ISBN: 978-93-90847-07-5

11. An Overview of Multivariate Analysis

Chandra Prakash Saini

Assistant Professor of Chemistry
G. H. S. Govt. P. G. College,
H. Sujangarh, Churu.

Abstract:

Multivariate means involving multiple dependent variables resulting in one outcome. This explains that the majority of the problems in the real world are Multivariate. For example, we cannot predict the weather of any year based on the season. There are multiple factors like pollution, humidity, precipitation, etc. Here, we will introduce you to multivariate analysis, its history, and its application in different fields. Multivariate analysis is used widely in many industries, like healthcare. In the recent event of COVID-19, a team of data scientists predicted that Delhi would have more than 5lakh COVID-19 patients by the end of July 2020. This analysis was based on multiple variables like government decision, public behaviour, population, occupation, public transport, healthcare services, and overall immunity of the community.

Keywords: Multivariate, Analysis, Multivariate Analysis,

11.1 Introduction:

11.2 The History of Multivariate Analysis:

In 1928, Wishart presented his paper. The Precise distribution of the sample covariance matrix of the multivariate normal population, which is the initiation of MVA. In the 1930s, R.A. Fischer, Hotelling, S.N. Roy, and B.L. Xu et al. made a lot of fundamental theoretical work on multivariate analysis. At that time, it was widely used in the fields of psychology, education, and biology. In the middle of the 1950s, with the appearance and expansion of computers, multivariate analysis began to play a big role in geological, meteorological.

Medical and social and science. From then on, new theories and new methods were proposed and tested constantly by practice and at the same time, more application fields were exploited. With the aids of modern computers, we can apply the methodology of multivariate analysis to do rather complex statistical analyses.

11.2.1 Definition:

Multivariate analysis in a broad sense is the set of statistical methods aimed simultaneously analyse datasets. That is, for each individual or object being studied, analysed several variables. The essence of multivariate thinking is to expose the inherent structure and meaning revealed within these sets if variables through application and interpretation of various statistical methods. Suppose a project has been assigned to you to predict the sales of the company. You cannot simply say that 'X' is the factor which will affect the sales.

There are two determining factors that have to take into account when doing a multivariate approach [1]: (I) the multidimensional nature of the data matrix and (II) the purpose of trying it, preserving its complex structure. This is based on the belief that the variables are interrelated, so that only the set of the same test may provide a better understanding of the studied object obtaining information univariate and bivariate statistical methods are unable to achieve. The joint treatment of the variables will faithfully reflect the reality of the problem addressed [2].

11.2.2 Types of Multivariate Methods: Multivariate methods can be classified based on the types of variables [3] (Figure 11.1):

