Multivariate Analysis for the Biobehavioral and Social Sciences

A Graphical Approach

Bruce L. Brown, Suzanne B. Hendrix, Dawson W. Hedges, Timothy B. Smith

WILEY

WWW.
LINK AVAILABLE

Table of Contents

MULTIVARIATE ANALYSIS FOR THE BIOBEHAVIORAL AND SOCIAL SCIENCES

A Graphical Approach

BRUCE L. BROWN
SUZANNE B. HENDRIX
DAWSON W. HEDGES
TIMOTHY B. SMITH

WILEY

A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.

Published simultaneously in Canada.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care

Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data:

Brown, Bruce (Bruce L.)

Multivariate analysis for the biobehavioral and social sciences / Bruce L. Brown, Suzanne B. Hendrix, Dawson W. Hedges, Timothy B. Smith.

p. cm.

Includes bibliographical references and index.

ISBN 978-0-470-53756-5

1. Social sciences–Statistical methods. 2. Multivariate analysis. I. Title.

HA29.B82857 2012

300.1'519535–dc23

2011018420

ISBN 978-1-118-13159-6 (ePDF)

ISBN 978-1-118-13161-9 (ePub)

ISBN 978-1-118-13160-2 (Mobi)

PREFACE

The plan of this book is unusual. It is unusual in that we have elevated graphics to the status of an equal partner in the data analysis process. Our intent is to demonstrate the centrality of good graphics to the scientific process, to provide a graphical concomitant for each of the classical multivariate statistical methods presented, and to demonstrate the superiority of graphical expressions in clarifying and laying bare the meaning in the data.

The plan of the book is also unusual in the pedagogical approach taken. The first three chapters are all preparatory: giving an overview of multivariate methods (Chapter 1), reviewing the fundamental principles of elementary statistics as “habits” that are necessary preparation for understanding multivariate methods (Chapter 2), and then introducing matrix algebra (Chapter 3).

The most unusual aspect of the book, however, is the six methods chapters (4, 5, 6, 7, 8, and 9)—the core of the book. We introduce each with a published paper that is in some way exemplary as a “research-publication case study.” We first showcase the method in a strong piece of published work to demonstrate to the student the practical value of that method. The next section answers the question “How do you do that?” It is our intent to answer that question fully with a complete but simplified demonstration of the mathematics and the concepts of the method. This is a unique feature of the book. There are books that are fully enabling mathematically, and there are also books that are highly accessible to the beginning student, but this book is unique in combining these two characteristics by the use of simplest case demonstrations.

The next step in each chapter is to demonstrate how the analysis of the data is accomplished using one of the commonly used statistical packages, such as Stata®, SAS®, or SPSS®. One of the major tasks in demonstrating statistical packages is that of instructing the student in the reading of output. The simplest case demonstration of the full computational process is an effective way to deal with that aspect. After carrying out the full analysis with simplest case data, and then using Stata, SAS, or SPSS to analyze the same simple set of data, the meaning of each of the parts in the computer output becomes obvious and clear.

Although this book covers many of the commonly used multivariate methods and builds upon them graphically, this approach is also applicable to a wide variety of additional methods, both modern regression methods and also extensions of multivariate methods. Factor analysis has grown into structural equations modeling, MANOVA and ANOVA have grown into multilevel, hierarchical, and mixed models, and general linear models have grown into generalized linear models that can deal with a broad variety of data types, including categorical. All of these can be supplemented and improved upon with graphics.

The focus of this book on the application of graphics to the classical methods is, we believe, the appropriate beginning given the relative simplicity of the fundamental multivariate methods. The principles the student learns here can then more easily be expanded in future texts to the full power of advanced derivative methods. It is curious that the rather simple multiple regression model is the foundation of many if not most of the higher-level developments. We have closed the book with a simple presentation of multiple regression in Chapter 9, both as a look backward and also as a look forward. It is a basic example of the application of matrix methods to multiple variables, but also as a prelude to the higher-level methods.

