

Ira H. Bernstein

Applied Multivariate Analysis

With Calvin P. Garbin
and Gary K. Teng



Springer-Verlag

Ira H. Bernstein

Applied Multivariate Analysis

With Calvin P. Garbin
and Gary K. Teng

With 37 Illustrations



Springer-Verlag
New York Berlin Heidelberg
London Paris Tokyo

Ira H. Bernstein
Department of Psychology
University of Texas at Arlington
Arlington, Texas 76019
USA

Gary K. Teng
Technical Evaluation and Management
Systems, Inc. (TEAMS[®])
Dallas, Texas 75240
USA

Calvin P. Garbin
Department of Psychology
University of Nebraska at Lincoln
Lincoln, Nebraska 68588
USA

Library of Congress Cataloging-in-Publication Data
Bernstein, Ira H.

Applied multivariate analysis.

Bibliography: p.

Includes index.

1. Multivariate analysis. I. Garbin, Calvin P.

II. Teng, Gary K. III. Title.

QA278.B457 1987 519.5'35 87-15174

© 1988 by Springer-Verlag New York Inc.

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer-Verlag, 175 Fifth Avenue, New York, New York 10010, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden. The use of general descriptive names, trade names, trademarks, etc. in this publication, even if the former are not especially identified, is not to be taken as a sign that such names, as understood by the Trade Marks and Merchandise Marks Act, may accordingly be used freely by anyone.

Typeset by Asco Trade Typesetting Ltd., Hong Kong.

Printed and bound by R.R. Donnelley & Sons, Harrisonburg, Virginia.

Printed in the United States of America.

9 8 7 6 5 4 3 2 1

ISBN 0-387-96542-4 Springer-Verlag New York Berlin Heidelberg

ISBN 3-540-96542-4 Springer-Verlag Berlin Heidelberg New York

To Linda, Cari, Dina
and
the memory of Jum C. Nunnally

Preface

Like most academic authors, my views are a joint product of my teaching and my research. Needless to say, my views reflect the biases that I have acquired. One way to articulate the rationale (and limitations) of my biases is through the preface of a truly great text of a previous era, Cooley and Lohnes (1971, p. v). They draw a distinction between mathematical statisticians whose intellect gave birth to the field of multivariate analysis, such as Hotelling, Bartlett, and Wilks, and those who chose to “concentrate much of their attention on methods of analyzing data in the sciences and of interpreting the results of statistical analysis . . . (and) . . . who are more interested in the sciences than in mathematics, among other characteristics.”

I find the distinction between individuals who are temperamentally “mathematicians” (whom philosophy students might call “Platonists”) and “scientists” (“Aristotelians”) useful as long as it is not pushed to the point where one assumes “mathematicians” completely disdain data and “scientists” are never interested in contributing to the mathematical foundations of their discipline. I certainly feel more comfortable attempting to contribute in the “scientist” rather than the “mathematician” role.

As a consequence, this book is primarily written for individuals concerned with data analysis. However, as noted in Chapter 1, true expertise demands familiarity with both traditions.

One consequence of my bias is that even though I have a great love for data, I have long since learned not to worship a particular data set, i.e., I believe in sampling error and replication. I especially believe in Henry Kaiser’s (1970) aphorism “It don’t make no nevermind” about highly elaborate weighting schemes that more often than not prove less useful than simpler ones such as weighting variables equally. (Kaiser, by the way, shows the limitations of the “rule of thumb” distinction between Mathematician/Platonists and Scientist/Aristotelians; bright people do both as his numerous other contributions such as the varimax rotation and alpha factoring attest.) If you are unsure as to what I mean by “magical equations,” pick up almost any scholarly journal containing path analyses and confirmatory factor analyses. Similarly, it should not be too difficult to locate a book in which the author attempts

to use the force of copyright law to protect the weighting scheme used in an equation.

The major reason that I refer to Cooley and Lohnes (1971) as belonging to a prior generation is that it was written before the ascendancy of the major computer packages. I assume you will be using these packages. At one point in writing this book, I thought of offering a “translation” table to explain the printouts that are provided by the packages. The reason that I did not devote more space to the topic is the rate at which revisions of the packages make such information obsolete.