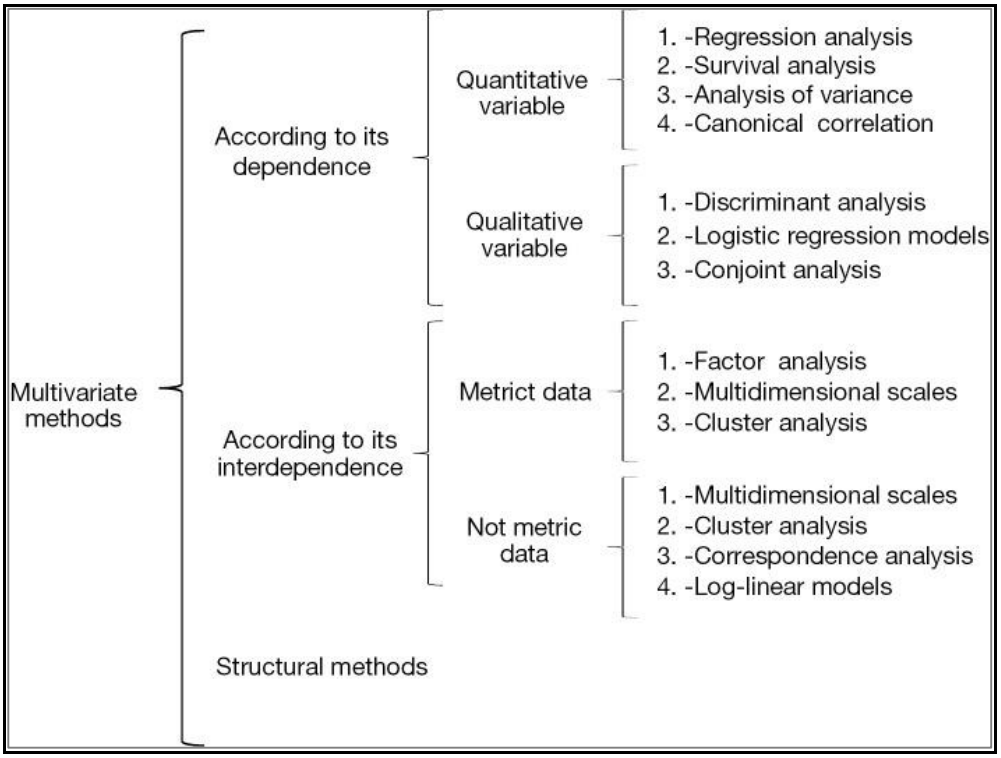


Figure 11.1: Classification of Multivariate Methods.

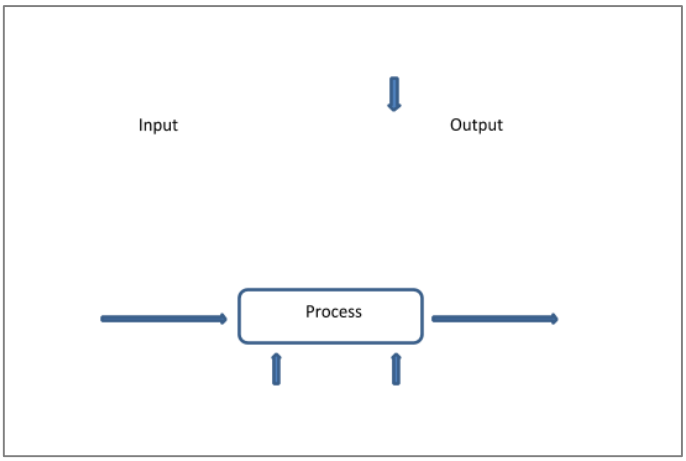
We know that there are multiple aspects or variables which will impact sales. To analyse the variables that will impact sales majorly, can only be found with multivariate analysis. And in most cases, it will not be just one variable.

Like we know, sales will depend on the category of product, production capacity, geographical location, marketing effort, presence of the brand in the market, competitor analysis, cost of the product, and multiple other variables. Sales is just one example; this study can be implemented in any section of most of the fields.

As per the Data Analysis study by Murtaza Haider of Ryerson University on the coast of the apartment and what leads to an increase in cost or decrease in cost, is also based on multivariate analysis. As per that study, one of the major factors was transport infrastructure.

People were thinking of buying a home at a location which provides better transport, and as per the analysing team, this is one of the least thought of variables at the start of the study. But with analysis, this came in few final variables impacting outcome. Multivariate analysis is part of exploratory data analysis. Based on MVA, we can visualize the deeper insight of multiple variables. There are more than 20 different methods to perform multivariate analysis and which method is best depends on the type of data and the problem you are trying to solve. [4]

Multivariate Analysis (MVA) is a Statistical procedure for analysis of data involving more than one type of measurement or observation. It may also mean solving problems where more than one dependent variable is analysed simultaneously with other variables.



There are three categories of analysis to be aware of:

- **Univariate analysis**, which looks at just one variable
- **Bivariate analysis**, which analyses two variables
- **Multivariate analysis**, which looks at more than two variables

As you can see, multivariate analysis encompasses all statistical techniques that are used to analyse more than two variables at once. The aim is to find patterns and correlations between several variables simultaneously—allowing for a much deeper, more complex understanding of a given scenario than you'll get with bivariate analysis. [5]

There are several steps to teaching how to write about multivariate analysis in graduate coursework or for dissertation writers. First, assign readings that cover key principles about statistical research writing, such as Miller (2005), Treiman (2009), or other books or articles on writing or professional research practice. Second, in lecture, briefly cover the principles and associated skills for writing about multivariate analysis, followed by in-class demonstration using such as the "poor/better/best" technique (shown below) to show students examples of how to translate abstract writing principles into concrete sentences or paragraphs; see Miller (2005) or Miller, England, Treiman and Wu (2009). Third, reinforce those concepts by assigning students to apply them to their own work or to evaluating existing published work, using one of several types of exercises, such as those shown below. Fourth, have the students use checklists such as those at the end of each chapter in The Chicago Guide to Writing about Multivariate Analysis (Miller, 2005) to plan and evaluate their work. [6-8].

