Journal of Advances and Scholarly Researches in Allied Education
Vol. XV, Issue No. 6, August-2018, ISSN 2230-7540

A Study the Review of Multivariate Statistical Techniques in Biostatistics

Jagmohan Singh Dhakar[1]*, Dr. Sudesh Kumar[2]

[1] Research Scholar of Sunrise University

[2] Associate Professor, Sunrise University, Alwar, Rajasthan

Abstract - Biostatistics is largely concerned with Mathematicians & Statisticians who work in the biological sciences. Biometrical procedures are used frequently by biologists & physicians. Multivariate statistical analysis concerns in understanding different aims and background of each of the different forms of multivariate regression models and how they relate to each other. Any data study including the description of the connection between a response variable and one or more explanatory factors has used regression methods. There exists sufficient literature on applications of Multivariate Statistical tools in Biostatistics. Presently, it has been observed from the literature on Multivariate statistical techniques that Multivariate Logistic regression analysis and Poisson regression analysis have become, in many applications, the standard methods of analysis in Biostatistics.

Keywords - Biostatistics, ANOVA, MANOVA & MANCOVA, Multivariate Statistical Techniques

----------X----------

INTRODUCTION

Biostatistics is largely concerned with Mathematicians & Statisticians who work in the biological sciences. Biometrical procedures are used frequently by biologists & physicians. Biometrics is a discipline of statistics that involves the use of various computational and scientific approaches to biological research challenges. Biostatisticians collect data on a variety of variables in biological and medical research investigations on a regular basis. These multivariate data are described & analysed using multivariate statistical techniques. Multivariate statistical approaches can be thought of as a generalisation of univariate methods. Researchers in the biological & medical sciences must use multivariate statistical tools to analyse correlations amongst multiple variables, which are intrinsically challenging to apply.

Understanding the various goals & backgrounds of each of the several types of multivariate regression models, as well as how they connect to one another, is the focus of multivariate statistical analysis. Any data study including the description of the connection between a response variable and one or more explanatory factors has used regression methods. There is a substantial amount of literature on the use of multivariate statistical tools in biostatistics.

Multivariate analysis is naturally challenging for investigators in the biological and medical sciences to comprehend because of the interactions between multiple variables. In order to make inferences using multivariate statistical approaches, more mathematics is required than in a univariate environment. In many applications, Multivariate Logistic regression analysis and Poisson regression analysis have become the standard methods of analysis in Biostatistics, according to the literature on Multivariate statistic approaches.

MULTIPLE VARIABLES STATISTICAL ANALYSIS: ASSUMPTION

Most multivariate statistical analyses are based on 3 basic assumptions, but some contain additional assumptions.

Normality: Multivariate normality refers to the sample's normal distribution over all possible combinations of observations & variables. In the absence of multivariate normality, each variable's univariate normality is not a guarantee of its multivariate normality. All binary combinations of variables should meet the bivariate normality requirement, according to Mertler&Vannatta (2005), who state that each variable has a normal distribution and that linear combinations of variables should as well. Residues should have a normal distribution and be independent of one another if multivariate normality is respected. Based on whether or not the variables are grouped, multivariate normality assumption is tested. If the data are not grouped, the normality is taken to be a normal distribution for all variables and their residuals. In other words, a normal distribution is assumed for each variable and a linear & homogenous relationship between them is

anticipated. Aggregated data is regarded to have a normally distributed sample distribution of the variable means. Sample size must be large enough for the central limit theorem to hold regardless of how many variables are accessible. In order to ensure that the variables are normal, statistical analysis or graphical representations are applied. These two characteristics of normalcy are called "skewedness" and "kurtosis," respectively. The symmetry of the distribution is described by the term "skewness," whereas the term "kurtosis" describes the central point of the distribution. Dispersion of right-to-left deviations is known as skewness. Kurtosis is a measure of the central tendency of a distribution to skew either upwards or downwards.

Linearity: Linearity refers to linear relationship between 2 variables. Because linear combinations of variables underlie multivariate analysis, the assumption of linearity is particularly critical in these studies (Tabachnic 1996).

Change in a continuous variable's scores can be observed by looking at changes in other variables. In univariate analysis, we call this assumption homogeneity of variance; in multivariate analysis, we call it homogeneity of variance. The assumption of normality is linked to the assumption of homogeneity. Because the binary combinations of variables employed in multivariate analysis must be homogeneous, they must be normal (Tabachnic 1996).

