MULTIVARIATE DATA ANALYSIS

Fifth Edition

JOSEPH F. HAIR, JR.
Louisiana State University

ROLPH E. ANDERSON
Drexel University

RONALD L. TATHAM
Burke Marketing Research

WILLIAM C. BLACK
Louisiana State University

PRENTICE HALL
Upper Saddle River, New Jersey 07458

Acquisitions Editor: Whitney Blake
Assistant Editor: John Larkin
Editorial Assistant: Rachel Falk
Vice President/Editorial Director: James Boyd
Marketing Manager: John Chillingworth
Marketing Director: Brian Kibby
Production Editor: Aileen Mason
Production Coordinator: Carol Samet
Managing Editor: Dee Josephson
Associate Managing Editor: Linda DeLorenzo
Manufacturing Supervisor: Arnold Vila
Manufacturing Manager: Vincent Scelta
Design Manager: Pat Smythe
Interior Design: Brian Deep
Cover Design: Jayne Conte
Illustrator (Interior): ElectraGraphics, Inc.
Composition: York Graphic Services, Inc.
Cover Photo: William Whitehurst/The Stock Market

Library of Congress Cataloging-in-Publication Data
Multivariate data analysis / Joseph F. Hair, Jr. . . . [et al.].
p. cm.
Rev. ed. of: Multivariate data analysis with readings. 4th ed. c1995.
Includes bibliographical references (p. –) and index.
ISBN 0-13-894858-5
1. Multivariate analysis. I. Hair, Joseph F. II. Multivariate data analysis with readings.
QA278.M85 1998
519.5'35—dc21 97-47031
CIP

Prentice-Hall International (UK) Limited, London
Prentice-Hall of Australia Pty. Limited, Sydney
Prentice-Hall Canada, Inc., Toronto
Prentice-Hall Hispanoamericana, S.A., Mexico
Prentice-Hall of India Private Limited, New Delhi
Prentice-Hall of Japan, Inc., Tokyo
Prentice-Hall Asia Pte. Ltd., Singapore
Editora Prentice-Hall do Brasil, Ltda., Rio de Janeiro

Printed in the United States of America

15 14 13 12

CHAPTER 2

Examining Your Data

LEARNING OBJECTIVES

Upon completing this chapter, you should be able to do the following:

- Select the appropriate graphical method to examine the characteristics of the data or relationships of interest.
- Understand the different types of missing data processes.
- Assess the type and potential impact of missing data.
- Explain the advantages and disadvantages of the approaches available for dealing with missing data.
- Identify univariate, bivariate, and multivariate outliers.
- Test your data for the assumptions underlying most multivariate techniques.
- Determine the best method of data transformation given a specific problem.
- Understand how to incorporate nonmetric variables as metric variables.

CHAPTER PREVIEW

This chapter reviews and describes the methods currently available to examine data. Data examination is a time-consuming, but necessary, step that is sometimes overlooked by researchers. Careful analysis of data leads to better prediction and more accurate assessment of dimensionality. The introductory section of this chapter offers a summary of various graphical techniques available to the researcher as a means of representing data. These techniques provide the researcher with a set of simple yet comprehensive ways to examine both the individual variables and the relationships among them. Other important concerns to the researcher

when examining data are how to assess and overcome pitfalls resulting from the research design and data collection. Specifically, this chapter addresses the evaluation of missing data, identification of outliers, and testing of the assumptions underlying most multivariate techniques. Missing data are a nuisance to researchers and may result from data entry errors or from the omission of answers by respondents. Classification of missing data and the processes, or reasons, underlying their presence are discussed in this chapter. Outliers, or extreme responses, may unduly influence the outcome of any multivariate analysis. For this reason, methods to assess their impact are discussed. Finally, the assumptions underlying most multivariate analyses are reviewed. Before applying any multivariate technique, the researcher must assess the fit of the sample data with the statistical assumptions underlying that multivariate technique. For example, researchers wishing to apply regression analysis (chapter 4) would be particularly interested in assessing the assumptions of normality, homoscedasticity, independence of error, and linearity. Each of these issues should be addressed to some extent for each application of a multivariate technique.

In addition, this chapter introduces the researcher to methods of incorporating nonmetric variables in applications that require metric variables through the creation of a special type of metric variable known as dummy variables. The applicability of using dummy variables varies with each data analysis project.

KEY TERMS

Before starting the chapter, review the key terms to develop an understanding of the concepts and terminology used. Throughout the chapter the key terms appear in **boldface.** Other points of emphasis in the chapter are *italicized.* Also, cross-references within the Key Terms appear in *italics.*

All-available approach *Imputation method* for missing data that computes values based on all available valid observations.

Boxplot Method of representing the distribution of a variable. A box represents the major portion of the distribution, and the extensions—called whiskers—reach to the extreme points of the distribution. Very useful in making comparisons of one or more variables across groups.

Censored data Observations that are incomplete in a systematic and known way. One example occurs in the study of causes of death in a sample in which some individuals are still living. Censored data are an example of *ignorable missing data.*

Comparison group The category of a nonmetric variable that receives all zeros when *indicator coding* is used or all minus ones (−1s) when *effects coding* is used in creating *dummy variables.*

Complete case approach Approach for handling missing data that computes values based on data from only complete cases, that is, cases with no *missing data.*

Data transformations A variable may have an undesirable characteristic, such as nonnormality, that detracts from its use in a multivariate technique. A transformation, such as taking the logarithm or square root of the variable, creates a transformed variable that is more suited to portraying the relationship. Transformations may be applied to either the dependent or independent variables, or both. The need and specific type of transformation may

be based on theoretical reasons (e.g., transforming a known nonlinear relationship) or empirical reasons (e.g., problems identified through graphical or statistical means).

Dummy variable Special metric variable used to represent a single category of a nonmetric variable. To account for L levels of an nonmetric variable, $L - 1$ dummy variables are needed. For example, gender is measured as male or female and could be represented by two dummy variables (X_1 and X_2). When the respondent is male, $X_1 = 1$ and $X_2 = 0$. Likewise, when the respondent is female, $X_1 = 0$ and $X_2 = 1$. However, when $X_1 = 1$, we know that X_2 must equal 0. Thus we need only one variable, either X_1 or X_2, to represent the variable gender. If a nonmetric variable has three levels, only two dummy variables are needed. We always have one dummy variable less than the number of levels for the nonmetric variable.

Effects coding Method for specifying the reference category for a set of *dummy variables* where the reference category receives a value of minus one (−1) across the set of dummy variables. With this type of coding, the dummy variable coefficients represent group deviations from the mean of all groups. This is in contrast to *indicator coding*.

Heteroscedasticity See *homoscedasticity*.

Histogram Graphical display of the distribution of a single variable. By forming frequency counts in categories, the shape of the variable's distribution can be shown. Used to make a visual comparison to the *normal distribution*.

Homoscedasticity When the variance of the error terms (ϵ) appears constant over a range of predictor variables, the data are said to be homoscedastic. The assumption of equal variance of the population error E (where E is estimated from ϵ) is critical to the proper application of linear regression. When the error terms have increasing or modulating variance, the data are said to be *heteroscedastic*. Analysis of *residuals* best illustrates this point.

Ignorable missing data *Missing data process* that is explicitly identifiable and/or is under the control of the researcher. Ignorable missing data do not require a remedy because the missing data are explicitly handled in the technique used.

Imputation methods Process of estimating the *missing data* of an observation based on valid values of the other variables. The objective is to employ known relationships that can be identified in the valid values of the sample to assist in representing or even estimating the replacements for missing values.

Indicator coding Method for specifying the reference category for a set of *dummy variables* where the reference category receives a value of zero across the set of dummy variables. The dummy variable coefficients represent the category differences from the reference category. Also see *effects coding*.

Kurtosis Measure of the peakedness or flatness of a distribution when compared with a normal distribution. A positive value indicates a relatively peaked distribution, and a negative value indicates a relatively flat distribution.

Linearity Used to express the concept that the model possesses the properties of additivity and homogeneity. In a simple sense, linear models predict values that fall in a straight line by having a constant unit change (slope) of the dependent variable for a constant unit change of the independent variable. In the population model $Y = b_0 + b_1X_1 + E$, the effect of a change of 1 in X_1 is to add b_1 (a constant) units of Y.

Missing at random (MAR) Classification of *missing data* applicable when missing values of Y depend on X, but not on Y. When missing data are MAR,

observed data for Y are a truly random sample for the X values in the sample, but not a random sample of all Y values due to missing values of X.

Missing completely at random (MCAR) Classification of *missing data* applicable when missing values of Y are not dependent on X. When missing data are MCAR, observed values of Y are a truly random sample of all Y values, with no underlying process that lends bias to the observed data.

Missing data Information not available for a subject (or case) about whom other information is available. Missing data often occur when a respondent fails to answer one or more questions in a survey.

Missing data process Any systematic event external to the respondent (such as data entry errors or data collection problems) or any action on the part of the respondent (such as refusal to answer a question) that leads to *missing data.*

Multivariate graphical display Method of presenting a multivariate profile of an observation on three or more variables. The methods include approaches such as glyphs, mathematical transformations, and even iconic representations (e.g., faces).

Normal distribution Purely theoretical continuous probability distribution in which the horizontal axis represents all possible values of a variable and the vertical axis represents the probability of those values occurring. The scores on the variable are clustered around the mean in a symmetrical, unimodal pattern known as the bell-shaped, or normal, curve.

Normal probability plot Graphical comparison of the form of the distribution to the *normal distribution.* In the normal probability plot, the normal distribution is represented by a straight line angled at 45 degrees. The actual distribution is plotted against this line, so that any differences are shown as deviations from the straight line, making identification of differences quite apparent and interpretable.

Normality Degree to which the distribution of the sample data corresponds to a *normal distribution.*

Outlier An observation that is substantially different from the other observations (i.e., has an extreme value). At issue is its representativeness of the population.

Residual Portion of a dependent variable not explained by a multivariate technique. Associated with dependence methods that attempt to predict the dependent variable, the residual represents the unexplained portion of the dependent variable. Residuals can be used in diagnostic procedures to identify problems in the estimation technique or to identify unspecified relationships.

Scatterplot Representation of the relationship between two metric variables portraying the joint values of each observation in a two-dimensional graph.

Skewness Measure of the symmetry of a distribution; in most instances the comparison is made to a *normal distribution.* A positively skewed distribution has relatively few large values and tails off to the right, and a negatively skewed distribution has relatively few small values and tails off to the left. Skewness values falling outside the range of -1 to $+1$ indicate a substantially skewed distribution.

Stem and leaf diagram A variant of the *histogram* which provides a visual depiction of the variable's distribution as well as an enumeration of the actual data values.

Variate Linear combination of variables formed in the multivariate technique by deriving empirical weights applied to a set of variables specified by the researcher.

Introduction

The tasks involved in examining your data may seem mundane and inconsequential but they are an essential part of any multivariate analysis. Multivariate techniques place tremendous analytical power in the researcher's hands, but they also place a greater burden on the researcher to ensure that the statistical and theoretical underpinning on which they are based are also supported. By examining the data before the application of a multivariate technique, the researcher gains several critical insights into the characteristics of the data. First and foremost, the researcher attains a basic understanding of the data and relationships between variables. Multivariate techniques place great demands on the researcher to understand, interpret, and articulate results based on relationships that are ever increasing in complexity. Knowledge of variable interrelationships can aid immeasurably in the specification and refinement of the multivariate model as well as provide a reasoned perspective for interpretation of the results. Second, multivariate techniques demand much more from the data they are to analyze. The statistical power of the multivariate techniques requires larger data sets and more complex assumptions than encountered with univariate analyses. The analytical sophistication needed to ensure that the statistical requirements are met has forced the researcher to use a series of data examination techniques that in many instances match the complexity of the multivariate techniques. Moreover, the effects of missing data, which by definition are not directly represented in the results, can nevertheless be substantial in their impact on the nature and character of the results. The purpose of this chapter is to provide an overview of the available data examination techniques, ranging from the simple process of visual inspection of graphical displays to the multivariate statistical methods involved in handling missing data and testing the assumptions underlying all multivariate methods.

Both novice and experienced researchers may be tempted to skim or even skip this chapter to spend more time in gaining knowledge of a multivariate technique(s). Although the time, effort, and resources devoted to the data examination process may seem almost wasted because often no corrective action is warranted, the researcher should view these techniques as an "investment in multivariate insurance." Even though a technique may estimate properly and obtain results, the "hidden" problems arising from issues in the chapter can lead to potentially catastrophic problems. These problems can be avoided by following these analyses each and every time a multivariate technique is applied. These efforts will more than pay for themselves in the long run, as the occurrence of one serious and possibly fatal problem will make a convert of any researcher. We encourage you to embrace these techniques before problems raised during analysis force you to do so.

This chapter addresses four separate phases of examining your data: (1) a graphical examination of the nature of the variables in the analysis and the relationships that form the basis of multivariate analysis; (2) an evaluation process for understanding the impact missing data can have on the analysis, plus alternatives for retaining cases with missing data in the analysis; (3) the techniques best suited for identifying outliers, those cases that may distort the relationships by their uniqueness on one or more of the variables under study; and (4) the analytical methods necessary to assess the ability of the data to meet the statistical assumptions specific to many multivariate techniques. The chapter concludes by

introducing a technique for incorporating nonmetric variables when metric variables are required. A set of replacement metric variables are created to represent the categories of the nonmetric variables.

Graphical Examination of the Data

As discussed earlier, the use of multivariate techniques places an increased burden on the researcher to understand, evaluate, and interpret the more complex results. One aid in these tasks is a thorough understanding of the basic characteristics of the underlying data and relationships. When univariate analyses are considered, the level of understanding is fairly simple. But as the researcher moves to more complex multivariate analyses, the need and level of understanding increase dramatically. This section reviews some of the graphical methods available to assist in gaining a basic understanding of the characteristics of the data, particularly in a multivariate sense.

The advent and widespread use of statistical programs designed for the personal computer has led to increased access to such methods. Most statistical programs have comprehensive modules of graphical techniques available for data examination that are augmented with more detailed statistical measures of data description. The following sections detail some of the more widely used techniques for examining the characteristics of the distribution, bivariate relationships, group differences, and even multivariate profiles.

The Nature of the Variable: Examining the Shape of the Distribution

The starting point for understanding the nature of any variable is to characterize the shape of its distribution. A number of statistical measures are discussed in a later section on normality, but many times the researcher can gain an adequate perspective on the variable through a **histogram**. A histogram is a graphical representation of a single variable that represents the frequency of occurrences (data values) within data categories. The frequencies are plotted to examine the shape of the distribution of responses. If the integer values ranged from one to ten, the researcher could construct a histogram by counting the number of responses that were a one, a two, and so on. For continuous variables, categories are formed within which the frequency of data values are tabulated. For example, the responses for X_1 from the database introduced in chapter 1 are represented in Figure 2.1. Categories with midpoints of 0.0, .5, 1.0, 1.5, . . . , 6.0 are used. The height of the bars represents the frequencies of data values within each category. If examination of the distribution is to assess its normality (see section on testing assumptions for details on this issue), the normal curve can be superimposed on the distribution as well, as was done in Figure 2.1. The histogram can be used to examine any type of metric variable, from original values to residuals from a multivariate technique.

A variant of the histogram is the **stem and leaf diagram**, which presents the same graphical picture as the histogram but also provides an enumeration of the actual data values. The stem and leaf diagram in Figure 2.2 is composed of *stems* and *leaves*. The stem is the root value, to which the leaves are added. For example, in Figure 2.2, the first stem is 0.0. To this is added the leaf of 0, resulting in a

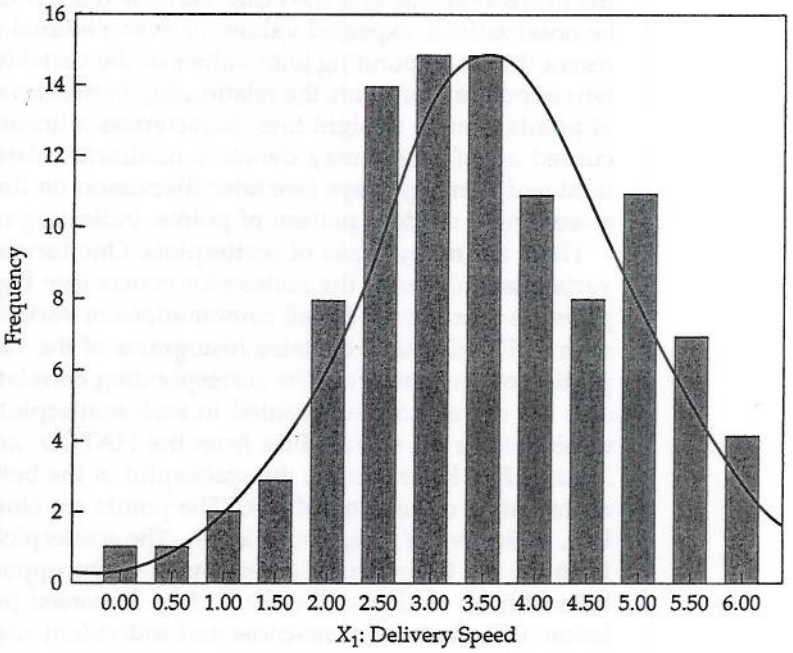


FIGURE 2.1 Graphical Representation of a Univariate Distribution: The Histogram

value of 0. In the next stem, a value of .6 is added to the stem of 0, resulting in a value of .6. If the frequencies of X_1 are compiled, 0.0 and .6 are the first two values. At the other end of the figure, the stem is 6.0. It is associated with two leaves (0 and 1), representing the values 6.0 and 6.1. These are the two highest values for X_1. The stem and leaf diagram provides a general shape of the distribution, as found with the histogram, as well as providing the actual data values.