The one feature that I hope sets this book apart from the many other excellent books in the field is my use of the computer as a device to teach you about data structures. It is all too easy to see only the data analytic features of computer packages, as wonderful as they are. (Consider that in almost no time at all I can enter the coding to do all variants on a particular analysis that would have been prohibitive in terms of time when I was a student.) Much less apparent to the student is the way that a computer can *generate* data to conform to a model or deviate in specified ways. Scholars have long used Monte Carlo and related simulation approaches; I feel that it is sufficiently easy for students at this level, or even at a more introductory level, to perform simulations so that they should be part of all students’ training.

The material in this book is designed to be covered in a standard one semester graduate course. It is assumed that a student has had a conventional first semester graduate class in statistics. I have, however, devoted a chapter to material that is largely a summary of this core material (Chapter 2). Of course, I have also tried to include useful information that would make the book valuable beyond the formal confines of the course (*lagniappe*, as one would say in the language of those to my Southeast).

Acknowledgments

I would like to thank Springer-Verlag for their faith in this project. Robert Widner’s comments on various points of the text were most highly appreciated. Various ideas were outgrowths of conversations I had with various departmental colleagues, especially Jim Nairne, Jim Erickson, Jim Baerwaldt, Bill Ickes, Paul Paulus, Duane Martin, and the ophthalmologists at the Glaucoma Associates of Texas. I also am grateful to Melvin Pierce, Tom Kennedy, Ed Homko, Michael Griffin, and others at the University of Texas Academic Computer center, Rory Gresham of Seabrook Computers, and Sara Kapka of IBM for their technical assistance. The oblique, but nonetheless vital contributions of Emma Peel, Ferdinand LaMenthe, John B. Gillespie, John Coltrane, Charles C. Parker, Jr., Larson E. Whipsnade, Jonathan Steed, John Sheridan, and James Cullum and associates are also noted. Finally, the wisdom of a comment by my long-term friend and colleague, Professor Stanley

Coren of the University of British Columbia that “you never finish writing a book, they just take it away from you” is hereby duly noted.

Arlington, Texas
May 1987

IRA H. BERNSTEIN

Contents

Preface	vii
CHAPTER 1 INTRODUCTION AND PREVIEW	1
Chapter Overview.	1
Multivariate Analysis: A Broad Definition	2
Multivariate Analysis: A Narrow Definition	6
Some Important Themes	8
Obtaining Meaningful Relations	8
Selecting Cutoffs	9
Questions of Statistical Inference.	9
Outliers.	10
The Importance of Theory	11
Problems Peculiar to the Analysis of Scales.	13
The Role of Computers in Multivariate Analysis.	14
Multivariate Analysis and the Personal Computer	15
Choosing a Computer Package.	17
Problems in the Use of Computer Packages	19
The Importance of Matrix Procedures	20
CHAPTER 2 SOME BASIC STATISTICAL CONCEPTS.....	22
Chapter Overview.	22
Univariate Data Analysis	23
Frequency Distributions	23
Normal Distributions	23
Standard Normal Distributions.	23
Parameters and Statistics	26
Locational Parameters and Statistics.	26
Measures of Variability.	26
A Note on Estimation.	27
Binary Data and the Binomial Distribution	29
Data Transformation	30
Bivariate Data Analysis	31
Characteristics of Bivariate Relationships	31
Bivariate Normality	33