11.2.3 Multivariate Analysis Example:

Wells et al.[9] published in New England Journal of Medicine a study were they hypothesized that a computed tomographic (CT) metric of pulmonary vascular disease [pulmonary artery enlargement, as determined by a ratio of the diameter of the pulmonary artery to the diameter of the aorta (PA: A ratio) of >1] would be associated with previous severe COPD exacerbations . A univariate logistic regression was used to determine the associations between patient characteristics (including the PA: A ratio) and the occurrence of a severe exacerbation of COPD in the year before enrollment. Variables showing a univariate association with severe exacerbations (at $P<0.10$) were included in stepwise backward multivariate logistic models to adjust for confounders. These models included also variables previously reported to be independently associated with acute exacerbations of COPD in the ECLIPSE study as gastro-esophageal reflux disease (GERD), lower values for the forced expiratory volume in 1 second (FEV1), a history of acute exacerbations of COPD within the previous year, increased white-cell count, and decreased quality of life as measured by the St. George's Respiratory Questionnaire (SGRQ) score (which ranges from 0 to 100, with higher scores indicating worse quality of life and with a minimal clinically important difference of 4 points). Authors found significant univariate associations between severe exacerbations and younger age, black race, use of supplemental oxygen, congestive heart failure, sleep apnea, thromboembolic disease, GERD, asthma, chronic bronchitis, employment in a hazardous job. Thanks to the development of a multivariate model, it will not only let to handle many covariates, it will let to asses potential confounders and also test for interaction or effect modification. Multiple logistic-regression analyses showed continued significant independent associations between severe exacerbations and younger age, lower FEV1 values, higher score on the SGRQ, and a PA: A ratio of more than 1.

11.2.4 Variables in Multivariate Analysis Research Methodology

Before we describe the various multivariate techniques, it seems appropriate to have a clear idea about the term, 'variables' used in the context of multivariate analysis. Many variables used in multivariate analysis can be classified into different categories from several points of view. Important ones are as under: [10]

a. Explanatory variable and criterion variable: If X may be considered to be the cause of Y, then X is described as explanatory variable (also termed as causal or independent variable) and Y is described as criterion variable (also termed as resultant or dependent variable). In some cases both explanatory variable and criterion variable may consist of a set of many variables in which case set (X1, X2, X3, …., Xp) may be called a set of explanatory variables and the set (Y1, Y2, Y3, …., Yq) may be called a set of criterion variables if the variation of the former may be supposed to cause the variation of the latter as a whole. In economics, the explanatory variables are called external or exogenous variables and the criterion variables are called endogenous variables. Some people use the term external criterion for explanatory variable and the term internal criterion for criterion variable.
b. Observable variables and latent variables: Explanatory variables described above are supposed to be observable directly in some situations, and if this is so, the same are termed as observable variables. However, there are some unobservable variables which may influence the criterion variables. We call such unobservable variables as latent variables.

c. Discrete variable and continuous variable: Discrete variable is that variable which when measured may take only the integer value whereas continuous variable is one which, when measured, can assume any real value (even in decimal points).
d. Dummy variable (or Pseudo variable): This term is being used in a technical sense and is useful in algebraic manipulations in context of multivariate analysis. We call Xi (i = 1, …., m) a dummy variable, if only one of Xi is 1 and the others are all zero.

11.2.5 Characteristics of Multivariate Analysis Techniques:

Multivariate analysis techniques are largely empirical and deal with the reality; they possess the ability to analyse complex data. Accordingly in most of the applied and behavioural researches, we generally resort to multivariate analysis techniques for realistic results. Besides being a tool for analysing the data, multivariate techniques also help in various types of decision-making. For example, take the case of college entrance examination wherein a number of tests are administered to candidates, and the candidates scoring high total marks based on many subjects are admitted.

This system, though apparently fair, may at times be biased in favour of some subjects with the larger standard deviations. Multivariate techniques may be appropriately used in such situations for developing norms as to who should be admitted in college. We may also cite an example from medical field. Many medical examinations such as blood pressure and cholesterol tests are administered to patients.

Each of the results of such examinations has significance of its own, but it is also important to consider relationships between different test results or results of the same tests at different occasions in order to draw proper diagnostic conclusions and to determine an appropriate therapy. Multivariate techniques can assist us in such a situation. In view of all this, we can state that "if the researcher is interested in making probability statements on the basis of sampled multiple measurements, then the best strategy of data analysis is to use some suitable multivariate statistical technique."

