age, income, residence, level of education, marital status, occupation, gender, smoking habit, physical labor, food habit, habituated in processed food, restaurant food, etc. These factors mentioned here are interrelated and some the factors depend on income. Again, income depends on level of education and profession. If it is expected that blood sugar level (y mmol /l) of any person, children or adolescent or adult depends on age (x_1 years) , height (x_2 meter), weight (x_3 kg) food habit (x_4, taking more protein = 3, more rice = 2 more sugar product = 1 and healthy food = 0) and family income (x_5 in thousand taka), multiple regression analysis of y on x_1 , x_2 , x_3 , x_4 and x_5 can be performed , where the multiple regression model is given by

$$y = B_0 + B_1x_1 + B_2x_2 + \ldots\ldots\ldots\ldots + B_5x_5 + e$$

In general y depends on k (= 5) explanatory variables x's, e is a random component which is inserted in the model to study the impacts of other variables which are not included in the model. The objective of the study is to estimate the parameters B_i's and to test the significance of these parameters. Under usual assumptions, the analysis can be done. Using a part of the data of Atika and Bhuyan [1] the analytical results of regression were shown in Tables 1 & 2. The analysis was done using 125 observations. It was observed that blood sugar level significantly dependent on income. Here for the conclusion made was dependent on the assumption that the explanatory variables (x's) were independent and the random component (e) was normally and independently distributed with mean zero and with common variance (Tables 1 & 2).

Table 1: Estimate of Coefficients with Significance Level.

Estimate of Coefficients with Significance Level			
Variable	Estimates of Coefficients	T	P-Value
Constant	3.211	7.076	0.000
Age	0.294	1.285	0.201
Height	-0.002	-0.475	0.636
Weight	-0.065	-0.597	0.552
Income	-0.254	-4.194	0.000
Food habit	0.013	0.206	0.837

Table 2: Anova Table.

Anova Table					
Sources of Variation	d.f	S.S	M.S = S.S/D.F	F	p-value
Regression	5	14.179	2.835	6.440	0.000
Residual	119	52.349	0.440	-	
Total	124				

Logistic Regression Analysis

Let us consider that the prevalence of diabetes (y = 1 for diabetic patient, y = 0 non-diabetic person) depends on some of the variables discussed above. Here dependent variable is a binary variable (indicator variable) instead of a continuous variable. Thus, usual multiple regression analysis is not suitable to study the effect of explanatory variables on the dependent variable. To overcome the problem, Logistic regression analysis is to be done. As an example, using the same set of data as mentioned above, the logistic regression analysis was performed, and the results were presented in Table 3. This analysis also indicated that the income and height of the respondent had significant impacts on prevalence of diabetes (Table 3). As a further example of logistic regression, let us consider the analytical results of data on smoking habit of students of American International University - Bangladesh [3,4], where smoking habit [yes=1, no=0] was considered as dependent variable and age of students, father's education, mother's education, family income, residential origin and knowledge of health hazard of students were considered as dependent variable. The analytical results were shown in Table 4. The analytical results showed that smoking habit was significantly influenced due to the variable age of students, residential origin and knowledge regarding health hazard of tobacco smoking. The example is a case of binary logistic regression, where dependent variable is classified into two classes. The dependent variable can also be classified into several classes and we can do the similar analysis.

Table 3: Results of logistic regression analysis.

Results of logistic Regression Analysis					
Variable	Estimate of coefficient	Wald statistic	p-value	d.f	Exp (B)
Constant	-1.108	0.535	0.464	1	0.330
Height	0.018	4.84	0.028	1	1.018
Income	-0.615	20.168	0.000	1	0.54
Food habit	-0.278	0.215	0.643	1	0.758

Table 4.