We are grateful to those exemplary researchers and quantitative methods scholars whose work we have built upon. We are grateful to our students from whom we have learned, many of whom appear in this book. Most of all, we are grateful to our families who have been supportive and patient through this process.

BRUCE L. BROWN

SUZANNE HENDRIX

DAWSON W. HEDGES

TIMOTHY B. SMITH

CHAPTER ONE

OVERVIEW OF MULTIVARIATE AND REGRESSION METHODS

1.1 INTRODUCTION

More information about human functioning has accrued in the past five decades than in the preceding five millennia, and many of those recent gains can be attributed to the application of multivariate and regression statistics. The scientific experimentation that proliferated during the 19th century was a remarkable advance over previous centuries, but the advent of the computer in the mid-20th century opened the way for the widespread use of complex analytic methods that exponentially increased the pace of discovery. Multivariate and regression methods of data analysis have completely transformed the bio-behavioral and social sciences.

Multivariate and regression statistics provide several essential tools for scientific inquiry. They allow for detailed descriptions of data, and they identify patterns impossible to discern otherwise. They allow for empirical testing of complex theoretical propositions. They enable enhanced prediction of events, from disease onset to likelihood of remission. Stated simply, multivariate statistics can be applied to a broad variety of research questions about the human condition.

Given the widespread application and utility of multivariate and regression methods, this book covers many of the statistical methods commonly used in a broad range of bio-behavioral and social sciences, such as psychology, business, biology, medicine, education, and sociology. In these disciplines, mathematics is not typically a student's primary focus. Thus, the approach of the book is *conceptual*. This does not mean that the mathematical account of the methods is compromised, just that the mathematical developments are employed in the service of the conceptual basis for each method. The math is presented in an accessible form, called *simplest case*. The idea is that we seek a demonstration for each method that uses the *simplest case* we can find that has all the key attributes of the full-blown cases of actual practice. We provide exercises that will enable students to learn the simplified case thoroughly, after which the focus is expanded to more realistic cases.

We have learned that it is possible to make these complex mathematical concepts accessible and enjoyable, even to those who may see themselves as nonmathematical. It is possible with this *simplest-case* approach to teach the underlying conceptual basis so thoroughly that some students can perform many multivariate and regression analyses on simple "student-accommodating" data sets from memory, without referring to written formulas. This kind of deep conceptual acquaintance brings the method up close for the student, so that the meaning of the analytical results becomes clearer.

This first chapter defines *multivariate data analysis methods* and introduces the fundamental concepts. It also outlines and explains the structure of the remaining chapters in the book. All analysis method chapters follow a common format. The main body of each chapter starts with an example of the method, usually from an article in a prominent journal. It then explains the rationale for each method and gives complete but simplified numerical demonstrations of the various expressions of each method using *simplest-case data*. At the end of each chapter is the section entitled *Study Questions*, which consists of three types: *essay questions, calculation questions*, and *data-analysis questions.* There is a complete set of answers to all of these questions available electronically on the website at https://mvgraphics.byu.edu.

1.2 MULTIVARIATE METHODS AS AN EXTENSION OF FAMILIAR UNIVARIATE METHODS

The term *multivariate* denotes the analysis of multiple dependent variables. If the data set has only one dependent variable, it is called *univariate*. In elementary statistics, you were probably introduced to the two-way analysis of variance (ANOVA) and learned that any ANOVA that is two-way or higher is referred to as a *factorial* model. *Factorial* in this instance means having multiple *independent* variables or factors. The advantage of a factorial ANOVA is that it enables one to examine the interaction between the independent variables in the effects they exert upon the dependent variable.

Multivariate models have a similar advantage, but applied to the multiple dependent variables rather than independent variables. Multivariate methods enable one to deal with the *covariance* among the dependent variables in a way that is analogous to the way factorial ANOVA enables one to deal with interaction.

Fortunately, many of the multivariate methods are straightforward extensions of the corresponding univariate methods (Table 1.1). This means that your considerable investment up to this point in understanding univariate statistics will go a long way toward helping you to understand multivariate statistics. (This is particularly true of Chapters 7, 8, and 9, where the *t*-tests are extended to multivariate *t*-tests, and various ANOVA models are extended to corresponding multiple ANOVA [MANOVA] models.) Indeed, one can think of multivariate statistics in a simplified way as just the same univariate methods that you already know (*t*-test, ANOVA, correlation/regression, etc.) rewritten in *matrix algebra* with the matrices extended to include multiple dependent variables.

