

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

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*

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.

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.

¹ See Chapter 2, Section 2.1, for a review of the properties of a ratio scale and also of the other three types of scales.

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

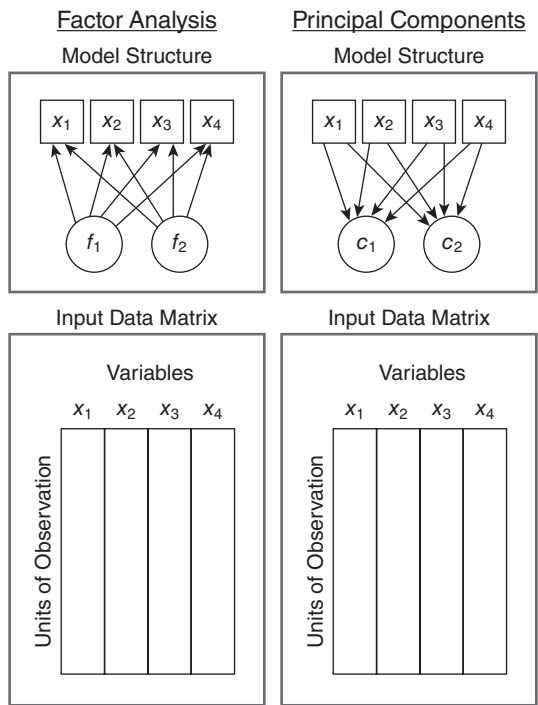
experimental group and a control group, and with two or more dependent variables measured on each observation unit. For example, the effects of a certain medication could be assessed by applying it to 12 patients selected at random (the experimental group) and not applying it to the other 12 patients (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

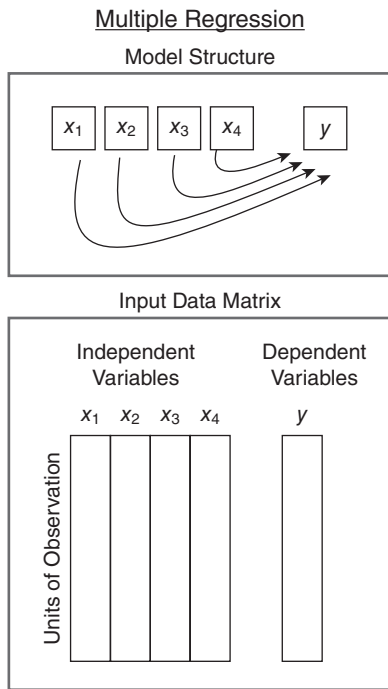
observations and columns representing variables, and the two methods are seen to be identical in this aspect.



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”).



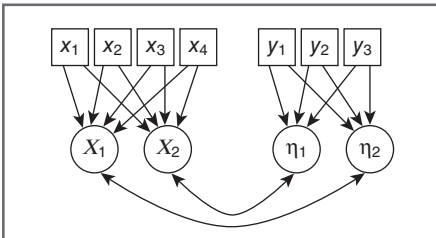
Canonical correlation is similar to multiple regression (and the multiple correlation coefficient on which multiple regression is based), but it deals with two sets of multiple variables rather than one. As such, it fits the second type of data set structure explained above, *a single sample with two 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.

Canonical Correlation

Model Structure

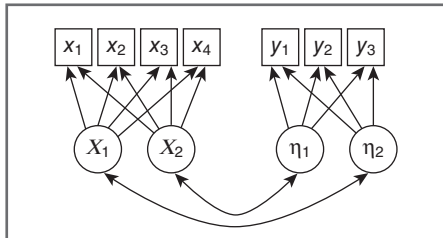


Input Data Matrix

The diagram illustrates a two-way layout. On the left, a vertical label "Units of Observation" is positioned next to a large rectangular area divided into four vertical columns. Above these columns are the labels x_1 , x_2 , x_3 , and x_4 , collectively labeled as the "X set". To the right of this is another large rectangular area divided into three vertical columns, labeled y_1 , y_2 , and y_3 above them, collectively labeled as the "Y set".

