2 Examining Your Data

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.
- Assess the type and potential impact of missing data.
- Understand the different types of missing data processes.
- 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

Data examination is a time-consuming, but necessary, initial step in any analysis that researchers often overlook. Here the researcher evaluates the impact of missing data, identifies outliers, and tests for the assumptions underlying most multivariate techniques. The objective of these data examination tasks is as much to reveal what is not apparent as it is to portray the actual data, because the "hidden" effects are easily overlooked. For example, the biases introduced by nonrandom missing data will never be known unless explicitly identified and remedied by the methods discussed in a later section of this chapter. Moreover, unless the researcher reviews the results on a case-by-case basis, the existence of outliers will not be apparent, even if they substantially affect the results. Violations of the statistical assumption may cause biases or nonsignificance in the results that cannot be distinguished from the true results.

Before we discuss a series of empirical tools to aid in data examination, the introductory section of this chapter offers (1) a summary of various graphical techniques available to the researcher as a means of representing data and (2) some new measures of association to complement the traditional correlation coefficient. These graphical techniques provide the researcher with a set of simple yet comprehensive ways to examine both the individual variables and the relationships among them. They are not meant to replace the empirical tools, but rather provide a complementary means of portraying the data and its relationships. As you will see, a histogram can graphically show the shape of a data distribution, just as we can reflect that same distribution with skewness and kurtosis values.

The empirical measures quantify the distribution's characteristics, whereas the histogram portrays them in a simple and visual manner. Likewise, other graphical techniques (i.e., scatterplot and boxplot) show relationships between variables represented by the correlation coefficient and means difference test, respectively.

The additional measures of association are attempts to overcome one of the primary limitations of the Pearson correlation coefficient—the requirement of a linear relationship. These measures attempt to measure dependence, whether it is in a linear or nonlinear form. While they may be less useful in many of our traditional statistical techniques, they still provide the researcher with methods to identify relationships not previously discoverable and then the researcher can decide how to integrate them into the analysis.

With the graphical techniques and association measures addressed, the next task facing the researcher is how to assess and overcome pitfalls resulting from the research design (e.g., questionnaire design) and data collection practices. Specifically, this chapter addresses the following:

- Evaluation of missing data
- Identification of outliers
- Testing of the assumptions underlying most multivariate techniques
- Transforming data for either improved statistical properties or interpretability.

Missing data are a nuisance to researchers and primarily result from errors in data collection/data entry or from the omission of answers by respondents. Classifying missing data and the reasons underlying their presence are addressed through a series of steps that not only identify the impacts of the missing data, but that also provide remedies for dealing with it in the analysis. *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 *statistical 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 5) 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 several methods of data transformations. These range from incorporating nonmetric variables in applications that require metric variables through the creation of a special type of metric variable known as *dummy* variables to such techniques as binning and logarithmic transformations to represent specific types of relationships. While many forms of transformations are associated with meeting specific statistical properties, they also represent unique ways to modify the character of the data to enhance interpretation or applicability to a specific research question.

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 and key term cross-references are *italicized.*

All-available approach *Imputation* method for missing data that computes values based on all-available valid observations, also known as the pairwise approach.

Binning Process of categorizing a metric variable into a small number of categories/bins and thus converting the variable into a nonmetric form.

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. This method is useful in making comparisons of one or more metric variables across groups formed by a nonmetric variable.

Cardinality The number of distinct data values for a variable.

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.*

Centering A variable *transformation* in which a specific value (e.g., the variable mean) is subtracted from each observation's value, thus improving comparability among variables.

Cold deck imputation *Imputation* method for *missing data* that derives the imputed value from an external source (e.g., prior studies, other samples).

Comparison group See *reference category.*

Complete case approach Approach for handling *missing data* that computes values based on data from complete cases, that is, cases with no missing data. Also known as the *listwise deletion* approach.

Curse of dimensionality The problems associated with including a very large number of variables in the analysis. Among the notable problems are the distance measures becoming less useful along with higher potential for irrelevant variables and differing scales of measurement for the variables.

Data management All of the activities associated with assembling a dataset for analysis. With the arrival of the larger and diverse datasets from *Big Data*, researchers may now find they spend a vast majority of their time on this task rather than analysis.

Data quality Generally referring to the accuracy of the information in a dataset, recent efforts have identified eight dimensions that are much broader in scope and reflect the usefulness in many aspects of analysis and application: completeness, availability and accessibility, currency, accuracy, validity, usability and interpretability, reliability and credibility, and consistency.

Data transformations A variable may have an undesirable characteristic, such as non-normality, 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), empirical reasons (e.g., problems identified through graphical or statistical means) or for interpretation purposes (e.g., standardization).

dCor A newer measure of association that is distance-based and more sensitive to nonlinear patterns in the data.

Dichotomization Dividing cases into two classes based on being above or below a specified value.

Dummy variable Special metric variable used to represent a single category of a nonmetric variable. To account for L levels of a 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. The omitted category is termed the *reference category.*

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, which is in contrast to *indicator coding.*

Elasticity Measure of the ratio of percentage change in Y for a percentage change in X. Obtained by using a log-log transformation of both dependent and independent variables.

EM *Imputation* method applicable when *MAR* missing data processes are encountered which employs maximum likelihood estimation in the calculation of imputed values.

Extreme groups approach Transformation method where observations are sorted into groups (e.g., high, medium and low) and then the middle group discarded in the analysis.

Heat map Form of scatterplot of nonmetric variables where frequency within each cell is color-coded to depict relationships.

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.*

Hoeffding's D New measure of association/correlation that is based on distance measures between the variables and thus more likely to incorporate nonlinear components.

Homoscedasticity When the variance of the error terms (e) 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 e) is critical to the proper application of many multivariate techniques. When the error terms have increasing or modulating variance, the data are said to be *heteroscedastic.* Analysis of *residuals* best illustrates this point.

Hot deck imputation *Imputation* method in which the *imputed* value is taken from an existing observation deemed similar.

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 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.*

Ipsatizing Method of transformation for a set of variables on the same scale similar to centering, except that the variable used for centering all of the variables is the mean value for the observation (e.g., person-centered).

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 to Y.

Listwise deletion See *complete case approach.*

Mean substitution *Imputation* method where the mean value of all valid values is used as the imputed value for *missing data.*

MIC (mutual information correlation) New form of association/correlation that can represent any form of dependence (e.g., circular patterns) and not limited to just linear relationships.

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 random sample of the Y values, but the missing values of Y are related to some other observed variable (X) in the sample. For example, assume two groups based on gender have different levels of missing data between male and female. The data is MAR if the data is missing at random within each group, but the levels of missing data depend on the gender.

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, for example, 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.*

Missingness The absence or presence of *missing data* for a case or observation. Does not relate directly to how that missing data value might be *imputed.*

Multiple imputation. *Imputation* method applicable to *MAR* missing data processes in which several datasets are created with different sets of imputed data. The process eliminates not only bias in imputed values, but also provides more appropriate measures of standard errors.

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) on one or more characteristics (variables). At issue is its representativeness of the population.

Reference category The category of a nonmetric variable that is omitted when creating *dummy variables* and acts as a reference point in interpreting the dummy variables. In *indicator coding,* the reference category has values of zero (0) for all dummy variables. With *effects coding,* the reference category has values of minus one (-1) for all dummy variables.

Regression imputation *Imputation* method that employs regression to estimate the *imputed value* based on valid values of other variables for each observation.

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.

Response surface A transformation method in which a form of polynomial regression is used to represent the distribution of an outcome variable in an empirical form that can be portrayed as a surface.

Robustness The ability of a statistical technique to perform reasonably well even when the underlying statistical assumptions have been violated in some manner.

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.

Standardization Transformation method where a variable is *centered* (i.e., variable's mean value subtracted from each observation's value) and then "standardized" by dividing the difference by the variable's standard deviation. Provides a measure that is comparable across variables no matter what their original scale.

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 underpinnings on which they are based also are supported. By examining the data before the application of any 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 greater demands on the researcher to understand, interpret, and articulate results based on relationships that are more complex than encountered before. A thorough knowledge of the 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, the researcher ensures that the *data underlying the analysis meet all of the requirements for a multivariate analysis*. Multivariate techniques demand much more from the data in terms of larger datasets and more complex assumptions than encountered with univariate analyses. Missing data, outliers, and the statistical characteristics of the data are all much more difficult to assess in a multivariate context. Thus, the analytical sophistication needed to ensure that these requirements are met forces the researcher to use a series of data examination techniques that are as complex as the multivariate techniques themselves.

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). The time, effort, and resources devoted to the data examination process may seem almost wasted because many times no corrective action is warranted. The researcher should instead view these techniques as "*investments in multivariate insurance*" that ensure the results obtained from the multivariate analysis are truly valid and accurate. Without such an "investment" it is quite easy, for example, for several unidentified outliers to skew the results, for missing data to introduce a bias in the correlations between variables, or for non-normal variables to invalidate the results. And yet the most troubling aspect of these problems is that they are "hidden," because in most instances the multivariate techniques will go ahead and provide results. Only if the researcher has made the "investment" will the potential for catastrophic problems be recognized and corrected *before* the analyses are performed. 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; 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 that arise during analysis force you to do so.