The basic objective underlying multivariate techniques is to represent a collection of massive data in a simplified way. In other words, multivariate techniques transform a mass of observations into a smaller number of composite scores in such a way that they may reflect as much information as possible contained in the raw data obtained concerning a research study.

Thus, the main contribution of these techniques is in arranging a large amount of complex information involved in the real data into a simplified visible form. Mathematically, multivariate techniques consist in "forming a linear composite vector in a vector subspace, which can be represented in terms of projection of a vector onto certain specified subspaces." For better appreciation and understanding of multivariate techniques, one must be familiar with fundamental concepts of linear algebra, vector spaces, orthogonal and oblique projections and univariate analysis.

Even then before applying multivariate techniques for meaningful results, one must consider the nature and structure of the data and the real aim of the analysis. We should also not forget that multivariate techniques do involve several complex mathematical computations and as such can be utilized largely with the availability of computer facility. [11]

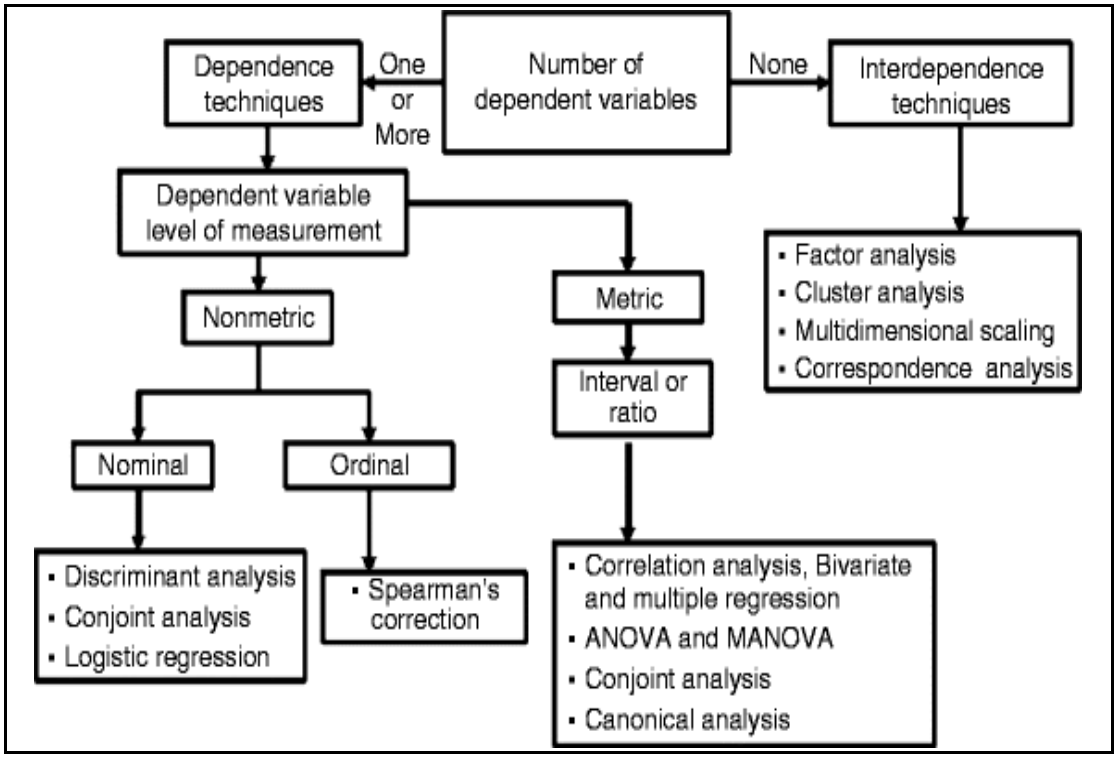
11.2.6 Stages of Realization of a Multivariate Analysis:

The steps (I) to perform a multivariate analyse can be summarized in:

I. State the objectives of the analysis. Define problem in its conceptual terms, objectives and multivariate techniques that are going to be employed.
II. Design analysis. To determine the sample size and estimation techniques those are going to be employed.
III. Decide what to do with the missing data.
IV. Perform the analysis. Identify outliers and influential observations whose influence on the estimates and goodness of fit should be analysed.
V. Interpret the results. These interpretations can lead to redefine the variables or the model which can return back to steps (III) and (IV).
VI. Validate the results. At this point, we must establish the validity of the results obtained by analysing other results obtained with the sample is generalized to the population from which it comes.