Measures of Bivariate Relation	38
Range Restriction	39
Pearson Correlation Formulas in Special Cases	40
Non-Pearson Estimates of Pearson Correlations.	40
The Eta-Square Measure.	41
Phi Coefficients with Unequal Probabilities	42
Sampling Error of a Correlation	42
The Z' Transformation	42
Linear Regression	43
The Geometry of Regression	43
Raw-Score Formulas for the Slope	45
Raw-Score Formulas for the Intercept.	45
Residuals	45
The Standard Error of Estimate	46
Why the Term "Regression"?	48
A Summary of Some Basic Relations.	49
Statistical Control: A First Look at Multivariate Relations.	49
Partial and Part Correlation	49
Statistical versus Experimental Control.	53
Multiple Partialling.	53
Within-Group, Between-Group, and Total Correlations	54
 CHAPTER 3 SOME MATRIX CONCEPTS	 57
Chapter Overview.	57
Basic Definitions.	59
Square Matrices	61
Transposition	64
Matrix Equality	64
Basic Matrix Operations.	65
Matrix Addition and Subtraction	65
Matrix Multiplication	66
Correlation Matrices and Matrix Multiplication	68
Partitioned Matrices and Their Multiplication	69
Some Rules and Theorems Involved in Matrix Algebra.	69
Products of Symmetric Matrices	70
More about Vector Products	71
Exponentiation.	71
Determinants	72
Matrix Singularity and Linear Dependency.	73
Matrix Rank.	74
Matrix "Division"	75
The Inverse of a 2×2 Matrix	75
Inverses of Higher-Order Matrices	75
Recalculation of an Inverse Following Deletion of Variable(s).	76
An Application of Matrix Algebra	77
More about Linear Combinations	79
The Mean of a Linear Combination	80
The Variance of a Linear Combination	80
Covariances between Linear Combination	81

The Correlation between Two Different Linear Combinations	82
Correlations between Linear Combinations and Matrix Notation	82
The “It Don’t Make No Nevermind” Principle	83
Eigenvalues and Eigenvectors	85
A Simple Eigenanalysis	86
Eigenanalysis of Gramian Matrices	88

CHAPTER 4 MULTIPLE REGRESSION AND CORRELATION—PART 1. BASIC CONCEPTS 89

Chapter Overview	89
Assumptions Underlying Multiple Regression	91
The Multivariate Normal Distribution	93
A Bit of the Geometry of Multiple Regression	93
Basic Goals of Regression Analysis	94
The Case of Two Predictors	96
A Visual Example	96
A Note on Suppressor Variables	98
Computational Formulas	99
Raw-Score Formulas	100
Other Equations for R^2	100
Determining the Relative Importance of the Two Predictors	101
Bias in Multiple Correlation	102
The Case of More Than Two Predictors	103
Checking for Multicollinearity	104
Another Way to Obtain R^2	104
Residuals	104
Inferential Tests	106
Testing R	106
Testing Beta Weights	107
Testing the Uniqueness of Predictors	108
Evaluating Alternative Equations	108
Cross Validation	109
Computing a Correlation from <i>a priori</i> Weights	110
Testing the Difference between R^2 and r^2 Derived from <i>a priori</i> Weights	110
Hierarchical Inclusion of Predictors	111
Stepwise Inclusion of Predictors	111
Other Ways to Handle Multicollinearity	113
Comparing Alternative Equations	113
Example 1—Perfect Prediction	114
Example 2—Imperfect Prediction plus a Look at Residuals	116
Example 3—Real Personality Assessment Data	117
Alternative Approaches to Data Aggregation	118

CHAPTER 5 MULTIPLE REGRESSION AND CORRELATION—PART 2. ADVANCED APPLICATIONS 121

Chapter Overview	121
Nonquantitative Variables	123

Dummy Coding	123
Effect Coding	124
Orthogonal Coding	126
The Simple Analysis of Variance (ANOVA)	127
Fixed Effects versus Random Effects	127
The Simple ANOVA as a Formal Model	128
Results of Regression ANOVAs	131
Multiple Comparisons	133
Orthogonal versus Nonorthogonal Contrasts	133
Planned versus Unplanned Comparisons	134
Individual Alpha Levels versus Groupwise Alpha Levels	135
Evaluation of Quantitative Relations	135
Method I	136
Method II	138
The Two-Way ANOVA	138
Equal- N Analysis	142
Unequal- N Analysis	143
Fitting Parallel Lines	145
Simple Effect Models	148
The Analysis of Covariance (ANCOVA)	150
Effects of the ANCOVA on the Treatment Sum of Squares	151
Using Dummy Codes to Plot Group Means	152
Repeated Measures, Blocked and Matched Designs	153
Higher-Order Designs	156
 CHAPTER 6 EXPLORATORY FACTOR ANALYSIS.....	157
Chapter Overview	157
The Basic Factor Analytic Model	161
The Factor Equation	161
The Raw-Score Matrix	161
Factor Scores and the Factor Score Matrix	161
Pattern Elements and the Pattern Matrix	161
Error Scores and Error Loadings	162
The Covariance Equation	162
The First Form of Factor Indeterminacy	163
An Important Special Case	164
Common Uses of Factor Analysis	164
Orthogonalization	165
Reduction in the Number of Variables	166
Dimensional Analysis	166
Determination of Factor Scores	166
An Overview of the Exploratory Factoring Process	168
Principal Components	168
The Eigenvectors and Eigenvalues of a Gramian Matrix	170
A Note on the Orthogonality of PCs	172
How an Eigenanalysis Is Performed	173
Determining How Many Components to Retain	174
Other Initial Component Solutions	175