Frequency	Stem and Leaf		
1.00	0	*	0
1.00	0	.	6
3.00	1	*	013
7.00	1	.	6688999
12.00	2	*	001333444444
10.00	2	.	5566788899
18.00	3	*	000001111233444444
10.00	3	.	5666777889
10.00	4	*	001122233
10.00	4	.	556778999
11.00	5	*	00112223344
5.00	5	.	55689
2.00	6	*	01

Stem width: 1.0
Each leaf: 1 case (s)

Valid cases: 100.0 Missing cases: .0 Percent missing: .0

FIGURE 2.2 Stem and Leaf Plot of X_1 (Delivery Speed)

Examining the Relationship between Variables

Whereas examining the distribution of a variable is essential, many times the researcher is also interested in examining relationships between two or more variables. The most popular method for examining bivariate relationships is the **scatterplot,** a graph of data points based on two variables. One variable defines the horizontal axis and the other variable defines the vertical axis. Variables may be observations, expected values, or even **residuals.** The points in the graph represent the corresponding joint values of the variables for any given case. The pattern of points represents the relationship between variables. A strong organization of points along a straight line characterizes a linear relationship or correlation. A curved set of points may denote a nonlinear relationship, which can be accommodated in many ways (see later discussion on linearity). Or there may be only a seemingly random pattern of points, indicating no relationship.

There are many types of scatterplots. One format particularly suited to multivariate techniques is the scatterplot matrix (see Figure 2.3), in which the scatterplots are represented for all combinations of variables in the lower portion of the matrix. The diagonal contains histograms of the variables. Included in the upper portion of the matrix are the corresponding correlations so that the reader can assess the correlation represented in each scatterplot. Figure 2.3 presents the scatterplots for a set of variables from the HATCO database (X_1, X_2, X_3, X_4, X_5, X_6, X_7, and X_9). For example, the scatterplot in the bottom left corner (X_1 versus X_9) represents a correlation of .676. The points are closely aligned around a straight line, indicative of a high correlation. The scatterplot in the leftmost column, third from the top (X_1 versus X_4) demonstrates the opposite, an almost total lack of relationship as evidenced by the widely dispersed pattern of points and the correlation .050. Scatterplot matrices and individual scatterplots are now available in all popular statistical programs. A variant of the scatterplot is discussed in the following section on outlier detection, where an ellipse representing a specified confidence interval for the bivariate normal distribution is superimposed to allow for outlier identification.

Examining Group Differences

The researcher is also faced with understanding the extent and character of differences between two or more groups for one or more metric variables, such as found in discriminant analysis, analysis of variance, and multivariate analysis of variance. In these cases, the researcher needs to understand how the values are distributed for each group and if there are sufficient differences between the groups to support statistical significance. Another important aspect is to identify **outliers** that may become apparent only when the data values are separated into groups. The method used for this task is the **boxplot,** a pictorial representation of the data distribution. The upper and lower boundaries of the box mark the upper and lower quartiles of the data distribution. Thus, the box length is the distance between the 25th percentile and the 75th percentile, so that the box contains the middle 50 percent of the data values. The asterisk (*) inside the box identifies the median. If the median lies near one end of the box, skewness in that direction is indicated. The larger the box, the greater the spread of the observations. The lines extending from each box (called *whiskers*) represent the distance to the smallest and the largest observations that are less than one quartile range from the box. Outliers are observations that range between 1.0 and 1.5 quartiles away from the

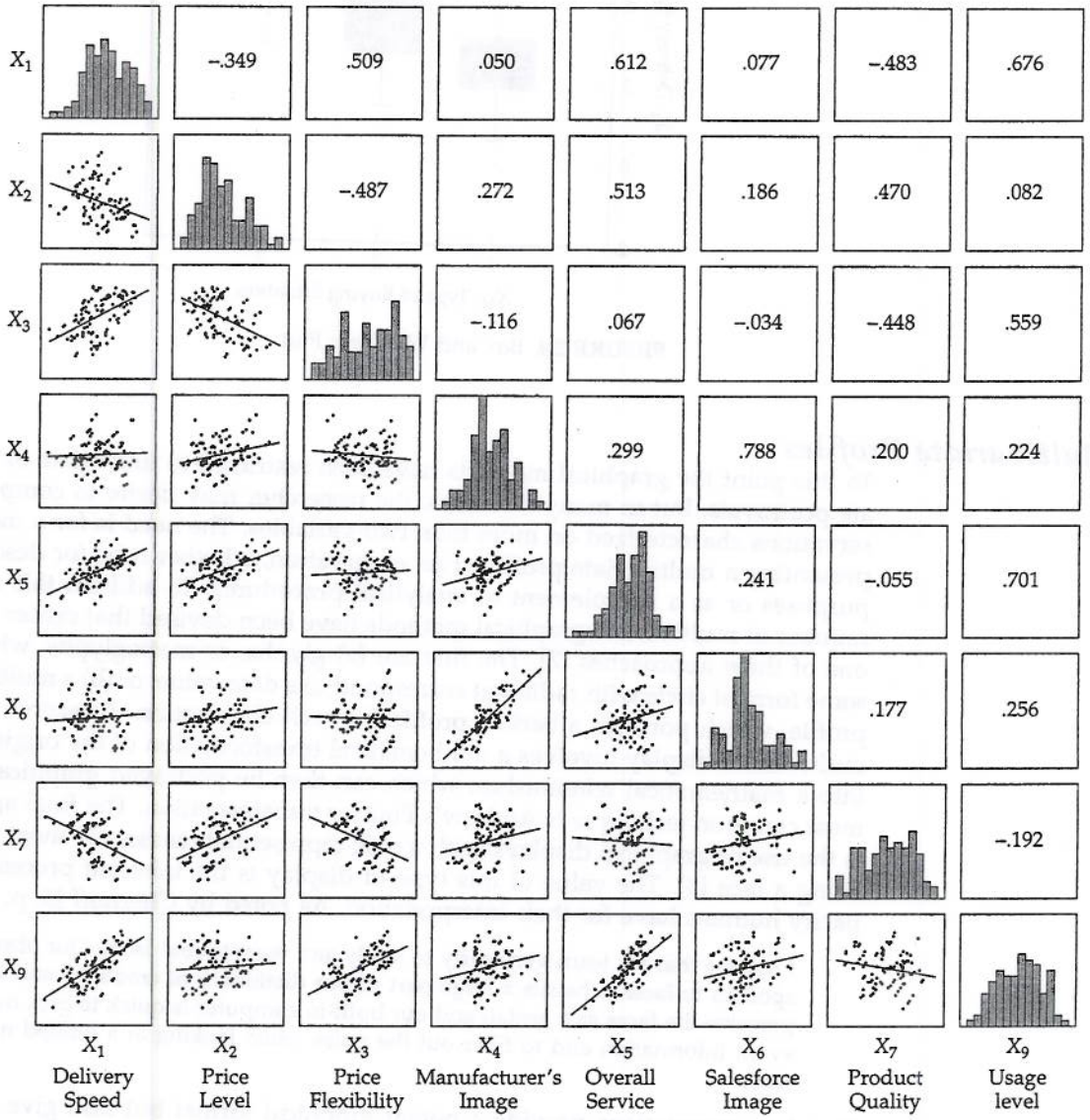


FIGURE 2.3 Scatterplot Matrix of Metric Variables
Note: Values above the diagonal are bivariate correlations, with corresponding scatterplots below the diagonal. Diagonal portrays the distribution of each variable.

box. Extreme values are those observations greater than 1.5 quartiles away from the end of the box.

Figure 2.4 (p. 44) shows the boxplots for X_1 (delivery speed) for each group of X_{14} (type of buying situation) from the HATCO database. The three groups have markedly different boxplots, indicating differences among the groups in terms of perceptions of delivery speed. The boxplot for the first type of buying situation also indicates that an outlier exists. The researcher should examine this observation and then consider the possible remedies. The remedies available for outliers are discussed in more detail later.

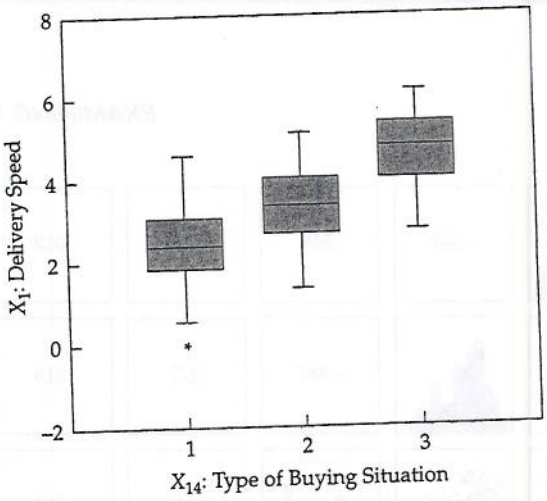


FIGURE 2.4 Box and Whiskers Plot

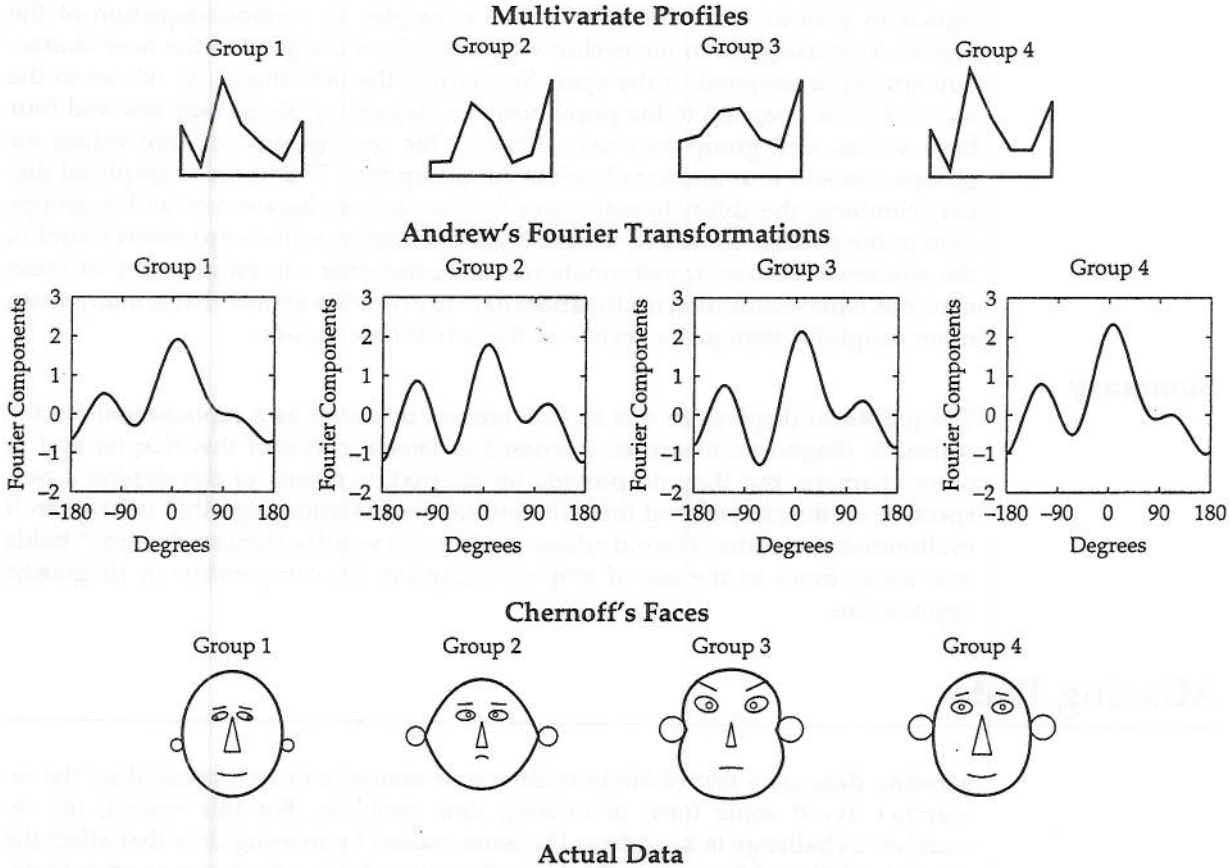
Multivariate Profiles

To this point the graphical methods have been restricted to univariate or bivariate portrayals. But in many instances, the researcher may desire to compare observations characterized on more than two variables. The need is for a means of presenting a multivariate profile of an observation, whether it be for descriptive purposes or as a complement to analytical procedures. To address this need, a number of multivariate graphical methods have been devised that center around one of three approaches [7]. The first are (a) glyphs, or metroglyphs, which are some form of circle with radii that correspond to a data value; or (b) a multivariate profile, which portrays a barlike profile for each observation. A second form of multivariate display involves a mathematical transformation of the original data into a mathematical relationship, which can then be portrayed graphically. The most common technique is Andrew's Fourier transformation. The final approach is the use of graphical displays with iconic representativeness, the most popular being a face [3]. The value of this type of display is the inherent processing capacity humans have for their interpretation. As noted by Chernoff [3, p. 9]:

> I believe that we learn very early to study and react to real faces. Our library of responses to faces exhausts a large part of our dictionary of emotions and ideas. We perceive the faces as a gestalt and our built-in computer is quick to pick out the relevant information and to filter out the noise when looking at a limited number of faces.

Facial representations provide a potent graphical format but also give rise to a number of considerations that impact the assignment of variables to facial features, unintended perceptions, and the quantity of information that can actually be accommodated. Discussion of these issues is beyond the scope of this text, and interested readers are encouraged to review them before attempting to use these methods [10, 11].

Figure 2.5 contains illustrations of three types of **multivariate graphical displays** produced using SYSTAT, which are also available in several other personal computer-based statistical programs. The upper portion of Figure 2.5 contains ex-



Actual Data

Group	X_1	X_2	X_3	X_4	X_5	X_6	X_7
1	4.794	1.622	8.267	4.717	3.222	2.067	5.044
2	2.011	2.133	6.544	5.267	2.039	2.672	8.483
3	3.700	4.158	6.008	6.242	3.900	3.233	8.258
4	4.809	1.510	9.319	5.690	3.148	3.195	6.981

FIGURE 2.5 Examples of Multivariate Graphical Displays

amples of each type of multivariate graphical display: profiles, Fourier transformations, and iconic faces. Data values for four observations on seven variables are contained in a table at the bottom of the figure. In this instance, the data are profiles of four customer groups for the seven performance factors from the HATCO database. From the actual data values in the table, similarities and differences are difficult to distinguish, even to the extent that there may not be any differences. The objective of the multivariate profiles is to portray the data in a manner that enables each identification of differences and similarities. The first display in Figure 2.5 contains multivariate profiles, which show the leftmost portion is lowest for group two, and highest for groups one and four. This pattern corresponds to the values of X_1, and comparisons can be made between groups on a single

variable or across variables for a single group. The second type of multivariate graphical display in the figure is Andrew's Fourier transformation, which represents the data values by a mathematical expression. Although comparisons on a single value are more difficult, this form of graphical display provides a single representation for generalized comparison and grouping of observations. Finally, iconic symbols (Chernoff's faces) were constructed with the seven variables assigned to various facial features. In this example, X_1 controls depiction of the mouth, X_2 is assigned to the eyebrow features, X_3 is assigned to the nose characteristics, X_4 is assigned to the eyes, X_5 controls the face shape, X_6 relates to the ear, and X_7 is assigned to the pupil position. Regarding X_1, groups one and four have smiles, and group two has a frown. This corresponds to large values for groups one and four and small values for group two. This form of graphical display combines the ability to make specific comparisons between or within groups seen in the profiles as well as the more generalized overall comparisons found in the Andrew's Fourier transformations. The researcher can employ any of these methods when examining multivariate data to provide a format that is many times more insightful than just a review of the actual data values.

Summary

The graphical displays in this section are not intended as a replacement for the statistical diagnostic measures discussed in later sections of this chapter and in other chapters. But they do provide an alternative means of developing a perspective on the character of the data and the interrelationships that exist, even if multivariate in nature. The old adage "a picture is worth a thousand words" holds true many times in the use of graphical displays for comparative or diagnostic applications.

Missing Data

Missing data are a fact of life in multivariate analysis; in fact, rarely does the researcher avoid some form of missing data problem. For this reason, the researcher's challenge is to address the issues raised by missing data that affect the generalizability of the results. To do so, the researcher's primary concern is to determine the reasons underlying the missing data, with the extent of missing data being a secondary issue in most instances. This need to focus on the reasons for missing data comes from the fact that the researcher must understand the processes leading to the missing data in order to select the appropriate course of action.

A **missing data process** is any systematic event external to the respondent (such as data entry errors or data collection problems) or action on the part of the respondent (such as refusal to answer) that leads to missing values. The effects of some missing data processes are known and directly accommodated in the research plan. But others, particularly those based on actions by the respondent, are rarely known. When the missing data processes are unknown, the researcher attempts to identify any patterns in the missing data that would characterize the missing data process. In doing so, the researcher asks such questions as: (1) Are the missing data scattered randomly throughout the observations or are distinct patterns identifiable? and (2) How prevalent are the missing data? If patterns are

found and the extent of missing data is sufficient to warrant action, then it is assumed that some missing data process is in operation. Any statistical results based on these data would be biased to the extent that the variables included in the analysis are influenced by the missing data process. The concern for understanding the missing data processes is similar to the need to understand the causes of nonresponse in the data collection process. For example, are those individuals who did not respond different from those who did? If so, do these differences have any impact on the analysis, the results, or their interpretation? Concerns similar to these also arise from missing responses for individual variables.