Table 1.1 Overview of Univariate and Multivariate Statistical Methods

Description and Number of Predictor (Independent) Variables	**Univariate Method**	**Multivariate Method**
	One quantitative outcome (dependent) variable	**Multiple quantitative outcome (dependent) variables**
No predictor variable	—	Factor analysis
		Principal component analysis
		Cluster analysis
One categorical predictor variable, two levels	*t* tests	Hotelling's T^2 tests
	z tests	Profile analysis using Hotelling's T^2
One categorical predictor, variable, three or more levels	ANOVA, one-way models	MANOVA, one-way models
Two or more categorical predictor variables	ANOVA, factorial models	MANOVA, factorial models
Categorical predictor(s) with one or more quantitative control variables	ANCOVA, one-way or factorial models	MANCOVA, one-way or factorial models
One quantitative predictor variable	Bivariate regression	Multivariate regression
Two or more quantitative predictor variables	Multiple regression	Multivariate multiple regression Canonical correlation*

Matrix algebra is a tool for more efficiently working with data matrices. Many of the formulas you learned in elementary statistics (variance, covariance, correlation coefficients, ANOVA, etc.) can be expressed much more compactly and more efficiently with matrix algebra. Matrix multiplication in particular is closely connected to the calculation of variances and covariances in that it directly produces sums of squares and sums of products of input vectors. It is as if matrix algebra were invented specifically for the calculation of covariance structures. Chapter 3 provides an introduction to the fundamentals of matrix algebra. Readers unfamiliar with matrix algebra should therefore carefully read Chapter 3 prior to the other chapters that follow, since all are based upon it.

The second prerequisite for understanding this book is a *knowledge of elementary statistical methods:* the normal distribution, the binomial distribution, confidence intervals, *t*-tests, ANOVA, correlation coefficients, and regression. It is assumed that you begin this course with a fairly good grasp of basic statistics. Chapter 2 provides a review of the fundamental principles of elementary statistics, expressed in matrix notation where applicable.

1.3 MEASUREMENT SCALES AND DATA TYPES

Choosing an appropriate statistical method requires an accurate categorization of the data to be analyzed. The four kinds of measurement scales identified by S. Smith Stevens (1946) are nominal, ordinal, interval, and ratio. However, there are almost no examples of interval data that are not also ratio, so we often refer to the two collectively as an interval/ratio scale. So, effectively, we have only three kinds of data: those that are categorical (nominal), those that are ordinal (ordered categorical), and those that are fully quantitative (interval/ratio). As we investigate the methods of this book, we will discover that ordinal is not a particularly meaningful category of data for multivariate methods. Therefore, from the standpoint of data, the major distinction will be between those methods that apply to fully quantitative data (interval/ratio), those that apply to categorical data, and those that apply to data sets that have both quantitative and categorical data in them.

Factor analysis (Chapter 4) is an example of a method that has only quantitative variables, as is multiple regression. Log-linear models (Chapter 9) are an example of a method that deals with data that are completely categorical. MANOVA (Chapter 8) is an example of an analysis that requires both quantitative and categorical data; it has categorical independent variables and quantitative dependent variables.

Another important issue with respect to data types is the distinction between discrete and continuous data. Discrete data are whole numbers, such as the number of persons voting for a proposition, or the number voting against it. Continuous data are decimal numbers that have an infinite number of possible points between any two points. In measuring cut lengths of wire, it is possible in principal to identify an infinitude of lengths that lie between any two points, for example, between 23 and 24 inches. The number possible, in practical terms, depends on the accuracy of one's measuring instrument. Measured length is therefore continuous. By extension, variables measured in biomedical and social sciences that have multiple possible values along a continuum, such as oxytocin levels or scores on a measure of personality traits, are treated as continuous data.