Factor Definition and Rotation	176
Factor Definition	177
Simple Structure	178
PC versus Simple Structure	179
Graphic Representation	180
Analytic Orthogonal Rotation	182
Oblique Rotations	184
Reference Vectors	185
Analytic Oblique Rotation	187
The Common Factor Model	187
An Example of the Common Factor Model	190
A Second Form of Factor Indeterminacy	192
Factor Scores	193
"Exact" Procedures	193
Estimation Procedures	194
Approximation Procedures	195
Addendum: Constructing Correlation Matrices with a Desired Factor Structure	195
 CHAPTER 7 CONFIRMATORY FACTOR ANALYSIS	198
Chapter Overview	198
Comparing Factor Structures	202
Similarity of Individual Factors versus Similarity of the Overall Solution	202
Case I—Comparing Alternate Solutions Derived from the Same Data . . .	203
Case II—Comparing Solutions Obtained from the Same Subjects but on Different Variables	205
Case III—Comparing Solutions with the Same Variables but on Different Individuals; Matrix Information Available	206
Case IV—Comparing Solutions with the Same Variables but Different Individuals; Matrix Information Unavailable	207
Case V—Factor Matching	207
Oblique Multiple Groups Tests of Weak Structure	209
Basic Approach	209
Evaluating the Substantive Model	210
Alternative Models	212
Performing the Actual OMG Analysis	214
Computational Steps	214
OMG Common Factor Solutions	216
A Numerical Example	216
LISREL Tests of Weak Substantive Models	219
Specification of Weak Models	219
Estimation	221
Identification	222
Assessment of Fit	222
Numerical Examples	223
LISREL Tests of Strong Substantive Models	225
Causal Models and Path Analysis	227

Fundamental Path Analytic Concepts	228
Example I: One Exogenous and Two Endogenous Variables	231
Example II: Two Exogenous and One Endogenous Variables	233
Example III: One Exogenous and Two Endogenous Variables	233
Decomposing Observed Correlations	234
Causal Models and LISREL	236
A Formal Separation of Exogenous and Endogenous Variables	236
The Confirmatory Factor Model for Exogenous Variables	237
The Confirmatory Factor Model for Endogenous Variables	237
The Structural Model	238
Addendum: A Program to Obtain Oblique Multiple Groups Solutions.	239
 CHAPTER 8 CLASSIFICATION METHODS—PART 1. FORMING DISCRIMINANT AXES	 246
Chapter Overview	246
Discriminant Analysis with Two Groups and Two Predictors	247
Graphic Representation	248
One Way to Obtain Discriminant Weights	250
Another Way to Obtain Discriminant Weights	251
Computation	252
A Third Way to Obtain Discriminant Weights	256
Eigenanalysis of $\mathbf{W}^{-1}\mathbf{B}$ versus Eigenanalysis of $\mathbf{S}\mathbf{P}_w^{-1}\mathbf{S}\mathbf{P}_b$	257
Strength of Relation	257
Discriminant Scores, Discriminant Means, and the Mahalanobis Distance Measure	258
The Discriminant Structure	259
Interpreting a Discriminant Axis	259
Discriminant Analysis with Two Predictors and Three Groups	260
Extracting Multiple Discriminant Axes	262
Inferring Concentration from the Eigenanalysis	263
Numerical Examples	263
Depiction of Group Differences in the Diffuse Example	269
Varimax Rotation	271
Use of Prior Weights	271
Heteroscedasticity	272
Discriminant Analysis—The General Case	274
Components as Discriminant Variables	274
Hypothesis Testing	274
Confirmatory Factor Analysis	275
Stepwise Discriminant Analysis	275
 CHAPTER 9 CLASSIFICATION METHODS—PART 2. METHODS OF ASSIGNMENT	 276
Chapter Overview	276
The Equal Variance Gaussian Model	279
Decision Rules and Cutoff Rules	280
Other Properties	282