The Challenge of Big Data Research Efforts

As first discussed in Chapter 1, the age of "Big Data" is impacting all aspects of both academic and practitioner research efforts. One area most impacted is data examination, where the issues addressed in this chapter have substantial impact on the ultimate success or failure of the research effort. While many researchers are still able to operate within the domain of small, tightly controlled datasets of 50 to 100 variables from a selected sample, many others are facing the task of dealing with widely disparate data sources (e.g., customer-level data from firms, social media data, locational data) that change the entire nature of the dataset. And while many researchers, particularly in

the academic domain, may wish to avoid these complications, the trend is inevitable towards more integration [16]. While it is beyond the scope of this chapter to address all of the issues arising from the emergence of Big Data, there are two general topics regarding the data itself that are faced by researchers in all areas.

DATA MANAGEMENT

Perhaps the most daunting challenge when venturing into the world of Big Data is the fundamental task of data management—assembling a dataset for analysis. So many times researchers focus on the type of technique to be used, but are then faced with the reality of the data available, its format and structure. A common axiom among Big Data researchers is that 80 percent or more of the project time is spent on data management. This task, many times referred to as data wrangling which perhaps describes it best metaphorically, is the principle task of data fusion [14], which is becoming more commonplace as researchers in all areas attempt to combine data from multiple sources. The creation in 1988 of the DAMA, the Data Management Association International, signaled the emergence of a new field within analytics.

The issues associated with using and combining data from multiple sources become complex very quickly. Even what seems like a simple merging of customer data from a firm database with survey data becomes more complex with issues of identity management, extracting the appropriate customer data from the firm's databases, matching timeframes and a host of other issues. And this is relatively simple compared to merging structured data (e.g., survey or customer data) with unstructured data, such as social media posts that involve text mining. And these issues do not even include the technical challenges facing Big data users today in terms of data storage, retrieval and processing. The end result is that researchers in all fields are going to have become "data managers" as much as data analysts in the near future as the need and availability of disparate sources of data increases.

DATA QUALITY

Once the dataset is finally assembled, the task is far from complete. For example, having a value for a variable does not mean that analysis is ready to begin, since the issue of data quality must be addressed first. Indeed, data quality is multi-faceted and the era of Big Data has forced an examination of what is actually meant by the term. Researcher long accustomed to dealing with their own domain of data and its characteristics are now having to reconcile the notion of data quality, and are finding it quite challenging [51]. Yet this is an issue of extreme importance to the entire research community [77]. Recent research has attempted to synthesize this diverse set of characteristics of data quality and has tentatively identified eight dimensions [49]: (1) completeness, (2) availability and accessibility, (3) currency, (4) accuracy, (5) validity, (6) usability and interpretability, (7) reliability and credibility, and (8) consistency. While the operational definitions of these elements may differ by research domain and even type of data, today's researchers must always consider this more comprehensive perspective for any data source being used in research today. The emergence of books such as the Bad Data Handbook [63] and Data Cleaning Techniques [19] are both encouraging (since the topics are addressed), but also discouraging because the topic deserves such prominence.

The topic of data quality is not solely a challenge for Big Data users, but many data sources require close examination. Several disparate examples illustrate the varying challenges faced as the types of data expand. One example is Mechanical Turk, a service offered by Amazon, that unfortunately has become quite widely used across a number of research domains. While it offers quick access to a wide range of respondents, its data quality is often compromised by a number of factors including the type of participants providing data [89]. And even when the "quality" of the respondents from other sources is controlled, the widespread use of online data collection formats has necessitated measures of consistency indices and multivariate outlier analysis to identify "carelessness" in the data provided [64]. Finally, a widely used source of customer information comes from "data brokers" or "data aggregators" who collect extensive data from numerous sources on a household level and then sell this information to a wide array of users—firms, government entities and even other data brokers. While quality control measures may be performed, many times even the very nature of the data can be problematic. One example is the widespread use of binary measures of "interest" which provide a simple indication if that household exhibits a particular interest (e.g., outdoor activities, political activism, consumer electronics). The list is almost endless and provides firms potential targeting information

about each unit. Yet there is a fundamental question—the data is coded as a one or blank. Now for those interested in a specific topic, a value of one indicates potential interest. But what are we to make of the blank—does it equate to "No Interest" or "Missing." These are not equivalent for the researcher's purposes, but may still be beyond the ability of the data source to distinguish.

SUMMARY

The era of Big Data and the use of data from many sources present researchers with many challenges well before the first analysis is performed. As we discussed earlier, data examination is somewhat like insurance—an investment that hopefully pays off with better results. As the nature, scope and volume of data expands, this should point the researcher toward more data examination versus the too often tendency to be overwhelmed by the scale of the problem and thus move forward without data examination. The issues covered in this chapter are as applicable to "Big Data" as "small data" and researchers today have a wider array to techniques and measures to address their data examination requirements.

Preliminary Examination of the Data

As discussed earlier, the use of multivariate techniques places an increased burden on the researcher to understand, evaluate, and interpret complex results. This complexity requires 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. As the researcher moves to more complex multivariate analyses, however, the need and level of understanding increase dramatically and require even more powerful empirical diagnostic measures. The researcher can be aided immeasurably in gaining a fuller understanding of what these diagnostic measures mean through the use of graphical techniques, portraying the basic characteristics of individual variables and relationships between variables in a simple "picture." For example, a simple scatterplot represents in a single picture not only the two basic elements of a correlation coefficient, namely the type of relationship (positive or negative) and the strength of the relationship (the dispersion of the cases), but also a simple visual means for assessing linearity that would require a much more detailed analysis if attempted strictly by empirical means. Correspondingly, a boxplot illustrates not only the overall level of differences across groups shown in a *t*-test or analysis of variance, but also the differences between pairs of groups and the existence of outliers that would otherwise take more empirical analysis to detect if the graphical method was not employed. The objective in using graphical techniques is not to replace the empirical measures, but to use them as a complement to provide a visual representation of the basic relationships so that researchers can feel confident in their understanding of these relationships.

But what is to be done when the relationships are nonlinear or more pattern-based? The traditional measures of correlation based on a linear relationship are not adequate to identify these types of relationships without substantial data transformations. Recent research in the area of data science has developed new measures of dependence that can assess not only linear relationships, but nonlinear patterns of many types. In these situations the researcher can at least be aware of their existence and then decide how they are to be included in the analysis.

The advent and widespread use of statistical programs have increased access to such methods. Most statistical programs provide comprehensive modules of graphical techniques available for data examination that are augmented with more detailed statistical measures of data description, including these new measures of association. 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.

UNIVARIATE PROFILING: 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 of the variable through a **histogram.** A histogram is a graphical representation of a single

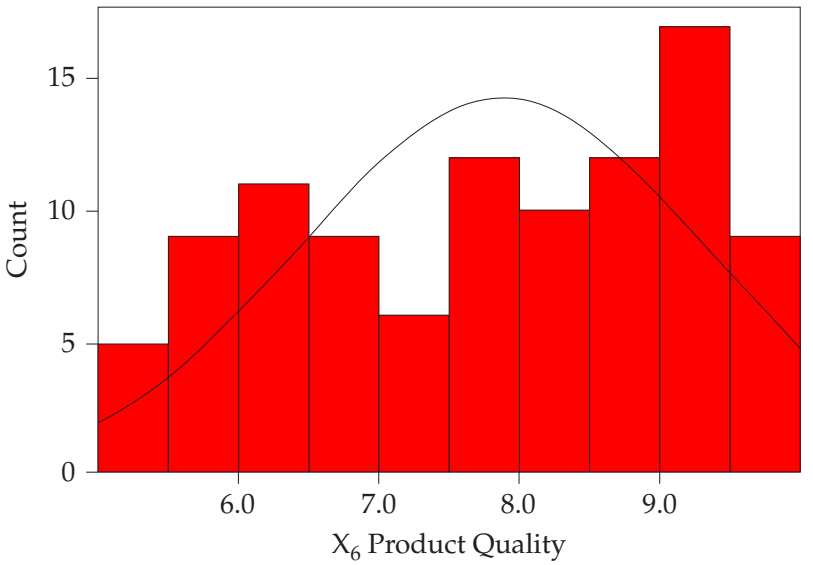


Figure 2.1
Graphical Representation of Univariate Distribution

variable that represents the frequency of occurrences (data values) within data categories. The frequencies are plotted to examine the shape of the distribution of values. If the integer values ranged from 1 to 10, the researcher could construct a histogram by counting the number of responses for each integer value. For continuous variables, categories are formed within which the frequency of data values is tabulated. 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 to assess the correspondence of the actual distribution to the desired (normal) distribution. The histogram can be used to examine any type of metric variable.

For example, the responses for X_6 from the database introduced in Chapter 1 are represented in Figure 2.1. The height of the bars represents the frequencies of data values within each category. The normal curve is also superimposed on the distribution. As will be shown in a later section, empirical measures indicate that the distribution of X_6 deviates significantly from the normal distribution. But how does it differ? The empirical measure that differs most is the kurtosis, representing the peakedness or flatness of the distribution. The values indicate that the distribution is flatter than expected. What does the histogram show? The middle of the distribution falls below the superimposed normal curve, while both tails are higher than expected. Thus, the distribution shows no appreciable skewness to one side or the other, just a shortage of observations in the center of the distribution. This comparison also provides guidance on the type of transformation that would be effective if applied as a remedy for non-normality. All of this information about the distribution is shown through a single histogram.