MULTIVARIATE STATISTICAL TECHNIQUES: CATEGORIZATION & SELECTION

To make the analysis of complicated data sets easier, multivariate statistical approaches are applied (okluk et al., 2010). Data sets with several independent and dependent variables can be studied using these strategies. Mertler and Vannatta (2005) point out that the simultaneous study of all variables' correlations is a significant advantage. As a result, using univariate statistical approaches to evaluate these associations simultaneously is not possible. Scientific research is far too complicated to be clarified by a single factor. While many aspects influence the problematic while answering a research topic, and the problem to be solved should be assessed in light of these numerous factors. As a result of the limitations of univariate statistics, multivariate statistical studies were developed. As a result, research yields more objective and consistent results, as the presumed limits in univariate statistics are removed.

Analytical methods based on the classification of measured & latent variables as causal & relational variables are called SEMs (correlation-based). SEM is based on the study of hypothesis testing in relation to theoretically built structural models. The foundation of these structural models is a network of causal relationships between variables. Causal links are described using regression equations. Causal equations can be made easier to grasp through the use of schematic representations. When used to the social sciences, behavioural sciences, educational sciences, business, marketing, or health sciences, SEM is a statistical method that relies on a causal and relational characterization of variables that can be observed and those that cannot be (Raykov&Marcoulides, 2000). SEM's broad use now is largely due to the fact that direct and indirect effects between observable & unobserved variables may be examined in a single model (Bryne, 2010). Due to its ability to account for both visible and unobservable effects in one model, SEM has become extremely popular in recent years. It's known as multiple regression analysis (SEM). As a member of the linear model family, SEM modelling can describe complex systems with simultaneous and linked linkages, and it may model interactions amongst non-observable variables. SEM focuses on the relationships between variables and their causes. Thus, it's widely utilized in the social &behavioural sciences (Pang, 1996). When a model is employed to analyse a hypothetical situation, it can generate fictional or relevant data called SEM. There are a lot of models based on real or speculative ideas (Raykov&Marcoulides, 2000, p. 6-7). Research environments are explained and defined by these concepts. SEM is one of a kind because it enables for extremely accurate modelling of measurement error. The SEM can be used to test a theory once it has been created about a circumstance. It's known as validation in SEM applications. Structural models are used in the same way in constructed validity. Measurement methods used in these applications are examined to determine how much of an unobservable variable is recorded (Raykov&Marcoulides, 2000).

MANOVA & MANCOVA

MANOVA is a method for simultaneously revealing the associations between two or more metric dependent variables & large number of category independent factors. This is a more advanced variant of a one-variable variance analysis. After the experiment, ANCOVA can be utilized in the final part of the MANOVA to lessen the impact of the uncontrolled metric independent variable on the dependent variable. This is the same as decreasing the influence of the third variable on the bivariate correlation Data sets with two or more variables with a normal distribution & common variables can be utilised to test hypotheses (Ünlükaplan, 2008).

Multidimensional scaling is a technique that can be used instead of factor analysis. This strategy aids the researcher in explaining the similarities and contrasts between the units or items observed. This technique is used to show the important structures that lie behind the dimensions. Variables and relationships between variables are used in factor analysis. Multidimensional scaling analysis, on the other hand, uses similarities or contrasts between units to

graphically characterise things of smaller scale (Arc, 2001).

Correspondence analysis: At its most basic level, correspondence analysis is a cross-table that expresses two category variables. The nonmetric data is subsequently converted to metric data. After that, it performs variable reduction (akin to factor analysis) and creates conceptual shapes (similar to multidimensional analysis). Customers' brand preferences & demographic data (gender, income groups, business, and so on) are, for example, first stated as a cross table. The similarity & distinguishing characteristics of trademarks are stated in two or three-dimensional forms using correspondence analysis. Brands that are similar are brought together. Similarly, the distance between the demographic variable and the brand determines the discriminative aspects of client brand perception. Similarity analysis is a multivariate analysis method that can be used to analyse models with nonmetric variables that are difficult to measure using other methods (Yener, 2007).

There are several ways to reduce the amount of connected data structures to a lower number of independent data structures using factor analysis. By collecting variables assumed to explain or contribute to an occurrence or cause, this method seeks to find commonalities among them (Exploratory factor analysis). Factor analysis is a technique for sorting variables that have an impact on a structure. Structural validity can be determined by using methods like confirmatory factor analysis (zdamar, 2002). Data can be transformed and reduced via factor analysis, a type of multivariate statistical analysis often utilized in marketing research (Kinnear & Taylor, 1996). Factor analysis is a method of constructing general variables called factors by combining a large number of variables that share a strong association with each other. Reduce the amount of variables and organise them into categories by figuring out how they are related (Kalayc, 2010).

