Multivariate Statistical Analysis

Mushtak A.K. Shiker
Mathematical Department
Education College for Pure Science – Babylon University

Abstract

Multivariate Analysis contain many Techniques which can be used to analyze a set of data. In this paper we deal with these techniques with its useful and difficult. And we provide an executive understanding of these multivariate analysis techniques, resulting in an understanding of the appropriate uses for each of them, which help researchers to understanding the types of research questions that can be formulated and the capabilities and limitations of each technique in answering those questions. Then we gave an application of one of these techniques .

Historical View

In 1889 Galton gave us the normal distribution, this statistical methods used in traditional, established the correlation coefficient and linear regression. In thirty's of 20th century Fisher proposed analysis of variance and discriminant analysis, SS Wilkes developed the multivariate analysis of variance, and H. Hotelling determined principal component analysis and canonical correlation. Generally, in the first half of the 20th century, most of the theory of multivariate analysis has been established. 60 years later, with the development of computer science, psychology, and multivariate analysis methods in the study of many other disciplines have been more widely used. Programs like SAS and SPSS, once restricted to mainframe utilization, are now readily available in Windows-based.The marketing research analyst now has access to a much broader array of sophisticated techniques in which to explore the data. The challenge becomes knowing which technique to select, and clearly understanding its strengths and weaknesses.

Introduction :

Multivariate statistical analysis is the use of mathematical statistics methods to study and solve the problem of multi-index theory and methods. The past 20 years, with the computer application technology and the urgent need for research and production, multivariate statistical analysis techniques are widely used in geology, meteorology, hydrology, medicine, industry, agriculture and economic and many other fields, has become to solve practical problems in effective way. Simplified system architecture to explore the system kernel, can use principal component analysis, factor analysis, correspondence analysis and other methods, a number of factors in each variable to find the best subset of information from a subset of the description contained in the results of multivariable systems and the impact of various factors on the system.

In multivariate analysis, controlling for the prediction of the model has two categories. One is the prediction model, often using multiple linear regression, stepwise regression analysis, discriminant analysis or stepwise regression analysis of double screening modeling. The other is a descriptive model, commonly used cluster analysis modeling techniques.

In multivariate analysis system, the system requires a similar nature of things or phenomena grouped together, to identify the links between them and the inherent regularity, many previous studies are mostly qualitative treatment by a single factor, so the results do not reflect the general characteristics of the system. For example numerical classification, general classification model constructed using cluster analysis and discriminant analysis techniques.

In which fields we can use the multivariate analysis?

Multivariate techniques are used to study datasets in consumer and market research, quality control and quality assurance, process optimization and process control, and research and development. These techniques are particularly important in social science research because social researchers are generally unable to use randomized laboratory experiments, like those used in medicine and natural sciences. Here multivariate techniques can statistically estimate relationships between different variables, and correlate how important each one is to the final outcome and where dependencies exist between them.

Why we use multivariate techniques?

Because most data analysis tries to answer complex questions involving more than two variables, these questions are best addressed by multivariate statistical techniques. There are several different multivariate techniques to choose from, based on assumptions about the nature of the data and the type of association under analysis. Each technique tests the theoretical models of a research question about associations against the observed data. The theoretical models are based on facts plus new hypotheses about plausible associations between variables. Multivariate techniques allow researchers to look at relationships between variables in an overarching way and to quantify the relationship between variables. They can control association between variables by using cross tabulation, partial correlation and multiple regressions, and introduce other variables to determine the links between the independent and dependent variables or to specify the conditions under which the association takes place. This gives a much richer and realistic picture than looking at a single variable and provides a powerful test of significance compared to univariate techniques.

What is the difficult of using multivariate techniques?

Multivariate techniques are complex and involve high level mathematics that require a statistical program to analyze the data. These statistical programs are generally expensive. The results of multivariate analysis are not always easy to interpret and tend to be based on assumptions that may be difficult to assess. 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 a statistician to make sense of the output.

How to choose the appropriate methods to solve practical problems?