The impact of missing data is detrimental not only through its potential "hidden" biases of the results but also in its practical impact on the sample size available for analysis. For example, if remedies for missing data are not applied, any observation with missing data on any of the variables will be excluded from the analysis. In many multivariate analyses, particularly survey research applications, missing data may eliminate so many observations that what was an adequate sample is reduced to an inadequate sample. In such situations, the researcher must either gather additional observations or find a remedy for the missing data in the original sample. Although finding a remedy for missing data is the most practical solution, few guidelines exist pertaining to the diagnosis and remedy of missing data. For this reason, the following sections discuss the different types of missing data processes, methods to identify the nature of the missing data processes, and available remedies for accommodating missing data into multivariate analyses.

A Simple Example of a Missing Data Analysis

Table 2.1 (p. 48) contains a simple example of missing data among 20 cases. As typical of many data sets, particularly in survey research, the number of missing data vary widely among both cases and variables. In this example, we can see that all of the variables (V_1 to V_5) have some missing data, with V_3 missing over one-half (55 percent) of all values. Three cases (3, 13, and 15) have more than 50 percent missing data and only five cases have complete data. Overall, 22 percent of the data values are missing. If a multivariate analysis was run that required complete data, the data would be reduced to only five cases, too few for any type of analysis. This level of reduction in available cases is not uncommon in many applications.

More sophisticated remedies for missing data will be discussed in detail in later sections, but an obvious option is the elimination of variables and/or cases. In our example, assuming that the conceptual foundations of the research are not altered substantially by the deletion of a variable, eliminating V_3 is one approach to reducing the number of missing data. By just eliminating V_3, seven additional cases, for a total of 12, now have complete information. If the three cases (3, 13, 15) with exceptionally high numbers of missing data are also eliminated, the total number of missing data is now reduced to only five instances, or 7.4 percent of all values. These five missing data, however, are all present in V_4, and we must look for any patterns among these data as well. By comparing the cases with missing data for V_4 with those having valid V_4 values, we see a pattern emerge with respect to V_2. The five cases with missing values for V_4 also have the five lowest values for V_2, indicating that missing data for V_4 are strongly associated with lower scores on V_2. This systematic association between missing and valid data

TABLE 2.1 Hypothetical Example of Missing Data

						Missing Data by Case	
Case ID	V_1	V_2	V_3	V_4	V_5	*Number*	*Percent*
1	1.3	9.9	6.7	3.0	2.6	0	0
2	4.1	5.7			2.9	2	40
3		9.9		3.0		3	60
4	.9	8.6		2.1	1.8	1	20
5	.4	8.3		1.2	1.7	1	20
6	1.5	6.7	4.8		2.5	1	20
7	.2	8.8	4.5	3.0	2.4	0	0
8	2.1	8.0	3.0	3.8	1.4	0	0
9	1.8	7.6		3.2	2.5	1	20
10	4.5	8.0		3.3	2.2	1	20
11	2.5	9.2		3.3	3.9	1	20
12	4.5	6.4	5.3	3.0	2.5	0	9
13					2.7	4	80
14	2.8	6.1	6.4		3.8	1	20
15	3.7			3.0		3	60
16	1.6	6.4	5.0		2.1	1	20
17	.5	9.2		3.3	2.8	1	20
18	2.8	5.2	5.0		2.7	1	20
19	2.2	6.7		2.6	2.9	1	20
20	1.8	9.0	5.0	2.2	3.0	0	0
MISSING DATA BY VARIABLE						TOTAL MISSING VALUES	
Number	2	2	11	6	2	Number: 23	
Percent	10	10	55	30	10	Percent: 23	

directly impacts any analysis in which V_4 and V_2 are both included. In this instance, the researcher must always scrutinize results including either V_4 and V_2 for the possible impact of this missing data process on the results.

Understanding the Reasons Leading to Missing Data

Before any missing data remedy can be implemented, the researcher must first diagnose and understand the missing data processes underlying the missing data. Sometimes these processes are under the control of the researcher and can be explicitly identified. In such instances, the missing data are termed **ignorable,** which means that specific remedies for missing data are not needed because the allowances for missing data are inherent in the technique used [9].

Ignorable Missing Data

One example of an ignorable missing data process is the "missing data" of those observations in a population that are not included when taking a sample. The purpose of multivariate techniques is to generalize from the sample observations to the entire population, which is really an attempt to overcome the missing data of observations not in the sample. The researcher makes this missing data ignorable by using probability sampling to select respondents. Probability sampling allows the researcher to specify that the missing data process leading to the omitted observations is random and that the missing data can be accounted for as sampling error in the statistical procedures. Thus, the "missing data" of the nonsampled observations is ignorable.

Another instance of ignorable missing data occurs when the data are censored. **Censored data** are observations not complete because of their stage in the missing data process. A typical example is an analysis of the causes of death. Respondents who are still living cannot provide complete information (i.e., cause or time of death) and are thus censored. Another interesting example of censored data is found in the attempt to estimate the heights of the U.S. general population based on the heights of armed services recruits (as cited in [9]). The data are censored because in certain years the armed services had height restrictions that varied in level and enforcement. Thus, the researchers are faced with the task of estimating the heights of the entire population when it is known that certain individuals (i.e., all those below the height restrictions) are not included in the sample. In both instances the researcher's knowledge of the missing data process allows for the use of specialized methods, such as event history analysis, to accommodate censored data [9].

The justification for designating missing data as ignorable is that the missing data process is operating at random (i.e., the observed values are a random sample of the total set of values, observed and missing) or explicitly accommodated in the technique used. However, in most instances the missing data process is not explicitly addressed by the techniques used. Thus, the researcher must assess the extent and impact of the missing data to determine whether they are due to a random process or, if not, whether they are amenable to one of the available remedies.

Other Types of Missing Data Processes

Missing data can occur for many reasons and in many situations. One type of missing data process that may occur in any situation is due to procedural factors, such as errors in data entry that create invalid codes, disclosure restrictions (e.g., small counts in U.S. census data), failure to complete the entire questionnaire, or even the morbidity of the respondent. In these situations, the researcher has little control over the missing data processes, but some remedies may be applicable if the missing data are found to be random. Another type of missing data process occurs when the response is inapplicable, such as questions regarding the years of marriage for adults who have never been married. Again, the analyses can be specifically formulated to accommodate these respondents.

Other missing data processes may be less easily identified and accommodated. Most often these are related directly to the respondent. One example is the refusal to respond to certain questions. This is common in questions of a sensitive nature (e.g., income or controversial issues) or when the respondent has no opinion or insufficient knowledge to answer the question. The researcher should anticipate these problems and attempt to minimize them in the research design and data collection stages of the research. However, they still may occur, and the researcher must now deal with the resulting missing data. But all is not lost. When the missing data occur in a random pattern, remedies may be available to mitigate their effect.

Examining the Patterns of Missing Data

To decide whether a remedy for missing data can be applied, the researcher must first ascertain the degree of randomness present in the missing data. Assume for the purposes of illustration that two variables (X and Y) are collected. X has no missing data, but Y does have some missing data. If a missing data process is found between X and Y where there are significant differences in the values of X

between cases for Y with valid and missing data, then the missing data are not at random. Any analysis must explicitly accommodate the missing data process between X and Y or else bias is introduced into the results.

Missing data are termed **missing at random (MAR)** if the missing values of Y depend on X, but not on Y. By this we mean that the observed Y values represent a random sample of the actual Y values for each value of X, but the observed data for Y do not necessarily represent a truly random sample of all Y values. Even though the missing data process is random in the sample, its values are not generalizable to the population. For example, assume that we know the gender of respondents (the X variable) and are asking about household income (the Y variable). We find that the missing data are random for both males and females but occur at a much higher frequency for males than females. Even though the missing data process is operating in a random manner, any remedy applied to the missing data will still reflect the missing data process because gender affects the ultimate distribution of the household income values.

A higher level of randomness is termed **missing completely at random (MCAR).** In these instances the observed values of Y are truly a random sample of all Y values, with no underlying process that lends bias to the observed data. In our earlier example, this would be shown by the fact that the missing data for household income were randomly missing in equal proportions for both males and females. If this is the form of the missing data process, any of the remedies can be applied without making allowances for the impact of any other variable or missing data process.

Diagnosing the Randomness of the Missing Data Process

As previously noted, the researcher must ascertain whether the missing data process occurs in a completely random manner. Three methods are available for this diagnosis. The first assesses the missing data process of a single variable Y by forming two groups—observations with missing data for Y and those with valid values of Y. Statistical tests are then performed to determine whether significant differences exist between the two groups on other variables of interest. Significant differences indicate the possibility of a nonrandom missing data process. Let us use our earlier example of household income and gender. We would first form two groups of respondents, those with missing data on the household income question and those who answered the question. We would then compare the percentages of gender for each group. If one gender (e.g., males) was found in greater proportion in the missing data group, we would suspect a nonrandom missing data process. If the variable we were comparing was metric (e.g., an attitude or perception) instead of categorical (gender), then *t* tests could be performed. The researcher should examine a number of variables to see whether any consistent pattern emerges. Remember that some differences will occur by chance, but any series of differences may indicate an underlying nonrandom pattern.

A second approach utilizes dichotomized correlations to assess the correlation of missing data for any pair of variables. For each variable, valid values are represented by the value of one, and missing data are replaced by the value of zero. These missing value indicators for each variable are then correlated. The correlations indicate the degree of association between the missing data on each variable pair. Low correlations denote randomness in the missing data for each pair of variables. Although no strict guidelines exist for identifying the level of correlation needed to indicate a nonrandom missing data process, statistical signifi-

cance tests of the correlations provide a conservative estimate of the degree of randomness. If randomness is indicated for all variable pairs, then the researcher can assume that the missing data can be classified as MCAR. If significant correlations exist between some pairs of variables, then the researcher may have to assume that the data are only MAR, and these relationships must be accommodated in any remedies that are applied.

Finally, an overall test of randomness can be performed that determines whether the missing data can be classified as MCAR. This test analyzes the pattern of missing data on all variables and compares it with the pattern expected for a random missing data process. If no significant differences are found, the missing data can be classified as MCAR. If significant differences are found, however, the researcher must use the approaches described previously to identify the specific missing data processes that are nonrandom.

Approaches for Dealing with Missing Data

The approaches or remedies for dealing with missing data can be classified into one of four categories based on the randomness of the missing data process and the method used to estimate the missing data [9]. If nonrandom or MAR missing data processes are found, the researcher should apply only one remedy—the specifically designed modeling approach [9]. Application of any other method introduces bias into the results. Only if the researcher determines that the missing data process can be classified as MCAR can the approaches discussed in the following sections be used.

However, researchers often make the assessment for randomness of the missing data process before applying one of these missing data remedies. And even if the remedy is appropriate, the researcher must note the specific impact on the results associated with that remedy. Too often a remedy is applied without an assessment of the missing data processes, the appropriateness of the selected remedy, or its consequences. Thus, the researcher never realizes the effects because they are hidden in the overall results.

Use of Observations with Complete Data Only

The simplest and most direct approach for dealing with missing data is to include only those observations with complete data, also known as the **complete case approach.** This method is available in all statistical programs and is the default method in many programs. Yet the complete case approach should be used only if the missing data are MCAR, because missing data that are not MCAR have nonrandom elements that bias the results. Thus, even though only valid observations are used, the results are not generalizable to the population. Moreover, in many situations, the resulting sample size is reduced to an inappropriate size. The complete case approach is best suited for instances in which the extent of missing data is small, the sample is sufficiently large to allow for deletion of the cases with missing data, and the relationships in the data are so strong as to not be affected by any missing data process.

Delete Case(s) and/or Variable(s)

Another simple remedy for missing data is to delete the offending case(s) and/or variable(s). In this approach, the researcher determines the extent of missing data on each case and variable and then deletes the case(s) or variable(s) with

excessive levels. In many cases where a nonrandom pattern of missing data is present, this may be the most efficient solution. The researcher may find that the missing data are concentrated in a small subset of cases and/or variables, with their exclusion substantially reducing the extent of the missing data. Again, no firm guidelines exist on the necessary level for exclusion, but any decision should be based on both empirical and theoretical considerations. If missing values are found for what will be a dependent variable in the proposed analysis, the case is usually excluded. This avoids any artificial increases in the explanatory power of the analysis, which can occur when the researcher first estimates the missing data for the dependent variable by one of the imputation processes described next and then uses the estimated values in the analysis of the dependence relationship. If a variable other than a dependent variable has missing values and is a candidate for deletion, the researcher should be sure that alternative variables, hopefully highly correlated, are available to represent the intent of the original variable. The researcher must always consider the gain of eliminating a source of missing data versus the deletion of a variable in the multivariate analysis.

Imputation Methods

A third category of remedies for handling missing data is through one of the many **imputation methods**. Imputation is the process of estimating the missing value based on valid values of other variables and/or cases in the sample. The objective is to employ known relationships that can be identified in the valid values of the sample to assist in estimating the missing values. However, the researcher should carefully consider the use of imputation in each instance because of its potential impact on the analysis [6]:

> The idea of imputation is both seductive and dangerous. It is seductive because it can lull the user into the pleasurable state of believing that the data are complete after all, and it is dangerous because it lumps together situations where the problem is sufficiently minor that it can be legitimately handled in this way and situations where standard estimators applied to the real and imputed data have substantial biases.

The methods discussed in this section are used primarily with metric variables for two reasons. First, estimates of the missing data for metric variables can be made with such values as a mean of all valid values. Second, nonmetric variables require an estimate of a specific value rather than an estimate on a continuous scale. It is very different to estimate a missing value for a metric variable, such as an attitude or perception—even income—than it is to estimate the respondent's gender when missing. Thus, nonmetric variables are typically not filled by the imputation process, but require either the specific modeling approach discussed in the next section or are left as missing.

Imputation methods can be defined as one of two types: (1) use of all of the available information from a subset of cases to generalize to the entire sample, or (2) methods of estimating replacement values for the missing data, which are then analyzed by standard multivariate techniques. The following discussion will describe the various options within each type and their advantages and disadvantages.

Using All Available Information as the Imputation Technique

The first type of imputation method does not actually replace the missing data, but instead imputes the distribution characteristics (e.g., means or standard deviations) or relationships (e.g., correlations) from all available valid values. Known

as the **all-available approach,** this method (the PAIRWISE option in SPSS, and the CORPAIR, COVPAIR, or ALLVALUE options in BMDP) is primarily used to estimate correlations and maximize the pairwise information available in the sample. The distinguishing characteristic of this approach is that each correlation for a pair of variables is based on a potentially unique set of observations and the number of observations used in the calculations can vary for each correlation. The imputation process occurs not by replacing the missing data on the remaining cases, but instead by using the obtained correlations as representative for the entire sample. This approach can be compared to the complete-case approach discussed earlier, which uses data only from observations that have no missing data. Either approach can introduce bias if the missing data process is not MCAR.

Even though the all-available method maximizes the data utilized and overcomes the problem of missing data on a single variable eliminating a case from the entire analysis, several problems can also arise from this approach. First, correlations may be calculated that are "out of range" and inconsistent with the other correlations in the correlation matrix. Any correlation between X and Y is constrained by their correlation to a third variable Z, as shown in the following formula:

$$\text{Range of } r_{xy} = r_{xz}r_{yz} \pm \sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}$$

The correlation between X and Y can range only from $+1$ to -1 if both X and Y have zero correlation with all other variables in the correlation matrix. Yet rarely are the correlations with other variables zero. As the correlations with other variables increase, the range of the correlation between X and Y decreases. This increases the potential for the correlation in a unique set of cases to be inconsistent with correlations derived from other sets of cases. For example, if X and Y have correlations of .6 and .4, respectively, with Z, then the possible range of correlation between X and Y is $.24 \pm .73$, or from $-.49$ to .97. Any value outside this range is mathematically inconsistent, yet may occur if the correlation is obtained with a differing number and set of cases for the two correlations in the all-available approach.

An associated problem is that the eigenvalues in the correlation matrix can become negative, thus altering the variance properties of the correlation matrix. Although the correlation matrix can be adjusted to eliminate this problem (e.g., the ALLVALUE option in BMDP), many procedures do not include this adjustment process. In extreme cases, the estimated variance/covariance matrix is not positive definite. All of these problems must be considered when selecting the all-available approach.

The Replacement of Missing Data

The second form of imputation involves replacing missing values with estimated values based on other information available in the sample. There are many options, varying from the direct substitution of values to estimation processes based on relationships among the variables. The following discussion focuses on the more widely used methods, although many other forms of imputation are available [9].

Case Substitution In this method, observations with missing data are replaced by choosing another nonsampled observation. A common example is to replace a sampled household that cannot be contacted or that has extensive missing data with another household not in the sample, preferably very similar to the original

observation. This method is most widely used to replace observations with complete missing data, although it can be used to replace observations with lesser amounts of missing data as well.

Mean Substitution One of the more widely used methods, mean substitution replaces the missing values for a variable with the mean value of that variable based on all valid responses. In this manner, the valid sample responses are used to calculate the replacement value. The rationale of this approach is that the mean is the best single replacement value. This approach, although it is used extensively, has three disadvantages. First, it makes the variance estimates derived from the standard variance formulas invalid by understating the true variance in the data. Second, the actual distribution of values is distorted by substituting the mean for the missing values. Third, this method depresses the observed correlation because all missing data will have a single constant value. It does have the advantage, however, of being easily implemented and providing all cases with complete information.