Comparable units are found by using a variety of statistical methods, including cluster analysis. Clustering is a technique for classifying and dividing objects into subgroups according to their shared characteristics. It is constructed on similarities between individuals or things, taking into account all attributes, and grouping comparable individuals together into the same groups or clusters, so that a new individual can be estimated to belong to the same group as others. Cluster analysis is a three-step process. Finding commonalities among the variables is the first stage in estimating the number of groups in the data set. The second step is the clustering of variables. This is the final stage, which is to group the variables (Yener, 2007).

LITERATURE REVIEW

Advanced & Multivariate Statistical Methods, Seventh Edition by Craig A. Mertler et al. (2017) delivers conceptual and practical information about multivariate statistical approaches to students who do not require technical and/or mathematical experience in these methods. There are three main goals for this text. The primary goal is to make multivariate statistical methods more conceptually understandable by minimising the technical character of the presentation and focusing on their practical applications. The second goal is to equip students with the knowledge and abilities they'll need to interpret research articles that use multivariate statistical approaches. Finally, the third goal of AMSM is to prepare graduate students to use multivariate statistical methods to analyse their own or their institutions' quantitative data.

S. Hajduk (2017) The growing importance of mature smart cities can be seen all around the world right now. The majority of smart city ideas concentrate on hard areas like communication and technology infrastructure. Scientists underline the importance of considering residents' social capital and expertise. Smart cities place a premium on data transparency and openness. Citizens, businesses, and visitors in mature smart cities can access real-time evidence and information. Bottom-up management and civil administration are hallmarks of smart cities. The goal of this article is to use the ISO 37120 standard to evaluate the urban smartness of a few European cities. The Multidimensional Statistical Analysis (MSA) was one of the study approaches used. The author attempted to fill knowledge gaps and evaluate the maturity of smart cities by using statistical analysis of European smart cities with the implemented ISO 37120 standard. According to the findings of the study, the smart city concept is a practical strategy that contributes to urban sustainability. The author also discovered that whereas urban sustainability frameworks include a large number of indicators that measure environmental sustainability, smart city frameworks do not include environmental indicators while focusing on social and economic factors.

Benita Percival et al. (2017) Significant historically-developed composites of sophisticated methods of statistical analysis & analytical/bioanalytical chemistry have been critical to the understanding and interpretation of the relevance of obtained results in research and industry, with applications in a wide range of fields, including biomedical sciences, healthcare, & conservation biology. Multicomponent nuclear magnetic resonance (NMR) analysis is utilised as a paradigm in this paper to show how advanced statistical tools,

both univariate and multivariate, can be employed to undertake complex spectrum dataset analyses in metabolomic applications and deliver valuable, validated results. Statistical applications can be successfully applied to spectral and chromatographic datasets gathered in a variety of domains, including disease diagnosis, disease stratification, and prognostics, as well as the production of pseudo-two-dimensional spectra, as in STOCSY-type techniques. NMR datasets can be used to develop sound associations if proper standard operating procedures are followed and cautious experimental design is taken into account. In both multivariate and univariate meanings, statistical approaches can help discriminate between the metabolic patterns of different disease classifications and disease phases. Machine learning augments statistical tools and aids in the knowledge of metabolite clustering.

Attila Csala et al. (2017) The state-of-the-art multivariate statistical approaches for high-dimensional multisetomics data analysis are covered in this chapter. Recent biotechnological advancements have enabled large-scale measurement of diverse biomolecular data spread over many omics domains, such as genotypic and phenotypic data. A new research path is to use an integrated method to study different data sources in order to better model and understand the underlying biology of complicated illness states. However, there are no comprehensive analytic approaches that can manage both the bulk and complexity of such data while also accounting for the hierarchical structure. This chapter provides an overview of some recent advances in multivariate techniques for high-dimensional omics data analysis, with a focus on two well-known multivariate methods, canonical correlation analysis (CCA) and redundancy analysis (RDA). In the realm of omics data analysis, penalised variants of CCA are common, and there has been new work on multisetpenalised RDA that is relevant to multisetomics data. These methods are discussed in terms of how they address the statistical issues that come with high-dimensional multisetomics data processing and how they contribute to our understanding of the human condition in terms of health and disease. Also covered are the present issues in the field of omics data analysis that need to be addressed.