BIVARIATE PROFILING: 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 metric variables. One variable defines the horizontal axis and the other variable defines the vertical axis. Variables may be any metric value. 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 the 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 a seemingly random pattern of points may indicate no relationship.

Of the many types of scatterplots, one format particularly suited to multivariate techniques is the scatterplot matrix, 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. Scatterplot matrices and individual scatterplots are now available

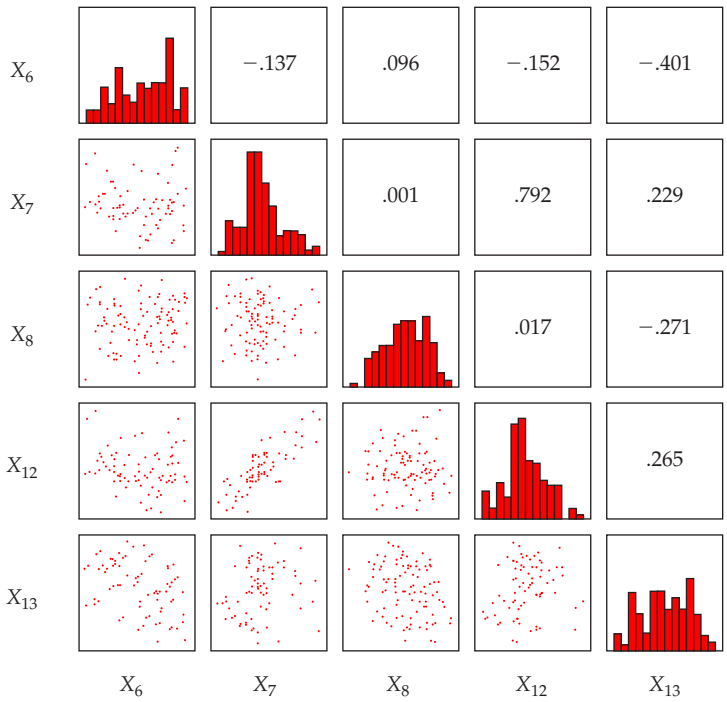


Figure 2.2
Bivariate Profiling of Relationships Between Variables: Scatterplot Matrix of Selected Metric Variables (X_6, X_7, X_8, X_{12}, and X_{13})

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.

Figure 2.2 presents the scatterplots for a set of five variables from the HBAT database (X_6, X_7, X_8, X_{12}, and X_{13}). For example, the highest correlation can be easily identified as between X_7 and X_{12}, as indicated by the observations closely aligned in a well-defined linear pattern. In the opposite extreme, the correlation just above (X_7 versus X_8) shows an almost total lack of relationship as evidenced by the widely dispersed pattern of points and the correlation .001. Finally, an inverse or negative relationship is seen for several combinations, most notably the correlation of X_6 and X_{13} $(-.401)$. Moreover, no combination seems to exhibit a nonlinear relationship that would not be represented in a bivariate correlation.

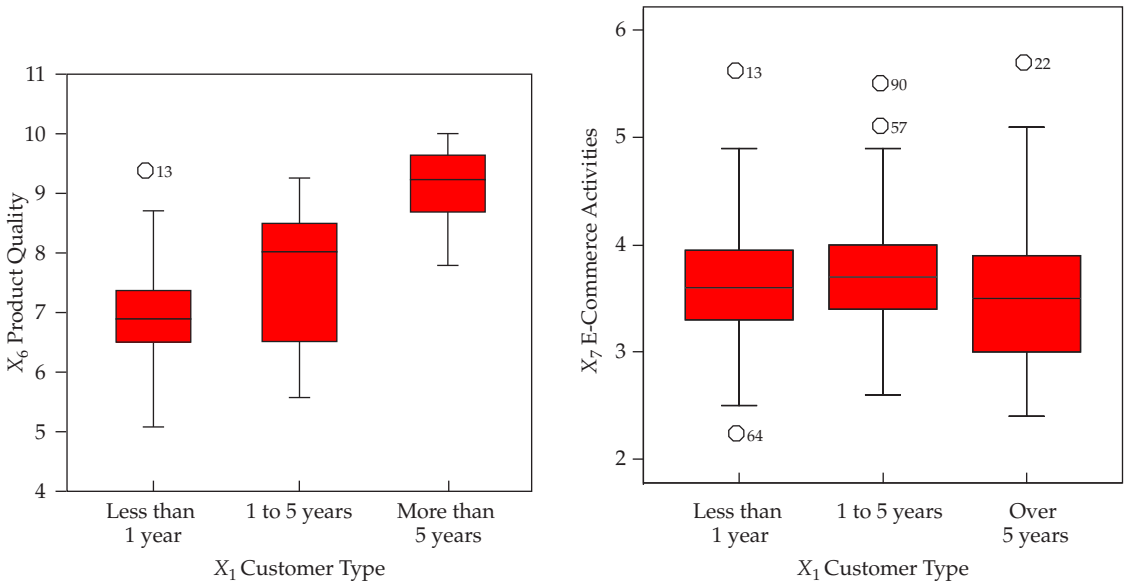
The scatterplot matrix provides a quick and simple method of not only assessing the strength and magnitude of any bivariate relationship, but also a means of identifying any nonlinear patterns that might be hidden if only the bivariate correlations, which are based on a linear relationship, are examined.

BIVARIATE PROFILING: EXAMINING GROUP DIFFERENCES

The researcher also is faced with understanding the extent and character of differences of one or more metric variables across two or more groups formed from the categories of a nonmetric variable. Assessing group differences is done through univariate analyses such as *t*-tests and analysis of variance and the multivariate techniques of discriminant analysis and multivariate analysis of variance. Another important aspect is to identify outliers (described in more detail in a later section) that may become apparent only when the data values are separated into groups.

The graphical method used for this task is the **boxplot**, a pictorial representation of the data distribution of a metric variable for each group (category) of a nonmetric variable (see example in Figure 2.3). First, the upper and lower quartiles of the data distribution form the upper and lower boundaries of the box, with the box length being the distance between the 25th percentile and the 75th percentile. The box contains the middle 50 percent of the data values and the larger the box, the greater the spread (e.g., standard deviation) of the observations. The median is depicted by a solid line within the box. If the median lies near one end of the box, skewness in the opposite direction is indicated. The lines extending from each box (called *whiskers*) represent the distance to the smallest and the largest

Figure 2.3 Bivariate Profiling of Group Differences: Boxplots of X_6 (Product Quality) and X_7 (E-Commerce Activities) with X_1 (Customer Type)



observations that are less than one quartile range from the box. Outliers (observations that range between 1.0 and 1.5 quartiles away from the box) and extreme values (observations greater than 1.5 quartiles away from the end of the box) are depicted by symbols outside the whiskers. In using boxplots, the objective is to portray not only the information that is given in the statistical tests (Are the groups different?), but also additional descriptive information that adds to our understanding of the group differences.

Figure 2.3 shows the boxplots for X_6 and X_7 for each of the three groups of X_1 (Customer Type). Before examining the boxplots for each variable, let us first see what the statistical tests tell us about the differences across these groups for each variable. For X_6, a simple analysis of variance test indicates a highly significant statistical difference (F value of 36.6 and a significance level of .000) across the three groups. For X_7, however, the analysis of variance test shows no statistically significant difference (significance level of .419) across the groups of X_1.

Using boxplots, what can we learn about these same group differences? As we view the boxplot of X_6, we do see substantial differences across the groups that confirm the statistical results. We can also see that the primary differences are between groups 1 and 2 versus group 3. Essentially, groups 1 and 2 seem about equal. If we performed more statistical tests looking at each pair of groups separately, the tests would confirm that the only statistically significant differences are group 1 versus 3 and group 2 versus 3. Also, we can see that group 2 has substantially more dispersion (a larger box section in the boxplot), which prevents its difference from group 1. The boxplots thus provide more information about the extent of the group differences of X_6 than just the statistical test.

For X_7, we can see that the three groups are essentially equal, as verified by the nonsignificant statistical test. We can also see a number of outliers in each of the three groups (as indicated by the notations at the upper portion of each plot beyond the whiskers). Although the outliers do not impact the group differences in this case, the researcher is alerted to their presence by the boxplots. The researcher could examine these observations and consider the possible remedies discussed in more detail later in this chapter.

MULTIVARIATE PROFILES

To this point the graphical methods have been restricted to univariate or bivariate portrayals. In many instances, however, the researcher may desire to compare observations characterized on a multivariate profile, whether it be for descriptive purposes or as a complement to analytical procedures. To address this need, a number of **multivariate**

graphical displays center around one of three types of graphs [34]. The first graph type is a direct portrayal of the data values, either by (a) glyphs, or metroglyphs, which are some form of circle with radii that correspond to a data value; or (b) multivariate profiles, which portray a barlike profile for each observation. A second type 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 of this type is Andrew's Fourier transformation [5]. The final approach is the use of graphical displays with iconic representativeness, the most popular being a face [17]. The value of this type of display is the inherent processing capacity humans have for their interpretation. As noted by Chernoff [17]:

> 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 affect 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 [87, 88].