Canonical Correlation as a Double Factor Analysis

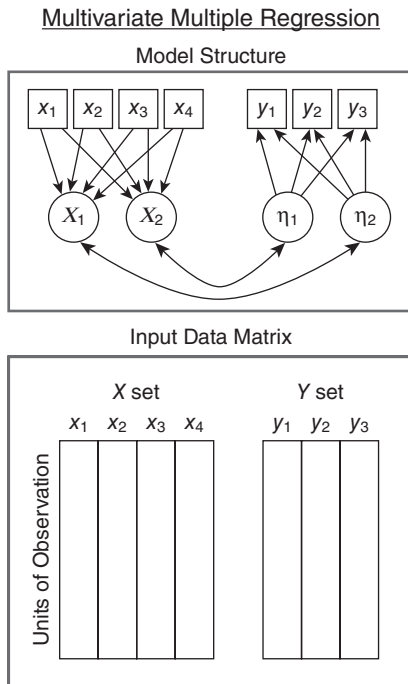
Model Structure



Input Data Matrix

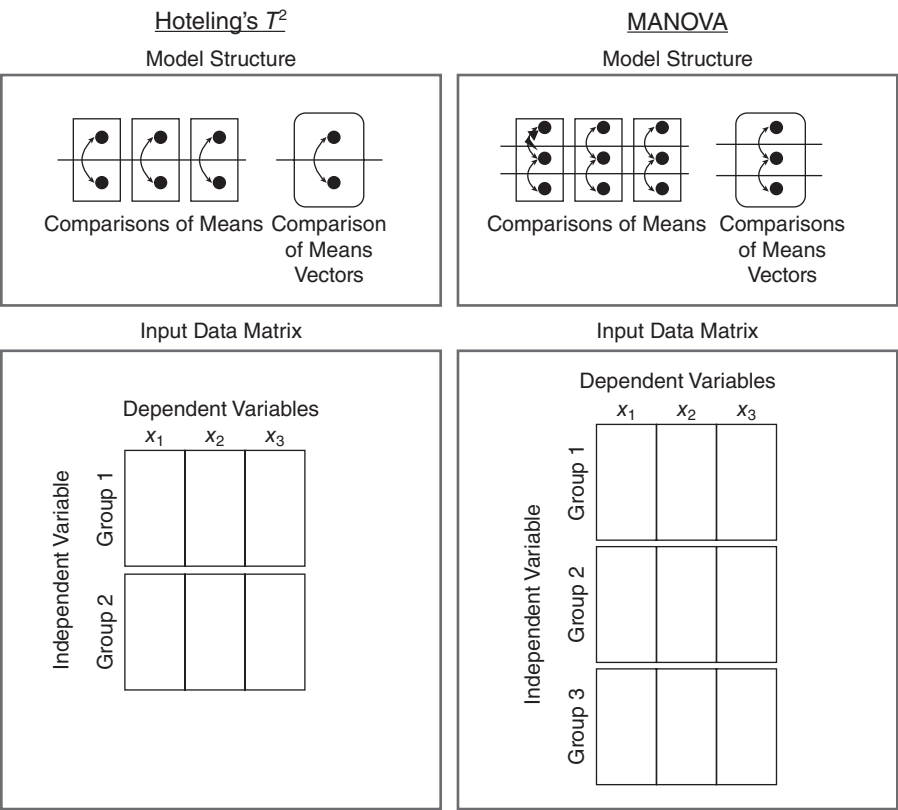
The diagram shows two vertical rectangular blocks representing data matrices. The left block is labeled 'X set' and has four columns labeled x_1 , x_2 , x_3 , and x_4 . The right block is labeled 'Y set' and has three columns labeled y_1 , y_2 , and y_3 . To the left of both blocks, the text 'Units of Observation' is written vertically, indicating that the rows represent individual units across which the variables are measured.

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 Pearson product moment correlation coefficient between X and Y is the same as that between Y and X), but simple bivariate regression is not. The regression equation is used for predicting Y from X but not usually for predicting X from Y .

