....Statistics for Economic Development

STAT DAY 2009

26th March 2009

Platinum Partners

QATAR

NSB National Savings Bank

Print Media Partner

The Nation on Sunday

Electronic Media Partner

ITN Sri Lanka

Organized by Stat Circle - University of Colombo

What is Multivariate Analysis?

by

Miss. H. Thilaksha Tharanganie
Statistics Special, 2004/2005
Department of Statistics, University of Colombo

Introduction

Multivariate Data Analysis refers to statistical techniques used for analyzing data that arises from multiple measurements. This essentially models reality as most products or decisions involve more than one variable. When information is stored as tables containing multiple measurements, Multivariate Analysis can be used to analyze the information in a meaningful way.

Objectives of using multivariate analysis

1. **Data reduction or structural simplification** – replace the variables with fewer components so that they explain the most of the variability of the data hence interpretation will become easier.
2. **Sorting and grouping** – Grouping or clustering the similar experimental units.
3. **Investigation of correlations among variables** – The relationships (whether dependent or independent) among variables are of interest.
4. **Prediction** – Forecast the values of one or more variables on the basis of observations on the other variables.
5. **Hypothesis testing** - To validate assumptions or to reinforce prior convictions, specific statistical hypotheses in terms of the parameters of multivariate populations are tested.

Types of multivariate techniques

- **Principal Components Analysis** – Involves a mathematical procedure that transforms a set of correlated response variables into a smaller, uncorrelated set called "Principal Components".
- **Factor Analysis** – Describes the covariance among the many variables in terms of a few underlying but unobservable random quantities called "factors". This analysis can be considered an extension of principal component analysis.
- **Cluster Analysis** – Grouping or clustering the experimental units according to some similarity measure with respect to performance. This analysis is more primitive in that no assumptions are made concerning the number of groups or group structure.

- **Discrimination and Classification Analysis** - Separating distinct sets of observations (discrimination / separation) and allocating new observations to previously defined groups (classification / allocation).
- **Canonical Correlation Analysis** – Identify the correlations between two sets of linear combination of the variables. The pairs of linear combinations are called "canonical variables" and their correlations are called "canonical correlations".
- **Multiple Regression Analysis or Partial Least Squares (PLS)** - Use one set of variables to predict another using the relationships among variables, for the purpose of optimization. Then find out which variables are important in the relationship.
- **MANOVA (Multivariate ANalysis Of VAriance)** - Multivariate version of the univariate ANOVA. This tests whether several samples have the same mean. If the MANOVA shows significant overall difference between groups, the analysis can proceed by pairwise comparisons.

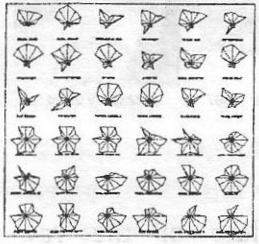
Graphical displays of multivariate data

Following techniques are some of the graphical displays for multivariate data and they display the relationships of three or more variables in one plot.

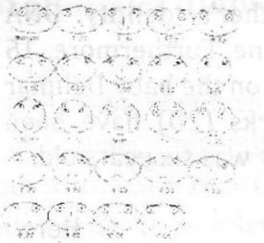
- **Star plots** – "Detecting clusters". Each observation is represented as a star-shaped figure with one ray for each variable and the length of each ray is made proportional to the size of that variable. Then stars can be grouped according to their similarities.
- **Chernoff Faces** – "Detecting clusters". Code each variable in some feature of a face (e.g. length of nose, curvature of mouth) and they can handle up to 18 variables. Display each observation as a face, and then look for grouping similar faces.
- **Scatterplot matrix** – "Plotting all pairs of variables" and it is a useful way to display multivariate observations with any number of variables.
- **Biplots** – "Plotting observations and variables together". Hence, their joint relationships can be depicted.
- **Glyph plots** – "Adding more variables to scatter plot". These plots are useful for 3-5 variables, but they do not generalize easily to an arbitrary number of variables.
- **Growth Curves** – "Points can be plotted and connected by lines to produce a curve". These curves are used in general; when the repeated measurements of the same characteristic on the same unit are available.

- **Chi-square plots or Gamma Q-Q plots** – "Assessing multivariate normality and detecting outliers". A systematic deviation from a straight line, suggests the lack of normality. Potential outliers appear as points in the upper right which are substantially above the line.

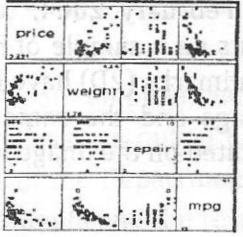
Some examples of graphical displays for multivariate data:



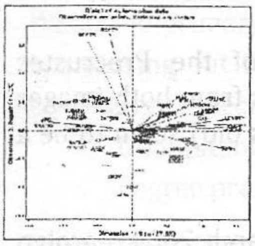
Star plot



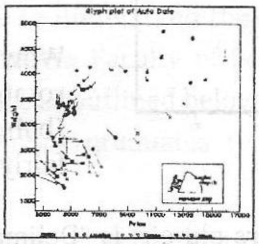
Chernoff faces



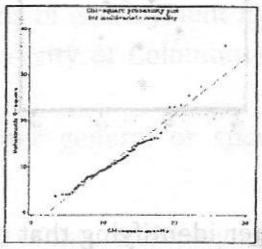
Scatterplot matrix



Biplot



Glyph plot



Chi-square plot

Multivariate applications

- Quality control and quality assurance across a range of industries such as food and beverage, paint, pharmaceuticals, chemicals, energy, telecommunications, etc
- Consumer and market research
- Process optimization and process control
- Research and development
- Genetic experiments

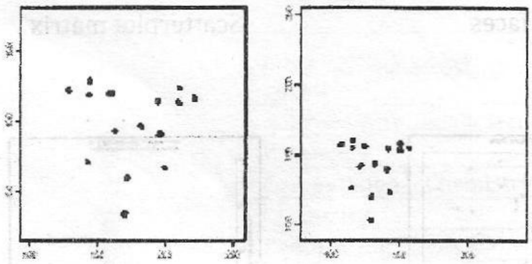
A Real Life Application of a Multivariate Analysis

A mother's instinct

December 1997 January 2004

"Delimar Vera", a 10 day old baby thought to have died in a fire in December 1997 was actually kidnapped by a woman. The girl's real mother saw the girl and recognized her as her own at a birthday party when she was six years old in January 2004.

In February 2004, to prove mother's identity, DNA tests on a sample of hair were done. Furthermore, 15 landmarks (2D) have been located on the baby Delimar image and the same 15 landmarks (2D) have been located on the image when Delimar was six years old.



Here are the plots of the raw landmark coordinates (left is the baby picture; right is the picture from six years old). Quite clearly the scales for the two images are not the same.

Wow! The plots of the Procrustes rotated coordinates from both images (baby and six years old) seem to be a fairly close match.

After identifying that this six years old girl is "Delimar Vera", in March 2004, Carolyn Correa, who had raised Delimar for six years, was charged for kidnapping.

References :

Johnson,R.A. , & Wichern,D.W. *Applied Multivariate Statistical Analysis.* Prentice Hall, 1992.

http://www.math.yorku.ca/SCS/sugi/sugi16-paper.html

http://www.nickfieller.staff.shef.ac.uk/sheff-only/mvalectures.pdf