David Núñez-Alonso, et al. (2017) From 2010 to 2017, 22 monitoring stations in Madrid city and province were used to report on the distribution of pollutants in the city and province. Air pollution data was interpreted and modelled using statistical methods. The data includes yearly average nitrogen oxide, ozone, and particle matter (PM10) concentrations gathered in Madrid and its suburbs, which is one of Europe's largest metropolitan areas and whose air quality has not been adequately researched. A map of the distribution of these contaminants was created to demonstrate the relationship between them as well as the region's population. Correlation analysis, PCA, and cluster analysis (CA) were used in a multivariate analysis to establish a correlation between different contaminants. The findings allowed separate monitoring stations to be classified based on each of the four pollutants, exposing information about their sources and methods, displaying their spatial distribution, and monitoring their levels according to the legislation's average yearly restrictions. The conclusion generated from the multivariate analysis indicating NO2 levels surpassing the yearly limit in the centre, south, and east of the Madrid province was also corroborated by the development of contour maps using the geostatistical approach of ordinary kriging.

AvijitHazra et al. (2017) Multivariate analysis is a statistical approach that examines three or more variables in connection to the subjects under inquiry at the same time in order to discover or clarify links between them. Dependence techniques, which look at the correlation between one or more dependent variables and their independent predictors, & interdependence techniques, which don't make that distinction and treat all variables equally in their search for underlying relationships, are two types of these techniques. A situation in which a single numerical dependent variable is to be predicted from many numerical independent variables is modelled by multiple linear regression. While the outcome variable is dichotomous, logistic regression is utilised. The log-linear technique can be used to evaluate cross-tabulations with more than two variables because it models count data. An expansion of ANOVA, analysis of covariance incorporates an extra independent variable of interest, the covariate, into the study. It attempts to determine if a difference exists after "controlling" for the effects of a covariate on the numerical dependent variable of interest. When many numerical dependent variables must be included in the study, MANOVA is a multivariate extension of ANOVA. Psychometrics, social sciences, & market research are the most common uses of interdependence methods. Exploratory factor analysis and principal component analysis are two similar techniques that aim to extract a smaller number of composite factors or components from a larger number of metric variables that are linearly connected to the original variables. Cluster analysis seeks to find reasonably homogeneous groupings called clusters in a large number of examples without any prior knowledge of the groups. Because multivariate analysis is so computationally demanding, most academics have avoided employing it on a regular basis. With the increased availability and sophistication of statistical software, researchers should no longer

be hesitant to investigate the applications of multivariate approaches to real-world data sets.

Dr. Sateesh Kumar Ojha et al. (2016) If variables are not correctly examined, we often get misled conclusions in research. All of the latent and observable variables must be correctly understood in order for management decisions to be relevant and effective in various functional areas of management. The purpose of this work is to look into the usage of various multivariate tools for analysing in management research, whether they are applied or basic. Data comes from both original and secondary sources. The first step is observing various research articles published in the proceedings of various conferences. The secondary section contains a variety of multivariate analysis-related papers. The investigation uncovered the reasons behind the lack of use of such research techniques. According to the preliminary findings, the majority of studies do not make extensive use of such analytical methods. The main cause for not applying proper design is carelessness in design while addressing the design aspect.

PAL IY, O. et al. (2016) Recent developments in high-throughput molecular analysis tools have resulted in an avalanche of studies yielding large-scale ecological data sets. In the field of microbial ecology, where novel experimental methodologies offered in-depth analyses of the composition, functions, and dynamic changes of complex microbial communities, a notable effect was achieved. Because even a single high-throughput experiment generates a vast amount of data, multivariate analysis is a strong statistical tool for analysing and interpreting enormous data sets. There are many different multivariate approaches available, and it's not always clear which one should be used on a given data set. The most extensively used multivariate statistical approaches, such as exploratory, interpretative, and discriminatory procedures, are described and compared in this paper. We discuss the methods' limits and assumptions, as well as examples of how they've been used in recent studies to get insight into the microbial world's ecology. Finally, we make recommendations for optimal method selection depending on the study question and data set format.

NoemíMengual-Macenlle et al. (2015) Multivariate analysis is based on the simultaneous observation and analysis of many statistical result variables. The technique is used in design and analysis to conduct trade studies across various dimensions while accounting for the impact of all variables on the responses of interest. To examine enormous databases and increasingly complicated data, multivariate approaches were developed. We should employ multivariate statistical methods since modelling is the greatest way to reflect knowledge of reality. Multivariate approaches are used to evaluate data sets in parallel, i.e., the examination of several variables for each person or object being examined. Always keep in mind that all variables must be treated in a way that truly reflects the reality of the problem at hand. There are three forms of multivariate analysis, each of which should be used depending on the variables being studied: dependent, interdependent, and structural techniques. To summarise, multivariate approaches are great for analysing huge data sets and determining cause and effect correlations between variables; we can utilise a variety of study types.