1. The problem needs to be taken into account. Problem can be integrated on a variety of statistical methods used for analysis. For example, a prediction model can be based on the biological, ecological principles, to determine the theoretical models and experimental design; based on test results, test data collection; preliminary extraction of data; and then apply statistical analysis methods (such as correlation analysis, and gradually regression analysis, principal component analysis) to study the correlation between the variables, select the best subset of variables; on this basis construct forecasting models, the final diagnosis of the model and optimization, and applied to actual production.

Multivariate analysis, taking into account the multiple response variables of statistical analysis methods.

Its main contents include two mean vectors of hypothesis testing, multivariate analysis of variance, principal component analysis, factor analysis, cluster analysis and model-related analysis.

2. So , Over several multivariate analysis methods they have advantages and limitations, each method has its specific assumptions, conditions and data requirements, such as normality, linearity, and the same variance and so on. Therefore, in the application of multivariate analysis methods, should be examined in the planning stage to determine the theoretical framework to determine what data to collect, how to collect, and how to analyze data.

Linear model approach

Multivariate analysis of commonly used methods include three categories:

1. Multivariate analysis of variance, multiple regression analysis and analysis of covariance, known as the linear model approach to research to determine the independent variables and the relationship between the dependent variable.

2. Discriminant function analysis and poly-type of analysis to study the classification of things.

3. Principal component analysis, canonical correlation and factor analysis to study how a combination of factors with less instead of more of the original number of variables.

Multivariate analysis of variance in total variance in accordance with its source (or experimental design) is divided into several parts, which test the various factors on the dependent variable and the interaction between the factors of statistical methods. For example, in the analysis of 2 × 2 factorial design data, the total variance can be divided into two factors belong to two groups variation, the interaction between two factors, and error (ie, variation within the group) and four parts, then interaction between group variation and the significance of the F test .

The advantage of multivariate analysis of variance can be tested simultaneously in one study, a number of factors with multiple levels of each of the dependent variable and the interaction between various factors. Limits its application is a sample of each level of each factor must be independent random samples, the repeated observations of the data follow a normal distribution, and the population variances are equal.

Some Types of Multivariate Analysis Techniques :

1. Multiple regression analysis :

Multiple regression analysis used to assess and analyze a number of independent variables with the dependent variable linear function of the relationship between the statistical methods.

A dependent variable y and independent variables $x_1, x_2, \ldots x_m$ is a linear regression relationship:

$$y = \alpha + \beta_1 x_1 + \ldots + \beta_m x_m + \varepsilon \quad \ldots\ldots\ldots \quad (1)$$

Where α , $\beta_1, \ldots, \beta_m$ are parameters to be estimated, ε is a random variable that error. Obtained through experiments $x_1, x_2, \ldots, x_m$ of several sets of data and the corresponding y values.

Multiple regression analysis has the advantage of a phenomenon can be described quantitatively between certain factors and a linear function. The known values of each variable into the regression equation can be obtained estimates of the dependent variable (predictor), which can effectively predict the occurrence and development of a phenomenon. It can be used for continuous variables, it also can be used for dichotomous variables.

In this technique we consider the linear relationship between one or more y's (the dependent or response variables) and one or more x's (the independent or predictor variables).We will use a linear model to relate the y's to the x's and will be concerned with estimation and testing of the parameters in the model. One aspect of interest will be choosing which variables to include in the model if this is not already known.

We can distinguish three cases according to the number of variables:

1. Simple linear regression: one y and one x. For example, suppose we wish to predict college grade point average (GPA) based on an applicant's high school GPA.
2. Multiple linear regression: one y and several x's. For example, we could attempt to improve our prediction of college GPA by using more than one independent variable, such like high school GPA, standardized test scores (such as ACT or SAT), or rating of high school.
3. Multivariate multiple linear regression: several y's and several x's. In the preceding illustration, we may wish to predict several y's (such as number of years of college the person will complete GPA in the sciences, arts, and humanities).