Cold Deck Imputation In this method, the researcher substitutes a constant value derived from external sources or previous research for the missing values. This is similar in nature to the mean substitution method, differing only in the source of the substitution value. Cold deck imputation has the same disadvantages as the mean substitution method, and the researcher must be sure that the replacement value from an external source is more valid than an internally generated value, such as the mean. This method can provide the researcher with the option of replacing the missing data with a value that may be deemed more valid than the mean of the sample.

Regression Imputation In this method, regression analysis (described in chapter 4) is used to predict the missing values of a variable based on its relationship to other variables in the data set. Although it has the appeal of using relationships already existing in the sample as the basis of prediction, this method also has several disadvantages. First, it reinforces the relationships already in the data. As the use of this method increases, the resulting data become more characteristic of the sample and less generalizable. Second, unless stochastic terms are added to the estimated values, the variance of the distribution is understated. Third, this method assumes that the variable with missing data has substantial correlations with the other variables. If these correlations are not sufficient to produce a meaningful estimate, then other methods, such as mean substitution, are preferable. Finally, the regression procedure is not constrained in the estimates it makes. Thus, the predicted values may not fall in the valid ranges for variables (e.g., a value of 11 may be predicted for a 10-point scale), thereby requiring some form of additional adjustment. Even with all of these potential problems, the regression method of imputation holds promise in those instances for which moderate levels of widely scattered missing data are present and for which the relationships between variables are sufficiently established so that the researcher is confident that using this method will not impact the generalizability of the results.

Multiple Imputation The final imputation method is actually a combination of several methods. In this approach, two or more methods of imputation are used to derive a composite estimate—usually the mean of the various estimates—for

the missing value. The rationale of this approach is that the use of multiple approaches minimizes the specific concerns with any single method and the composite will be the best possible estimate. The choice of this approach is primarily based on the trade-off between the researcher's perception of the potential benefits versus the substantially higher effort required to make and combine the multiple estimates.

Model-Based Procedures

The final set of procedures explicitly incorporates the missing data into the analysis, either through a process specifically designed for missing data estimation or as an integral portion of the standard multivariate analysis. The first approach involves maximum likelihood estimation techniques that attempt to model the processes underlying the missing data and to make the most accurate and reasonable estimates possible [9]. One example is the EM approach in SPSS. It is an iterative two-stage method (the E and M stages) in which the E-stage makes the best possible estimates of the missing data and the M-stage then makes estimates of the parameters (means, standard deviations, or correlations) assuming the missing data were replaced. The process continues going through the two stages until the change in the estimated values is negligible and they replace the missing data.

The second approach involves the inclusion of missing data directly into the analysis, defining observations with missing data as a select subset of the sample. This approach is most applicable for dealing with missing values on the independent variables of a dependent relationship. Its premise has best been characterized by Cohen and Cohen [4, p. 299]:

> We thus view missing data as a pragmatic fact that must be investigated, rather than a disaster to be mitigated. Indeed, implicit in this philosophy is the idea that like all other aspects of sample data, missing data are a property of the population to which we seek to generalize.

When the missing values occur on a nonmetric variable, the researcher can easily define those observations as a separate group and then include them in any analysis, such as ANOVA, MANOVA, or even discriminant analysis. When the missing data are present on a metric independent variable in a dependence relationship, a procedure has been developed to incorporate the observations into the analysis while maintaining the relationships among the valid values [4]. This procedure is best illustrated in the context of regression analysis, although it can be used in other dependence relationships as well. The first step is to code all observations with missing data as dummy variables (where the cases with missing data receive a value of one and the other cases have a value of zero). The missing values are then imputed by the mean substitution method. Finally, the relationship is estimated by normal means. The dummy variable represents the difference for the dependent variable between those observations with missing data and those observations with valid data. The test of the dummy variable coefficient assesses the statistical significance of this difference. The coefficient of the original variable represents the relationship for all cases with nonmissing data. This method allows the researcher to retain all the observations in the analysis for purposes of maintaining the sample size. It also provides a direct test for the differences between the two groups along with the estimated relationship between the dependent and independent variables.

An Illustration of Missing Data Diagnosis

To illustrate the process of diagnosing the patterns of missing data and the application of possible remedies, a new data set is introduced (see appendix A for a complete listing of the observations). This data set was collected during the pretest of the questionnaire used to collect the data described in chapter 1. The pretest involved 70 individuals and collected responses on all 14 variables. In the course of pretesting, however, missing data occurred. The following sections detail the diagnosis of the extent of missing data in the data set and the analyses available for selecting and applying the various missing data remedies available in most statistical programs. A number of software programs are adding analyses of missing data, among them BMDP and SPSS. The analyses described next can all be replicated by data manipulation and conventional analysis. Examples are provided in appendix A.

Examining the Patterns of Missing Data

Table 2.2 contains the descriptive statistics for the observations with valid values, including the percentage of cases with missing data on each variable. Six cases were eliminated from the analysis owing to missing data on more than half of the variables of interest. The extent of missing data for the remaining 64 observations ranges from a high of 30 percent of the cases for X_1 to a low of a single case (1.6 percent) for X_6. For the variables with the higher levels of missing data (X_1, X_2, and X_3), the levels are not so excessive that they automatically dictate exclusion of the variable. Given the integral role these variables are expected to play in the various multivariate analyses, all efforts should be made to retain them in the analysis.

Summary Statistics of Pretest Data

One factor that could alleviate some of the high levels of missing data for certain variables is the elimination of cases from the analysis. To determine whether the missing data are concentrated on a select set of cases, Table 2.3 provides a graphical display of the missing data for each case that has missing data. Except for the six cases already eliminated because of extremely high levels of missing data, we see that no other cases have a disproportionate number of missing values. In fact, of the 38 cases with missing data, only four cases have more than two missing values.

TABLE 2.2 Summary Statistics of Pretest Data

	Number of Cases with Valid Data	*Mean*	*Standard Deviation*	*Missing Data*	
				Number	*Percent*
X_1 Delivery speed	45	4.0133	.9664	19	29.7
X_2 Price level	54	1.8963	.8589	10	15.6
X_3 Price flexibility	50	8.1300	1.3194	14	21.9
X_4 Manufacturer image	60	5.1467	1.1877	4	6.3
X_5 Overall service	59	2.8390	.7541	5	7.8
X_6 Salesforce image	63	2.6016	.7192	1	1.6
X_7 Product quality	60	6.7900	1.6751	4	6.3
X_9 Usage level	60	45.9667	9.4204	4	6.3
X_{10} Satisfaction level	60	4.7983	.8194	4	6.3

Note: Six of the original 70 cases had more than 50 percent missing data and were excluded from the analysis. All analyses are based on the remaining 64 cases. Twenty-six cases had no missing data.

TABLE 2.3 Graphical Display of Missing Data

Case	Number of Missing Values	Variables: Missing Data X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_9	X_{10}
202	2	S		S						
203	2		S					S		
204	3	S		S						S
205	1			S						
207	3	S		S						S
213	2		S	S						
216	2	S				S				
218	2	S				S				
219	2							S	S	
220	1		S							
221	3	S		S				S		
222	2			S		S				
224	3	S	S						S	
225	2			S	S					
227	2		S						S	
228	2	S			S					
229	1					S				
231	1							S		
232	2	S	S							
235	2						S			S
237	1		S							
238	1	S								
240	1	S								
241	2			S		S				
244	1								S	
246	1				S					
248	2	S	S							
249	1		S							
250	2	S		S						
253	1	S								
255	2	S		S						
256	1	S								
257	2		S	S						
259	1	S								
260	1	S								
267	2			S	S					
268	1									S
269	2	S		S						

Legend: S = a missing value

Table 2.4 (p. 58) portrays the patterns of missing data. The most prevalent pattern is the missing data for X_1 found in six cases, with the second most common pattern being missing data for X_1 and X_3 in four cases. All of the remaining cases exhibit patterns that are essentially unique or shared with only a very small number of cases. As this analysis demonstrates, no patterns occur with a frequency that suggests an underlying missing data process. Thus, no case or set of cases with a missing data pattern can be eliminated that would markedly improve the missing data problem.

TABLE 2.4 Tabulated Missing Data Patterns

	Missing Data Patterns[a]								
Number of Cases	X_6	X_{10}	X_4	X_7	X_9	X_5	X_2	X_3	X_1
26								X	
1								X	X
4									X
6									X
1			X						
1			X					X	
2			X			X		X	
2						X			
1						X			X
2							X		X
2							X		
3							X	X	
2				X			X		
1				X					
1				X	X				
1					X				
1					X		X		
1					X		X		X
1		X							
1	X	X						X	X
2		X						X	X
1				X					

[a]Variables are sorted on missing patterns.

Diagnosing Randomness of the Missing Data

The next step is an empirical examination of the patterns of missing data to determine whether the missing data are distributed randomly across the cases and the variables. The first test for assessing randomness is to compare the observations with and without missing data for each variable on the other variables. For example, the observations with missing data on X_1 are placed in one group and those observations with valid responses for X_1 are placed in another group. Then, these two groups are compared to identify any differences on the remaining metric variables (X_2 through X_{10}). Once comparisons have been made on all of the variables, new groups are formed based on the missing data for the next variable (X_2) and the comparisons are performed again on the remaining variables. This process continues until each variable (X_1 through X_{10}) has been examined for any differences. The objective is to identify any systematic missing data process that would be reflected in patterns of significant differences.

Table 2.5 contains the results for this analysis of the 64 remaining observations from the pretest sample. The first noticeable pattern of significant t values occurs for X_9, for which six of the nine comparisons found significant differences between the two groups. However, the impact of these differences is marginal as the number of cases with missing data on X_9 ranged from only three to five. X_7 exhibited a pattern of differences similar to X_9 with four significant differences, but small groups of cases with missing data. This analysis indicates that although

TABLE 2.5 Assessing the Randomness of Missing Data through Group Comparisons of Observations with Missing versus Valid Data

Groups Formed by Missing Data on:		X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_9	X_{10}
X_1	*t*		−.3	1.3	2.2	2.6	1.9	−1.1	2.6	2.1
	Significance		.763	.223	.033	.017	.065	.273	.017	.049
	Number present	45	38	38	42	42	44	42	42	43
	Number missing	0	16	12	18	17	19	18	18	17
	Mean (present)	4.01	1.87	8.27	5.34	3.02	2.71	6.61	48.17	4.95
	Mean (missing)		1.95	7.68	4.69	2.39	2.36	7.20	40.83	4.42
X_2	*t*	−.5		.7	−2.2	−4.2	−2.4	−1.2	−1.1	−1.2
	Significance	.646		.528	.044	.001	.034	.260	.318	.233
	Number present	38	54	42	50	49	53	51	52	50
	Number missing	7	0	8	10	10	10	9	8	10
	Mean (present)	3.97	1.90	8.18	4.99	2.70	2.51	6.68	45.46	4.75
	Mean (missing)	4.23		7.86	5.94	3.50	3.11	7.40	49.25	5.02
X_3	*t*	.4	1.4		1.1	2.0	.2	.0	1.9	.9
	Significance	.693	.180		.286	.066	.818	.965	.073	.399
	Number present	38	42	50	48	47	49	47	46	48
	Number missing	7	12	0	12	12	14	13	14	12
	Mean (present)	4.03	1.98	8.13	5.24	2.95	2.61	6.80	47.02	4.84
	Mean (missing)	3.90	1.60		4.79	2.42	2.56	6.77	42.50	4.62
X_4	*t*	−.2	2.6	−.3		.2	1.4	1.5	.2	−2.4
	Significance	.882	.046	.785		.888	.249	.197	.830	.064
	Number present	42	50	48	60	55	59	56	56	56
	Number missing	3	4	2	0	4	4	4	4	4
	Mean (present)	4.01	1.94	8.12	5.15	2.84	2.63	6.83	46.02	4.76
	Mean (missing)	4.07	1.33	8.35		2.80	2.25	6.20	45.25	5.38
X_5	*t*	−.1	−.3	.8	.4		−.9	−.4	.5	.6
	Significance	.900	.749	.502	.734		.423	.696	.669	.605
	Number present	42	49	47	55	59	58	55	55	55
	Number missing	3	5	3	5	0	5	5	5	5
	Mean (present)	4.01	1.89	8.20	5.16	2.84	2.58	6.76	46.18	4.82
	Mean (missing)	4.10	1.98	7.10	5.04		2.86	7.14	43.60	4.56
X_7	*t*	3.0	.9	.2	−2.1	.9	−1.5		.5	.4
	Significance	.036	.440	.864	.118	.441	.193		.658	.704
	Number present	42	51	47	56	55	59	60	57	56
	Number missing	3	3	3	4	4	4	0	3	4
	Mean (present)	4.07	1.92	8.14	5.07	2.86	2.58	6.79	46.14	4.81
	Mean (missing)	3.27	1.50	8.00	6.18	2.55	2.90		42.67	4.70
X_9	*t*	6.1	−1.4	2.2	−1.1	−.9	−1.8	1.7		1.6
	Significance	.000	.384	.101	.326	.401	.149	.128		.155
	Number present	42	52	46	56	55	59	57	60	56
	Number missing	3	2	4	4	4	4	3	0	4
	Mean (present)	4.08	1.85	8.26	5.11	2.82	2.57	6.82	45.97	4.82
	Mean (missing)	3.10	3.00	6.63	5.62	3.08	3.03	6.30		4.47
X_{10}	*t*	1.7	.8	−2.1	2.5	2.7	1.3	.9	2.4	
	Significance	.249	.463	.235	.076	.056	.302	.409	.066	
	Number present	43	50	48	56	55	60	56	56	60
	Number missing	2	4	2	4	4	3	4	4	0
	Mean (present)	4.03	1.92	8.09	5.23	2.89	2.62	6.83	46.43	4.80
	Mean (missing)	3.55	1.60	9.20	3.95	2.08	2.17	6.30	39.50	

Each cell contains six values: 1) *t* value for comparison of the column variable between group a (observations with valid data on the row variable) and group b (observations with missing data on the row variable); 2) significance of *t* value for group comparisons; 3) & 4) number of observations for group a (valid data) and group b (missing data); 5) & 6) mean of variable for group a (valid data) and group b (missing data)

Interpretation of the table:

The upper right cell indicates that the *t* value for the comparison of X_{10} between group a (valid data) and group b (missing data) on X_1 is 2.1, which has a significance level of .049. The sample sizes of group a and group b are 43 and 17, respectively. Finally, the mean of X_{10} for group a (valid data) is 4.95, whereas the mean for group b (missing data) is 4.42.

significant differences can be found due to the missing data on two variables (X_7 and X_9), the small number of cases involved makes this of marginal concern. If later tests of randomness indicate a nonrandom pattern of missing data, these results would then provide a starting point for possible remedies.

A second test for randomness involves the use of correlations between dichotomous variables. The dichotomous variables are formed by replacing valid values with a value of one and missing data with a value of zero. The resulting correlations between the dichotomous variables indicate the extent to which missing data are related in pairs of variables. Low correlations indicate low association between the missing data process for those two variables. Table 2.6 contains the correlations between the nine dichotomized metric variables. Review of the values indicates that only one correlation is in the moderate range (X_6 and X_{10} have a correlation of .488). This suggests that the missing data process influencing X_{10} corresponds to the missing data process affecting X_6. However, given the absence of any other correlations with even moderate values, the researcher can be assured that no single missing data process is significantly affecting a substantial number of variables.

The final test is an overall test of the missing data for being missing completely at random (MCAR). The test makes a comparison of the actual pattern of missing data with what would be expected if the missing data were totally randomly distributed. In this instance, as shown in Table 2.6, the significance level of the

TABLE 2.6 Assessing the Randomness of Missing Data through Dichotomized Variable Correlations and the Multivariate Test for Missing Completely at Random (MCAR)

	X_1 *Delivery Speed*	X_2 *Price Level*	X_3 *Price Flexibility*	X_4 *Manufacturer Image*	X_5 *Overall Service*	X_6 *Salesforce Image*	X_7 *Product Quality*	X_9 *Usage Level*	X_{10} *Satisfaction Level*
X_1	1.000								
	45								
X_2	0.003	1.000							
	38	54							
X_3	0.235	−0.020	1.000						
	38	42	50						
X_4	−0.026	−0.111	0.176	1.000					
	42	50	48	60					
X_5	0.066	−0.125	0.128	−0.075	1.000				
	42	49	47	55	59				
X_6	−0.082	−0.054	−0.067	−0.033	−0.037	1.000			
	44	53	49	59	58	63			
X_7	−0.026	0.067	0.020	−0.067	−0.075	−0.033	1.000		
	42	51	47	56	55	59	60		
X_9	−0.026	0.244	−0.137	−0.067	−0.075	−0.033	0.200	1.000	
	42	52	46	56	55	59	57	60	
X_{10}	0.115	−0.111	0.176	−0.067	−0.075	0.488*	−0.067	−0.067	1.000
	43	50	48	56	55	60	56	56	60

Little's MCAR Test: Chi-square: 174.464
Degrees of freedom: 159
Probability: .190

Interpretation:
First value in the table represents the correlation between the dichotomized variables, where cases with a valid value receive a 1 and missing data receive a 0. The second value, below the correlation, represents the number of cases having valid data on both variables in that specific correlation pair.
*Significant at the .05 level.

MCAR tests was .190, indicating that the missing data process can be considered to be MCAR. As a result, the researcher may employ any of the remedies for missing data, because no potential biases exist in the patterns of missing data.