All categorical data are by definition discrete. It is not possible for data to be both categorical and also continuous. Quantitative data, on the other hand, can be either continuous or discrete. Most measured quantities, such as height, width, length, and weight, are both continuous and also fully quantitative (interval/ratio). There are also, however, many other examples of data that are fully quantitative and yet discrete. For example, the count of the number of persons in a room is discrete, because it can only be a whole number, but it is also fully quantitative, with interval/ratio properties. If there are 12 persons in one room and twenty-four in another, it makes sense to say that there are twice as many persons in the second room. Counts of number of persons therefore have interval/ratio properties.[1]

When all the variables are measured on the same scale, we refer to them as *commensurate.* When the variables are measured with different scales, they are *noncommensurate.* An example of commensurate data would be width, length, and height of a box, each one measured in inches. An example of noncommensurate would be if the width of the box and its length were measured in inches, but the height was measured in centimeters. (Of course, one could make them commensurate by transforming all to inches or all to centimeters.) Another example of noncommensurate variables would be IQ scores and blood lead levels. Variables that are not commensurate can always be made so by standardizing them (transforming them into *Z*-scores or percentiles). A few multivariate methods, such as profile analysis (associated with Chapter 7 in connection with Hotelling's T^2), or principal component analysis of a covariance matrix (Chapter 4) require that variables be commensurate, but most of the multivariate methods do not require this.

1.4 FOUR BASIC DATA SET STRUCTURES FOR MULTIVARIATE ANALYSIS

Multivariate and regression data analysis methods can be creatively applied to a wide variety of types of data set structures. However, four basic types of data set structures include most of the multivariate and regression data sets that will be encountered. These four basic types of data fit almost all of the statistical methods introduced in this book.

FOUR BASIC TYPES OF DATA SET STRUCTURE

Type 1: Single sample with multiple variables measured on each sampling unit.

Possible methods include factor analysis, principal component analysis, cluster analysis, and confirmatory factor analysis.

Type 2: Single sample with two sets of multiple variables (an *X* set and a *Y* set) measured on each sampling unit.

Possible methods include canonical correlation, multivariate multiple regression, and structural equations modeling.

Type 3: Two samples with multiple variables measured on each sampling unit.

Possible methods include Hotelling's T^2 test, discriminant analysis, and some varieties of classification analysis.

Type 4 More than two samples with multiple variables measured on each sampling unit.

Possible methods include MANOVA, multiple discriminant analysis, and some varieties of classification analysis.

The first type of data set structure is *a single sample with multiple variables measured on each sampling unit*. An example of this kind of data set would be the scores of 300 people on seven psychological tests. Multivariate methods that apply to this kind of data are discussed in Chapter 4 and include *principal component analysis*, *factor analysis*, and *confirmatory factor analysis*. These methods provide answers to the question, "What is the covariance structure of this set of multiple variables?"

The second type of data set structure is *a single sample with two sets of multiple variables (an* X *set and a* Y *set) measured on each unit*. An example of data of this kind would be a linked data set of mental health inpatients' records, with the *X* set of variables consisting of several indicators of physical health (e.g., blood serum levels), and the *Y* set of variables consisting of several indicators of neurological functioning (e.g., results of testing). Multivariate methods that can be applied to this kind of data include *canonical correlation* (Chapter 6) and *multivariate multiple regression* (Chapter 9). These methods provide answers to the question, "What are the linear combinations of variables in the *X* set and in the *Y* set that are maximally predictive of the other set?" Another method that can be used with a single sample with two sets of multiple variables would be SEM, *structural equations modeling*. However, SEM can also be applied when there are *more than two* sets of multiple variables. In fact, it can handle any number of sets of multiple variables. It is the general case of which these other methods are special cases, and as such it has a great deal of potential analytical power.