11.2.7 Classification Chart of Multivariate Techniques:

Selection of the appropriate multivariate technique depends upon-



a. Are the variables divided into independent and dependent classification?
b. If yes, how many variables are treated as dependents in a single analysis?
c. How are the variables, both dependent and independent measured?

Multivariate analysis technique can be classified into two broad categories viz., [12] this classification depends upon the question: are the involved variables dependent on each other or not?

11.2.8 The Objective of Multivariate Analysis:

1. **Data reduction or structural simplification**: This helps data to get simplified as possible without sacrificing valuable information. This will make interpretation easier.
2. **Sorting and grouping**: When we have multiple variables, Groups of "similar" objects or variables are created, based upon measured characteristics.
3. **Investigation of dependence among variables**: The nature of the relationships among variables is of interest. Are all the variables mutually independent or are one or more variables dependent on the others?
4. **Prediction Relationships between variables**: must be determined for the purpose of predicting the values of one or more variables based on observations on the other variables.
5. **Hypothesis construction and testing**. Specific statistical hypotheses, formulated in terms of the parameters of multivariate populations, are tested. This may be done to validate assumptions or to reinforce prior convictions.

11.3 Important Methods of Factor Analysis – Research Methodology:

There are several methods of factor analysis, but they do not necessarily give same results. As such factor analysis is not a single unique method but a set of techniques. Important methods of factor analysis are:

1. The centroid method;
2. The principal components method;
3. The maximum likelihood method.

Before we describe these different methods of factor analysis, it seems appropriate that some basic terms relating to factor analysis be well understood. [13, 14]

A. Factor: A factor is an underlying dimension that account for several observed variables. There can be one or more factors, depending upon the nature of the study and the number of variables involved in it.

B. Factor-Loadings: Factor-loadings are those values which explain how closely the variables are related to each one of the factors discovered. They are also known as factor-variable correlations. In fact, factor-loadings work as key to understanding what the factors mean. It is the absolute size (rather than the signs, plus or minus) of the loadings that is important in the interpretation of a factor.

C. Communality (h2): Communality, symbolized as h2, shows how much of each variable is accounted for by the underlying factor taken together. A high value of communality means that not much of the variable is left over after whatever the factors represent is taken into consideration. It is worked out in respect of each variable as under: h2 of the ith variable = (ith factor loading of factor A) 2 + (ith factor loading of factor B) 2 + …

D. Eigen Value (or Latent Root): When we take the sum of squared values of factor loadings relating to a factor, then such sum is referred to as Eigen Value or latent root.

Eigen value indicates the relative importance of each factor in accounting for the particular set of variables being analysed.

E. Total Sum of Squares: When Eigen values of all factors are totalled, the resulting value is termed as the total sum of squares. This value, when divided by the number of variables (involved in a study), results in an index that shows how the particular solution accounts for what all the variables taken together represent. If the variables are all very different from each other, this index will be low. If they fall into one or more highly redundant groups, and if the extracted factors account for all the groups, the index will then approach unity.

F. Rotation: Rotation, in the context of factor analysis, is something like staining a microscope slide. Just as different stains on it reveal different structures in the tissue, different rotations reveal different structures in the data. Though different rotations give results that appear to be entirely different, but from a statistical point of view, all results are taken as equal, none superior or inferior to others. However, from the standpoint of making sense of the results of factor analysis, one must select the right rotation. If the factors are independent orthogonal rotation is done and if the factors are correlated, an oblique rotation is made. Communality for each variables will remain undisturbed regardless of rotation but the Eigen values will change as result of rotation.

G. Factor Scores: Factor score represents the degree to which each respondent gets high scores on the group of items that load high on each factor. Factor scores can help explain what the factors mean. With such scores, several other multivariate analyses can be performed. We can now take up the important methods of factor analysis.

11.4 Advantages and Disadvantages of Multivariate Analysis:

A. Advantages:

The main advantage of multivariate analysis is that since it considers more than one factor of independent variables that influence the variability of dependent variables, the conclusion drawn is more accurate. The conclusions are more realistic and nearer to the real-life situation.

B. Disadvantages:

The main disadvantage of MVA includes that it requires rather complex computations to arrive at a satisfactory conclusion. Many observations for a large number of variables need to be collected and tabulated; it is a rather time-consuming process.

Key Takeaways and Further Reading

In this post, we've learned that multivariate analysis is used to analyse data containing more than two variables. To recap, here are some key takeaways:

The aim of multivariate analysis is to find patterns and correlations between several variables simultaneously Multivariate analysis is especially useful for analysing complex datasets, allowing you to gain a deeper understanding of your data and how it relates to real-world scenarios.

