A HANDBOOK OF

Multivariate Data Analysis Using R

A.K.Sheik Manzoor
Ganesh Kumar R

Scan to Preview

SULTAN CHAND & SONS

HANDBOOK OF MULTIVARIATE DATA ANALYSIS USING R

Dr. A.K. Sheik Manzoor
*Professo*r,
Department of
Management Studies,
College of Engineering,
Anna University

Dr. Ganesh Kumar R.
Supply Chain Consultant

SULTAN CHAND & SONS®
Educational Publishers
New Delhi

SULTAN CHAND & SONS®

Educational Publishers

23, Daryaganj, New Delhi-110002

Phones : 011-23281876, 23266105, 41625022 (*Showroom & Shop*)
011-23247051, 40234454 (*Office*)

E-mail : sultanchand74@yahoo.com; info@sultanchandandsons.com

Fax : 011-23266357; Website : www.sultanchandandsons.com

First Edition: 2024

ISBN: 978-93-91820-75-6 (TC–1308)

Price: 225.00

EVERY GENUINE COPY OF THIS BOOK HAS A HOLOGRAM

In our endeavour to protect you against counterfeit/fake books, we have pasted a copper hologram over the cover of this book. The hologram displays the full visual image, unique 3D multi-level, multi-colour effects of our logo from different angles when tilted or properly illuminated under a single light source, such as 3D depth effect, kinetic effect, pearl effect, gradient effect, trailing effect, emboss effect, glitter effect, randomly sparking tiny dots, micro text, laser numbering, *etc*.

A fake hologram does not display all these effects.

Always ask the bookseller to put his stamp on the first page of this book.

Author's Acknowledgement: The writing of a Textbook always involves creation of a huge debt towards innumerable author's and publications. We owe our gratitude to all of them. We acknowledge our indebtedness in extensive footnotes throughout the book. If, for any reason, any acknowledgement has been left out we beg to be excused. We assure to carry out correction in the subsequent edition, as and when it is known.

Printed at : Himani Print Solution, Badarpur, New Delhi-110044

Preface

This book is addressed at Academic as well as Industry Researchers who use Statistical Analyses for their Research. Many times, researchers have to spend a lot of money to purchase proprietary softwares, especially academic researchers. They often have a need of financial support for their researches and frequently face tight financial situations. This book may especially benefit these researchers. This book performs Statistical analyses using a software tool called R. This software tool is open-source, which means it is not proprietary and freely downloadable.

The Statistical Analyses supported by R include a spectrum of techniques, right from univariate and bivariate Statistical techniques to Multivariate Statistical Techniques and then, even Big Data Analytics. In fact, learning R can take a researcher far into the Analytics highway. Apart from researchers in academics and industries, this book will also be useful for Analytics Professionals and of course, Bachelors and Masters Degree students.

The need for such a book was felt after several researchers facing financial difficulties during their research period were observed, as mentioned before. A result of this causes in researchers ending up with incomplete knowledge of analysis techniques and concepts involved in the analysis techniques having very much a cyclical effect on each other. This many times hampers the research careers of the researchers. This book is very much intended to fill this void and provide the researchers the necessary boost towards their careers.

This book discusses starting with analyses of Assumptions of Multivariate Techniques and discusses analyses of some of the most frequently used Multivariate Techniques namely Multiple Regression, Discriminant Analysis, Logistic Regression,

MANOVA, Conjoint Analysis, Cluster Analysis, Multidimensional Scaling, Correspondence Analysis, Exploratory Factor Analysis, Confirmatory Factor Analysis and Structural Equations Modelling (SEM) using R.

Dr. A.K. Sheik Manzoor

Dr. Ganesh Kumar R.

Contents

1. Introduction to Multivariate Data Analysis 1
2. Assessing the Characteristics of Data 5
   - 2.1. Normality 5
   - 2.2. Linearity 10
   - 2.3. Homogeneity 13
   - 2.4. Independence 21
3. Multiple Linear Regression 25
   - 3.1. Regression with Dummy Variable 28
4. Discriminant Analysis 35
5. Logistic Regression 45
6. MANOVA 53
7. Conjoint Analysis 61
8. Cluster Analysis 69
9. Multidimensional Scaling 79
10. Correspondence Analysis 89
11. Exploratory Factor Analysis 99
12. Confirmatory Factor Analysis 107
13. Structural Equations Modelling 115

Glossary 129

CHAPTER 1

Introduction to Multivariate Data Analysis

Multivariate data analysis refers to the use of more than two variables in the data, whereas univariate data analysis refers to the use of single variable in the analysis of data and bivariate data analysis refers to the use of two variables in the analysis of data. The reason why raw data collected from observations or other sources are analysed is to gain meaningful information from the data and use them for managerial decision-making. An example could be the data collected to record the number of cars of a company sold during different months of a year. The raw numbers are just data. When these indicate the number of cars sold during different months of a year, they become information and provide inferences, for example, let us assume that the sales of cars peak during April and May, because say, people like to spend more during these months. This inference suggests management of the respective car company to stock more car units and in turn produce more car units to source the demand during this period.

When the relationship between variables is fixed, it is referred to as a deterministic relationship. For example, consider some relationship like $4x + 3y = 20$. When the relationship between variables is defined to be based on probability, it is referred to as a stochastic or probabilistic relationship. The right-hand side value of 20 in the previous example in a probabilistic relationship may be expected to lie within a range of values based on probability. Such relationships are stochastic or probabilistic relationships. The word "random" in statistics refers to a probabilistic relationship as against its common meaning in language.