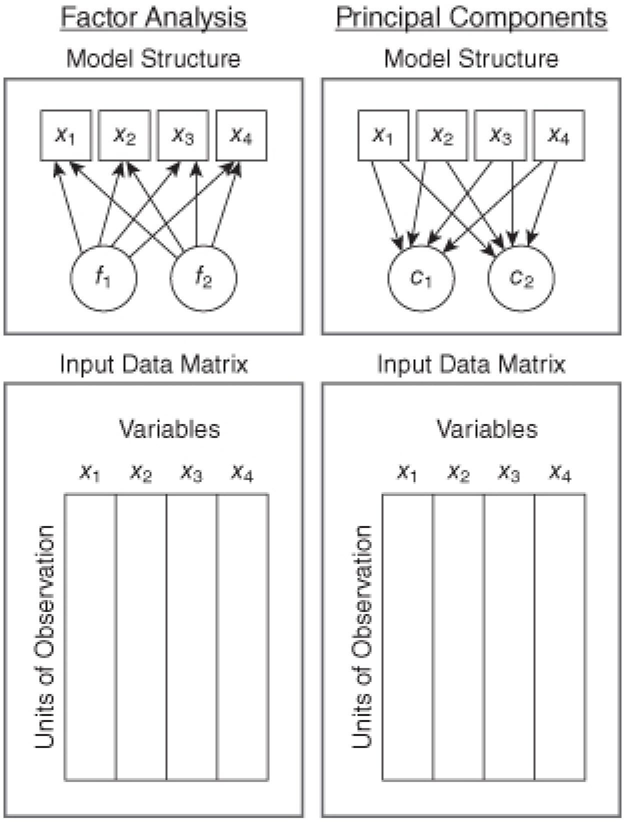
The third type of data set structure is *two samples with multiple variables measured on each unit*. An example would be a simple experiment with an experimental group and a control group, and with two or more dependent variables measured on each observation unit. For example, the effects of a certain medication could be assessed by applying it to 12 patients selected at random (the experimental group) and not applying it to the other 12 patients (the control group), using multiple dependent variable measurements (such as scores on several tests of patient functioning). Multivariate methods that can be applied to this kind of data are *Hotelling's* T^2 *test* (Chapter 7), *profile analysis*, *discriminant analysis* (Chapter 7), and some varieties of *classification analysis*. The Hotelling's T^2

test is the multivariate analogue of the ordinary *t*-test, which applies to two-sample data when there is only one dependent variable. The Hotelling's T^2 test extends the logic of the t test to compare two groups and analyze statistical significance holistically for the combined set of multiple dependent variables. The T^2 test answers the question, "Are the vectors of means for these two samples significantly different from one another?" Discriminant analysis and other classification methods can be used to find the optimal linear combination of the multiple dependent variables to best separate the two groups from one another.

The fourth type of data set structure is similar to the third but extended to three or more samples (with multiple dependent variables measured on each of the units of observation). For example, the same test of the effects of medication on hospitalized patients could be done with two types of medication plus the control group, making three groups to be compared simultaneously and multivariately. The major method here is *MANOVA*, or *multivariate ANOVA* (Chapter 8), which is the multivariate analog of ANOVA. In fact, for every ANOVA model (two-way, three-way, repeated measures, etc.), there exists a corresponding MANOVA model. MANOVA models answer all the same questions that ANOVA models do (significance of main effects and interactions), but holistically within multivariate spaces rather than just for a single dependent variable. *Multiple discriminant analysis* and *classification analysis* methods can also be applied to multivariate data having three or more groups, to provide a spatial representation that optimally separates the groups.

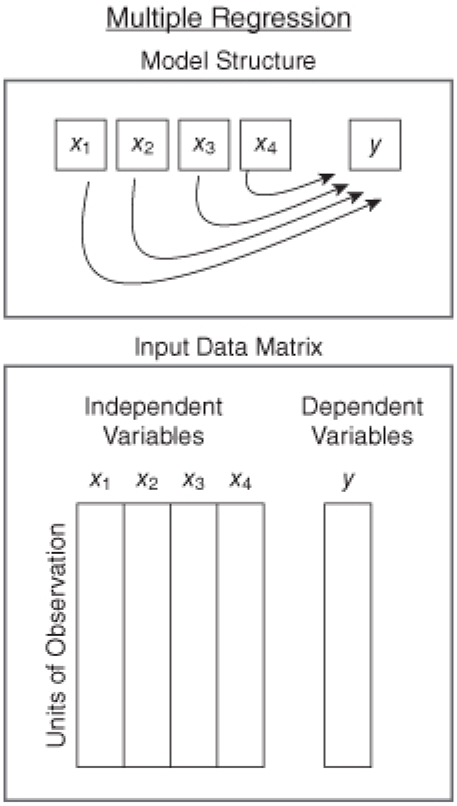
1.5 PICTORIAL OVERVIEW OF MULTIVARIATE METHODS