2. Logistic Regression Analysis :

Sometimes referred to as "choice models," this technique is a variation of multiple regression that allows for the prediction of an event. It is allowable to utilize nonmetric (typically binary) dependent variables, as the objective is to arrive at a probabilistic assessment of a binary choice. The independent variables can be either discrete or continuous. A contingency table is produced, which shows the classification of observations as to whether the observed and predicted events match.The sum of events that were predicted to occur which actually did occur and the events that were predicted not to occur which actually did not occur, divided by the total number of events, is a measure of the effectiveness of the model. This tool helps predict the choices consumers might make when presented with alternatives.

3. Multivariate Analysis of Variance (MANOVA) :

This technique examines the relationship between several categorical independent variables and two or more metric dependent variables. Whereas analysis of variance (ANOVA) assesses the differences between groups (by using T tests for 2 means and F tests between 3 or more means), MANOVA
examines the dependence relationship between a set of dependent measures across a set of groups. Typically this analysis is used in experimental design, and usually a hypothesized relationship between dependent measures is used. This technique is slightly different in that the independent
variables are categorical and the dependent variable is metric. Sample size is an issue, with 15-20 observations needed per cell. However, too many observations per cell (over 30) and the technique loses its practical significance. Cell sizes should be roughly equal, with the largest cell having less than 1.5 times the observations of the smallest cell. That is because, in this technique, normality of the dependent variables is important. The model fit is determined by examining mean vector equivalents across groups. If there is a significant difference in the means, the null hypothesis can be rejected and treatment differences can be determined.

4. Factor Analysis :

When there are many variables in a research design, it is often helpful to reduce the variables to a smaller set of factors. This is an independence technique, in which there is no dependent variable. Rather, the researcher is looking for the underlying structure of the data matrix. Ideally, the independent variables are normal and continuous, with at least 3 to 5 variables loading onto a factor. The sample size should be over 50 observations, with over 5 observations per variable.

There are two main factor analysis methods: common factor analysis, which extracts factors based on the variance shared by the factors, and principal component analysis, which extracts factors based on the total variance of the factors. Common factor analysis is used to look for the latent (underlying)

factors, where as principal components analysis is used to find the fewest number of variables that explain the most variance. The first factor extracted explains the most variance. Typically, factors are extracted as long as the eigenvalues are greater than 1.0 or the screw test visually indicates how many factors to extract .

In factor analysis we represent the variables $y_1, y_2, \ldots, y_p$ as linear combinations of a few random variables $f_1, f_2, \ldots, f_m$ ($m < p$) called factors. The factors are underlying constructs or latent variables that "generate" the y's. Like the original variables, the factors vary from individual to individual; but unlike the variables, the factors cannot be measured or observed.

The goal of factor analysis is to reduce the redundancy among the variables by using a smaller number of factors.

5. Discriminant function analysis (description of group separation) :

Discriminant function analysis is used to determine the classification of individual statistical methods. The basic principle is: According to two or more samples of known classes of observational data to identify one or several linear discriminant function and discriminant index, then determine the discriminant function based on another indicator to determine which category an individual belongs.

There are two major objectives in separation of groups:

1. Description of group separation, in which linear functions of the variables (discriminant functions) are used to describe or elucidate the differences between two or more groups.The goals of descriptive discriminant analysis include identifying the relative contribution of the p variables to separation of the groups and finding the optimal plane on which the points can be projected to best illustrate the configuration of the groups.
2. Prediction or allocation of observations to groups, in which linear or quadratic functions of the variables (classification functions) are employed to assign an individual sampling unit to one of the groups. The measured values in the observation vector for an individual or object are evaluated by the classification functions to find the group to which the individual most likely belongs.

Discriminant analysis is not only used for continuous variables and by means of number theory can be used for qualitative data, It helps to objectively determine the classification criteria. However, discriminant analysis can only be the case for the category have been identified. When the class itself is uncertain, we use the pre-separation of the first category with discriminant analysis or with cluster analysis.

6. Cluster analysis :

In cluster analysis we search for patterns in a data set by grouping the (multivariate) observations into clusters. The goal is to find an optimal grouping for which the observations or objects within each cluster are similar, but the clusters are dissimilar to each other. We hope to find the natural groupings in the data, groupings that make sense to the researcher.