Remedies for Missing Data

As discussed earlier, numerous remedies are available for dealing with missing data. In this instance, several of the remedies have definite disadvantages. If the complete case approach is taken, the sample size is reduced to 26 observations, barely sufficient for even the simplest univariate analyses, much less multivariate applications. Our earlier examination of the patterns of the missing data demonstrated that there was not a small set of cases that could be deleted and thereby markedly reduce the extent of missing data. Moreover, the only viable alternative in eliminating a variable is the elimination of X_1, which has missing data on almost 30 percent of the cases. But even if X_1 were deleted, all of the cases with missing data would still have at least one other variable with missing data. Even eliminating X_1 is a relatively ineffective approach for creating more observations with complete data on all variables.

The remaining option is to employ some form of imputation to estimate replacement values for the missing values. The first possibility is to use only observations with complete data or using all available information to estimate the correlations. The advantage of the complete information approach is that it maintains consistency in the correlation matrix. However, it may also reduce the number of observations used to such a small subset of the sample (28 cases) that the resulting correlations used for imputation differ markedly from those obtained using all available information. The all-available approach maximizes the number of cases used in calculating the correlations, but may introduce inconsistencies in the calculated correlations. A third option is to use a mean substitution for all the missing data and then calculate the correlations.

Table 2.7 (p. 62) contains the correlations obtained from the all-available, complete information, and mean substitution approaches. In most instances the correlations are similar, but there are several substantial differences. First, there is a consistency between the correlations obtained with the all-available and mean substitution approaches. The differences occur among the correlations obtained with the complete information approach. Second, the notable differences are concentrated in the correlations of X_1 and X_{10} with X_4, X_5, X_6, and X_7. These differences may indicate the impact of a missing data process, but this was not detected by the earlier diagnostic measures. Although the researcher has no proof of greater validity for either approach, these results demonstrate the marked differences sometimes obtained between the approaches. Whichever approach is chosen, the researcher should examine the correlations obtained by alternative methods to understand the range of possible values.

The researcher may also select a specific estimation approach for estimating replacement values for the missing data. Table 2.8 (p. 63) contains the results of employing the mean substitution regression and EM approaches for missing value imputation. These results include the means and standard deviations obtained after missing values are replaced by the imputed data. As seen in the earlier comparisons of correlations, some differences can be detected, but no consistent pattern emerges. For variables X_1 and X_2, there are marked differences in the estimated values. In general, for the remaining variables the estimates are very similar, if not identical. Thus, the researcher does not have a definitive indication of which

TABLE 2.7 Comparison of Correlations Obtained with the All-Available (Pairwise), Complete Case (Listwise), and Mean Substitution Approaches

	X_1 *Delivery Speed*	X_2 *Price Level*	X_3 *Price Flexibility*	X_4 *Manufacturer Image*	X_5 *Overall Service*	X_6 *Salesforce Image*	X_7 *Product Quality*	X_9 *Usage Level*	X_{10} *Satisfaction Level*
X_1	1.000								
	1.000								
	1.000								
X_2	−.479	1.000							
	−.502	1.000							
	−.349	1.000							
X_3	.416	−.357	1.000						
	.429	−.294	1.000						
	.329	−.289	1.000						
X_4	−.099	.299	−.065	1.000					
	−.245	.320	−.061	1.000					
	−.086	.245	−.057	1.000					
X_5	.366	.440	.047	.432	1.000				
	.566	.421	.157	.046	1.000				
	.232	.382	.042	.422	1.000				
X_6	.031	.260	−.035	.810	.344	1.000			
	−.094	.356	−.066	.804	.213	1.000			
	.027	.219	−.032	.769	.323	1.000			
X_7	−.138	.348	−.358	.398	.066	.402	1.000		
	−.416	.354	−.230	.382	−.150	.529	1.000		
	−.106	.310	−.297	.374	.061	.395	1.000		
X_9	.376	.149	.601	.223	.712	.268	−.202	1.000	
	.599	.048	.648	.191	.683	.301	−.099	1.000	
	.265	.134	.503	.216	.656	.260	−.195	1.000	
X_{10}	.514	−.184	.702	.378	.533	.233	−.256	.669	1.000
	.549	−.278	.725	.170	.304	.064	−.405	.566	1.000
	.381	−.173	.626	.344	.477	.229	−.250	.647	1.000

Interpretation: The top value is the correlation obtained with a pairwise or all-available approach, the second value is the correlation obtained with a listwise or complete information approach, and the third value is the correlation obtained with mean substitution. Sample sizes for the all-available information approach varied; the actual sample sizes are listed in Table 2.5. A sample size of 26 was used for the complete information correlations; there were no missing data after mean substitution, so the sample size for this approach was 64.

approach is appropriate. Instead the researcher must coalesce the missing data patterns with the strengths and weaknesses of each approach and then select the most appropriate method. In the instance of differing estimates, the more conservative approach of combining the estimates into a single estimate (the multiple imputation approach) may be the most appropriate choice. Whichever approach is used, the data set with replacement values should be saved for further analysis.

A Recap of the Missing Value Analysis

Our evaluation of the issues surrounding missing data in the pretest data can be summarized in four conclusions:

1. *The missing data process is MCAR.* All of the diagnostic techniques support the conclusion that no systematic missing data process exists, making the missing data MCAR (missing completely at random). Such a finding provides two ad-

TABLE 2.8 Results of the Regression and EM Imputation Methods

	Estimated Means								
Imputation Methods	X_1 *Delivery Speed*	X_2 *Price Level*	X_3 *Price Flexibility*	X_4 *Manufacturer Image*	X_5 *Overall Service*	X_6 *Salesforce Image*	X_7 *Product Quality*	X_9 *Usage Level*	X_{10} *Satisfaction Level*
EM	3.71	2.03	8.11	5.15	2.82	2.60	6.84	45.85	4.77
Regression	3.84	1.96	8.10	5.15	2.81	2.59	6.88	45.77	4.77
	Estimated Standard Deviations								
Imputation Methods	X_1 *Delivery Speed*	X_2 *Price Level*	X_3 *Price Flexibility*	X_4 *Manufacturer Image*	X_5 *Overall Service*	X_6 *Salesforce Image*	X_7 *Product Quality*	X_9 *Usage Level*	X_{10} *Satisfaction Level*
EM	1.15	1.00	1.27	1.16	.75	.71	1.68	9.29	.82
Regression	.99	.83	1.26	1.15	.75	.72	1.69	9.18	.82

vantages to the researcher. First, there should be no "hidden" impact on the results that need to be considered when interpreting the results. Second, any of the imputation methods can be applied as remedies for the missing data. Their selection need not be based on their ability to handle nonrandom processes, but instead on the applicability of the process and its impact on the results.

2. *Imputation is the most logical course of action.* Given the minimal benefit of deleting cases and variables, the researcher is precluded from the simple solution of deleting cases or variables. Moreover, the complete case method would result in an inadequate sample size. Some form of imputation is therefore needed to maintain an adequate sample size for any multivariate analysis.
3. *Imputed correlations differ across techniques.* When estimating correlations among the variables in the presence of missing data, the researcher can choose from the three most common techniques: the complete information method, the all-available information method, and the mean substitution method. The researcher is faced in this situation, however, with differences in the results among these three methods. The all-available information and mean substitution approaches lead to generally consistent results, although the mean substitution values are generally somewhat lower in magnitude. Notable differences are found between these two approaches and the complete information approach. While the complete information approach would seem the most "safe" and conservative, in this case it is not recommended due to the small sample used (only 26 observations) and its marked differences from the other two methods. The researcher should choose, if necessary, among the two other approaches.
4. *Multiple methods for replacing the missing data are available and appropriate.* As mentioned above, mean substitution is one acceptable means of generating replacement values for the missing data. The researcher also has available the regression and EM imputation methods, each of which give reasonably consistent estimates for most variables. The presence of three acceptable methods also allows the researcher to combine the three estimates into a single composite, hopefully mitigating any effects strictly due to one of the methods.

In conclusion, the analytical tools and the diagnostic processes presented in the earlier section have provided an adequate basis for understanding and accommodating the missing data found in the pretest data. As this example demonstrates, the researcher need not fear that missing data will always preclude a multivariate analysis or always limit the generalizability of the results. Instead, the possibly "hidden" impact of missing data can be identified and actions taken to minimize the effect of missing data on the analyses performed.

Summary

The procedures available for handling missing data are varied in form, complexity, and intent. The researcher must always be prepared to assess and deal with missing data, as it is frequently encountered in multivariate analysis. The decision to use only observations with complete data may seem to be conservative and "safe," but as the preceding discussion illustrated, there are inherent limitations and biases in this and the other approaches. The researcher has no single method best suited in every situation, but instead must make a reasoned judgment of the situation, considering all of the factors described above.

Outliers

Outliers are observations with a unique combination of characteristics identifiable as distinctly different from the other observations. Outliers cannot be categorically characterized as either beneficial or problematic, but instead must be viewed within the context of the analysis and should be evaluated by the types of information they may provide. When beneficial, outliers—although different from the majority of the sample—may be indicative of characteristics of the population that would not be discovered in the normal course of analysis. In contrast, problematic outliers are not representative of the population, are counter to the objectives of the analysis, and can seriously distort statistical tests. Owing to the variability in the impact of outliers, it is imperative that the researcher examine the data for the presence of outliers to ascertain their type of influence. The reader is also referred to the discussions in chapter 4 and the appendix to that chapter, which relate to the topic of influential observations. In these discussions, outliers are placed in a framework particularly suited for assessing the influence of individual observations and determining whether this influence is helpful or harmful.

Why do outliers occur? Outliers can be classified into one of four classes. The first class arises from a procedural error, such as a data entry error or a mistake in coding. These outliers should be identified in the data cleaning stage, but if overlooked, they should be eliminated or recorded as missing values. The second class of outlier is the observation that occurs as the result of an extraordinary event, which then is an explanation for the uniqueness of the observation. The researcher must decide whether the extraordinary event should be represented in the sample. If so, the outlier should be retained in the analysis; if not, it should be deleted. The third class of outlier comprises extraordinary observations for which the researcher has no explanation. Although these are the outliers most likely to be omitted, they may be retained if the researcher feels they represent a valid segment of the population. The fourth and final class of outlier contains observations that fall within the ordinary range of values on each of the variables

but are unique in their combination of values across the variables. In these situations, the researcher should retain the observation unless specific evidence is available that discounts the outlier as a valid member of the population.

The following sections detail the methods used in detecting outliers in univariate, bivariate, and multivariate situations. Once the outliers have been identified, they may be profiled to aid in placing them into one of the four classes described above. Finally, the researcher must decide on the retention or exclusion of each outlier, judging not only from the characteristics of the outlier but also from the objectives of the analysis.

Detecting Outliers

Outliers can be identified from a univariate, bivariate, or multivariate perspective. The researcher should utilize as many of these perspectives as possible, looking for a consistent pattern across methods to identify outliers. The following discussion details the processes involved in each of the three perspectives.

Univariate Detection

The univariate perspective for identifying outliers examines the distribution of observations and selects as outliers those cases falling at the outer ranges of the distribution. The primary issue is establishing the threshold for designation of an outlier. The typical approach first converts the data values to standard scores, which have a mean of 0 and a standard deviation of 1. Because the values are expressed in a standardized format, comparisons across variables can be made easily. For small samples (80 or fewer observations), the guidelines suggest identifying those cases with standard scores of 2.5 or greater as outliers. When the sample sizes are larger, the guidelines suggest that the threshold value of standard scores range from 3 to 4. If standard scores are not used, then the researcher can identify cases falling outside the ranges of 2.5 versus 3 or 4 standard deviations, depending on the sample size. In either case, the researcher must recognize that a certain number of observations may occur normally in these outer ranges of the distribution. The researcher should strive to identify only those truly distinctive observations and designate them as outliers.

Bivariate Detection

In addition to the univariate assessment, pairs of variables can be assessed jointly through a scatterplot. Cases that fall markedly outside the range of the other observations can be noted as isolated points in the scatterplot. To assist in determining the expected range of observations, an ellipse representing a specified confidence interval (varying between 50 and 90 percent of the distribution) for a bivariate normal distribution can be superimposed over the scatterplot. This provides a graphical portrayal of the confidence limits and facilitates identification of the outliers. Another variant of the scatterplot is termed the influence plot. In this type of scatterplot, each point varies in size in relation to its influence on the relationship. These methods provide some assessment of the influence of each observation to complement the designation of cases as outliers.

Multivariate Detection

The third perspective for identifying outliers involves a multivariate assessment of each observation across a set of variables. Because most multivariate analyses involve more than two variables, the researcher needs a means to objectively

measure the multidimensional position of each observation relative to some common point. The Mahalanobis D^2 measure can be used for this purpose. Mahalanobis D^2 is a measure of the distance in multidimensional space of each observation from the mean center of the observations. It provides a common measure of multidimensional centrality and also has statistical properties that allow for significance testing. Given the nature of the statistical tests, it is suggested that a very conservative level, such as .001, be used as the threshold value for designation as an outlier.

Outlier Designation

When observations that are candidates for designation as an outlier have been identified by the univariate, bivariate, or multivariate methods, the researcher must then select observations that demonstrate real uniqueness in comparison with the remainder of the population. The researcher should refrain from designating too many observations as outliers and not succumb to the temptation of eliminating those cases not consistent with the remaining cases just because they are different.

Outlier Description and Profiling

Once the potential outliers have been identified, the researcher should generate profiles on each outlier observation and carefully examine the data for the variable(s) responsible for its being an outlier. In addition to this visual examination, the researcher can also employ multivariate techniques such as discriminant analysis or multiple regression to identify the differences between outliers and the other observations. The researcher should continue this analysis until satisfied with the aspects of the data that distinguish the outlier from the other observations. If possible the researcher should assign the outlier to one of the four classes described earlier.

Retention or Deletion of the Outlier

After the outliers have been identified, profiled, and categorized, the researcher must decide on the retention or deletion of each one. There are many philosophies among researchers as to how to deal with outliers. Our belief is that they should be retained unless there is demonstrable proof that they are truly aberrant and not representative of any observations in the population. But if they do represent a segment of the population, they should be retained to ensure generalizability to the entire population. As outliers are deleted, the researcher runs the risk of improving the multivariate analysis but limiting its generalizability. If outliers are problematic in a particular technique, many times they can be accommodated in the analysis in a manner in which they do not seriously distort the analysis.

An Illustrative Example of Analyzing Outliers

As an example of outlier detection, the observations of the HATCO database introduced in chapter 1 are examined here for outliers. The variables considered in the analysis are the metric variables X_1, X_2, X_3, X_4, X_5, X_6, X_7, and X_9. The outlier analysis includes univariate, bivariate, and multivariate diagnoses. If candidates for outlier designation are found, they are examined, and a decision on retention or deletion is made.

Univariate and Bivariate Detection

The first step is to examine the observations on each of the variables individually. Table 2.9 contains the observations with standardized variable values exceeding ±2.5. From this univariate perspective, a few observations exceed the threshold on a single variable, but no observation was a univariate outlier on more than one variable. For a bivariate perspective, scatterplots are formed for X_1, X_2, X_3, X_4, X_5, X_6, and X_7 versus X_9, one of the metric variables used as a dependent variable in many of the multivariate techniques. An ellipse representing the 90 percent confidence interval of a bivariate normal distribution is then superimposed on the scatterplot (see Figure 2.6, p. 68). The second part of Table 2.9 contains observations falling outside this ellipse. This is a 90 percent confidence interval; thus we would expect some observations normally to fall outside the ellipse. However, some observations (3, 5, 57, and 96) appear several times, perhaps indicating they are bivariate outliers.

Multivariate Detection

The final diagnostic method is to assess multivariate outliers with the Mahalanobis D^2 measure (see Table 2.10, p. 69). This evaluates the position of each observation compared with the center of all observations on a set of variables. In this case, all the metric variables were used for the evaluation of observations. As noted earlier, the statistical tests for significance with this measure should be very conservative (exceeding .001). With this threshold, two observations (22 and 55) are identified as significantly different. It is interesting that these observations were not seen in earlier univariate and bivariate analyses but appear only in the multivariate tests. This indicates they are not unique on any single variable but instead are unique in combination.