NEW MEASURES OF ASSOCIATION

Before discussing some new measures of association, we must also discuss being willing to use our existing set of measures when appropriate. Recent research using non-normal data compared 12 different measures, including Pearson correlation, Spearman's rank-order correlation, various transformations (e.g., nonlinear transformations or the Rank-Based Inverse Normal Transformation), and resampling approaches (e.g., the permutation test or bootstrapping measures) [11]. While the Pearson correlation worked fairly well, the Rank-Based Inverse Normal Transformation worked well in more situations across samples sizes and degrees of non-normality. The significance of these results is not to propose using a new method of association per se, but instead to expose researchers to the multiplicity of existing measures that may be more suited to the task of measuring association across a variety of situations.

The rapid development of data mining, particularly in the arena of Big Data, brought to attention the need for more sophisticated measures of association than the traditional correlation coefficient [71]. The capability for close examination of literally thousands of relationships to see if the correlation captured the relationship was impossible. What was needed were more "robust" measures of association/dependence which could assess the more complicated patterns that might be encountered. To this end, several new measures have been developed, including Hoeffding's D, dCor (the distance correlation) and MIC (mutual information correlation). **Hoeffding's D** is a nonparametric measure of association based on departures from independence and can work [45] with many types of data [45]. **dCor** is a distance-based measure of association which also is more sensitive to [84] nonlinear patterns in the data [84]. Perhaps the most interesting measure is **MIC (mutual information correlation)** which has been shown as capable of not only identifying nonlinear relationships, but a wide range of distinct patterns which are not of the traditional nonlinear type (e.g., two lines intersecting at an angle, a line and parabola, a pattern like the letter X and an ellipse, along with many others) [54]. Based on pattern matching, it provides a method for quickly scanning large amounts of data for these more atypical relationships that would otherwise go undetected.

SUMMARY

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. Moreover, the multivariate methods enable the researcher to use a single graphical portrayal to represent a large number of variables, instead of using a large number of the univariate or bivariate methods to portray the same number of variables. And an expanded set of measures of association may assist in identifying previously undiscovered relationships, especially as datasets increase in size.

Missing Data

Missing data, where valid values on one or more variables are not available for analysis, are a fact of life in multivariate analysis. In fact, rarely does the researcher avoid some form of missing data problem. 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 identify the patterns and relationships underlying the missing data in order to maintain as close as possible the original distribution of values when any remedy is applied.* The extent of missing data is a secondary issue in most instances, affecting the type of remedy applied. These patterns and relationships are a result of a **missing data process**, which is 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) that leads to missing values. The 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.

THE IMPACT OF MISSING DATA

The effects of some missing data processes are known and directly accommodated in the research plan, as will be discussed later in this section. More often, the missing data processes, particularly those based on actions by the respondent (e.g., non-response to a question or set of questions), are rarely known beforehand. To identify any patterns in the missing data that would characterize the missing data process, 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 distinct 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.

Why worry about the missing data processes? Can't the analysis just be performed with the valid values we do have? Although it might seem prudent to proceed just with the valid values, both substantive and practical considerations necessitate an examination of the missing data processes.

Practical Impact The *practical impact* of missing data is the reduction of 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. For example, it has been shown that if 10 percent of the data is randomly missing in a set of five variables, on average almost 60 percent of the cases will have at least one missing value [53]. Thus, when complete data are required, the sample is reduced to 40 percent of the original size. In such situations, the researcher must either gather additional observations or find a remedy for the missing data in the original sample.

Substantive Impact From a *substantive perspective,* any statistical results based on data with a nonrandom missing data process could be inaccurate. This inaccuracy can occur either in biased parameter estimates or inaccurate hypothesis tests due to incorrect standard errors or the reduction in statistical power [66]. But in both instances the missing data process "causes" certain data to be missing and these missing data lead to erroneous results. For example, what if we found that individuals who did not provide their household income tended to be almost exclusively those in the higher income brackets? Wouldn't you be suspect of the results knowing this specific group of people were excluded? Enders [31] provides an excellent discussion of the interplay of missing data and parameter estimates, while Newman and Cottrell [67] demonstrate that the impact of missing data on a correlation is a combination of its degree of nonrandom missingness with each variable and the relative variances of the missing versus complete cases. The effects of missing data are sometimes termed *hidden* due to the fact that we still get results from the analyses even without the missing data. The researcher could consider these biased results as valid unless the underlying missing data processes are identified and understood.

Need for Concern The concern for missing data processes is similar to the need to understand the causes of non-response in the data collection process. Just as we are concerned about who did not respond during data collection and any subsequent biases, we must also be concerned about the non-response or missing data among the collected

data. The researcher thus needs to not only remedy the missing data if possible, but also understand any underlying missing data processes and their impacts. Yet, too often, researchers either ignore the missing data or invoke a remedy without regard to the effects of the missing data. The next section employs a simple example to illustrate some of these effects and some simple, yet effective, remedies. Then, a four-step process of identifying and remedying missing data processes is presented. Finally, the four-step process is applied to a small data set with missing data.

RECENT DEVELOPMENTS IN MISSING DATA ANALYSIS

The past decade has seen a resurgence in interest in missing data analysis due not only to the increasing need in light of the new types of data being analyzed, but also from the expanded availability and improved usability of model-based methods of imputation, such as maximum likelihood and multiple imputation. Limited in use until this time due to a lack of understanding and the complexity of use, these methods, which might be termed "Missing Data Analysis 2.0," represent a new generation of approaches to addressing issues associated with missing data. As will be discussed in a later section, these methods provide alternatives to "traditional" methods which required that researchers make strict assumptions about the missing data and potentially introduced biases into the analysis when remedying missing data.

Accompanying this increased use of model-based approaches is a renewed interest among the academic community in missing data analysis. As a result, a series of excellent tests or chapters have emerged [31, 59, 4, 93, 2] along with interest in every academic discipline, including health sciences [93, 57]; education research [22, 15]; psychology [60, 32]; data management/data fusion [69]; genetics [24]; management research [66, 35] and marketing [56]. And while the primary focus of these efforts has been within the academic community, the practitioner sector, with the increased emphasis on analytics and Big Data, has also recognized the need for addressing these issues [9, 82, 86] and even proposed the use of data mining methods techniques such as CART for missing data analysis [41].

As a result of this increased interest, these model-based approaches have become widely available in the major software packages (e.g., SAS, IBM SPSS, and Stata) as well as R and other software platforms. Researchers now have at their disposal these more advanced methods which require fewer assumptions as to the nature of the missing data process and provide a means for imputation of missing values without bias.

A SIMPLE EXAMPLE OF A MISSING DATA ANALYSIS

To illustrate the substantive and practical impacts of missing data, Figure 2.4 contains a simple example of missing data among 20 cases. As is typical of many datasets, particularly in survey research, the number of missing data varies 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 more than one-half (55%) of all values. Three cases (3, 13, and 15) have more than 50 percent missing data and only five cases have complete data. Overall, 23 percent of the data values are missing.

Practical Impact From a *practical standpoint,* the missing data in this example can become quite problematic in terms of reducing the sample size. For example, if a multivariate analysis was performed that required complete data on all five variables, the sample would be reduced to only the five cases with no missing data (cases 1, 7, 8, 12, and 20). This sample size is too few for any type of analysis. Among the remedies for missing data that will be discussed in detail in later sections, 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.

Substantive Impact The *substantive impact,* however, can be seen in these five that are still missing data; all occur in V_4. By comparing the values of V_2 for the remaining five cases with missing data for V_4 (cases 2, 6, 14, 16, and 18) versus those cases having valid V_4 values, a distinct pattern emerges. The five cases with missing values for V_4 have the five lowest values for V_2, indicating that missing data for V_4 are strongly associated with lower scores on V_2.