Diagrammatic representations can help explain and differentiate among the various multivariate statistical methods. Several such methods are described pictorially in this section, starting with factor analysis (Chapter 4), a method that applies to the simplest of the four data set structures just described, *a single sample with multiple variables measured on each sampling unit or unit of observation.* Principal component analysis (Chapter 4) also applies to this simple data set structure. The ways in which these two methods differ will be more fully explained in Chapter 4, but one difference can be seen from the schematic diagram of each method given below. The bottom part of each figure shows the matrix organization of the input data, with rows representing observations and columns representing variables, and the two methods are seen to be identical in this aspect.



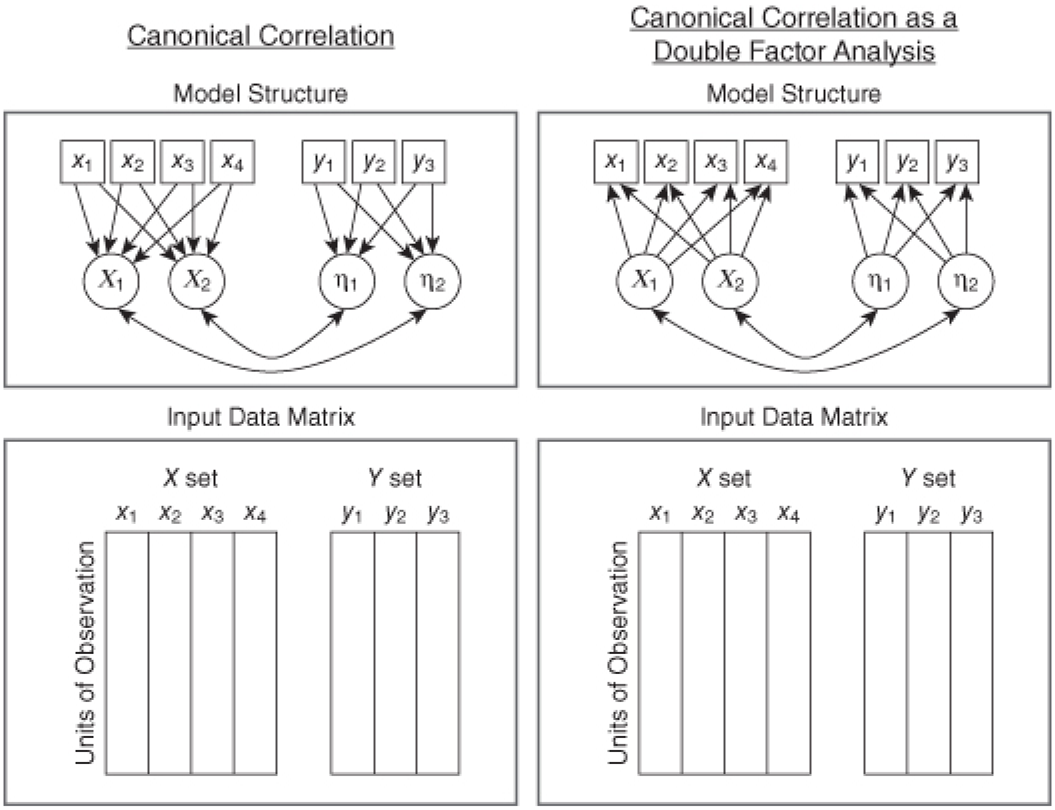
The top part of each figure shows the structure of the model, how the observed variables (x_1 through x_4 for this example) are related to the underlying *latent* variables, which are the factors (f_1 and f_2) for factor analysis, and the components (c_1 and c_2) for principal component analysis. As can be seen by the direction of the arrows, principal components are defined as *linear combinations* (which can be thought of as weighted sums) of the observed variables. However, in factor analysis, the direction is reversed. The observed variables are expressed as linear combinations of the factors. Another difference is that in principal component analysis, we seek to explain a large part of the total variance in the observed variables with the components, but in factor analysis, we seek to account for the covariances or correlations among the variables. (Note that latent variables are represented with circles, and manifest/observed variables are represented with squares, consistent with structural equation modeling notation.)