Retention or Deletion of the Outliers

As a result of these diagnostic tests, no observations seem to demonstrate the characteristics of outliers that should be eliminated. Each variable has some observations that are extreme, and they should be considered if that variable is used in an analysis. But no observations are extreme on a sufficient number of variables

TABLE 2.9 Identification of Univariate and Bivariate Outliers

Univariate Outliers Cases with Standardized Values (Z scores) Exceeding ± 2.5	
Variable	*Cases*
X_1	39
X_2	71
X_3	none
X_4	82
X_5	96
X_6	5, 42
X_7	none
X_9	none
X_{10}	none

Bivariate Outliers Cases Lying Outside the 90% Confidence Interval Ellipse	
X_9 *with*	*Cases*
X_1	1, 39, 95, 96
X_2	3, 49, 57, 7, 96, 97
X_3	11, 57, 96, 100
X_4	5, 22, 42, 50, 72, 82, 93, 96
X_5	3, 22, 39, 57, 71, 96
X_6	5, 7, 42, 82, 96
X_7	57, 58, 95, 96

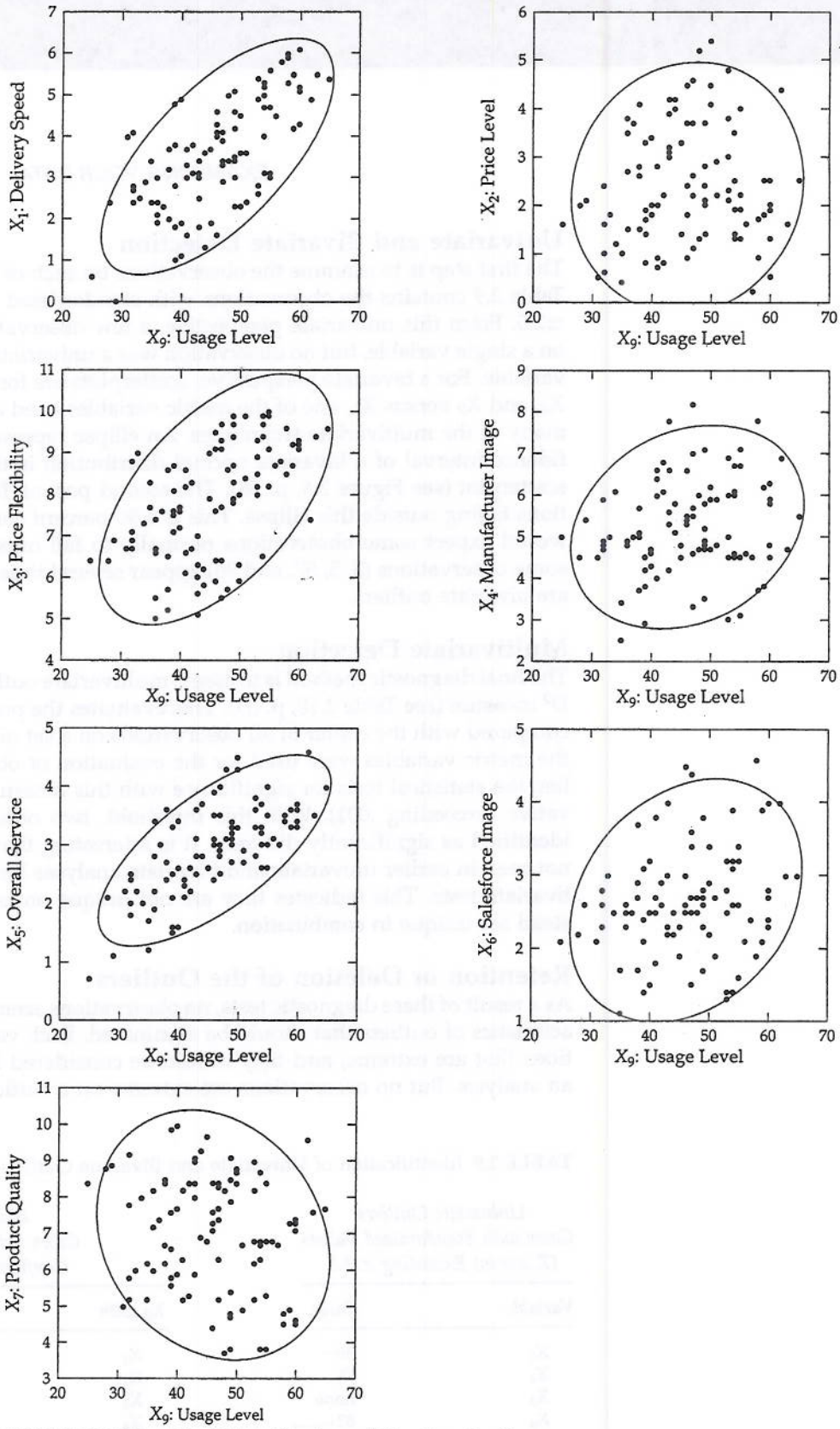


FIGURE 2.6 Graphical Identification of Bivariate Outliers

Chi Square

TABLE 2.10 Identification of Multivariate Outliers

Case Number	Mahalanobis D^2	D^2/df	df	Significance	Case Number	Mahalanobis D^2	D^2/df	df	Significance
1	7.031	1.004	7	0.4256	51	6.362	0.909	7	0.4982
2	6.691	0.956	7	0.4617	52	8.467	1.210	7	0.2932
3	7.567	1.081	7	0.3723	53	6.913	0.988	7	0.4380
4	7.103	1.015	7	0.4182	54	3.244	0.463	7	0.8615
5	12.870	1.839	7	0.0753	55	35.197	5.028	7	0.0000
6	.517	0.931	7	0.4809	56	3.082	0.440	7	0.8773
7	8.634	1.233	7	0.2800	57	10.488	1.498	7	0.1626
8	6.563	0.938	7	0.4758	58	5.265	0.752	7	0.6276
9	6.375	0.911	7	0.4967	59	4.348	0.621	7	0.7390
10	3.626	0.518	7	0.8217	60	7.012	1.002	7	0.4276
11	4.237	0.605	7	0.7522	61	13.001	1.857	7	0.0721
12	3.389	0.484	7	0.8468	62	5.798	0.828	7	0.5635
13	3.768	0.538	7	0.8061	63	3.322	0.475	7	0.8537
14	5.030	0.719	7	0.6563	64	6.926	0.989	7	0.4367
15	8.962	1.280	7	0.2554	65	11.683	1.669	7	0.1115
16	6.398	0.914	7	0.4942	66	2.109	0.301	7	0.9536
17	7.212	1.030	7	0.4071	67	4.382	0.626	7	0.7349
18	5.350	0.764	7	0.6173	68	5.925	0.846	7	0.5486
19	5.899	0.843	7	0.5516	69	4.878	0.697	7	0.6749
20	8.962	1.280	7	0.2554	70	5.057	0.722	7	0.6530
21	2.978	0.425	7	0.8870	71	8.294	1.185	7	0.3074
22	35.390	5.056	7	0.0000	72	10.095	1.442	7	0.1833
23	8.333	1.190	7	0.3042	73	5.887	0.841	7	0.5530
24	2.974	0.425	7	0.8874	74	5.363	0.766	7	0.6157
25	4.909	0.701	7	0.6711	75	6.471	0.924	7	0.4859
26	3.463	0.495	7	0.8391	76	4.925	0.704	7	0.6691
27	3.171	0.453	7	0.8687	77	5.847	0.835	7	0.5577
28	5.765	0.824	7	0.5674	78	7.522	1.075	7	0.3766
29	7.601	1.086	7	0.3691	79	12.279	1.754	7	0.0918
30	5.188	0.741	7	0.6370	80	2.270	0.324	7	0.9434
31	2.751	0.393	7	0.9071	81	4.943	0.706	7	0.6669
32	7.024	1.003	7	0.4264	82	14.118	2.017	7	0.0491
33	5.678	0.811	7	0.5778	83	6.837	0.977	7	0.4460
34	3.529	0.504	7	0.8321	84	2.366	0.338	7	0.9369
35	6.539	0.934	7	0.4784	85	3.016	0.431	7	0.8835
36	2.900	0.414	7	0.8941	86	3.493	0.499	7	0.8359
37	6.704	0.958	7	0.4603	87	3.354	0.479	7	0.8504
38	3.030	0.433	7	0.8823	88	2.417	0.345	7	0.9332
39	10.213	1.459	7	0.1768	89	6.011	0.859	7	0.5385
40	3.827	0.547	7	0.7995	90	4.860	0.694	7	0.6771
41	2.898	0.414	7	0.8943	91	3.763	0.538	7	0.8067
42	12.282	1.755	7	0.0917	92	5.841	0.834	7	0.5584
43	7.129	1.018	7	0.4156	93	14.328	2.047	7	0.0456
44	4.819	0.688	7	0.6821	94	5.407	0.772	7	0.6105
45	6.670	0.953	7	0.4640	95	7.391	1.056	7	0.3893
46	7.475	1.068	7	0.3811	96	16.708	2.387	7	0.0194
47	14.094	2.013	7	0.0495	97	8.195	1.171	7	0.3157
48	6.152	0.879	7	0.5221	98	4.990	0.713	7	0.6612
49	7.561	1.080	7	0.3729	99	5.587	0.798	7	0.5888
50	9.029	1.290	7	0.2506	100	4.704	0.672	7	0.6960

df = degrees of freedom

Mahalanobis D^2 value based on the following variables (X_1, X_2, X_3, X_4, X_5, X_6, and X_7). The D^2/df value is approximately distributed as a t value.

to be considered unrepresentative of the population. In all instances, the observations designated as outliers, even with the multivariate tests, seem similar enough to the remaining observations to be retained in the multivariate analyses. However, the researcher should always examine the results of each specific multivariate technique to identify observations that may become outliers in that particular application.

Testing the Assumptions of Multivariate Analysis

The final step in examining the data involves testing the assumptions underlying multivariate analysis. The need to test the statistical assumptions is increased in multivariate applications because of two characteristics of multivariate analysis. First, the complexity of the relationships, owing to the typical use of a large number of variables, makes the potential distortions and biases more potent when the assumptions are violated. This is particularly true when the violations compound to become even more detrimental than if considered separately. Second, the complexity of the analyses and of the results may mask the "signs" of assumption violations apparent in the simpler univariate analyses. In almost all instances, the multivariate procedures will estimate the multivariate model and produce results even when the assumptions are severely violated. Thus, the researcher must be aware of any assumption violations and the implications they may have for the estimation process or the interpretation of the results.

Assessing Individual Variables versus the Variate

Multivariate analysis requires that the assumptions underlying the statistical techniques be tested twice: first for the separate variables, akin to the tests of assumption for univariate analyses, and second for the multivariate model **variate,** which acts collectively for the variables in the analysis and thus must meet the same assumptions as individual variables. This chapter focuses on the examination of individual variables for meeting the assumptions underlying the multivariate procedures. Discussions in each chapter address the methods used to assess the assumptions underlying the variate for each multivariate technique.

Normality

The most fundamental assumption in multivariate analysis is **normality,** referring to the shape of the data distribution for an individual metric variable and its correspondence to the **normal distribution,** the benchmark for statistical methods. If the variation from the normal distribution is sufficiently large, all resulting statistical tests are invalid, as normality is required to use the F and t statistics. Both the univariate and the multivariate statistical methods discussed in this text are based on the assumption of univariate normality, with the multivariate methods also assuming multivariate normality. Univariate normality for a single variable is easily tested, and a number of corrective measures are possible, as shown later. In a simple sense, multivariate normality (the combination of two or more variables) means that the individual variables are normal in a univariate sense and

that their combinations are also normal. Thus, if a variable is multivariate normal, it is also univariate normal. However, the reverse is not necessarily true (two or more univariate normal variables are not necessarily multivariate normal). Thus a situation in which all variables exhibit univariate normality will help gain, although not guarantee, multivariate normality. Multivariate normality is more difficult to test, but some tests are available for situations in which the multivariate technique is particularly affected by a violation of this assumption. In this text, we focus on assessing and achieving univariate normality for all variables, and address multivariate normality only when it is especially critical. Even though large sample sizes tend to diminish the detrimental effects of nonnormality, the researcher should assess the normality for all variables included in the analysis.

Graphical Analyses of Normality

The simplest diagnostic test for normality is a visual check of the histogram that compares the observed data values with a distribution approximating the normal distribution (see Figure 2.1). Although appealing because of its simplicity, this method is problematic for smaller samples, where the construction of the histogram (e.g., the number of categories or the width of categories) can distort the visual portrayal to such an extent that the analysis is useless. A more reliable approach is the **normal probability plot,** which compares the cumulative distribution of actual data values with the cumulative distribution of a normal distribution. The normal distribution forms a straight diagonal line, and the plotted data values are compared with the diagonal. If a distribution is normal, the line representing the actual data distribution closely follows the diagonal.

Figure 2.7 (p. 72) shows several normal probability plots and the corresponding univariate distribution of the variable. One characteristic of the distribution's shape, the kurtosis, is reflected in the normal probability plots. **Kurtosis** refers to the "peakedness" or "flatness" of the distribution compared with the normal distribution. When the line falls below the diagonal, the distribution is flatter than expected. When it goes above the diagonal, the distribution is more peaked than the normal curve. For example, in the normal probability plot of a peaked distribution (Figure 2.7d), we see a distinct S-shaped curve. Initially the distribution is flatter, and the plotted line falls below the diagonal. Then the peaked part of the distribution rapidly moves the plotted line above the diagonal, and eventually the line shifts to below the diagonal again as the distribution flattens. A nonpeaked distribution has the opposite pattern (Figure 2.7c). Another common pattern is a simple arc, either above or below the diagonal, indicating the **skewness** of the distribution. A negative skewness (Figure 2.7e) is indicated by an arc below the diagonal, whereas an arc above the diagonal represents a positively skewed distribution (Figure 2.7f). An excellent source for interpreting normal probability plots, showing the various patterns and interpretations is Daniel and Wood [5]. These specific patterns not only identify nonnormality but also tell us the form of the original distribution and the appropriate remedy to apply.

Statistical Tests of Normality

In addition to examining the normal probability plot, one can also use statistical tests to assess normality. A simple test is a rule of thumb based on the skewness and kurtosis values (available as part of the basic descriptive statistics for a

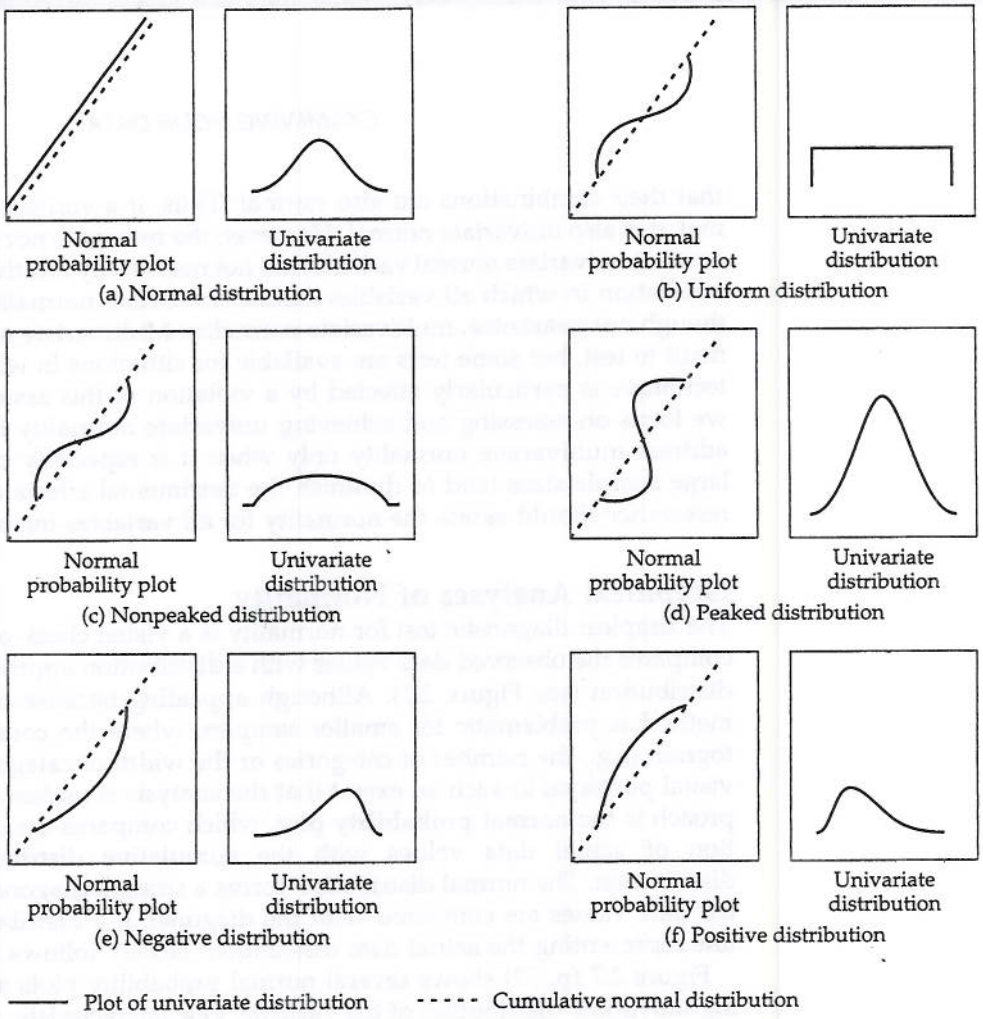


FIGURE 2.7 Normal Probability Plots and Corresponding Univariate Distributions

variable computed by all statistical programs). The statistic value (z) for the skewness value is calculated as:

$$z_{\text{skewness}} = \frac{\text{skewness}}{\sqrt{\frac{6}{N}}}$$

where N is the sample size. A z value can also be calculated for the kurtosis value using the following formula:

$$z_{\text{kurtosis}} = \frac{\text{kurtosis}}{\sqrt{\frac{24}{N}}}$$

If the calculated z value exceeds a critical value, then the distribution is nonnormal in terms of that characteristic. The critical value is from a z distribution, based on the significance level we desire. For example, a calculated value exceeding

± 2.58 indicates we can reject the assumption about the normality of the distribution at the .01 probability level. Another commonly used critical value is ±1.96, which corresponds to a .05 error level.

Specific statistical tests are also available in SPSS, SAS, BMDP, and most other programs. The two most common are the Shapiro-Wilks test and a modification of the Kolmogorov-Smirnov test. Each calculates the level of significance for the differences from a normal distribution. The researcher should always remember that tests of significance are less useful in small samples (fewer than 30) and quite sensitive in large samples (exceeding 1,000 observations). Thus, the researcher should always use both the graphical plots and any statistical tests to assess the actual degree of departure from normality.

Remedies for Nonnormality

A number of data transformations available to accommodate nonnorm
butions are discussed later in the chapter. This chapter confines the disc
univariate normality tests and transformations. However, when we exam
multivariate methods, such as multivariate regression or multivariate ar
variance, we discuss tests for multivariate normality as well. Moreov
times when nonnormality is indicated, it is actually the result of other
tion violations; therefore, remedying the other violations eliminates the
mality problem. For this reason, the researcher should perform norma
after or concurrently with analyses and remedies for other violations. (F
interested in multivariate normality, see references [8, 11].)