Figure 2.4
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	

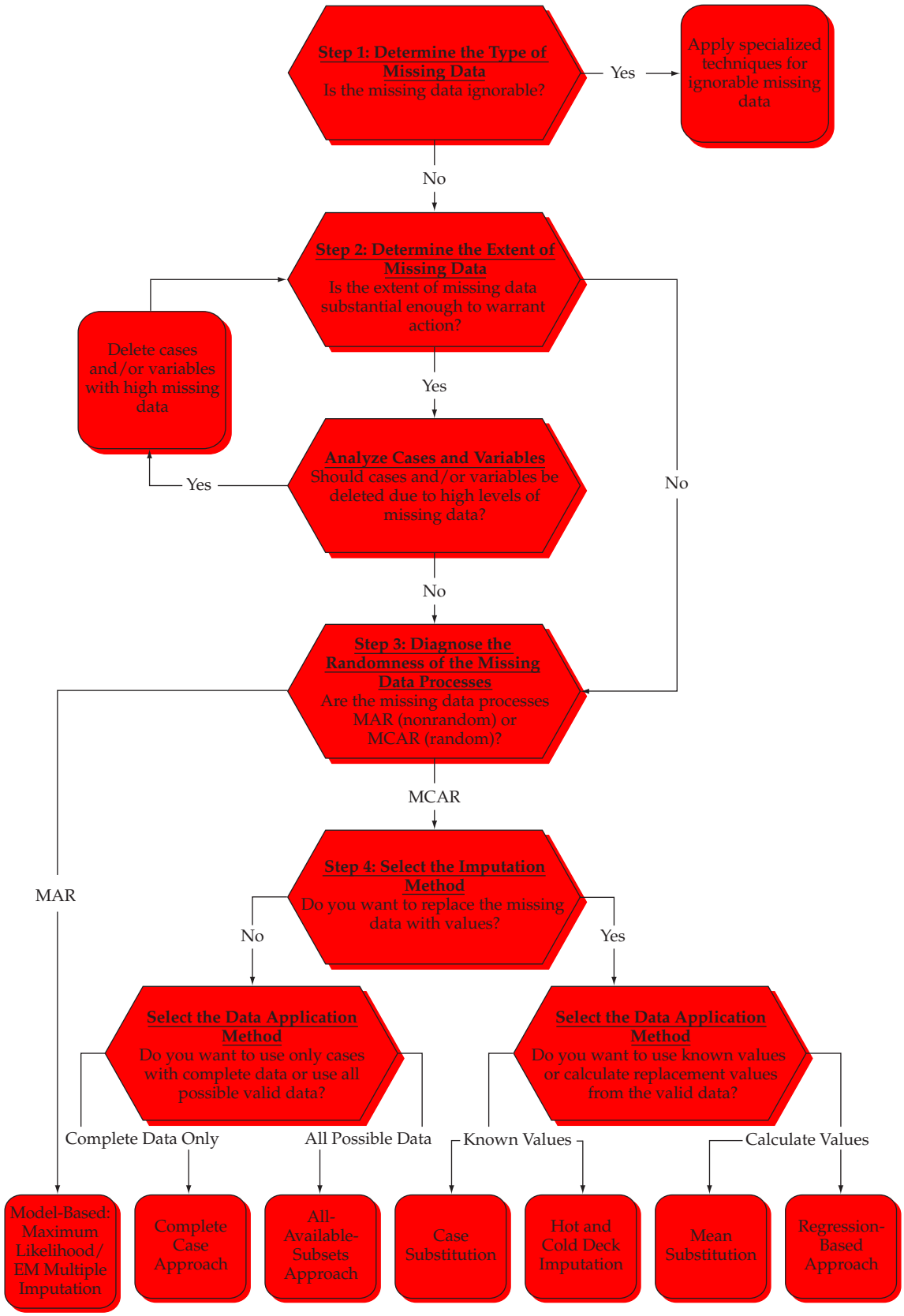
This systematic association between missing and valid data directly affects any analysis in which V_4 and V_2 are both included. For example, the mean score for V_2 will he higher if cases with missing data on V_4 are excluded $(\text{mean} = 8.4)$ than if those five cases are included $(\text{mean} = 7.8)$. In this instance, the researcher must always scrutinize results including both V_4 and V_2 for the possible impact of this missing data process on the results.

Overall Impact As we have seen in the example, finding a remedy for missing data (e.g., deleting cases or variables) can be a practical solution for missing data. Yet the researcher must guard against applying such remedies without diagnosis of the missing data processes. Avoiding the diagnosis may address the practical problem of sample size, but only cover up the substantive concerns. What is needed is a structured process of first identifying the presence of missing data processes and then applying the appropriate remedies. In the next section we discuss a four-step process to address both the practical and substantive issues arising from missing data.

A FOUR-STEP PROCESS FOR IDENTIFYING MISSING DATA AND APPLYING REMEDIES

As seen in the previous discussions, missing data can have significant impacts on any analysis, particularly those of a multivariate nature. Moreover, as the relationships under investigation become more complex, the possibility also increases of not detecting missing data processes and their effects. These factors combine to make it essential that any multivariate analysis begin with an examination of the missing data processes. To this end, a four-step process (see Figure 2.5) is presented, which addresses the types and extent of missing data, identification of missing data processes, and available remedies for accommodating missing data into multivariate analyses.

Figure 2.5
A Four-Step Process for Identifying Missing Data and Applying Remedies



Step 1: Determine the Type of Missing Data The first step in any examination of missing data is to determine the type of missing data involved. Here the researcher is concerned whether the missing data are part of the research design and under the control of the researcher or whether the "causes" and impacts are truly unknown. Also, researchers should understand the "levels" of missingness present in their data so that the most effective missing data strategies can be developed. Let's start with the missing data that are part of the research design and can be handled directly by the researcher.

IGNORABLE MISSING DATA Many times missing data are expected and part of the research design. In these instances, the missing data are termed **ignorable missing data**, meaning that specific remedies for missing data are not needed because the allowances for missing data are inherent in the technique used [62, 83]. 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. There are three instances in which a researcher most often encounters ignorable missing data.

A Sample as Missing Data The first example encountered in almost all surveys and most other datasets is the ignorable missing data process resulting from taking a sample of the population rather than gathering data from the entire population. In these instances, the missing data are 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 these missing data ignorable by using probability sampling to select respondents. Probability sampling enables 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 non-sampled observations are ignorable.

Part of Data Collection A second instance of ignorable missing data is due to the specific design of the data collection process. Certain non-probability sampling plans are designed for specific types of analysis that accommodate the nonrandom nature of the sample. Much more common are missing data due to the design of the data collection instrument, such as through skip patterns where respondents skip sections of questions that are not applicable.

For example, in examining customer complaint resolution, it might be appropriate to require that individuals make a complaint before asking questions about how complaints are handled. For those respondents not making a complaint, they do not answer the questions on the process and thus create missing data. The researcher is not concerned about these missing data, because they are part of the research design and would be inappropriate to attempt to remedy.

Censored Data A third type 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 [62]). The data are censored because in certain years the armed services had height restrictions that varied in level and enforcement. Thus, the researchers face 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 [62].

In each instance of an ignorable missing data process, the researcher has an explicit means of accommodating the missing data into the analysis. It should be noted that it is possible to have both ignorable and non-ignorable missing data in the same data set when two different missing data processes are in effect.

MISSING DATA PROCESSES THAT ARE NOT IGNORABLE Missing data that cannot be classified as ignorable occur for many reasons and in many situations. In general, these missing data fall into two classes based on their source: known versus unknown processes.

Known Processes Many missing data processes are *known* to the researcher in that they can be identified due to procedural factors, such as errors in data entry that create invalid codes, disclosure restrictions (e.g., small counts in US 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.

Unknown Processes These types of missing data processes are less easily identified and accommodated. Most often these instances are related directly to the respondent. One example is the refusal to respond to certain questions, which 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.

In most instances, the researcher faces a missing data process that cannot be classified as ignorable. Whether the source of this non-ignorable missing data process is known or unknown, the researcher must still proceed to the next step of the process and assess the extent and impact of the missing data.

LEVELS OF MISSINGNESS In addition to the distinction between ignorable and not ignorable missing data, the researcher should understand what forms of missing data are likely to impact the research. Note that the missing data process refers to whether a case has a missing value or not, but does not relate to the actual value that is missing. Thus, **missingness** is concerned with the absence or presence of a missing/valid value. Determining how that missing data value might be imputed is addressed once the type of missing data process is determined. Newman [66] proposed three levels of missingness described below that follow a hierarchical arrangement:

Item-level The level of missingness first encountered, this is when a value is not available (i.e., a respondent does not answer a question, a data field is missing in a customer record, etc.). This is the level at which remedies for missing data (e.g., imputation) are identified and performed.

Construct-level This level of missingness is when item-level missing data acts to create a missing value for an entire construct of interest. A common example is when a respondent has missing data on all of the items for a scale, although it could also apply to single-item scales as well. Since constructs are the level of interest in most research questions, the missing data become impactful on the results through its actions at the construct level.

Person-level this final level is when a participant does not provide responses to any part of the survey. Typically also known as non-response, it potentially represents influences from both characteristics of the respondent (e.g., general reluctance to participate) as well as possible data collection errors (e.g., poorly designed or administered survey instrument).

While most missing data analysis occurs at the item level, researchers should still be aware of the impact at the construct-level (e.g., the impact on scale scores when using only valid data [66]) and the factors impacting person-level missingness and how they might be reflected in either item-level and even construct-level missing data. For example, person-level factors may make individuals unresponsive to all items of a particular construct, so while we might think of them as item-level issues, they are actually of a different order.

Step 2: Determine the Extent of Missing Data Given that some of the missing data are not ignorable and we understand the levels of missingness in our data, the researcher must next examine the patterns of the missing data and determine the extent of the missing data for individual variables, individual cases, and even overall (e.g., by person). The primary issue in this step of the process is to *determine whether the extent or amount of missing data is low enough to not affect the results, even if it operates in a nonrandom manner.* If it is sufficiently low, then any of the approaches for remedying missing data may be applied. If the missing data level is not low enough, then we must first determine the randomness of the missing data process before selecting a remedy (step 3). The unresolved issue at this step is this question: What is low enough? In making the assessment as to the extent of missing data, the researcher may find that the deletion of cases and/or variables will reduce the missing data to levels that are low enough to allow for remedies without concern for creating biases in the results.