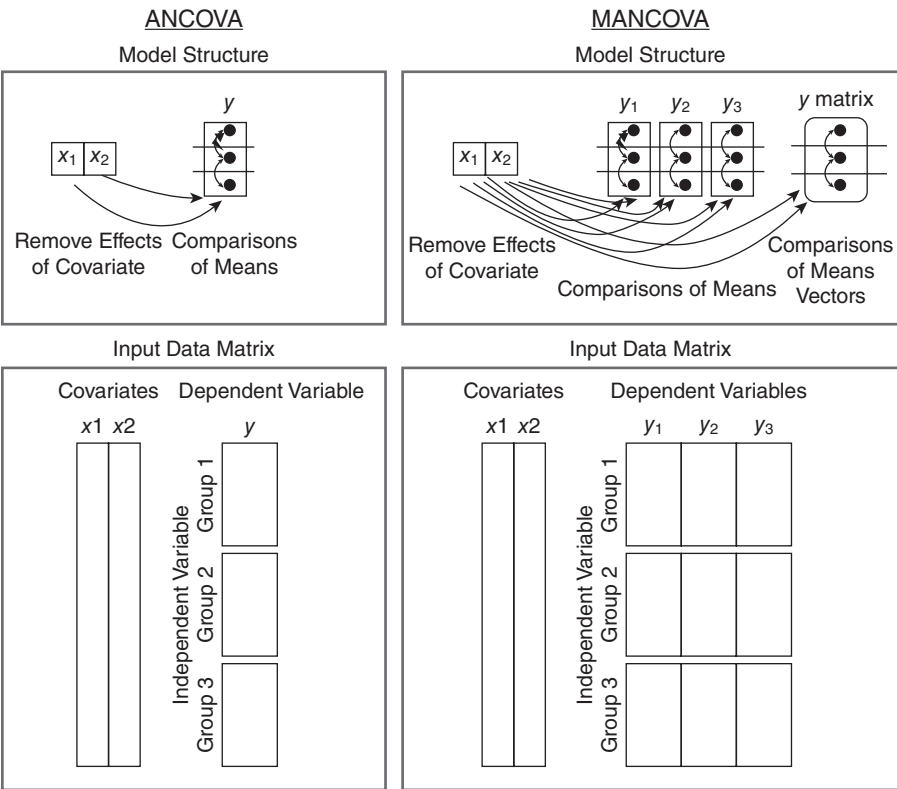


Another way to think of multivariate multiple regression is as the multivariate extension of multiple regression. Instead of predicting one dependent variable from a linear combination of independent variables, one predicts a set of multiple dependent variables from linear combinations of independent variables.

None of the methods discussed so far are specifically intended for data from true experimental designs. In fact, most multivariate methods are for correlational rather than experimental methods. However, a number of the multivariate methods are specifically designed to deal with truly experimental data having multiple dependent variables. These are the T^2 , MANOVA, ANCOVA, and MANCOVA methods presented in Chapters 7, 8, and 9 (one half of the methods chapters in this book). These methods fit the third and fourth types of data set structure discussed above, *two samples with multiple variables measured on each unit*, and *three or more samples with multiple variables measured on each unit*. These are illustrated in the two diagrams below. The two-sample type of data set can be analyzed with Hotelling's T^2 (Chapter 7), as shown in the diagram on the left, and data sets with three or more treatment groups require MANOVA (Chapter 8), as shown in the diagram on the right. In the same way that the t -test is a special case of ANOVA, the case restricted to two treatment groups, and the F -ratio of ANOVA is just the square of the corresponding t -value, Hotelling's T^2 is also a special case of MANOVA, and when there are only two groups, the same results will be obtained by using either method.