Homoscedasticity

Homoscedasticity is an assumption related primarily to dependence relationships between variables. It refers to the assumption that dependent variable(s) exhibit equal levels of variance across the range of predictor variable(s). Homoscedasticity is desirable because the variance of the dependent variable being explained in the dependence relationship should not be concentrated in only a limited range of the independent values. Although the dependent variables must be metric, this concept of an equal spread of variance across independent variables can be applied when the independent variables are either metric or nonmetric. With metric independent variables, the concept of homoscedasticity is based on the spread of dependent variable variance across the range of independent variable values, which is encountered in techniques such as multiple regression. The same concept also applies when the independent variables are nonmetric. In these instances, such as is found in ANOVA and MANOVA, the focus now becomes the equality of the variance (single dependent variable) or the variance/covariance matrices (multiple independent variables) across the groups formed by the nonmetric independent variables. The equality of variance/covariance matrices is also seen in discriminant analysis, but in this technique the emphasis is on the spread of the independent variables across the groups formed by the nonmetric dependent measure. In each of these instances, the purpose is the same: to ensure that the variance used in explanation and prediction is distributed across the range of values, thus allowing for a "fair test" of the relationship across all values of the nonmetric variables.

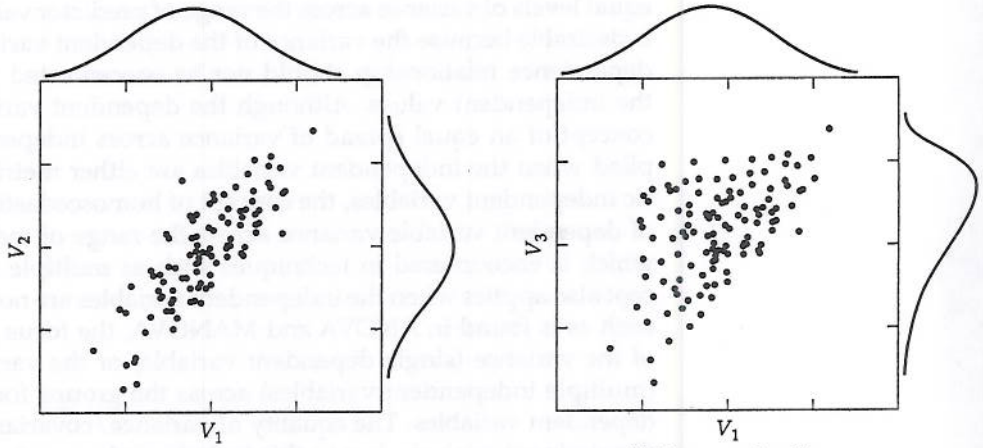
In most situations, we have many different values of the dependent variable at each value of the independent variable. For this relationship to be fully captured, the dispersion (variance) of the dependent variable values must be equal at each

value of the predictor variable. Most problems with unequal variances stem from one of two sources. The first source is the type of variables included in the model. For example, as a variable increases in value (e.g., units ranging from near zero to millions), there is a naturally wider range of possible answers for the larger values. The second source results from a skewed distribution that creates heteroscedasticity. In Figure 2.8a, the scatterplots of data points for two variables (V_1 and V_2) with normal distributions exhibit equal dispersion across all data values (i.e., homoscedasticity). However, in Figure 2.8b, we see unequal dispersion **(heteroscedasticity)** caused by skewness of one of the variables (V_3). For the different values of V_3, there are different patterns of dispersion for V_1. This will cause the predictions to be better at some levels of the independent variable than at others. Violating this assumption often makes hypothesis tests either too conservative or too sensitive.

The effect of heteroscedasticity is also often related to sample size, especially when examining the variance dispersion across groups. For example, in ANOVA or MANOVA, the impact of heteroscedasticity on the statistical test depends on the sample sizes associated with the groups of smaller and larger variances. In multiple regression analysis, similar effects would occur in highly skewed distributions where there were disproportionate numbers of respondents in certain ranges of the independent variable.

Graphical Tests of Equal Variance Dispersion

The test of homoscedasticity for two metric variables is best examined graphically. The most common application of this form of assessment occurs in multiple regression, which is concerned with the dispersion of the dependent variable across the values of the metric independent variables. Because the focus of regression analysis is on the regression variate, the graphical plot of residuals is used to reveal the presence of homoscedasticity (or its opposite, heteroscedasticity). The discussion of residual analysis in chapter 4 details these procedures. Boxplots work well to represent the degree of variation between groups formed by a categorical variable. The length of the box and the whiskers each portray the variation of data within that group.



(a) Homoscedasticity (b) Heteroscedasticity

FIGURE 2.8 Scatterplots of Homoscedastic and Heteroscedastic Relationships

Statistical Tests for Homoscedasticity

The statistical tests for equal variance dispersion relate to the variances within groups formed by nonmetric variables. The most common test, the Levene test, can be used to assess whether the variances of a single metric variable are equal across any number of groups. If more than one metric variable is being tested, so that the comparison involves the equality of variance/covariance matrices, the Box's *M* test is applicable. The Box's *M* test is available in both multivariate analysis of variance and discriminant analysis and is discussed in more detail in later chapters pertaining to these techniques.

Remedies for Heteroscedasticity

Heteroscedastic variables can be remedied through data transformations similar to those used to achieve normality. As mentioned earlier, many times heteroscedasticity is the result of nonnormality of one of the variables, and correction of the nonnormality also remedies the unequal dispersion of variance. A later section discusses data transformations of the variables to make all values have a potentially equal effect in prediction.

Linearity

An implicit assumption of all multivariate techniques based on correlational measures of association, including multiple regression, logistic regression, factor analysis, and structural equation modeling, is **linearity**. Because correlations represent only the linear association between variables, nonlinear effects will not be represented in the correlation value. This results in an underestimation of the actual strength of the relationship. It is always prudent to examine all relationships to identify any departures from linearity that may impact the correlation.

Identifying Nonlinear Relationships

The most common way to assess linearity is to examine scatterplots of the variables and to identify any nonlinear patterns in the data. An alternative approach is to run a simple regression analysis (the specifics of this technique are covered in chapter 4) and to examine the residuals. The residuals reflect the unexplained portion of the dependent variable; thus, any nonlinear portion of the relationship will show up in the residuals. The examination of residuals can also be applied to multiple regression, where the researcher can detect any nonlinear effects not represented in the regression variate. A more detailed discussion of residual analysis is included in chapter 4.

Remedies for Nonlinearity

If a nonlinear relationship is detected, the most direct approach is to transform one or both variables to achieve linearity. A number of available transformations are discussed later in this chapter. An alternative to data transformation is the creation of new variables to represent the nonlinear portion of the relationship. The process of creating and interpreting these additional variables, which can be used in all linear relationships, is discussed in chapter 4.

Absence of Correlated Errors

Predictions in any of the dependence techniques are not perfect, and we will rarely find a situation in which they are. However, we do attempt to ensure that any prediction errors are uncorrelated with each other. For example, if we found a pattern

that suggests every other error is positive while the alternative error terms are negative, we would know that some unexplained systematic relationship exists in the dependent variable. If such a situation exists, we cannot be confident that our prediction errors are independent of the levels at which we are trying to predict. Some other factor is affecting the results, but is not included in the analysis.

Identifying Correlated Errors

The most common violations of the assumption that errors are uncorrelated are due to the data collection process. Similar factors that affect one group may not affect the other. If the groups are analyzed separately, the effects are constant within each group and do not impact the estimation of the relationship. But if the observations from both groups are combined, then the final estimated relationship must be a "compromise" between the two actual relationships. This causes the results to be biased because an unspecified cause is impacting the estimation of the relationship.

To identify correlated errors, the researcher must first identify possible causes. In our earlier example, this would be the two separate groups in data collection. Once the potential cause is identified, the researcher could see if differences did exist between the groups. Finding differences in the prediction errors in the two groups would then be the basis for determining that an unspecified effect was "causing" the correlated errors.

Remedies for Correlated Errors

Correlated errors must be corrected by including the omitted causal factor into the multivariate analysis. In our earlier example, the researcher would add a variable indicating in which class the respondents were. The most common remedy is the addition of a variable(s) to the analysis that represents the omitted factor. The key task facing the researcher is not the actual remedy, but rather the identification of the unspecified effect and a means of representing it in the analysis.

Data Transformations

Data transformations provide a means of modifying variables for one of two reasons: (1) to correct violations of the statistical assumptions underlying the multivariate techniques, or (2) to improve the relationship (correlation) between variables. Data transformations may be based on reasons that are either "theoretical" (transformations whose appropriateness is based on the nature of the data) or "data derived" (where the transformations are suggested strictly by an examination of the data). Yet in either case the researcher must proceed many times by trial and error, monitoring the improvement versus the need for additional transformations.

All the transformations described here are easily carried out by simple commands in the popular statistical packages. We focus on transformations that can be computed in this manner, although more sophisticated and complicated methods of data transformation are available (e.g., see Box and Cox [2]).

Transformations to Achieve Normality and Homoscedasticity

Data transformations provide the principal means of correcting nonnormality and heteroscedasticity. In both instances, patterns of the variables suggest specific transformations. For nonnormal distributions, the two most common patterns are "flat" distributions and skewed distributions. For the flat distribution, the most

common transformation is the inverse (e.g., $1/Y$ or $1/X$). Skewed distributions can be transformed by taking the square root, logarithms, or even the inverse of the variable. Usually negatively skewed distributions are best transformed by employing a square root transformation, whereas the logarithm typically works best on positive skewness. In any instance, the researcher should apply all of the possible transformations and then select the most appropriate transformed variable.

Heteroscedasticity is an associated problem, and in many instances "curing" this problem will deal with normality problems as well. Heteroscedasticity is also due to the distribution of the variable(s). When examining the residuals of regression analysis for heteroscedasticity, we note that an indication of unequal variance is a cone-shaped distribution of the residuals (see chapter 4 for more specific details of the graphical analysis of residuals). If the cone opens to the right, take the inverse; if the cone opens to the left, take the square root. Some transformations can be associated with certain types of data. For example, frequency counts suggest a square root transformation; proportions are best transformed by the arcsin transformation ($X_{new} = 2 \arcsin \sqrt{X_{old}}$); and proportional change is best handled by taking the logarithm of the variable. In all instances, once the transformations have been performed, the transformed data should be tested to see whether the desired remedy was achieved.

Transformations to Achieve Linearity

There are numerous procedures for achieving linearity between two variables, but most simple nonlinear relationships can be placed in one of four categories (see Figure 2.9). In each quadrant, the potential transformations for both dependent

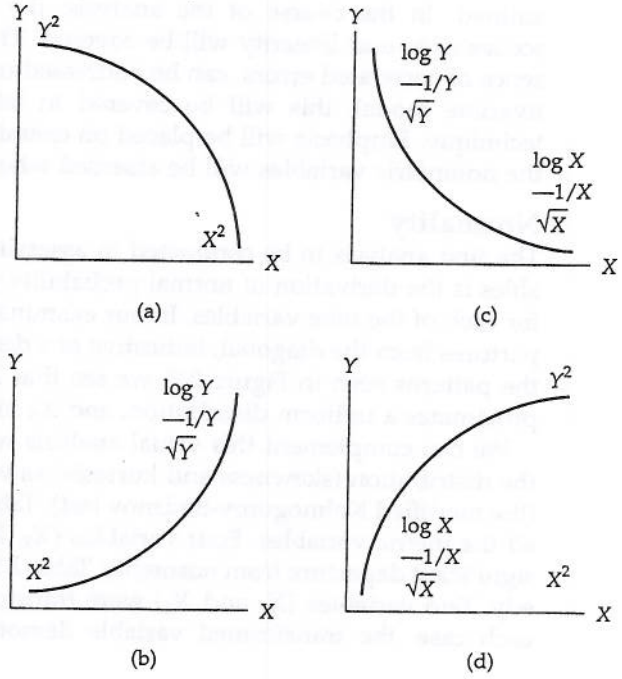


FIGURE 2.9 Selecting Transformations to Achieve Linearity

Source: F. Mosteller and J. W. Tukey, *Data Analysis and Regression.* Reading, Mass.: Addison-Wesley, 1977.

and independent variables are shown. For example, if the relationship looks like that in Figure 2.9a, then either variable can be squared to achieve linearity. When multiple transformation possibilities are shown, start with the top method in each quadrant and move downward until linearity is achieved. An alternative approach is to use additional variables, termed polynomials, to represent the nonlinear components. This method is discussed in more detail in chapter 4.

General Guidelines for Transformations

There are several points to remember when performing data transformations:

1. For a noticeable effect from transformations, the ratio of a variable's mean to its standard deviation should be less than 4.0.
2. When the transformation can be performed on either of two variables, select the variable with the smallest ratio from item 1.
3. Transformations should be applied to the independent variables except in the case of heteroscedasticity.
4. Heteroscedasticity can be remedied only by transformation of the dependent variable in a dependence relationship. If a heteroscedastic relationship is also nonlinear, the dependent variable, and perhaps the independent variables, must be transformed.
5. Transformations may change the interpretation of the variables. For example, transforming variables by taking their logarithm translates the relationship into a measure of proportional change (elasticity). Always be sure to explore thoroughly the possible interpretations of the transformed variables.

An Illustration of Testing the Assumptions Underlying Multivariate Analysis

To illustrate the techniques involved in testing the data for meeting the assumptions underlying multivariate analysis and to provide a foundation for use of the data in the subsequent chapters, the data set introduced in chapter 1 will be examined. In the course of the analysis, the assumptions of normality, homoscedasticity, and linearity will be covered. The fourth basic assumption, the absence of correlated errors, can be addressed only in the context of a specific multivariate model; this will be covered in later chapters for each multivariate technique. Emphasis will be placed on examining the metric variables, although the nonmetric variables will be assessed where appropriate.

Normality

The first analysis to be conducted in assessing the normality of the metric variables is the derivation of normal probability plots. Figure 2.10 contains the plots for each of the nine variables. In our examination of the graphs, we see some departures from the diagonal, indicative of a departure from normality. Referring to the patterns seen in Figure 2.7, we see that X_2 seems positively skewed, X_3 approximates a uniform distribution, and X_5 seems negatively skewed.

We can complement this visual analysis with statistics reflecting the shape of the distribution (skewness and kurtosis) as well as a statistical test for normality (the modified Kolmogorov-Smirnov test). Table 2.11 (p. 80) shows these values for all the metric variables. Four variables (X_2, X_3, X_4, and X_6) exhibit a statistically significant departure from normality. Table 2.11 also suggests the appropriate remedy. Two variables (X_2 and X_6) were transformed by taking the square root. In each case, the transformed variable demonstrated normality (see Table 2.11).