ASSESSING THE EXTENT AND PATTERNS OF MISSING DATA The most direct means of assessing the extent of missing data is by tabulating (1) the percentage of variables with missing data for each case and (2) the number of cases with missing data for each variable. This simple process identifies not only the extent of missing data, but any exceptionally high levels of missing data that occur for individual cases or observations. The researcher should look for any nonrandom

patterns in the data, such as concentration of missing data in a specific set of questions, attrition in not completing the questionnaire, and so on. Finally, the researcher should determine the number of cases with no missing data on any of the variables, which will provide the sample size available for analysis if remedies are not applied.

With this information in hand, the important question is: Is the missing data so high as to warrant additional diagnosis? At issue is the possibility that either ignoring the missing data or using some remedy for substituting values for the missing data can create a bias in the data that will markedly affect the results. Even though most discussions of this issue require researcher judgment, the two guidelines below apply:

- *10 percent or less generally acceptable*. Cases or observations with up to 10 percent missing data are generally acceptable and amenable to any imputation strategy. Notable exceptions are when nonrandom missing data processes are known to be operating and then they must be dealt with [62, 70].
- *Sufficient minimum sample.* Be sure that the minimum sample with complete data (i.e., no missing data across all the variables), is sufficient for model estimation.

If it is determined that the extent is acceptably low and no specific nonrandom patterns appear, then the researcher can employ any of the imputation techniques (step 4) without biasing the results in any appreciable manner. If the level of missing data is too high, then the researcher must consider specific approaches to diagnosing the randomness of the missing data processes (step 3) before proceeding to apply a remedy.

DELETING INDIVIDUAL CASES AND/OR VARIABLES Before proceeding to the formalized methods of diagnosing randomness in step 3, the researcher should consider the simple remedy of deleting offending case(s) and/or variable(s) with excessive levels of missing data. 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. Moreover, in many cases where a nonrandom pattern of missing data is present, this solution may be the most efficient. Again, no firm guidelines exist on the necessary level for exclusion (other than the general suggestion that the extent should be "large"), but any decision should be based on both empirical and theoretical considerations, as listed in Rules of Thumb 2-1.

Ultimately the researcher must compromise between the gains from deleting variables and/or cases with missing data versus the reduction in sample size and variables to represent the concepts in the study. Obviously, variables or

How Much Missing Data Is Too Much?

Missing data under 10 percent for an individual case or observation can generally be ignored, except when the missing data occurs in a specific nonrandom fashion (e.g., concentration in a specific set of questions, attrition at the end of the questionnaire, etc.) [62, 70].

The number of cases with no missing data must be sufficient for the selected analysis technique if replacement values will not be substituted (imputed) for the missing data.

Deletions Based on Missing Data

Variables with as little as 15 percent missing data are candidates for deletion [43], but higher levels of missing data (20% to 30%) can often be remedied.

Be sure the overall decrease in missing data is large enough to justify deleting an individual variable or case.

Cases with missing data for dependent variable(s) typically are deleted to avoid any artificial increase in relationships with independent variables.

When deleting a variable, ensure that alternative variables, hopefully highly correlated, are available to represent the intent of the original variable.

Always consider performing the analysis both with and without the deleted cases or variables to identify any marked differences.

cases with 50 percent or more missing data should be deleted, but as the level of missing data decreases, the researcher must employ more judgment and "trial and error." As we will see when discussing imputation methods, assessing multiple approaches for dealing with missing data is preferable.

Step 3: Diagnose the Randomness of the Missing Data Processes Having determined that the extent of missing data is substantial enough to warrant action, the next step is to ascertain the degree of randomness present in the missing data, which then determines the appropriate remedies available. Assume for the purposes of illustration that information on two variables (*X* and *Y*) is collected. *X* has no missing data, but *Y* has some missing data. A nonrandom missing data process is present between *X* and *Y* when significant differences in the values of *X* occur between cases that have valid data for *Y* versus those cases with missing data on *Y*. Any analysis must explicitly accommodate any nonrandom missing data process (i.e., missingness) between *X* and *Y* or else bias is introduced into the results.

LEVELS OF RANDOMNESS OF THE MISSING DATA PROCESS Missing data processes can be classified into one of three types [62, 31, 59, 93, 74]. Two features distinguish the three types: (a) the randomness of the missing values among the values of *Y* and (b) the degree of association between the missingness of one variable (in our example *Y*) and other observed variable(s) in the dataset (in our example *X*). Figure 2.6 provides a comparison between the various missing data patterns. Using Figure 2.6 as a guide, let's examine these three types of missing data processes.

Missing Data at Random (MAR) Missing data are termed **missing at random (MAR)** if the missing values of *Y* depend on *X*, but not on *Y*. In other words, 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. In Figure 2.6, the missing values of Y are random (i.e., spread across all values), but having a missing value on *Y* does relate to having low values of *X* (e.g., only values 3 or 4 of *X* correspond to missing values on *Y*). Thus, *X* is associated with the missingness of *Y*, but not the actual values of *Y* that are missing. Even though the missing data process is random in the sample, its values are not generalizable to the population. Most often, the data are missing randomly within subgroups, but differ in levels between subgroups. The researcher must determine the factors determining the subgroups and the varying levels between groups.

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

Figure 2.6
Missing Data Processes: MCAR, MAR and MNAR

Complete Data		Missing Data Process for *Y*		
X	***Y***	**MCAR:**	**MAR:**	**MNAR:**
3	9	9	Missing	9
3	5	5	Missing	5
4	1	Missing	Missing	Missing
4	3	3	Missing	Missing
5	2	Missing	2	Missing
6	6	Missing	6	6
7	7	7	7	7
7	4	4	4	Missing
8	5	5	5	5
9	9	Missing	9	9
Characteristics of the Missing Data Process				
Pattern of missing values of *Y*		**Random: Across all values of *Y***	**Random: Across all values of *Y***	**Nonrandom: Only lowest values of *Y***
Relationship of *X* to missingness of *Y*		**No Across all values of *X***	**Yes Lowest values of *X***	**No Across all values of *X***

Adapted from [31].

higher frequency for males than females. Even though the missing data process is operating in a random manner within the gender variable, 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.

Most missing data processes are in some manner MAR, thus necessitating a thorough missing data analysis whenever missing data is present. In years past, MAR missing data processes presented a dilemma for the researcher as the available remedies typically resulted in some form of bias in the results. But recent development of the model-based methods of imputation have provided imputation options that can easily accommodate MAR missing data processes.

Missing Completely at Random (MCAR) 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 association to the other observed variables, characterized as "purely haphazard missingness" [31]. In Figure 2.6, the missing values of *Y* are random across all *Y* values and there is no relationship between missingness on *Y* and the *X* variable (i.e., missing *Y* values occur at all different values of *X*). Thus, MCAR is a special condition of MAR since the missing values of *Y* are random, but it differs in that there is no association with any other observed variable(s). This also means that the cases with no missing data are simply a random subset of the total sample. In simple terms, the cases with missing data are indistinguishable from cases with complete data, except for the presence of missing data.

From our earlier example, an MCAR situation would be shown by the fact that the missing data for household income were randomly missing in equal proportions for both males and females. In this missing data process, any of the remedies can be applied without making allowances for the impact of any other variable or missing data process.

Not Missing at Random (MNAR) The third type of missing data process is **missing not at random (MNAR)**, which as the name implies, has a distinct nonrandom pattern of missing values. What distinguishes MNAR from the other two types is that the nonrandom pattern is among the *Y* values and the missingness of the *Y* values may or may not be related to the *X* values. This is the most problematic missing data process for several reasons. First, it is generally undetectable empirically and only becomes apparent through subjective analysis. In Figure 2.6, all of the missing values of *Y* were the lowest values (e.g., values 1, 2, 3, and 4). Unless we knew from other sources, we would not suspect that valid values extended below the lowest observed value of 5. Only researcher knowledge of the possibility of values lower than 5 might indicate that this was a nonrandom process. Second, there is no objective method to empirically impute the missing values. Researchers should be very careful when faced with MNAR situations as biased results can be substantial and threats to generalizability are serious.

Referring back to our household income example, an MNAR process would be indicated if all individuals with high incomes, whether male or female, would not report their income level. Thus, all the observed values for income would be biased downwards since no high income values were in the dataset.

Defining The Type of Missing Data Process Two of the types exhibit levels of randomness for the missing data of *Y*. One type requires special methods to accommodate a nonrandom component (MAR) while the second type (MCAR) is sufficiently random to accommodate any type of missing data remedy [62, 31, 37, 79]. Although both types seem to indicate that they reflect random missing data patterns, only MCAR allows for the use of any remedy desired. The distinction between these two types is in the generalizability to the population in their original form. The third type, MNAR, has a substantive nonrandom pattern to the missing data that precludes any direct imputation of the values. Since MNAR requires subjective judgment to identify, researchers should always be aware of the types of variables (e.g., sensitive personal characteristics or socially desirable responses) that may fall into this type of missing data pattern.