The sources of data are basically classified into two types: primary data and secondary data. Primary data is collected by the researcher for his research for the first time. Secondary data refers to data that is collected from sources which have been used or published by other researchers as a part of their research.

Data also have different scales namely nominal, ordinal, interval and ratio. Nominal scale data cannot be used to perform arithmetic operations on them. These scales are just labels. For example, the numbers behind the shirts of sportsmen may be considered. These numbers can be used only as labels and cannot be used for arithmetic operations as discussed before. Ordinal scale indicates an order in the data. For example, rank 3 is better than rank 4 and rank 4 is better than rank 5. When the interval between any two adjacent ranks is same, then the scale may be considered interval scale. But the interval scale does not have an absolute zero. An example for interval scale is temperature. Here, zero-degree Celsius or zero-degree Fahrenheit does not indicate absence of temperature. Ratio data can be used to perform all arithmetic operations on them and they have an absolute zero. For example, zero patients during a period indicates absence of patients during the period. More details of the different scales of data can be referred from any standard course textbooks. Nominal and ordinal data come under the category of non-metric data. Interval and ratio data come under the category of metric data.

Now referring to the multivariate data analysis techniques, they can be classified into two types: dependency techniques and inter-dependency techniques. Dependency techniques are techniques where variables can be classified into dependent variables and independent variables. In the interdependency techniques, the variables are grouped such that there is no dependence relationship between the variables.

Dependency techniques include Multiple Linear Regression (MLR), Multiple Discriminant Analysis, Logistic Regression, MANOVA, Conjoint Analysis, Structural Equations Modelling (SEM), *etc.* Interdependency techniques consist of Exploratory Factor Analysis (EFA), Confirmatory Factor Analysis (CFA), Cluster Analysis, Multi-dimensional Scaling, Correspondence Analysis, *etc*.

Chapter 2 talks about assessing the characteristics of data required for multivariate data analysis like normality, linearity, homogeneity and independence.

Chapter 3 talks about Multiple Linear Regression (MLR). As discussed before, it is a dependency technique, where the dependent variable and independent variables are metric.

Chapter 4 talks about Multiple Discriminant Analysis, where the dependent variable is a categorical or non-metric variable with two or more categories and the independent variables are metric.

Chapter 5 talks about Logistic Regression in which dependent variable is a categorical variable with only two categories and the independent variables are metric.

Chapter 6 talks about MANOVA in which there are multiple dependent and independent variables. Dependent variables are metric, whereas independent variables are categorical (non-metric).

Chapter 7 talks about Conjoint Analysis in which there is one dependent variable and multiple independent variables. The dependent variable is metric or non-metric and the independent variables are non-metric.

Chapter 8 talks about Cluster Analysis which is an inter-dependency technique and groups respondents based on some measure of similarity.

Chapter 9 talks about Multi-dimensional Scaling. It is an inter-dependency technique and it reduces the dimensions of a variable to represent the variable in a two-dimensional representation which is easier to comprehend by human perception and can also be represented graphically.

Chapter 10 talks about Correspondence Analysis which groups non-metric variables into a two-dimensional solution based on the chi-square distances between the variables.

Chapter 11 and 12 talks about EFA and CFA respectively inter-dependency techniques.

Chapter 13 talks about Structural Equations Modelling (SEM) which models a set of equations. The dependent variable in one equation can become the independent variable in the next equation. Viewed in another way, there are dependence relationships between exogenous and endogenous constructs. A construct is a latent variable which cannot be measured directly and is measured by a set of indicators.

RStudio is an open source software for performing statistical analyses. It is a software which can be used for running commands as well as scripts for performing statistical analyses. This software is used in this book for performing the analyses and illustrating the multivariate techniques.

The primary focus here is on data analysis. Instead of encountering challenges associated with proprietary software, this book emphasizes the widespread adoption of open-source software within the research community. It encompasses a broad range of analytical methods, starting from multiple regression, discriminant analysis, logistic regression, and MANOVA, and progressing to various advanced multivariate techniques. It provides clear and easily comprehensible explanations for all twelve of these techniques, including the complex domain of structural equations modeling.

Dr. A.K. Sheik Manzoor is a Management Professor at Anna University College of Engineering, Guindy Campus in Chennai, India. He holds degrees in ME, MBA, and a PhD in Management from renowned institutions, and he has amassed over two decades of experience in research and teaching. His expertise primarily centers on guiding research scholars, and his academic pursuits predominantly focus on operations and systems.

Dr. Ganesh Kumar R serves as a Supply Chain Consultant, possessing qualifications including a BE, MBA, and PhD from well-regarded institutions. With a background in both the software industry and the education sector, he brings a wealth of experience to his career. His research interests are primarily centered around supply chain management and systems.

Sultan Chand & Sons
Publishers of Standard Educational Textbooks
23 Daryaganj, New Delhi-110002
Phones (S) : 011-23281876, 23266105, 41625022
(O): 011-23247051, 40234454
Email : sultanchand74@yahoo.com
info@sultanchandandsons.com

Dedicated to the Service of Education
SULTAN
1950

Scan to Visit Us

TC 1308

ISBN 978-93-91820-75-6
9 789391 820756