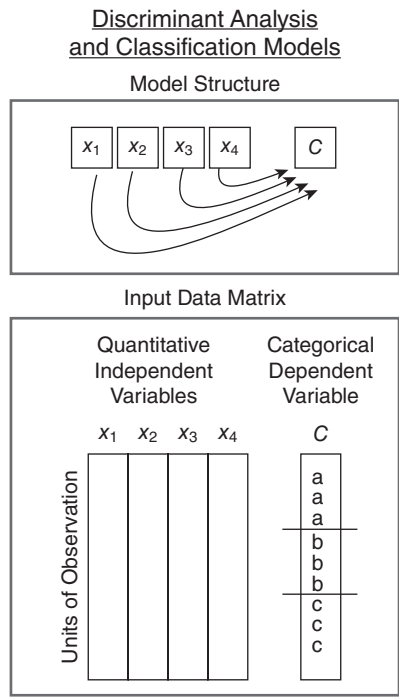
The simplest way to think of analysis of covariance (ANCOVA) is as an ANOVA calculated on the residuals from a regression analysis. That is, ANCOVA, like ANOVA provides tests of whether treatment effects are significant, but with the effects of one or more covariates “regressed out,” as shown in the diagram at the left below.



Strictly speaking ANCOVA is not a multivariate method, since there is only one dependent variable. The multivariate version of ANCOVA is *multivariate analysis of covariance* (MANCOVA), in which one essentially calculates a MANOVA with the effects of one or more covariates statistically controlled, as shown in the diagram at the right above. Both of these methods are presented in Chapter 9.

For the final diagram, we again draw upon the third and fourth types of data set structure, *two samples with multiple variables measured on each unit*, and *three or more samples with multiple variables measured on each unit*. These are the types for which Hotelling’s T^2 analysis and MANOVA are appropriate. However, the methods of discriminant analysis (Chapter 7) and classification analysis can also be used to good advantage with this kind of data structure,

even when it involves data from a true experimental design. The diagram for these methods applied to this kind of data structure is given below.



This looks very much like multiple regression, except that the dependent variable is categorical. Discriminant analysis asks the question “what is the best linear combination of a set of quantitative variables (X) to optimally separate categorical groups (C)?” In the usual way of using discriminant analysis and classification methods, the quantitative predictors of group membership would be thought of as independent variables, and the categories of group membership would be thought of as the dependent variable (which is how the diagram above is labeled). However, if one had MANOVA data from a true experiment, with the categories being treatment groups (the independent variable), then discriminant analysis could answer the question “what combination of the dependent variables best accounts for the significant multivariate effects of my experimental and control treatments?” This would reverse what is considered independent and what dependent variables.

The forgoing pictorial overview of methods includes most of the methods presented in this book. Notably absent are the methods of Chapter 5 (cluster analysis, multidimensional scaling, and multivariate graphics). These methods have much in common with factor analysis and principal components, and the diagram for cluster analysis would be similar to principal components.

However, cluster analysis creates taxonomic groupings rather than “factor loadings” (correlations between factors and observed variables), and it can be used to cluster both the variables and also the units of observation. Like factor analysis and principal component analysis, it can be applied to the first kind of data set structure, *a single sample with multiple variables measured on each sampling unit*.

Multidimensional scaling also applies to this kind of data set structure, but in a very unusual way, going directly to the latent variables without having any observed variables. It does this by inferring the latent variables from measured distances among the units of observation. Chapter 5, multivariate graphics, demonstrates how illuminating graphs can be constructed based upon the quantitative methods that apply to the first type of data set structure—factor analysis, principal component analysis, cluster analysis, and multidimensional scaling.

The log-linear methods presented in Chapter 9 are simple, involving two-way contingency tables like those analyzed with chi square. However, log-linear also can be used to deal with higher-order three-way, four-way, and in general “multiway” contingency tables in a more efficient manner. As such, it deals with data that are entirely categorical. However, the use of logarithms converts multiplicative relationships (the multiplication rule of probabilities) into additive relationships, and makes possible the full power of linear models (of the kind used in ANOVA and MANOVA) with categorical data. Generalized linear models use logarithmic (and other) linking functions to render categorical data amenable to linear models analysis. The second half of Chapter 9 introduces logistic regression and other generalized linear models, which can be used with any mix of categorical and quantitative variables. Diagrams for these methods would be very similar to those shown for various types of multiple and multivariate multiple regression.