So, cluster analysis is used to solve the problem of a statistical classification method. If a given observation of n objects, each object has p observed characteristics (variables), how they are clustered into several classes defined? If the object on the observed clustering, known as Q-type analysis; if the variables together class, called the R-type analysis, the basic principle of clustering is to make the internal differences of similar small, but large differences between categories. One of cluster analysis was is the hierarchical clustering method, for example, to n objects into k classes, the first n objects into a class of their own, a total of n classes, then calculate the pairwise kind of "distance" to find the nearest two classes, into a new class, then repeat this process step by step, until the date for k classes.

Cluster analysis differs fundamentally from classification analysis . In classification analysis, we allocate the observations to a known number of predefined groups or populations. In cluster analysis, neither the number of groups nor the groups themselves are known in advance. To group the observations into clusters, many techniques begin with similarities between all pairs of observations. In many cases the similarities are based on some measure of distance. Other cluster methods use a preliminary choice for cluster centers or a comparison of within- and between-cluster variability. It is also possible to cluster the variables, in which case the similarity could be a correlation.

We can search for clusters graphically by plotting the observations. If there are only two variables (p = 2), we can do this in a scatter plot. For p > 2, we can plot the data in two dimensions using principal components or biplots .

Cluster analysis has also been referred to as classification, pattern recognition (specifically, unsupervised learning), and numerical taxonomy. The techniques of cluster analysis have been extensively applied to data in many fields, such as medicine, psychiatry, sociology, criminology, anthropology, archaeology, geology, geography, remote sensing, market research, economics, and engineering.

When the sample size is large, the first n samples can be divided into k classes, and then gradually modified in accordance with the principles of a best until the classification is reasonable so far.

Cluster analysis is based on the relationship between the individual or the number of variables to classify, strong objectivity, but a variety of clustering methods can only be achieved under certain conditions, local optimum, the final clustering result is established, the experts still need to identification. Necessary to compare several different methods to choose a more in line with professional requirements of the classification results.

For example , the data matrix cab be written as :

$$y = \begin{pmatrix} y_1' \\ y_2' \\ \vdots \\ y_n' \end{pmatrix} = \left(y_{(1)}, y_{(2)}, \ldots, y_{(p)}\right) \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

where y_i' is a row (observation vector) and $y_{(j)}$ is a column (corresponding to a variable).

7. Multidimensional Scaling (MDS)

MDS is transforming consumer judgments of similarity into distances represented in multidimensional space. This is a decompositional approach that uses perceptual mapping to present the dimensions. As an exploratory technique, it is useful in examining unrecognized dimensions
about products and in uncovering comparative evaluations of products when the basis for comparison is unknown.

The multidimensional scaling is called a dimension reduction technique, In this scale we begin with the distances δ_{ij} between each pair of items. We wish to represent the n items in a low-dimensional coordinate system, in which the distances d_{ij} between items closely match the original distances δ_{ij}, that is, $d_{ij} \cong \delta_{ij}$ for all i, j.

The final distances d_{ij} are usually Euclidean. The original distances δ_{ij} may be actual measured distances between observations y_i and y_j in p dimensions, such as :

$$\delta_{ij} = [(y_i - y_j)'(y_i - y_j)]^{1/2} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (3)$$

On the other hand, the distances δ_{ij} may be only a proximity or similarity based on human judgment, for example, the perceived degree of similarity between all pairs of brands of a certain type of appliance .

The goal of multidimensional scaling is a plot that exhibits information about how the items relate to each other or provides some other meaningful interpretation of the data. For example, the aim may be seriation or ranking; if the points lie close to a curve in two dimensions, then the ordering of points along the curve is used to rank the points.

8. Principal component analysis :

In principal component analysis, we seek to maximize the variance of a linear combination of the variables. For example, we might want to rank students on the basis of their scores on achievement tests in English, mathematics, reading, and so on. An average score would provide a single scale on which to compare the students, but with unequal weights of these subjects we can spread the students out further on the scale and obtain a better ranking.