There are two types of multivariate analysis techniques: Dependence techniques, which look at cause-and-effect relationships between variables, and interdependence techniques, which explore the structure of a dataset.

Key multivariate analysis techniques include multiple linear regression, multiple logistic regression, MANOVA, factor analysis, and cluster analysis—to name just a few.

11.5 Limitations of Multivariate Analysis:

Multivariate techniques are complex and involve high level mathematics that require a statistical program to analyse the data. These statistical programs can be expensive for an individual to obtain. One of the biggest limitations of multivariate analysis is that statistical modeling outputs are not always easy for students to interpret. For multivariate techniques to give meaningful results, they need a large sample of data; otherwise, the results are meaningless due to high standard errors. Standard errors determine how confident you can be in the results, and you can be more confident in the results from a large sample than a small one. Running statistical programs is fairly straightforward but does require statistical training to make sense of the data. [15]

11.6 Significance for Usability:

As a quantitative method, multivariate analysis is one of the most effective methods of testing usability. At the same time, it is very complex and sometimes cost-intensive. Software can be used to help, but the tests as such are considerably more complex than A/B tests in terms of study design. The decisive advantage lies in the number of variables that can be considered and their weighting as a measure of the significance of certain variables.

Even four different versions of an article's headline can result in completely different click rates. The same applies to the design of buttons or the background colour of the order form. In individual cases, it is therefore worth considering from a multivariate perspective also financially, especially for commercially oriented websites, such as online shops or websites, which are to be amortized through advertising.[16]

11.7 Application of Multivariate Analysis:

- For developing taxonomies or systems of classification
- To investigate useful
- ways to conceptualize or group items
- To generate hypotheses
- To test hypotheses

Finds application in biology, medicine, psychology, neuroscience, market research, educational research, climatology, petroleum geology, crime analysis etc.

11.8 Conclusion:

This Chapter provides a brief overview of the importance of using multivariate studies in the health sciences and the different types of existing methods and their application depending on the type of variables to deal with. In addition, it described the steps to follow to design a multivariate study.

11.9 References:

1. Gil Pascual JA. Methods worth of research in education. Multivariate Analysis 2003.
2. Grimm LG, Yarnold PR. Eds. Reading and understanding multivariate statistics. Washington, DC: American Psychological Association Washington, 2011.
3. Rencher AC, Christensen WF. Eds. Methods of Multivariate Analysis. New Jersey: Wiley, 2012.
4. Berg, B.L. (2012). *Research Methods Qualitative for the Social Science*. (8th Ed). Long Beach: Allyn and Bacon.
5. Cohen, L et al (2007) *Research Methods in Education*. London. Routledge.
6. Miller, Jane E.2005. The Chicago Guide to Writing about Multivariate Analysis. The Chicago Guides to Writing, Editing and Publishing. University of Chicago Press.
7. Miller, Jane E., Paula England, Donald J. Treiman, and Lawrence Wu. "Learning How to Write about Multivariate Analysis for Sociologists: Integrating Concepts from Statistics, Research Methods, and Writing Courses." Working paper, 2009.
8. Treiman, Donald J. 2009. Quantitative Data Analysis: Doing Social Research to Test Ideas. San Francisco: Jossey-Bass.
9. Wells JM, Washko GR, Han MK, et al. Pulmonary arterial enlargement and acute exacerbations of COPD. N Engl J Med 2012; 367:913-21.
10. Creswell J.W. (1998) *Qualitative Inquiry and research design*. Thousand Oaks. Sage Publications.
11. Parahoo. K. (2006). Nursing Research - Principles, Process and Issues. (2r.sedn).Hounds mill: Palgrave.
12. Patrick, Li & Munro. S. (2004). The literature review: demystifying the literature search. Diabetes Educ. 30(1):30-8.
13. Polit, D.F, & Beck, C.T. (2006). Essentials of Nursing Research: Methods, Appraisal and Utilization. (C'edn.). Philadelphia: Lippincott, Williams & Wilkins.
14. Sharma, Y.K. (2011). Elements of Educational Research. New Delhi: Kanishka Publishers.
15. Alvin C. Rencher.2002: "Methods of Multivariate Analysis ", A. John Wiley & sons, inc. publication, Second Edition.
16. Ding Shi-Sheng 1981: "Multiple analysis method and its applications", Jilin People's Publishing House, Changchun.