Multiple regression, also referred to as OLS or “ordinary least-squares regression,” is probably the simplest of the methods presented in this book, but in its many variations, it is also the most ubiquitous. It is the foundation for understanding a number of the other methods, as it is the basis for the general linear model. ANOVA is a special case of multiple regression (multiple regression with categorical dummy variables as the predictor variables, the *X* variables in the diagram below), and when data are unbalanced (unequal cell sizes), multiple regression is by far the most efficient way to analyze the data (as will be demonstrated in Chapter 9). Logistic regression and the generalized linear model (Chapter 9) are adaptations of multiple regression to deal with a wide variety of data types, categorical as well as quantitative. Multilevel linear models, mixed models, and hierarchical linear models are high-level derivatives of regression. The simple data set structure of OLS regression consists of merely several independent variables (also referred to as “predictor variables”) being used to predict one dependent variable (also referred to as the “criterion variable”).



Canonical correlation is similar to multiple regression (and the multiple correlation coefficient on which multiple regression is based), but it deals with two sets of multiple variables rather than one. As such, it fits the second type of data set structure explained above, *a single sample with two sets of multiple variables (an* X *set and a* Y *set) measured on each unit*. Multiple regression gives the correlation coefficient between the best possible linear combination of a group of *X* variables and a single *Y* variable. Canonical correlation, by extension, gives the correlation coefficient between two linear combinations, one on the *X* set of multiple variables and one on the *Y* set of multiple variables. In other words, latent variables are extracted from both the *X* set of variables and the *Y* set of variables to fit the criterion that the correlation between the corresponding latent variables in the *X* set and the *Y* set is maximal. It is like a double multiple regression that is recursive, where the best possible linear combination of *X* variables for predicting *Y* variables is obtained, and also vice versa. This is shown in the diagram on the left below. To return to the example given above for this kind of linked multivariate data set, the canonical correlation of the mental health inpatient data set described would give the best possible linear combination of blood serum levels for predicting neurological functioning, but since it is recursive (bidirectional), it also gives the best possible combination of neurological functioning for predicting blood serum levels.

A slight change in the way the analysis is conceived and the calculations are performed turns canonical correlation into a double factor analysis, as shown in the diagram at the right below. The main difference here is theoretical, in how the latent variables (the linear combinations of observed variables) are interpreted. In the application of canonical correlation as a double factor analysis shown below, the interpretation is that the observed variables are in fact combinatorial expressions of the underlying latent variables, labeled here with the Greek letters chi (χ), for the latent variables for the *X* set, and nu (η) for the latent variables for the *Y* set. The concepts and mathematics for canonical correlation are presented in Chapter 6.



Another method closely related to canonical correlation and multiple regression is multivariate multiple regression, as shown in the diagram. This is essentially the same computational machinery as canonical correlation, except that the latent variables are not recursive. That is, the *X* set is thought of as being predictive of the *Y* set, but not vice versa. This is shown in the diagram by the arrows only going one way. An example of this would be predicting a *Y* set of mutual-fund performance variables from an *X* set of market index variables. The *X* set of variables on the left are combined together into the left-hand latent variables labeled as χ_1 and χ_2. These are the linear combinations of market indices that are most predictive of performance on the entire set of mutual funds as a whole, but this is mediated through the right-hand latent variables η_1 and η_2, which are combined together to predict the performance on each of the mutual funds, the Y variables. This is analogous to the way that simple bivariate correlation is recursive (the