Alexis Diamondet al. (2013) This work describes genetic matching, a multivariate matching method that use an evolutionary search process to calculate the weights assigned to each covariate. This method's limiting instances include propensity score matching and Mahalanobis distance matching. Certain concerns that all matching algorithms must deal with are made transparent by the algorithm. If the selection on observables assumption holds true, simulation experiments demonstrate that the approach improves covariate balancing and may reduce bias. Then, in the LaLonde (1986) debate, we give a reanalysis of a number of data sets.

GianpaoloReboldi et al. (2013) When referring to a multivariable analysis, the term'multivariate analysis' is frequently used. The term'multivariate', on the other hand, refers to a statistical analysis having numerous results. Multivariable analysis, on the other hand, is a statistical technique for identifying the relative contributions of different factors to a single event or outcome. The focus of this article is on analyses that take into account several factors. In contrast to a univariable (or'simple') analysis, which only considers one predictor variable, a multivariable (or 'complex') analysis considers many predictor variables. We go over the fundamentals of multivariable analyses, including what assumptions they are based on and how to interpret and assess them.

Saed-Moucheshi et al. (2013) The majority of scientists base their choices on the analysis of data gathered from research studies. Almost all data in research is plentiful, but it is of little use until it is summarised using specific procedures and acceptable interpretations are generated. The data collection may contain a large number of observations that stand out and cannot be explained by any easy explanation. A multivariate statistical technique is a type of

statistics that involves observing and analysing multiple statistical variables at the same time. With examples from agriculture and plant science, we hope to clarify how multivariate statistical methods like multiple regression analysis, PCA, FA, clustering analysis, and CC can be used to explain relationships among different variables and make decisions for future work.

CONCLUSION

MANOVA is a method for simultaneously revealing the associations between two or more metric dependent variables & large number of category independent factors. This is a more advanced variant of a one-variable variance analysis. Multivariate statistical techniques, especially MANOVA & MANCOVA are essential techniques to analyst various advanced research problem pertaining biological data. Most of Biostatisticians apply MANOVA and MANCOVA techniques mom than any other statistical tools in Biometrics.The literature on Biostatistics focuses only on practical applications of ANOVA and ANCOVA techniques and entirely ignores theoretical aspects of MANOVA and MANCOVA techniques.

REFERENCES

1. Abizadeh, S. and Basilevsky, A. (1986), “Socioeconomic classification of countries maximum likelihood factor analysis technique”, Soc. Sci. Res, 15; 97-112.
2. Adamo, J.M. (2001), “Data Mining for association rules and sequential pattern”, springer, New York.
3. Afifi, A.A and Azen, S,P. (1972), “statistical analysis: A computer oriented approach”, Acedemic press. inc. New york
4. Afsarinejad, K. (1990), “Repeated measurements designs- A Review”, Communications in statistics-Theory and Methods, 19, 3985-4028.
5. Afsarinejad. K and hedayat, A.S. (2002), “Repeated measurements designs for a model with self and simple mixed carry over effects”. J. statist. Plann. Inf, 106, 449-459.
6. Agresti, A (2002), “Categorical data analysis”, Second edition, wiley Inc. New York.
7. Agresti, A (2010), “analysis of ordinal categorical data”, Second edition, Wiley Inc. New York
8. Aitchison,J (1983), “Principal components analysis of compositional data”, Biometrika, 70; 57-65.
9. Akaike, H. (1978). A Bayesian analysis of the minimum AIC procedure, Annals of the Institute of Statistical Mathematics, Series A, **30**, 9–14.
10. Akaike, H. (1987), “Factor analysis and AIC”, Psychometrika, 52; 317-332.
11. Albayrak, A. (2006). Applied Multivariate Statistical Techniques, First Edition, Ankara, Turkey: Asil Publishing and Distributing.
12. Aldrin, M. (1996). Moderate projection pursuit regression for multivariate response data, Computational Statistics & Data Analysis, 21, 501–531.
13. Alexis Diamond, Jasjeet S. Sekhon; Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies. *The Review of Economics and Statistics* 2013; 95 (3): 932–945. doi: https://doi.org/10.1162/REST_a_00318

Corresponding Author

Jagmohan Singh Dhakar*

Research Scholar of Sunrise University