DIAGNOSTIC TESTS FOR LEVELS OF RANDOMNESS As previously noted, the researcher must ascertain whether the missing data process occurs in a completely random manner (MCAR) or with some relationship to other variables (MAR). When the dataset is small, the researcher may be able to visually see such patterns or perform a set of simple

calculations (such as in our simple example at the beginning of the chapter). However, as sample size and the number of variables increases, so does the need for empirical diagnostic tests. Some statistical programs add techniques specifically designed for missing data analysis (e.g., Missing Value Analysis in IBM SPSS), which generally include one or both diagnostic tests.

t Tests of Missingness The first diagnostic 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. The researcher can create an indicator variable with a value of 1 if there is a missing value for Y and a zero if Y has a valid value. Thus, the indicator value just measures missingness—presence or absence. Statistical tests are then performed between the missingness indicator and other observed variables—t tests for metric variables and chi-square tests for nonmetric variables. Significant differences between the two groups indicates a relationship between missingness and the variable being tested—an indication of a MAR missing data process.

Let us use our earlier example of household income and gender, plus a measure of life satisfaction. We would first form two groups of respondents, those with missing data on the household income question and those who answered the question. First, we would compare the percentages of gender for each group. If one gender (e.g., males) was found in greater proportion in the missing data group (i.e., a significant chi-square value), we would suspect a nonrandom missing data process. If the variable being compared is metric (e.g., life satisfaction) instead of categorical (gender), then t tests are performed to determine the statistical significance of the difference in the variable's mean between the two groups. The researcher should examine a number of variables to see whether any consistent pattern emerges. Remember that some differences will occur by chance, but either a large number or a systematic pattern of differences may indicate an underlying nonrandom pattern.

Little's MCAR Test A second approach is an overall test of randomness that determines whether the missing data can be classified as MCAR [58]. 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.

Is it MAR or MCAR? As a result of these tests, the missing data process is classified as either MAR or MCAR, which then determines the appropriate types of potential remedies. In reality, a researcher is most likely faced with a combination of missing data processes within any given set of variables. The need for distinguishing between MCAR and MAR used to be more impactful when the imputation methods were limited and most of them created bias in the imputed values. But the emergence of the model-based methods that can provide unbiased imputed values for MCAR or MAR data has alleviated the necessity of these distinctions. It is still useful for the researcher to understand what types of missing data processes are operating in the data being analyzed and to also ensure that any variables involved in MAR relationships are included in the model-based methods.

Step 4: Select the Imputation Method At this step of the process, the researcher must select the approach used for accommodating missing data in the analysis. This decision is based primarily on whether the missing data are MAR or MCAR, but in either case the researcher has several options for imputation [42, 62, 73, 78, 31, 37, 21]. **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 [27]:

> 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.

All of the imputation methods discussed in this section are used primarily with metric variables; nonmetric variables are left as missing unless a specific modeling approach is employed [e.g., 92]. Nonmetric variables are not amenable to imputation because even though estimates of the missing data for metric variables can be made with such values as a mean of all valid values, no comparable measures are available for nonmetric variables. As such, nonmetric variables require an estimate of a specific value rather than an estimate on a continuous scale. It is 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.

In the following sections we divide the imputation techniques into two classes: those that require an MCAR missing data process and those appropriate when facing a MAR situation. We should note that most of the "traditional" imputation methods require MCAR and were generally applied whether the MCAR was indicated or not since there were no other options available. The advantages and disadvantages of these methods are discussed to provide a more reasoned approach to selecting one of these methods if necessary. But as is discussed in the methods suitable for MAR situations, these methods are generally preferable to all other methods and have become available in all of the major software packages. So even if a missing data process is MCAR there are substantial benefits from using the model-based methods discussed in the MAR section.

IMPUTATION OF MCAR USING ONLY VALID DATA If the researcher determines that the missing data process can be classified as MCAR, either of two basic approaches be used: using only valid data or defining replacement values for the missing data. We will first discuss the two methods that use only valid data, and then follow with a discussion of the methods based on using replacement values for the missing data.

Some researchers may question whether using only valid data is actually a form of imputation, because no data values are actually replaced. The intent of this approach is to represent the entire sample with those observations or cases with valid data. As seen in the two following approaches, this representation can be done in several ways. The underlying assumption in both is that the missing data are in a random pattern and that the valid data are an adequate representation.

Complete Case Approach 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, also known as the LISTWISE method in IBM SPSS, is available in all statistical programs and is the default method in many programs. Yet the complete case approach has two distinct disadvantages. First, it is most affected by any nonrandom missing data processes, because the cases with any missing data are deleted from the analysis. Thus, even though only valid observations are used, the results are not generalizable to the population. Second, this approach also results in the greatest reduction in sample size, because missing data on any variable eliminates the entire case. It has been shown that with only two percent randomly missing data, more than 18 percent of the cases will have some missing data. Thus, in many situations with even very small amounts of missing data, the resulting sample size is reduced to an inappropriate size when this approach is used. As a result, 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. But even in these instances, most research suggests avoiding the complete case approach if at all possible [e.g., 66].

Using All-Available Data The second imputation method using only valid data also 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 every valid value. For example, assume that there are three variables of interest (V_1, V_2, and V_3). To estimate the mean of each variable, all of the valid values are used for each respondent. If a respondent is missing data for V_3, the valid values for V_1 and V_2 are still used to calculate the means. Correlations are calculated in the same manner, using all valid pairs of data. Assume that one respondent has valid data for only V_1 and V_2, whereas a second respondent has valid data for V_2 and V_3. When calculating the correlation between V_1 and V_2, the values from the first respondent will be used, but not for correlations of V_1 and V_3 or V_2 and V_3. Likewise, the second respondent will contribute data for calculating the correlation of V_2 and V_3, but not the other correlations.

Known as the **all-available approach**, this method (e.g., the PAIRWISE option in SPSS) is primarily used to estimate correlations and maximize the pairwise information available in the sample. The distinguishing characteristic of this approach is that the characteristic of a variable (e.g., mean, standard deviation) or the correlation for a pair of variables is based on a potentially unique set of observations. It is to be expected that the number of observations used in the calculations will vary for each correlation. The imputation process occurs not by replacing the missing data, but instead by using the obtained correlations on just the cases with valid data as representative for the entire sample.

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 arise. First, correlations may be calculated that are "out of range" and inconsistent with the other correlations in the correlation matrix [65]. Any correlation between X and Y is constrained by their correlation to a third variable Z, as shown in the following formula:

$$Range\ of\ r_{XY} = r_{XZ}r_{YZ} \pm \sqrt{(1-r_{XZ}^2)(1-r_{YZ}^2)}$$

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, which 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, many procedures do not include this adjustment process. In extreme cases, the estimated variance/covariance matrix is not positive definite [53]. Finally, while the all-available approach does generate the distributional characteristics to allow for estimation of models (e.g., regression) it does not provide for any case-level diagnostics (e.g., residuals, influential cases) that may be useful in model diagnostics. All of these problems must be considered when selecting the all-available approach.

IMPUTATION OF MCAR BY USING KNOWN REPLACEMENT VALUES The second form of imputation for MCAR missing data processes involves replacing missing values with estimated values based on other information available in the sample. The principal advantage is that once the replacement values are substituted, all observations are available for use in the analysis. The options vary from the direct substitution of values to estimation processes based on relationships among the variables. This section focuses on the methods that use a known replacement value, while the following section addresses the most widely used methods that calculate a replacement value from the observations [62, 73, 78, 93, 22].

The common characteristic in methods is to identify a known value, most often from a single observation, that is used to replace the missing data. The observation may be from the sample or even external to the sample. A primary consideration is identifying the appropriate observation through some measure of similarity. The observation with missing data is "matched" to a similar case, which provides the replacement values for the missing data. The trade-off in assessing similarity is between using more variables to get a better "match" versus the complexity in calculating similarity.

Hot or Cold Deck Imputation In this approach, the researcher substitutes a value from another source for the missing values. In the "**hot deck**" method, the value comes from another observation in the sample that is deemed similar. Each observation with missing data is paired with another case that is similar on a variable(s) specified by the researcher. Then, missing data are replaced with valid values from the similar observation. Recent advances in computer software have advanced this approach to more widespread use [65]. "**Cold deck**" imputation derives the replacement value from an external source (e.g., prior studies, other samples). Here the researcher must be sure that the replacement value from an external source is more valid than an internally generated value. Both variants of this method provide the researcher with the option of replacing the missing data with actual values from similar observations that may be deemed more valid than some calculated value from all cases, such as the mean of the sample.

Case Substitution In this method, entire observations with missing data are replaced by choosing another non-sampled 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 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. At issue is the ability to obtain these additional observations not included in the original sample.

IMPUTATION OF MCAR BY CALCULATING REPLACEMENT VALUES The second basic approach involves calculating a replacement value from a set of observations with valid data in the sample. The assumption is that a value derived from all other observations in the sample is the most representative replacement value. These methods, particularly mean substitution, are more widely used due to their ease in implementation versus the use of known values discussed previously.

Mean Substitution One of the most widely used methods, **mean substitution** replaces the missing values for a variable with the mean value of that variable calculated from all valid responses. The rationale of this approach is that the mean is the best single replacement value. This approach, although it is used extensively, has several disadvantages. First, it understates the variance estimates by using the mean value for all missing 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. A variant of this method is group mean substitution, where observations with missing data are grouped on a second variable, and then mean values for each group are substituted for the missing values within the group. It is many times the default missing value imputation method due to its ease of implementation, but researchers should be quite cautious in its use, especially as the extent of missing data increases. The impact of substituting a single value will be demonstrated in the HBAT example that follows.