Why a Cutoff Rule?	283
Bayesian Considerations	283
Why Bayesian Considerations Are Important	284
Varying Base Rates	285
Bayes' Theorem in Binary Classification	286
Receiver (Relative) Operating Characteristic (ROC) Curves	286
Describing Accuracy of Classification (Sensitivity)	288
The Unequal Variance Gaussian Model	291
Other Signal Detection Models	293
Strategies for Individual Classification	295
Describing Performance	295
Bayesian Considerations with Multiple Groups	296
Alternative Strategies—An Overview	297
A Numerical Example	298
Calibration Study	298
The Context of Classification	300
Classification Based on Salient Variables	303
Homoscedasticity Assumed	303
Normality Assumed	303
Dealing with Heteroscedasticity	304
Discriminant Functions and Classification	305
Homoscedasticity Assumed	306
Normality Assumed	306
Dealing with Heteroscedasticity	307
Using Multiple Discriminant Scores	307
Normality and Multiple Discriminant Functions	308
Classification Based on Distance Measures	309
Simple Distance	309
Fisher's Classification Functions	310
Mahalanobis Distances	311
A Summary of Strategic Considerations in Classification	312

CHAPTER 10 CLASSIFICATION METHODS—PART 3. INFERENTIAL CONSIDERATIONS IN THE MANOVA

315

Chapter Overview	315
The Two-Group MANOVA and Hotelling's T^2	316
The Formal MANOVA Model	319
Hotelling's T^2	320
Post Hoc Comparisons	323
Application of Hotelling's T^2 to a Single Group	323
Testing for Equality of Means	324
Single Group MANOVA versus Repeated Measures ANOVA	326
Tests of Vector Means with Multiple Groups	326
The Fundamental Problem	327
Testing a Concentrated Structure	328
Testing a Diffuse Structure	328
Testing for Homogeneity of Covariance	331

The Simple MANOVA with Multiple Groups	332
SP and Variance–Covariance Matrices	333
Applying Box’s <i>M</i> Test	334
Specifying the Nature of the Group Differences	335
The Multivariate MANOVA	336
Terminology and Basic Logic	336
A Numerical Example	338
The <i>A</i> Effect	339
The <i>B</i> Effect	340
The <i>AB</i> Interaction	340
The MANCOVA	341
Numerical Example	342
 CHAPTER 11 PROFILE AND CANONICAL ANALYSIS	 345
Chapter Overview	345
Profile Similarity	346
Similarity of Scalars	346
Similarity of Vectors	347
Measuring Profile Elevation	349
Profile Similarity Based on Elevation Alone	351
Numerical Example	351
Profile Similarity Based on Shape Information	353
A Numerical Example	354
Correcting for Elevation	355
Correlations, Covariances, and Cross Products as Alternative Indices of Similarity	355
Simple and Hierarchical Clustering	357
Numerical Example	360
Canonical Analysis	363
Goals of Canonical Analysis	363
Basic Logic	363
Redundancy Analysis	365
Basic Matrix Operations	366
Statistical Inference	369
Numerical Example	370
Alternatives to Canonical Analysis	375
 CHAPTER 12 ANALYSIS OF SCALES	 376
Chapter Overview	376
Properties of Individual Items	380
Item Formats and Scoring	381
Correction for Guessing	381
Item Distributions	382
Relation of Items to the Scale as a Whole	383
Numerical Example	387
Test Reliability	388
The Logic of Internal Consistency Measures	389

1

Introduction and Preview

Chapter Overview

Chapter 1 contains six major topics:

I. **MULTIVARIATE ANALYSIS: A BROAD DEFINITION**—I begin by offering a very broad definition of multivariate analysis. The definition is: *an inquiry into the structure of interrelationships among multiple measures*.

