海外优秀数学类教材系列丛书

THOMSON

影印版

Applied Multivariate Methods for Data Analysts

应用多元统计分析方法

DALLAS E. JOHNSON

高等教育出版社
Higher Education Press

海外优秀数学类教材系列丛书

影印版

Applied Multivariate Methods for Data Analysts

应用多元统计分析方法

Dallas E. Johnson
Kansas State University

高等教育出版社
Higher Education Press

图字：01-2004-3218 号

Dallas E.Johnson

Applied Multivariate Methods for Data Analysts, first Edition

ISBN: 0-534-23796-7

981-265-506-9

图书在版编目(CIP)数据

应用多元统计分析方法 = Applied Multivariate Methods for Data Analysts/(美)约翰逊(Johnson, D. E.)著. —北京:高等教育出版社,2005. 6
(海外优秀数学类教材系列丛书)
ISBN 7-04-016545-7

Ⅰ. 应... Ⅱ. 约... Ⅲ. 多元分析—统计分析—分析方法—高等学校—教材—英文 Ⅳ. O212. 4

中国版本图书馆 CIP 数据核字(2005)第 044048 号

策划编辑 徐 可 **责任编辑** 徐 可 **封面设计** 王凌波 **责任印制** 陈伟光

出版发行	高等教育出版社	购书热线	010-58581118
社　　址	北京市西城区德外大街 4 号	免费咨询	800-810-0598
邮政编码	100011	网　　址	http://www.hep.edu.cn
总　　机	010-58581000		http://www.hep.com.cn
		网上订购	http://www.landraco.com
经　　销	北京蓝色畅想图书发行有限公司		http://www.landraco.com.cn
印　　刷	北京民族印刷厂		
开　　本	787×1092 1/16	版　　次	2005 年 6 月第 1 版
印　　张	36. 5	印　　次	2005 年 6 月第 1 次印刷
字　　数	650 000	定　　价	43. 30 元(含光盘)

Preface

I once attended a conference at which George Box stated that "Statistics is much too important to be left entirely to statisticians." A bit later, Walt Federer stated that "Science is much too important to be left entirely to scientists." Both of these famous statisticians were correct! Never before in the history of science and statistics has there been a greater need for interactions and collaborations between scientists and statisticians. This book helps to facilitate such collaborations and interactions. I have been fortunate in that I have had substantial contact with scientists during my tenure at Kansas State University. These collaborations have greatly influenced my approach to teaching multivariate methods. I believe that multivariate methods are too important to be taught only to statisticians.

Furthermore, I have been teaching public seminars and college courses in applied multivariate analyses for the last 20 years. In these seminars and courses, students have posed many important questions that multivariate methods can help answer. As data sets grow in size, multivariate methods become ever more useful. Today's technologies make it very easy to collect large amounts of data; multivariate methods are needed to determine whether such massive amounts of data actually contain information. It has been said that while it is easy to collect data, it is much harder to collect information. Multivariate methods can help determine whether there is information in data, and they can also help to summarize that information when it exists.

To date, textbooks have emphasized only the theory of multivariate methods or only the application of the methods. Readers were given information that was either too advanced to apply or too elementary to illustrate the power of the methods. This text has broken the mold by focusing on the why, when, how, and what of multivariate analyses and answering the following questions:

Why should multivariate methods be used?

When should they be used?

How can they be used?

And what has been learned by the application of the methods?

Ideally, users of this book will have had a previous course in statistics that included multiple regression. Some familiarity with matrix algebra is desirable, but not crucial. My approach assumes familiarity with most of the statistical concepts encountered in a first course in statistics, such as means and standaı deviations, correlations, p-values, hypothesis tests, and confidence limits.

While this text is loaded with examples using real data, several of tł exercises are directed at data sets that students are asked to provide fro their own experiences. I find that students enjoy working with their own dat So, when I teach multivariate methods, I require each student to provide data set for class use along with a description of the data's important featur and the reasons behind its being collected. These data sets are then place in a computer directory that every student in the class can access. I then u these data sets as much as possible when assigning exercises to the class. strongly encourage instructors who use this book to do the same.

Other unique features of this text include:

- annotated computer output, emphasizing SAS and SPSS
- extensive use of graphics to explain concepts
- data disk that contains data files from text discussion and exercise as well as computer commands used to create the analyses describe throughout the text

I owe much of the development of this text to those who have participate in my seminars and courses. From these "students," I learned about the needs, their concerns, and their abilities. In writing this text, I have tried address their needs and concerns, while recognizing their differing abilitie

Acknowledgments

I wish to express my appreciation to all who helped me with the developme of this text. I am particularly grateful to the students at Kansas State Universi and students who have taken public seminars through the Institute of Profe sional Education. These students have provided numerous valuable sugge tions that have greatly improved the content of this text. I would also like thank Ms. Carolyn Crockett and Mr. Alexander Kugushev for their valuab suggestions. I would like to thank the following reviewers for their helpf comments: Marcia Gumpertz, North Carolina State University; John E. He ett, University of Missouri, Columbia; Linda S. Hynan, Baylor Universit Dipak Jain, Northwestern University; Lincoln Moses, Stanford Universit Mack C. Shelley II, Iowa State University; Eric Smith, Virginia Polyte Institute; and Richard Sundheim, St. Cloud State University. I also thank M Jane Cox for her help in creating many of the formulas in this text. Finall I would like to thank my parents, Chet and Dorothy Johnson, for giving n the opportunity for furthering my education and my wife, Erma, for the he and support that she provided during this endeavor.