Independent Variables	Regression Coefficients	Wald Statistic	P-value
Constant	- 0.570	17.246	0.000
Age	-0.124	5.240	0.047
Father's education	0.000	0.000	0.999
Mother's education	-0.000	0.009	0.927
Family income	-0.010	0.000	0.999
Residential origin	0.363	4.042	0.044
Knowledge of health hazard	0.606	7.532	0.006

Canonical Correlation Analysis

In a separate study [5] it was observed that smoking habit and awareness regarding health hazard of tobacco smoking were significantly associated. Again, these two variables were related with other socioeconomic characteristics. Hence, Bhuyan and Urmi (2018) decided to observe the joint relationship of the variable's awareness of the health hazard of tobacco smoking with other socioeconomic variables. This was done by Canonical Correlation analysis, which is also a component of multivariate analysis (Table 5).

Table 5: Correlation Between X – SET and Y – Set variables.

Correlation Between X – SET and Y – Set variables		
Variable	X-Set	Y-Set
Age	-0.334	-0.282
Sex	0.847	-0.263
Marital status	-0.187	-0.791
Religion	0.002	-0.027
Father's education	0.204	-0.027
Mother's education	-0.135	-0.025
Father's occupation	-0.065	0.170
Mother's occupation	0.097	0.009
Family income	-0.199	-0.003

Factor Analysis

It is a multivariate technique to reduce the data. For example, a diabetic patient came to a doctor for treatment. Before start of treatment, doctor needs to know about the height, weight, BMI, along with other characteristics. But BMI depends on height and weight. Here BMI is a common factor. Thus, instead of observing height, weight and BMI, it is better to observe only BMI and BMI will help to provide a conclusive decision regarding decision the prevalence of diabetes. Here instead of studying many variables, some common factors can be identified for conclusion. The technique of selection of few factors for further analysis known as Factor analysis and it is a data reduction technique. The factors are selected in such a way that most (around 90%) of the variability of the data set is explained by the selected factors. One of the selection procedure is Principle Analysis and principle component analysis is another technique of interdependence analysis. As an example of the factor analysis, the data mentioned above had been used to select some the important factors to study the prevalence of diabetes. The factor analysis provided two important factors to study the variability in the data set of the prevalence of diabetes. The two factors explained 65% inherent variation in the data set. The first component indicated that the prevalence of diabetes was mainly for body weight followed by age and height. The analytical results were presented in the following Table 6.

Table 6: Selected factors for the study of prevalence of diabetes.

Selected factors for the study of prevalence of diabetes		
Variables	**Factor ,1**	**Factor, 2**
Blood sugar level	-0.126	0.861
Age	0.864	0.181
Height	0.834	0.064
Weight	0.886	0.284
Food habit	0.447	-0.613
Medicine used	-0.469	0.169

Discriminant Analysis

It is also a multivariate technique in which a set of data can be classified into several classes according to some indicator variable and mathematical method is applied to discriminate the sample units so that some important variables are identified for the discrimination of the group of observations. For example, let us consider that the sample units are classified as diabetic and non-diabetic. It was observed in some studies that [5,6] diabetic and non-diabetic people were significantly different due to socioeconomic variables and some variables were very important to discriminate the two groups. Bhuyan et al [6] have done such an analysis to discriminate the students of public and private universities in respect of some social characters. There are different mathematical steps to estimate the discriminant scores for the students. Later on the correlation coefficients of each variable and the discriminant scores are calculated to identify the important factors for two groups of students are discriminated. The correlation coefficients between variables and discriminate scores are shown in Table 7.

Table 7.

Variable	Correlation Coefficient	Variable	Correlation Coefficient
Father's education	-0.795	Mother's occupation	0.175
Mother's education	-0.703	Income	-0.104
Age	0.470	Awareness of health hazard	0.033
Father's occupation	0.397	Residence	0.47

The analysis provides information that public and private university students were significantly different in respect of their social background. Education of parent was very much influencing in discriminating the students of public and private universities. The second important factor is the residential origin followed by age of students. More urban students and students of higher ages are admitted in private universities. Smoking habit was not significantly different between two groups of students (r = 0.002) (Table 7). As a further example of discriminant analysis, the analysis presented by Fardus and Bhuyan [7] may be mentioned. In that paper, the diabetic patients of some urban and rural areas in Bangladesh were discriminated by the types of diabetes. Including one unknown type, the patients were classified into 4 types of diabetes and 3 significantly different discriminant functions were derived. The major cause of discrimination of the patients were studied by the correlation coefficients of the variables and the discriminant scores. The significant correlation coefficients were presented in the following Table 8. The first function discriminated well among the groups of patients and the variables age and education followed by residence were important to discriminate among patients of different types of diabetes. The second function discriminated well among the patients of different groups and the important variables age, income followed by education were identified for discrimination.