Structural equations modeling would be difficult to diagram, since it is a general and very malleable set of methods that can be applied in one way or another to most of the types of data set structure presented. The results of virtually all of the other data analysis methods can be obtained from an adaptation of structural equations modeling.

1.6 CORRELATIONAL VERSUS EXPERIMENTAL METHODS

Experimental and correlational studies differ both in the research designs employed and also in the kind of statistics that are used to analyze the data. They also differ in the kinds of questions that can be answered, and in the way they use random processes. Correlational research designs are usually based on **random selection** of subjects, often in a naturalistic setting where there is little control over the variables. Experimental research designs, on the other hand, usually involve tight experimental controls and **random assignment** of subjects to treatment groups to ensure comparability. The critical distinction

between experimental and nonexperimental designs is that in true experimental designs, the experimenter manipulates the independent variable by randomly assigning subjects to treatment groups and the control group. Experimental designs enable the researcher to make more definitive conclusions and to attribute causality, whereas the inferences in correlational research are more tenuous. The multivariate methods introduced in Chapter 7 (Hotelling's T-Squared), 8 (MANOVA), 9 (ANCOVA, MANCOVA, repeated measures MANOVA, logistic regression models, etc.) are applicable to the data obtained from true experimental designs. The methods in the remainder of the chapters are used primarily with data from correlational studies and therefore provide less definitive conclusions.

Many seem to believe that the only real disadvantage of nonexperimental studies is that one cannot attribute causality with a high degree of confidence. While this is indeed a serious problem with nonexperimental designs, there are other issues. On page 162 of his 1971 book, Winer makes the very important point that ANOVA was originally developed to be used within the setting of a true experiment where one has control over extraneous variables and subjects are assigned to treatment groups at random. The logic underlying the significance tests and the determination of probabilities of the Type I error is based upon the assumption that treatment effects and error effects are independent of one another. The only assurance one can have that the two are indeed independent is that subjects are assigned at random to treatment groups. In other words, when ANOVA is used to analyze data from a nonexperimental design, and subjects are not assigned at random to treatment groups, there is no assurance that treatment and error effects are independent of one another, and the logic underlying the determination of the probability of the Type I error breaks down. One is, in this case, using the *P*-values from an ANOVA metaphorically. Winer (1971, 162) concludes this section with the words "hence the importance of randomization in design problems."

1.7 OLD VERSUS NEW METHODS

Factor analysis is a very old multivariate method. It dates back to a turn-of-the-century paper by Spearman (1904) and an earlier one by Pearson (1901). As such, it can be thought of as the fundamental and quintessential correlational multivariate method. Similarly, MANOVA is the fundamental multivariate method for dealing with data from true experimental designs (where variables are manipulated under controlled conditions). MANOVA was a comparatively late development, with its advent in the mid-20th century. However, both of these can be thought of as the "old" multivariate methods, one for correlational data and one for experimental data. These were the major multivariate methods used and taught in the research methodology classes a generation ago.

The young up-and-coming generation of researchers is much more excited about later developments in methods for dealing with both experimental data and also correlational data. By far, the most influential new method for dealing with correlational data is *structural equations modeling* (SEM). It has an updated approach to factor analysis referred to as *confirmatory factor analysis*, but confirmatory factor analysis is only one of a wide variety of SEM models, and one of the simpler ones at that. SEM is general and highly malleable, such that virtually all of the other correlational multivariate methods can be thought of as special cases of it.

SEM concepts are also foundational for the second “new” method, *hierarchical linear models* (HLM). HLM can be used to deal with data from true experimental designs, but it can also be used with correlational studies. At this point, it is not yet strictly speaking a multivariate method, in that the major data analytic packages (SAS, Stata, and SPSS) do not have procedures for a multivariate HLM (although there are ways to finesse the packages to accomplish multivariate goals with HLM). Also, the method is new enough that some of the mathematics for multivariate applications for HLM have yet to be worked out. It does not, however, by any means replace MANOVA (as some have erroneously thought). It is applicable to univariate data involving repeated measures, where one has a “mixed” model, involving a combination of fixed (treatment) and random (repeated measures) variables.

