Chapter
03

Multivariate Statistical Methods

Somy Kuriakose
ICAR- Central Marine Fisheries Research Institute, Kochi

Introduction

Multivariate Analysis is concerned with statistical methods designed to elicit information from data sets which include measurements on many variables. These techniques have emerged as a powerful tool to analyse data represented in terms of many variables. The main reason being that a series of univariate analysis carried out separately for each variable may lead to incorrect interpretation of the result and the inferences drawn may be misleading. This is so because univariate analysis does not consider the inter-dependence among the variables. These techniques are used in analyzing social, psychological, medical and economic data, especially when the variables concerning research studies of these fields are supposed to be correlated with each other and when rigorous probabilistic models cannot be appropriately used. Applications of multivariate techniques in practice have been accelerated in modern times because of the advent of high speed electronic computers.

The objectives of scientific investigations for which multivariate methods are commonly used are

- *Data reduction or structural simplification.* The phenomenon being studied is represented as simply as possible without sacrificing valuable information.
- *Sorting and Grouping.* Groups of similar objects or variables are created based upon measured characteristics.
- *Investigation of the dependence among variables.* The nature of the relationships among variables is of interest. Are all variables mutually independent or are one or more variables dependent on others? If so, how?
- *Prediction.* Relationships between variables must be determined for the purpose of predicting the values on the basis of observation on the other variables.

- *Hypothesis testing.* Specific statistical hypotheses, formulated in terms of the parameters of multivariate populations are tested.

Multiple Linear Regression

Multiple regression is the most commonly utilized multivariate technique. It is a statistical technique that simultaneously develops a mathematical relationship between two or more independent variables and an interval scaled dependent variable. It examines the relationship between a single dependent variable and two or more independent variables. The technique relies upon determining the linear relationship with the lowest sum of squared variances.

Let x_1, x_2,...x_k be k independent variables assumed to be related to a response variable y. The classical linear regression model states that Y is composed of a mean, which depends in a continuous manner on x_i's and random error ε.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon.$$

The beta coefficients (weights) are the marginal impacts of each variable, and the size of the weight can be interpreted directly. β_0 is the y-intercept or constant, β_1 is the coefficient on the first predictor variable, β_2 is the coefficient on the second predictor variable, and so on. ε is the error term or the residual that can't be explained by the model. The estimates of β's represented by b0, b1, b2, ..., bk that minimize the squared deviations between the expected and observed values of *Y*.are obtained by least square approach. This gives us a regression equation used for prediction of

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x$$

The multiple regression model is based on the following assumptions:

- There is a linear relationship between the dependent variables and the independent variables
- The independent variables are not too highly correlated with each other
- y_i observations are selected independently and randomly from the population
- Residuals should be normally distributed with a mean of 0 and variance σ^2

Multiple regression is often used as a forecasting tool. The multiple regression model allows an analyst to predict an outcome based on information provided on multiple explanatory variables. Still, the model is not always perfectly accurate

as each data point can differ slightly from the outcome predicted by the model. The residual value, e, which is the difference between the actual outcome and the predicted outcome, is included in the model to account for such slight variations. How well the equation fits the data is expressed by R-squared or R^2 the "coefficient of determination." R-squared is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable. It is indicative of the level of explained variability in the data set and is used as a guideline to measure the accuracy of the model. R^2 can only be between 0 and 1, where 0 indicates that the outcome cannot be predicted by any of the independent variables and 1 indicates that the outcome can be predicted without error from the independent variables. One way of interpreting this figure is to say that the variables included in a given model explain approximately x% of the observed variation. So, if the R2 = 0.50, then approximately half of the observed variation can be adequately explained by the model.

Cluster Analysis

Cluster analysis is a multivariate method which aims to classify a sample of subjects (or objects) on the basis of a set of measured variables into a number of different groups such that similar subjects are placed in the same group. It is an exploratory data analysis tool which aims at sorting different objects into groups in a way that the degree of association between two objects is maximal if they belong to the same group and minimal otherwise. Cluster analysis can be used to discover structures in data without explaining why they exist. In cluster analysis, there is no prior information about the group or cluster membership for any of the objects. Cluster analysis has no mechanism for differentiating between relevant and irrelevant variables. Therefore the choice of variables included in a cluster analysis must be underpinned by conceptual considerations. There are a number of different methods that can be used to carry out a cluster analysis which are classified as follows:

Hierarchical methods

A hierarchical procedure in cluster analysis is characterized by the development of a tree like structure. A hierarchical procedure can be agglomerative or divisive.

Agglomerative methods in which subjects start in their own separate cluster. The two 'closest' (most similar) clusters are then combined and this is done repeatedly until all subjects are in one cluster. At the end, the optimum number of clusters is then chosen out of all cluster solutions. Agglomerative

methods in cluster analysis consist of linkage methods, variance methods, and centroid methods. Linkage methods in cluster analysis are comprised of single linkage, complete linkage, and average linkage.

Divisive methods in which all subjects start in the same cluster and the above strategy is applied in reverse until every subject is in a separate cluster.

Non-hierarchical methods (k-means clustering methods)

It follows a simple procedure of classifying a given data set into a number of clusters, "k," which is fixed in advance. This method will categorize the items into k groups of similarity using the Euclidean distance as measurement. The clusters are then positioned as points and all observations or data points are associated with the nearest cluster, computed, adjusted and then the process starts over using the new adjustments until a desired result is reached.

The choice of clustering procedure and the choice of distance measure are interrelated. The relative sizes of clusters in cluster analysis should be meaningful. The clusters should be interpreted in terms of cluster centroids. The variables on which the cluster analysis is to be done should be selected by keeping past research in mind. It should also be selected by theory, the hypotheses being tested, and the judgment of the researcher.