Table 8: Correlation Coefficients of Variables with Discriminant Scores.

Correlation Coefficients of Variables with Discriminant Scores			
Variables	**Functions**		
	1	**2**	**3**
Age	0.654*	-0.463	-7.629
Education	-0.425	0.142	0.619*
Residence	0.104	0.170	0.802*
Work type	0.040	-0.100*	0.052
Gender	0.215	0.380*	-0.014
Occupation	0.195*	-0.074	0.133
Income	0.317	-0.695*	0.167

Note: The largest absolute correlation coefficients.

The third function discriminated well among the patients of different types of diabetes and the variables age and residence were identified very important for discrimination. Further statistical analysis in Medical Science are investigation of association of two characteristics and hence to study more prevalence of a particular characteristic. As an example, let us consider the data of used by Urmi and Bhuyan [1], where the association of level of obesity and prevalence of diabetes were studied. The results are shown below. It was observed that the prevalence of diabetes and level of obesity were significantly associated (p-value =0.033). Form the study of odd ratio it was observed that the overweight and obese group had 69% more chance to be affected by diabetes than the non-obese group. The risk ratio for this group is 1.47 (Table 9). In the above analysis both the variables used are qualitative in nature. These variables do not follow normal distribution. Most of the test statistics are based on the assumption of normalty of the data. But the test is valid as it is a non-parametric test. Other non-parametric test are also used in the analysis of data of medical science. The study of health hazard and survival analysis are other two aspects of analysis of data related to medical science. In this note a short review of multivariate analysis was presented. For further analysis one can go through the books on applied multivariate analysis [8].

Table 9: Distribution of Persons According to Level of Obesity and Prevalence of Diabetes.

Distribution of Persons According to Level of Obesity and Prevalence of Diabetes					
Prevalence of diabetes	Underweight n %	Normal n %	Overweight n %	Obese n %	Total n %
Yes	109 23.4	23 16.8	16 38.1	3 17.6	151 22.8
No	357 76.6	114 83.2	26 61.9	14 82.4	511 77.2
Total	466 70.4	137 20.7	42 6.3	17 2.6	662 100.0

References

1. Urmi AF, Bhuyan KC (2018) Obesity in children and adolescents and the factors responsible for it : A case study among children of some affluent families. Int Diab Card Dis 3(1): 56-66.
2. Bhuyan KC, Fardus J, Urmi AF (2017) Discriminating students of public and privyate universities in respect of some social characters. Jour Stats Studies 34: 13-23.
3. Khatun M, Bhuyan KC (2014) Awareness of health hazard of tobacco consumption among students of American International University-Bangladesh. AJSE 13(1): 85-91.
4. Bhuyan KC, Mortuza A, Fardus J (2017) Socioeconomic factors associated with overweight and obesity: A case study among adult people of Bangladesh. AJSE 16(2): 119-124.
5. Bhuyan KC, Urmi AF (2018) Canonical correlation analysis to study the impacts of different social factors on awareness of health hazard of tobacco consumption and smoking habit. BJSTR 10(5).
6. Mortuza A, Bhuyan KC, Fardus J (2018) A study on identification of socioeconomic variables associated with non-communicable diseases among Bangladeshi adults. AASCIT 4(3): 24-29.
7. Fardus J, Bhuyan KC (2016) Discriminating diabetic patients of some rural and urban areas of Bangladesh : A discriminating analysis approach. EMBJ 11(19): 134-140.
8. Bhuyan KC (2004) Multivariate Analysis and its Applications, New Central Book Agency[P] Ltd India.

ISSN: 2574-1241

DOI: 10.26717/BJSTR.2018.11.002090

KC Bhuyan. Biomed J Sci & Tech Res