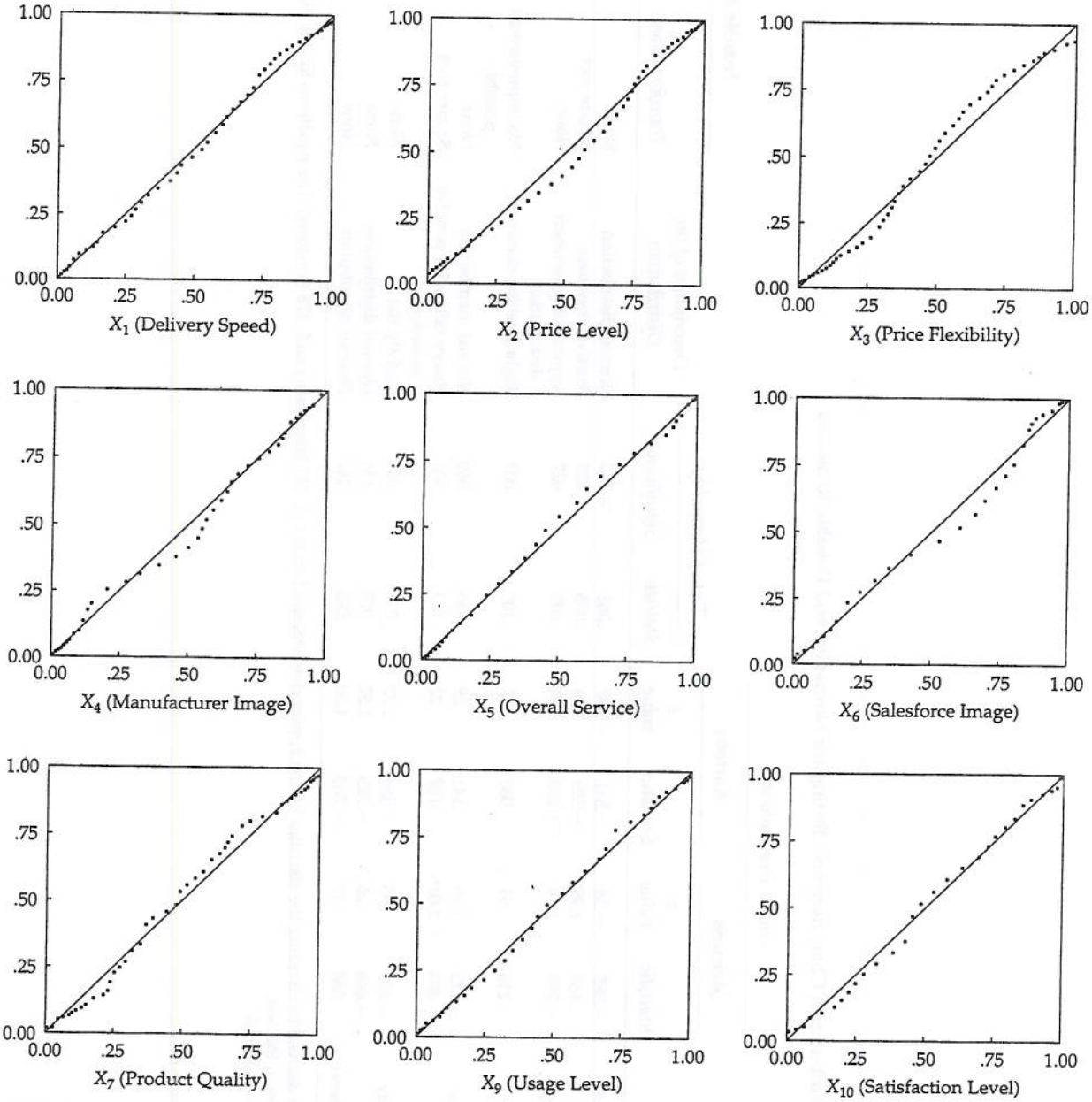


FIGURE 2.10 Normal Probability Plots of Metric Variables

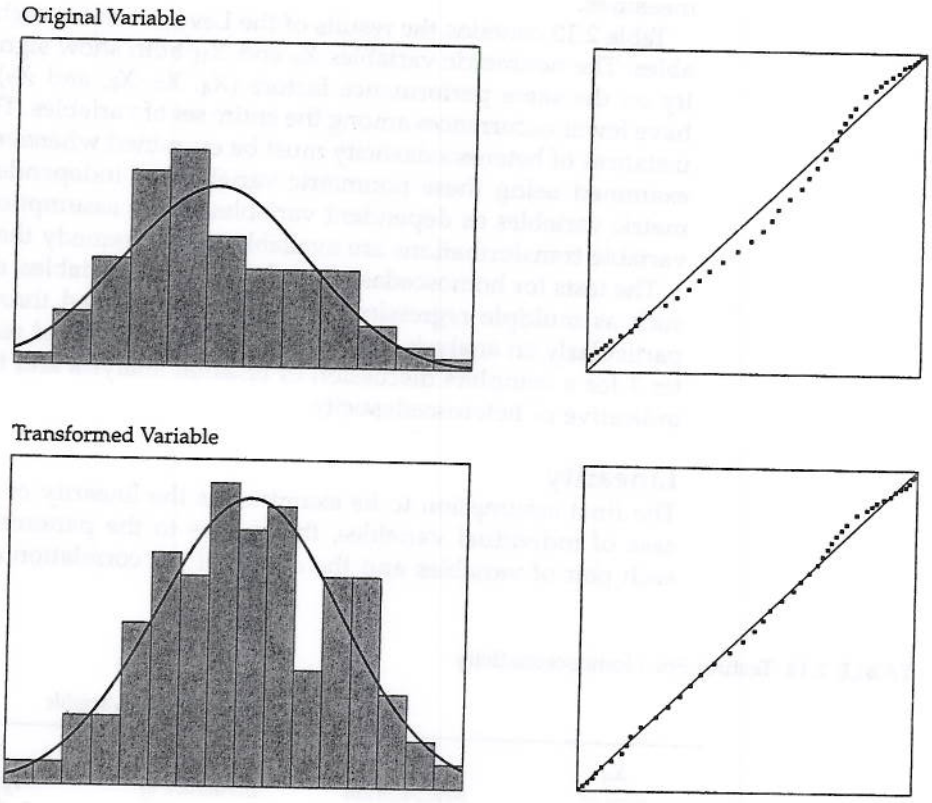
Figure 2.11 (p. 81) demonstrates the effect of the transformation on X_2 in achieving normality. The transformed X_2 appears markedly more normal in both graphical portrayals, and the statistical descriptors are also improved. The researcher should always examine the transformed variables as rigorously as the original variables in terms of their normality and distribution shape.

In the case of the two remaining variables (X_3 and X_4), none of the transformations could improve the normality. These variables will have to be used in their original form. In situations where the normality of the variables is critical, the

TABLE 2.11 Distributional Characteristics, Testing for Normality, and Possible Remedies

	Shape Descriptors[a]								
	Skewness		*Kurtosis*		*Test of Normality*			*Possible Remedies*	
Variable	*Statistic*	*z value*	*Statistic*	*z value*	*Statistic*	*Significance*	*Description of the Distribution*	*Transformation*	*Significance After Remedy*
X_1 Delivery speed	−.085	−.35	−.511	−1.07	.063	>.200	Normal distribution	None	
X_2 Price level	.469	1.95*	−.509	1.06	.095	.028	Positive skewness	Square root	>.200
X_3 Price flexibility	−.289	1.19	−1.073	2.24*	.095	.027	Approaching uniform distribution	None	
X_4 Manufacturer image	.218	.91	.085	.18	.107	.007	Slight positive skewness	No improvement possible	
X_5 Overall service	−.373	1.55	.141	.29	.085	.069	Normal distribution	None	
X_6 Salesforce image	.493	2.04*	.107	.22	.122	.001	Heavy tails with positive skewness	Square root	.032
X_7 Product quality	−.229	.95	−.850	1.77	.091	.041	Slightly flat	None	
X_9 Usage level	−.069	.26	−.725	1.52	.079	.131	Normal distribution	None	
X_{10} Satisfaction level	.089	.37	−.763	1.60	.078	.142	Normal distribution	None	

[a]The z values are derived by dividing the statistics by the appropriate standard errors of .241 (skewness) and .478 (kurtosis). The equations for calculating the standard errors are given in the text.
*Significant at the .05 level.



Distribution Characteristics Before and After Transformation

	Shape Descriptors[a]				Test of Normality	
	Skewness		Kurtosis			
Variable Form	Statistic	z Value	Statistic	z Value	Statistic	Significance
Original X_2	.469	1.95	−.509	1.06	.095	.028
Transformed X_2	−.106	.44	−.465	.97	.062	> .200

[a]The z values are derived by dividing the statistics by the appropriate standard errors of 0.241 (skewness) and 0.478 (kurtosis). The equations for calculating the standard errors are given in the text.

FIGURE 2.11 Transformation of X_2 (Price Level) to Achieve Normality

transformed variables can be used with the assurance that they meet the assumptions of normality. But the departures from normality were not so extreme in any of the original variables that they should never be used in any analysis in their original form. If the technique has a robustness to departures from normality, then the original variables may be preferred for the comparability in the interpretation phase.

Homoscedasticity

All statistical packages have tests to assess homoscedasticity on a univariate basis (e.g., the Levene test in SPSS) where the variance of a metric variable is compared across levels of a nonmetric variable. For our purposes, we examine each of the

metric variables across the five nonmetric variables in the data set. These are appropriate analyses in preparation for analysis of variance or multivariate analysis of variance, in which the nonmetric variables are the independent variables, or for discriminant analysis, in which the nonmetric variables are the dependent measures.

Table 2.12 contains the results of the Levene test for each of the nonmetric variables. The nonmetric variables X_8 and X_{11} both show significant heteroscedasticity on the same performance factors (X_4, X_5, X_6, and X_7), whereas X_{12} and X_{14} have fewer occurrences among the entire set of variables. The implications of these instances of heteroscedasticity must be examined whenever group differences are examined using these nonmetric variables as independent variables and these metric variables as dependent variables. If the assumption violations are found, variable transformations are available to help remedy the variance dispersion.

The tests for homoscedasticity of two metric variables, encountered in methods such as multiple regression, are best accomplished through graphical analysis, particularly an analysis of the residuals. The interested reader is referred to chapter 4 for a complete discussion of residual analysis and the patterns of residuals indicative of heteroscedasticity.

Linearity

The final assumption to be examined is the linearity of the relationships. In the case of individual variables, this relates to the patterns of association between each pair of variables and the ability of the correlation coefficient to adequately

TABLE 2.12 Testing For Homoscedasticity

	Nonmetric/Categorical Variable									
	X_8 *Size of Firm*		X_{11} *Specification Buying*		X_{12} *Structure of Procurement*		X_{13} *Type of Industry*		X_{14} *Type of Buying Situation*	
Metric Variable	*Levene Statistic*	*Sig.*	*Levene Statistic*	*Sig.*	*Levene Statistic*	*Sig.*	*Levene Statistic*	*Sig.*	*Levene Statistic*	*Sig.*
X_1 Delivery speed	.934	.336	.934	.336	.382	.538	.377	.540	.114	.892
X_2 Price level	1.582	.211	1.582	.211	13.761	.000	1.345	.249	8.081	.001
X_3 Price flexibility	1.194	.277	1.194	.277	4.765	.031	.192	.662	14.383	.000
X_4 Manufacturer image	6.549	.012	6.549	.012	.281	.597	.040	.842	2.030	.137
X_5 Overall service	7.819	.006	7.819	.006	5.141	.026	.003	.957	2.888	.060
X_6 Salesforce image	5.279	.024	5.279	.024	1.626	.205	.264	.609	1.735	.182
X_7 Product quality	8.748	.004	8.748	.004	4.129	.045	2.532	.115	2.051	.134
X_9 Usage level	1.377	.243	1.377	.243	1.575	.212	.091	.763	.056	.945
X_{10} Satisfaction level	.323	.571	.323	.571	.000	.986	.054	.817	3.302	.041

Note: Values represent the value and statistical significance (Sig.) of the Levene test assessing the variance dispersion of each metric variable across the levels of the nonmetric/categorical variables.

represent the relationship. If nonlinear relationships are indicated, then the researcher can either transform one or both of the variables to achieve linearity or create additional variables to represent the nonlinear components. For our purposes, we rely on the visual inspection of the relationships to determine whether nonlinear relationships are present. The reader can refer to Figure 2.3, the scatterplot matrix containing the scatterplot for all the metric variables in the data set. Examination of the scatterplots does not reveal any apparent nonlinear relationships. Thus, transformations are not deemed necessary. The assumption of linearity will also be checked for the entire multivariate model, as is done in the examination of residuals in multiple regression.

Summary

The series of graphical and statistical tests directed toward assessing the assumptions underlying the multivariate techniques revealed relatively little in terms of violations of the assumptions. Where violations were indicated, they were relatively minor and should not present any serious problems in the course of the data analysis. The researcher is encouraged always to perform these simple, yet revealing, examinations of the data to ensure that potential problems can be identified and resolved before the analysis begins.

Incorporating Nonmetric Data with Dummy Variables

A critical factor in choosing and applying the correct multivariate technique is the measurement properties of the dependent and independent variables. Some of the techniques, such as discriminant analysis or multivariate analysis of variance, specifically require nonmetric data as dependent or independent variables. But in many instances, metric variables must be used as independent variables, such as in regression analysis, discriminant analysis, and canonical correlation. Moreover, the interdependence techniques of factor and cluster analysis generally require metric variables. To this point, all discussions have assumed metric measurement for variables. But what can we do when the variables are nonmetric, with two or more categories? Are nonmetric variables, such as gender, marital status, or occupation, precluded from use in many multivariate techniques? The answer is no, and we now discuss how to incorporate nonmetric variables into many of these situations that require metric variables.

The researcher has available a method for using dichotomous variables, known as **dummy variables,** which act as replacement variables. A dummy variable is a dichotomous variable that represents one category of a nonmetric independent variable. Any nonmetric variable with k categories can be represented as $k - 1$ dummy variables. The following example will help clarify this concept.

First, assume we wish to include gender, which has two categories, female and male. We also have measured household income level by three categories (see Table 2.13, p. 84). To represent the nonmetric variable gender, we would create two new dummy variables (X_1 and X_2), as shown in Table 2.13. X_1 would represent those individuals who are female with a value of 1, and would give all males a value of 0. Likewise, X_2 would represent all males with a value of 1 and give females a value of 0. Both variables (X_1 and X_2) are not necessary, however,

TABLE 2.13 Representing Nonmetric Variables with Dummy Variables

Nonmetric Variable with Two Categories (Gender)		*Nonmetric Variable with Three Categories (Household Income Level)*	
Gender	*Dummy Variables*	*Household Income Level*	*Dummy Variables*
Female	$X_1 = 1$, else $X_1 = 0$	if <$15,000	$X_3 = 1$, else $X_3 = 0$
Male	$X_2 = 1$, else $X_2 = 0$	if ≥$15,000 & ≤$25,000	$X_4 = 1$, else $X_4 = 0$
		if >$25,000	$X_5 = 1$, else $X_5 = 0$

because when $X_1 = 0$, gender must be female by definition. Thus we need include only one of the variables (X_1 or X_2) to test the effect of gender.

Correspondingly, if we had also measured household income with three levels, as shown in Table 2.13, we would first define three dummy variables (X_3, X_4, and X_5). But as in the case of gender, we would not need the entire set of dummy variables, and instead use $k - 1$ dummy variables, where k is the number of categories. Thus, we would use two of the dummy variables to represent the effects of household income.

There are three ways to represent the household income levels with two dummy variables, as shown in Table 2.14. This form of dummy-variable coding is known as **indicator coding.** An important consideration in this form of dummy-variable coding is to remember the category that is omitted, known as the **comparison group.** This is the category that received all zeros for the dummy variables. For example, in regression analysis, the regression coefficients for the dummy variables represent deviations from the comparison group on the criterion variable. The deviations represent the differences between means for each group of respondents formed by a dummy variable and the comparison group. This form is most appropriate when there is a logical comparison group, such as in an experiment. In an experiment with a control group acting as the comparison group, the coefficients are the mean differences on the dependent variable for each treatment group from the control group. Any time dummy-variable coding is used, we must be aware of the comparison group and remember the impacts it has in our interpretation of the remaining variables.

An alternative method of dummy-variable coding is termed **effects coding.** It is the same as indicator coding except that the comparison group (the group that got all zeros in indicator coding) is now given the value of −1 instead of 0 for the dummy variables. Now the coefficients represent differences for any group from the mean of all groups rather than from the omitted group. Both forms of dummy-variable coding will give the same results; the only differences will be in the interpretation of the dummy-variable coefficients.

TABLE 2.14 Alternative Dummy Variable Coding Patterns for a Three-Category Nonmetric Variable

Household Income Level	*Pattern 1*		*Pattern 2*		*Pattern 3*	
	X_3	X_4	X_3	X_4	X_3	X_4
If <$15,000	1	0	1	0	0	0
If ≥$15,000 & ≤$25,000	0	1	0	0	1	0
If >$25,000	0	0	0	1	0	1

Dummy variables are used most often in regression and discriminant analysis, where the coefficients have direct interpretation. Their use in other multivariate techniques is more limited, especially for those that rely on correlational patterns, such as factor analysis, because the correlation of a binary variable is not well represented by the traditional Pearson correlation coefficient. However, special considerations can be made in these instances, as discussed in the appropriate chapters.

Summary

This chapter has provided the researcher with the necessary tools to examine and explore the nature of the data and the relationships among variables before the application of any of the multivariate techniques. Considerable time and effort can be expended in these activities, but the prudent researcher wisely invests the necessary resources to thoroughly examine the data to ensure that the multivariate methods are applied in appropriate situations and to assist in a more thorough and insightful interpretation of the results.

Questions

1. List potential underlying causes of outliers. Be sure to include attributions to both the respondent and the researcher.
2. Discuss why outliers might be classified as beneficial and as problematic.
3. Distinguish between data that are missing at random (MAR) and missing completely at random (MCAR). Explain how each type impacts the analysis of missing data.
4. Describe the conditions under which a researcher would delete a case with missing data versus the conditions under which a researcher would use an imputation method.
5. Evaluate the following statement: In order to run most multivariate analyses, it is not necessary to meet all the assumptions of normality, linearity, homoscedasticity, and independence.
6. Discuss the following statement: Multivariate analyses can be run on any data set, as long as the sample size is adequate.

References

1. Anderson, Edgar (1969), "A Semigraphical Method for the Analysis of Complex Problems." *Technometrics* 2 (August): 387–91.
2. Box, G. E. P., and D. R. Cox (1964), "An Analysis of Transformations." *Journal of the Royal Statistical Society* B (26): 211–43.
3. Chernoff, Herman. "Graphical Representation as a Discipline," in *Graphical Representation of Multivariate Data*, Peter C. C. Wang, ed. New York: Academic Press, pp. 1–11.
4. Cohen, Jacob, and Patricia Cohen (1983), *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 2d ed. Hillsdale, N.J.: Lawrence Erlbaum Associates.
5. Daniel, C., and F. S. Wood (1980), *Fitting Equations to Data*, 2d ed. New York: Wiley-Interscience.
6. Dempster, A. P., and D. B. Rubin (1983), "Overview," in *Incomplete Data in Sample Surveys: Theory and Annotated Bibliography*, vol. 2. Madow, Olkin, and Rubin, eds. New York: Academic Press.

7. Feinberg, Stephen (1979), "Graphical Methods in Statistics." *American Statistician* 33 (November): 165–78.
8. Johnson, R. A., and D. W. Wichern (1982), *Applied Multivariate Statistical Analysis*. Upper Saddle River, N.J.: Prentice-Hall.
9. Little, Roderick J. A., and Donald B. Rubin (1987), *Statistical Analysis with Missing Data*. New York: Wiley.
10. Wang, Peter C. C., ed. (1978), *Graphical Representation of Multivariate Data*. New York: Academic Press.
11. Weisberg, S. (1985), *Applied Linear Regression*. New York: Wiley.
12. Wilkinson, L. (1982), "An Experimental Evaluation of Multivariate Graphical Point Representations." In *Human Factors in Computer Systems: Proceedings*, New York: ACM Press, pp. 202–9.