Essentially, principal component analysis is a one-sample technique applied to data with no groupings among the observations and no partitioning of the variables into subsets y and x . All the linear combinations that we have considered previously were related to other variables or to the data structure. In regression, we have linear combinations of the independent variables that best predict the dependent variable(s), in canonical correlation, we have linear combinations of a subset of variables that maximally correlate with linear combinations of another subset of variables, and discriminant analysis involves linear combinations that maximally separate groups of observations. Principal components, on the other hand, are concerned only with the core structure of a single sample of observations on p variables. None of the variables is designated as dependent, and no grouping of observations is assumed.

Principal component analysis can help identify the main factors affecting the dependent variable, and can also be applied to other multivariate analysis methods in the resolution of the main components of these after regression analysis. Principal component analysis can also serve as the first step in factor analysis.

The disadvantage is that only involves a set of interdependencies between variables, to discuss the relationship between two variables required to use canonical correlation.

9. Correspondence Analysis:

This technique provides for dimensional reduction of object ratings on a set of attributes, resulting in a perceptual map of the ratings. However, it like MDS, both independent variables and dependent variables are examined at the same time. This technique is more similar in nature to factor analysis. It is a compositional technique, and is useful when there are many attributes and many companies. It is most often used in assessing the effectiveness of advertising campaigns. It is also used when the attributes are too similar for factor analysis to be meaningful. The main structural approach is the
development of a contingency table. This means that the model can be assessed by examining the Chi-square value for the model. Correspondence analysis is difficult to interpret, as the dimensions are a combination of independent and dependent variables.
It is a graphical technique for representing the information in a two-way contingency table, which contains the counts (frequencies) of items for a cross-classification of two categorical variables. With correspondence analysis, we construct a plot that shows the interaction of the two categorical variables along with the relationship of the rows to each other and of the columns to each other.

10. Canonical correlation analysis :

Canonical correlation is an extension of multiple correlation, which is the correlation between one y and several x's , It is often a useful complement to a multivariate regression analysis. It is working through the intervening comprehensive description of the typical correlation coefficient between two sets of multivariate random variable statistical methods.

Let x be a random variable, and let y be a random variable, how to describe the degree of correlation between them? This tedious does not reflect the nature of things. If we use of canonical correlation analysis, the basic procedure is, from a linear function of two variables each taking one of each form a pair, they should be the maximum correlation coefficient of a pair, known as the first pair of canonical variables, similarly we can also find the first two pairs, 3 pairs, ... , between these pairs of variables unrelated, the correlation coefficient for a typical variable called canonical correlation coefficient.

The resulting canonical correlation coefficient of more than the original sets of variables in any set number of variables . .

Canonical correlation analysis of two sets of variables contribute to comprehensive description of the typical relationship between them. The condition is that the two variables are continuous variables, the information must obey the multivariate normal distribution.

11.Conjoint analysis:

Conjoint analysis is often referred to "trade-off analysis," in that it allows for the evaluation of objects and the various levels of the attributes to be examined. It is both a compositional technique and a dependence technique, in that a level of preference for a combination of attributes and levels is developed. A part-worth, or utility, is calculated for each level of each attribute, and combinations of attributes at specific levels are summed to develop the overall preference for the attribute at each level. Models can be built which identify the ideal levels and combinations of attributes for products and services.

12.Structural Equation Modeling (SEM) :

Unlike the other multivariate techniques discussed, structural equation modeling (SEM) examines multiple relationships between sets of variables simultaneously. This represents a family of techniques.SEM can incorporate latent variables, which either are not or cannot be measured directly into the analysis. For example, intelligence levels can only be inferred, with direct measurement of variables like test scores, level of education, grade point average, and other related measures. These tools are often used to evaluate many scaled attributes or build summated scales.

13.Log – linear models :

This techniques deals with classification data . Where there is dependent variable and one or more than one independent variables . The observed data classified in contingency table , Then we make the log linear models depending on these data , and we find the expected values for each of these models , where each model have a formula to find these expected data , after that we calculate the (Pearson χ^2) and (Likelihood – Ration Statistic G^2) , to compare them with table χ^2 to know if the choosing model make a good represent to the data or not , that is if the calculated value of (χ^2 and G^2) less than the table χ^2 , that mean the calculated value is morale , and the model make a good represent to the data , and if not (that is the calculated value of (χ^2 and G^2) not less than the table χ^2) then we have to chose another model .