1.8 SUMMARY

Data are only as useful as the analyses performed with them. Increasingly, scientists have recognized that there are extremely few cases in which a single variable exerts sole influence on an isolated outcome. Typically, many factors influence one another, often in complex sequences. The world is multivariate.

Given the multivariate nature of biomedical and social sciences, advanced multivariate and regression methods are becoming increasingly utilized. This book covers those methods most commonly used in the research literature. Factor analysis, principal component analysis, and cluster analysis pertain to a single data set with multiple variables. Canonical correlation and multivariate multiple regression pertain to multivariate data broken down into distinct sets (i.e., classes of combinable variables in the case of canonical correlation, and predictors versus outcomes for multivariate multiple regression models). MANOVA, MANCOVA, and HLM involve continuous data distinguished by categories, and as such, these methods are essential to experimentation wherein groups or conditions are compared. Other statistics cover special cases, such as categorical outcomes (logistic and log-linear models). The common thread through all of these methods for quantitative and categorical outcomes is the general linear model.

The common features but also the clear differences between several multivariate methods have been represented diagrammatically in this chapter. These differences involve distinct configurations of the data and data types, with single-sample data sets of multiple variables being the simplest (as in factor analysis). Multiple-sample data sets (as in MANOVA) and prediction across two or more sets of multiple variables (as in canonical correlation or SEM) are some of the more complex configurations. The principles underlying all of these analyses are quite similar. Once the foundations are learned, different building blocks can be arranged to suit specific analytical purposes. It is also true that after the overall architectural design has been mastered, the student can more easily re-arrange building blocks as needed. The next chapter of this book describes the foundational building blocks of univariate statistics. Thereafter, the chapters progress systematically up through the specific to the most general cases of multivariate statistical methods, ending with methods for categorical outcomes. From simple to elegantly complex, multivariate methods provide the quantitative foundation for contemporary research in the biomedical and social sciences.

STUDY QUESTIONS

A. Essay Questions

1. Explain the difference between univariate statistical methods and multivariate statistical methods.
2. Explain the difference between factorial statistical methods and multivariate statistical methods. Can statistical methods be both factorial and also multivariate? Explain.
3. Discuss the statement that “most multivariate techniques were developed for use in nonexperimental research.”
4. Summarize the major kinds of data that are possible using the “four kinds of measurement scale” hypothesized by Stevens.
5. Explain the distinction between continuous and discrete data. Can data be both discrete and also interval/ratio? Explain. Can data be both continuous and also categorical? Explain.
6. There is a major difference between experimental and correlational research. Explain how research designs differ for these two. How do the statistical methods differ? How is randomization applied in each kind of research?
7. Evaluate the concept that although ANOVA methods were developed for experimental research, they can be applied to correlational data, that “the statistical methods ‘work’ whether or not the researcher manipulated the

independent variable.” You may wish to bring Winer’s (1971, 162) point about the assumption of independence of treatment effects and error into the discussion.

8. Chapter 1 states that the mathematical prerequisite for understanding this book is matrix algebra. Why is matrix algebra crucial to multivariate statistics?
9. Discuss the taxonomy of the four basic data set structures that are amenable to multivariate analysis; give examples of each and of multivariate methods that can be applied to each.

REFERENCES

- Pearson, K. 1901. On lines and planes of closest fit to systems of points in space. *Philosophy Magazine*, 2(6), 559–572.
- Spearman, C. 1904. General intelligence, objectively determined and measured. *American Journal of Psychology*, 15, 201–293.
- Stevens, S. S. 1946. On the theory of scales of measurement. *Science*, 103(2684), 677–680.
- Winer, B. J. 1971. *Statistical Principles in Experimental Design, Second Edition*. New York: McGraw-Hill.