II. **MULTIVARIATE ANALYSIS: A NARROW DEFINITION**—The broad definition is important because virtually all behavioral research deals with questions about structure. Unfortunately, it would subsume all of statistical analysis, making for quite a large textbook! Consequently, it is customary to define multivariate analysis in more limited terms: *the study of linear representations of relations among variables*. Contained within this narrow definition are several interrelated models that are discussed in later chapters: (1) *multiple regression* (Chapters 4 and 5), (2) *factor analysis* (Chapters 6 and 7), (3) *discriminant analysis, related classification techniques, and the multivariate analysis of variance and covariance* (Chapters 8–10), (4) *profile and canonical analysis* (Chapter 11), and (5) *analysis of scales* (Chapter 12).

III. **SOME IMPORTANT THEMES**—Models like factor analysis and multiple regression used to infer structure are all too often looked upon as if they were separate, unrelated procedures. In reality, certain themes recur regardless of the specific analytic procedure: (1) *equations that are the by-product of a particular analysis should be meaningful*; (2) *cutoffs need to be established defining “how high is high” and “how low is low” when one develops a prediction equation*; (3) *questions of statistical significance often arise*; (4) *outliers can render otherwise elegant analyses meaningless or, worse, highly misleading*; (5) *multivariate investigation should be guided by theory*; and (6) *particular problems are likely to arise when relations among individual items, as opposed to multi-item scales, are the unit of analysis*.

IV. **THE ROLE OF COMPUTERS IN MULTIVARIATE ANALYSIS**—A generation ago, many analyses that were conceived of were not practicable. The availability of computers and sophisticated computer pack-

ages now make such analyses routine. The increasing role of the personal computer is also noted.

V. CHOOSING A COMPUTER PACKAGE—There are several important computer packages now on the market. Some of the pros and cons of the three most popular ones—SAS, SPSSX, and BMDP—are discussed.

VI. PROBLEMS IN THE USE OF COMPUTER PACKAGES—Knowing how to use a computer package is important to multivariate analysis, but it is not sufficient. Four necessary ingredients to a successful analysis are considered: (1) *substantive knowledge of the research topic*; (2) *computer knowledge, including knowledge of the package*; (3) *empirical experience with various kinds of data*; and (4) *formal knowledge of the analytic procedures*.

This textbook is written for those who need to analyze complex behavioral data, whether in field settings such as clinical and social psychology, in applied settings like nursing and marketing research, or in experimental settings like learning and perception. Although I cannot escape some degree of formality, I am writing for students whose primary interests lie more in empirical phenomena as opposed to rigorous mathematical statistics.

Although I very much appreciate why examples related to your specific interests are important, I hope I can “wean” you from examples limited to your own content area as the textbook topics progress. One of the things you should note is that if you are a clinical psychology student working with personality test data, you will have problems that are abstractly *identical* to a market researcher working on a consumer survey. One of the reasons that I find quantitative applications so rewarding is that I can work in a variety of areas, as long as I have someone to provide me with a background of the underlying empirical issue or have that background myself. At any one time, I can and have been working on the issue of selecting police officers, studying patients attitudes toward health care, looking at choices of financial institutions, and evaluating different forms of eye surgery. What I learned in one setting applied to all. I hope you can share the enjoyment that I have been experiencing.

One of the decisions that any author of a quantitative textbook needs to confront is how to choose examples. Conceptually, it should not matter if examples come from clinical psychology, marketing, or visual psychophysics; yet it does matter to most students. What I will do, therefore, is to pick my examples to illustrate the relevance of the various multivariate models in different settings.

Multivariate Analysis: A Broad Definition

As with most topics, it is useful to try to define what is meant by the term “multivariate analysis” before proceeding too far into its details. In the broadest and most literal sense, it means *an inquiry into the structure of*