Application example :

We took our data from Ibn al-Haytham Hospital which is the largest hospitals in Iraq for eye disease, the sample size N = 1285 patients, here we used the Log – linear models technique.
Patients classified according to:
1. Type of disease(which contain two categories :which is considered here as a dependent variable :
   A - retinal detachment B - inflammation of the optic sticks
2. Age in years: It has been divided into three categories:
   A – 15-44 B – 44- 64 C – more than 64

The researcher neglected age groups under 15 years old, because injury to these two diseases in this age period have been nonexistent, according to the hospital. The data classified in two – way contingency table :

Table (1)

Two-way contingency table for opserved data of the patient

Age / Disease	15 - 44	45-64	More than 64
retinal detachment	91	350	162

inflammation of the optic sticks	90	387	205

Now , we find expected value under the independent model which it's formula is :

$$mij = \frac{(xi.)(x.j)}{N} \quad \text{.....} \quad (4)$$

We put these expected values in the next table :

Table (2)

Two-way contingency table for expected data of the patient

(independent model) Log mij = u + u_1i + u_2j		
expected values matrix		
85.93	345.84	172.23
96.06	391.15	194.79

Now – we calculate (χ^2 and G^2) where their formulas are :

$$\chi^2 = \sum_{i,j} \frac{(xij - mij)^2}{mij} \text{.....} (5)$$

$$G^2 = 2 \sum_{i,j} xij \; Log \; \frac{xij}{mij} \text{.....} \text{.....} (6)$$

The results are : χ^2 = 1.9171 and G^2 = 0.001

The degree of freedom for the independent model is 2 according to the degrees of freedom table :

Table (3)

Degrees of freedom for (saturated and unsaturated model) for two-way contingency tables

Terms of u	Degrees of freedom
U	1
u_1	r -1
u_2	c -1
u_{12}	(r – 1) (c – 1)
Sum	r c

So , with 2 degrees of freedom , under the morale = 0.05 , the value of table χ^2 = 5.99

Then by compare the calculated values of χ^2 and G^2 with the value of table χ^2 , we see that the calculated values less than table value , which mean that the calculated value morale and the independent model makes good represent for the data .

Conclusions

Any of above techniques has strengths points and weaknesses points , that is mean that the analyst must be carful when he use these techniques , where he must understood the strengths and the weaknesses of each one of these techniques . Each of multivariate techniques describe some types of data different from the other techniques . Statistical programs such like (spss , sas , and others) make it easy to run a procedure .

References :

1. Alvin C. Rencher.2002 : " Methods of Multivariate Analysis " , A. John Wiley & sons, inc. publication , Second Edition.
2. Bryant and Yarnold. 1994 : "Principal components analysis and exploratory and onfirmatory factor analysis". American Psychological Association Books. ISBN 978-1-55798-273-5 .
3. Ding Shi-Sheng . 1981: "Multiple analysis method and its applications", Jilin People's Publishing House, Changchun.
4. Feinstein, A. R. 1996 :"Multivariable Analysis". New Haven, C.T. Yale University Press.
5. Garson David . 2009 :"Factor Analysis". from Statnotes, Topics in Multivariate Analysis. Retrieved on April 13, from http://www2.chass.ncsu.edu/garson/pa765/statnote.htm.
6. Raubenheimer, J. E. 2004 : "An item selection procedure to maximize scale reliability and validity". South African Journal of Industrial Psychology, 30 (4), pages 59–64.
7. Satish Chandra & Dennis Menezes. 2001: " Applications of Multivariate Analysis in International Tourism Research: The Marketing Strategy Perspective of NTOs " Journal of Economic and Social Research 3(1), pages 77-98 .
8. Yoshimasa Tsuruoka , Junichi Tsujii and Sophia Ananiadou tochastic. 2009 : " Gradient Descent Training for 1-regularized Log-linear Models with Cumulative Penalty Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP" , pages 477–485,
   Suntec, Singapore.

9. Zhang Yao Ting, with Fang Kaitai. 1982: "Multivariate Statistical Analysis Introduction", Science Press, Beijing.