Dallas Johns

Contents

Applied Multivariate Methods

1

Multivariate data occur in all branches of science. Almost all data collected by today's researchers can be classified as multivariate data. For example, a marketing researcher might be interested in identifying characteristics of individuals that would enable the researcher to determine whether a certain individual is likely to purchase a specific product. Furthermore, a wheat breeder might be interested in more than just the yields of some new varieties of wheat. The wheat breeder may also be interested in these varieties' resistance to insect damage and drought. Finally, a social scientist might be interested in studying the relationships between teenage girls' dating behaviors and their fathers' attitudes. Each of these endeavors involves multivariate data.

To begin a discussion of multivariate data analysis methods, the concept of an experimental unit must be defined. An *experimental unit* is any object or item that can be measured or evaluated in some way. Measuring and evaluating experimental units is a principal activity of most researchers. Examples of experimental units include people, animals, insects, fields, plots of land, companies, trees, wheat kernels, and countries. *Multivariate data* result whenever a researcher measures or evaluates more than one attribute or characteristic of each experimental unit. These attributes or characteristics are usually called *variables* by statisticians.

The next section gives an overview of some multivariate methods that are discussed in this text.

1.1 An Overview of Multivariate Methods

Multivariate methods are extremely useful for helping researchers make sense of large, complicated, and complex data sets that consist of a lot of variables measured on large numbers of experimental units. The importance and usefulness of multivariate methods increase as the number of variables being measured and the number of experimental units being evaluated increase.

Often, the primary objective of multivariate analyses is to summarize large amounts of data by means of relatively few parameters. The underlying theme behind many multivariate techniques is simplification.

Multivariate analyses are often concerned with finding relationships among (1) the response variables, (2) the experimental units, and (3) both response variables and experimental units. One might say that relationships exist among the response variables when several of the variables really are measuring a common entity. For example, suppose one gives tests to third-graders in reading, spelling, arithmetic, and science. Individual students may tend to get high scores, medium scores, or low scores in all four areas. If this did happen, then these tests would be related to one another. In such a case, the common thing that these tests may be measuring might be "overall intelligence."

Relationships might exist between the experimental units if some of them are similar to each other. For example, suppose breakfast cereals are evaluated for their nutritional content. One might measure the grams of fat, protein, carbohydrates, and sodium in each cereal. Cereals would be related to eacl other if they tended to be similar with respect to the amounts of fat, protein carbohydrates, and sodium that are in a single serving of each cereal. On might expect sweetened cereals to be related to each other and high-fibe cereals to be related to each other. One might also expect sweetened cereal to be much different than high-fiber cereals.

Many multivariate techniques tend to be exploratory in nature rather tha confirmatory. That is, many multivariate methods tend to motivate hypothese rather than test them. Consider a situation in which a researcher may hav 50 variables measured on more than 2000 experimental units. Traditiona statistical methods usually require that a researcher state some hypothese collect some data, and then use these data to either substantiate or repudiat the hypotheses. An alternative situation that often exists is a case in which researcher has a large amount of data available and wonders whether ther might be valuable information in these data. Multivariate techniques are ofte useful for exploring data in an attempt to learn if there is worthwhile an valuable information contained in these data.

Variable- and Individual-Directed Techniques

One fundamental distinction between multivariate methods is that some ar classified as "variable-directed" techniques, while others are classified as "individual-directed" techniques.

Variable-directed techniques are those that are primarily concerned wit relationships that might exist among the response variables being measure Some examples of this type of technique are analyses performed on correlatio matrices, principal components analysis, factor analysis, regression analysi and canonical correlation analysis.

Individual-directed techniques are those that are primarily concerned wit relationships that might exist among the experimental units and/or individua being measured. Some examples of this type of technique are discrimina analysis, cluster analysis, and multivariate analysis of variance (MANOVA

Creating New Variables

We quite often find it useful to create new variables for each experiment unit so they can be compared to each other more easily. Many multivariat methods help researchers create new variables that have desirable propertie

Some of the multivariate techniques that create new variables are princip components analysis, factor analysis, canonical correlation analysis, canonic discriminant analysis, and canonical variates analysis.

Some brief overviews of the multivariate techniques that are considere in this book are given next.