Principal component analysis

Principal component analysis (PCA) is a technique used to emphasize variation and bring out strong patterns in a dataset. Principal Component Analysis (PCA) is a dimension-reduction tool that can be used to reduce a large set of variables to a small set that still contains most of the information in the large set. One way of reducing the number of variables is to consider some of the linear combinations of these variables only. Principal component analysis (PCA) is a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components. We can discard those linear combinations which have smaller variances and consider only those combinations which have high variances. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. Principal components are linear combinations of the statistical or random variable which have special properties in terms of the variances. For example, first PC is the normalized linear combination of the original variable with maximum variance. The second PC is the normalized linear combination,

which has the second maximum variance and uncorrelated with first PC .The total variance of the variables equals the total variance of the components.

Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components is less than or equal to the number of original variables.

Let X be the component vector with variance-covariance matrix $\sum$. Since we are interested in the variances and co-variances, we have suppose that E(X) = 0. Let β be the component vector such that $\beta'\beta = 1$ and $V(\beta) = \beta\sum'\beta$ is maximum. The vector of principal component is the solution of $(\sum - \lambda I) = 0$. Then the first principal component is $U_1 = \beta' X$ and the variance is the largest root of $\left|\sum - \lambda I\right| = 0$ and $V\left(\beta' X\right) = \lambda_1$.

The Eigen vectors of a square matrix are the non-zero vectors that, after being multiplied by the matrix, remain parallel to the original vector. For each Eigen vector, the corresponding Eigen vector is the factor by which the Eigen vector is scaled when multiplied by the matrix. The prefix Eigen is adopted from the German word "Eigen" for" self" in the sense of a characteristic description. The Eigen vectors are sometimes also called characteristic vectors. Similarly, the Eigen values are also known as characteristic values.

The mathematical expression of this idea is as follows; if a square matrix ,a non-zero vector v is an Eigen vector of A if there is scalar λ such that

$$Av=\lambda v$$

Then scalar is said to be the Eigen value of A corresponding to v. An Eigen space of A is the set of all Eigen vectors with the same Eigen value together with the zero vector. However, the zero vector is not an Eigen vector.

Steps in PCA

- Standardize the data.
- Perform Singular Vector Decomposition to get the Eigenvectors and Eigenvalues.
- Sort eigenvalues in descending order and choose the k- eigenvectors
- Construct the projection matrix from the selected k- eigenvectors.
- Transform the original dataset via projection matrix to obtain a k-dimensional feature subspace.

It is mostly used as a tool in exploratory data analysis and for making predictive models. Often its operation can be thought of as revealing the internal structure of the data in a way that best explain the variance in the data. If a multivariate data set is visualized as asset of coordinates in a high dimensional data space, principal component analysis can supply the user with a low-dimensional structure, a shadow of this object when viewed from its most informative view. This can be done by using only the first few principal components so that dimensionality of the transformed data is reduced. The principal component analysis is concerned with explaining the variance covariance structure through a few linear combinations of the original variables. Its general objectives are data reduction and interpretation.

Canonical Correlation

Canonical correlation analysis (CCA) is a way of measuring the linear relationship between two multidimensional variables. It is the multivariate extension of correlation analysis. It finds two bases, one for each variable, that are optimal with respect to correlations and, at the same time, it finds the corresponding correlations. The aim of canonical correlation analysis is to find the best linear combination between two multivariate datasets that maximizes the correlation coefficient between them. This is particularly useful to determine the relationship between criterion measures and the set of their explanatory factors. This technique involves, first, the reduction of the dimensions of the two multivariate datasets by projection, and second, the calculation of the relationship (measured by the correlation coefficient) between the two projections of the datasets.

It is the multivariate extension of correlation analysis. Ordinary correlation analysis is dependent on the coordinate system in which the variables are described. This means that even if there is a very strong linear relationship between two multidimensional signals, this relationship may not be visible in an ordinary correlation analysis if one coordinate system is used, while in another coordinate system this linear relationship would give a very high correlation. Canonical correlation analysis finds the coordinate system that is optimal for correlation analysis. CCA connects two sets of variables by finding linear combinations of variables that maximally correlate.

The major purposes of CCA are:

- Data reduction: explain covariation between two sets of variables using small number of linear combinations

- Data interpretation: find features (i.e., canonical variates) that are important for explaining covariation between sets of variables

The canonical correlation technique is to find several linear combinations of X variables and the same number of linear combination of Y variables in such as these linear combination best express the correlation between the two sets. These linear combinations are called the canonical variables. The correlation between the corresponding pairs of canonical variables is called canonical correlation.

Suppose we desire to examine the relationship between a set of variables x1, x2,· · · , xp and another set y1, y2,· · · , yq. And the sample means for all x and y variables are zero. The first step in canonical correlation is to form two linear combinations:

$$W_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p$$
$$V_1 = b_{11}y_1 + b_{12}y_2 + \cdots + b_{1q}y_q,$$

such that corr $(W_1, V_1) = C_1$ is maximum.

Then the second step is to identify another set of canonical variables

$$W_2 = a_{21}x_1 + a_{22}x_2 + \cdots + a_{2p}x_p$$
$$V_2 = b_{21}y_1 + b_{22}y_2 + \cdots + b_{2q}y_q$$

such that corr$(W_2, V_2) = C_2$ is maximum and corr$(W_1, W_2) = 0$, corr$(V_1, V_2) = 0$.

The process of extracting canonical variables can be repeated until the number of canonical variables equals the number of original variables or the number of classes minus one, whichever is smaller.