Regression Imputation In this method, **regression analysis** (described in Chapter 5) is used to predict the missing values of a variable based on its relationship to other variables in the dataset. First, a predictive equation is formed for each variable with missing data and estimated from all cases with valid data. Then, replacement values for each missing value are calculated from that observation's values on the variables in the predictive equation. Thus, the replacement value is derived based on that observation's values on other variables shown to relate to the missing value.

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. Fourth, the sample must be large enough to allow for a sufficient number of observations to be used in making each prediction. 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) and require 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 affect the generalizability of the results.

OVERVIEW OF MCAR IMPUTATION METHODS The methods for MCAR imputation are those most widely used in past research and are well known to researchers. They vary in their impact on the results (e.g., the variance reducing properties of listwise deletion or the potential statistical inconsistencies of the all-available approach), but they still provide unbiased results if the missing data process can be classified as MCAR. The widespread availability of these approaches makes them the method of choice in most research. But in many of these situations, testing for the assumption of MCAR was not made and little was achieved, thus resulting in potential biases in the final results. The objective of

most researchers in these instances was to select the least impactful approach from those available and attempt to control any impacts by reducing the extent of missing data. As we will discuss in the next section, the availability of model-based approaches that accommodate MAR missing data processes provides a new avenue for researchers to deal with the issues surrounding missing data.

IMPUTATION OF A MAR MISSING DATA PROCESS If a nonrandom or MAR missing data pattern is found, the researcher should apply only one remedy—the modeling approach specifically designed to deal with this [62, 31, 37, 59, 2]. Application of any other method introduces bias into the results. This set of procedures explicitly incorporates the MAR missing data process into the analysis and exemplifies what has been termed the "inclusive analysis strategy" [31] which also includes auxiliary variables into the missing data handling procedure, as will be discussed later. As a result, this set of procedures is comparable to "Missing Data 2.0" since it provides a greatly expanded and efficient manner for handling missing data. As noted by Allison [3] the limited use of MAR-appropriate methods is primarily because of lack of awareness of most researchers and the lack of software availability. But as these methods become available in all of the software platforms, their use should increase. Their inclusion in recent versions of the popular software programs (e.g., the Missing Value Analysis module of IBM SPSS and the PROC MI procedure in SAS) should increase its use. Comparable procedures employ structural equation modeling (Chapter 9) to estimate the missing data [6, 13, 28], but detailed discussion of these methods is beyond the scope of this chapter.

Maximum Likelihood and EM 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 [40, 62]. Maximum likelihood is not a technique, but a fundamental estimation methodology. However, its application in missing data analysis has evolved based on two approaches. The first approach is the use of maximum likelihood directly in the estimation of the means and covariance matrix as part of the model estimation in covariance-based SEM. In these applications missing data estimation and model estimation are combined in a single step. There is no imputation of missing data for individual cases, but the missing data process is accommodated in the "imputed" matrices for model estimation. The primary drawback to this approach is that imputed datasets are not available and it takes more specialized software to perform [3, 31].

A variation of this method employs maximum likelihood as well, but in an iterative process. The **EM** method [39, 31, 74] is a 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. One notable feature is that this method can produce an imputed dataset, although it has been shown to underestimate the standard errors in estimated models [31, 37].

Multiple Imputation The procedure of **multiple imputation** is, as the name implies, a process of generating multiple datasets with the imputed data differing in each dataset, to provide in the aggregate, both unbiased parameter estimates and correct estimates of the standard errors [75, 31, 32, 57, 35]. As we will see in the following discussion, multiple imputation overcomes the issues associated with MAR missing data processes while still generating complete data sets that can be used with conventional analytical techniques. The only additional condition is that after all of the datasets have been used to estimate models, the parameter estimates and standard errors must be combined to provide a final set of results. The result is a three-step process of multiple imputation:

1. *Generate a set of imputed datasets*. This stage is somewhat similar to some of the single imputation methods described earlier (i.e., stochastic regression imputation), but differs in both how many datasets are imputed (multiple versus only one) and how the model parameter estimates are generated for each dataset. The two most widely used methods for generating the model parameter estimates are forms of Bayesian estimation, either the Markov chain Monte Carlo (MCMC) method [80] or the fully conditional specification (FCS) method [79]. The objective in each method is to provide a set of imputed values that capture not only the "true" imputed values, but also their variability. The FCS has some advantages in terms of the nature of the imputed values, but both methods are widely employed.

A primary consideration in any model-based approach is what variables to be included in the multiple imputation procedure? Obviously, all of the variables to be used in any of the subsequent analyses, including those that have complete data, should be included, as well as outcome variables. Any variable that is thought to be part of the MAR process has to be included or the missing data process risks taking on qualities associated with MNAR [31]. So it is important that all of the associations in the dataset be represented in the imputation process. Some researchers wonder if it is appropriate if the outcome measure is included, but multiple imputation does not make any distinction between the roles of the variables [59]. The number of variables included in the analysis has little impact on the imputation process, so there is no need for variable selection before the imputation proceeds. As we will discuss in the next section, auxiliary variables provide additional information to the imputation process even if not used in subsequent analyses.

A common question is "How many datasets should be generated?" While theoretically the optimum number would be an infinite number, practical considerations have shown that a minimum of five imputed datasets will suffice [75, 80, 79]. Recent research has indicated, however, that a considerably larger number of datasets, as large as 20 datasets, will provide safeguards against some potential issues [31, 38]. While this is not a trivial consideration if the datasets being analyzed are extremely large in size, today's computational power makes this less of an issue than in years past. And the number of imputed datasets does not impact the combination of results in any substantive fashion.

2 *Estimate the model.* This second step is estimation of the desired analysis. One of the primary advantages of multiple imputation is that it can be used with almost all of the most widely used linear models – t tests, the general linear model (e.g., multiple regression and ANOVA/MANOVA) and the generalized linear model (e.g., logistic regression). Thus, the researcher can perform any of these methods as they would a single dataset with no missing data. Each imputed dataset is analyzed separately, comparable to what is performed in IBM SPSS (SPLIT FILES) or SAS (BY command) with an imputation variable defining the separate imputed datasets.

3 *Combining results from multiple imputed datasets.* The final step is to combine the results from the individual imputed datasets into a final set of "combined" results. Procedures are available (e.g., PROC MIANALYZE in SAS) for combining the parameter estimates so that the researcher now has the unbiased estimate of the parameters and the standard errors.

An additional consideration that is particularly useful in multiple imputation is inclusion of auxiliary variables. **Auxiliary variables** are variables that will not be used in the analysis, but in some way may relate to missingness and thus be representative of the MAR process [21]. And while some research has suggested that in a few instances irrelevant variables may cause issues [85], the general consensus is that there is little harm caused by including a wide array of auxiliary variables, even if they ultimately have little impact [21]. An interesting approach to reduce the number of auxiliary variables while still incorporating their effects is to employ principal components analysis and include the component scores as the auxiliary variables [46]. No matter the approach used, the inclusion of auxiliary variables provides the researcher additional variables, outside those in the analysis, to represent and then accommodate the MAR process.

Maximum Likelihood versus Multiple Imputation The choice between the two approaches is more a matter of researcher preference, as both methods provide equal results in large sample (e.g., $>$ 200 for simpler models) situations where equivalent variables and models of imputation are used (e.g., where auxiliary variables used in both). Some researchers may prefer maximum likelihood because of its integrated nature (e.g., imputation of dataset and estimation in single step), but others may desire an imputed dataset that has limitations in maximum likelihood. On the other hand, multiple imputation may be preferred due to it use of a wide range of conventional techniques and seemingly straightforward approach, but it still has issues such as the results vary each time it is performed since the imputed datasets are randomly generated and there is not a single dataset for which additional diagnostics (e.g., casewise analysis) is performed.

SUMMARY The range of possible imputation methods varies from the conservative (complete data method) to those that attempt to replicate the MAR missing data as much as possible (e.g., model-based methods like maximum likelihood/EM and multiple imputation). What should be recognized is that each method has advantages and disadvantages, such that the researcher must examine each missing data situation and select the most appropriate imputation method. Figure 2.7 provides a brief comparison of the imputation method, but a quick review shows that no single

Figure 2.7
Comparison of Imputation Techniques for Missing Data

Imputation Method	*Advantages*	*Disadvantages*	*Best Used When:*
Methods for MCAR Missing Data Processes			
Imputation Using Only Valid Data			
Complete Data	Simplest to implement Default for many statistical programs	Most affected by nonrandom processes Greatest reduction in sample size Lowers statistical power	Large sample size Strong relationships among variables Low levels of missing data
All Available Data	Maximizes use of valid data Results in largest sample size possible without replacing values	Varying sample sizes for every imputation Can generate "out of range" values for correlations and eigenvalues	Relatively low levels of missing data Moderate relationships among variables
Imputation Using Known Replacement Values			
Case Substitution	Provides realistic replacement values (i.e., another actual observation) rather than calculated values	Must have additional cases not in the original sample Must define similarity measure to identify replacement case	Additional cases are available Able to identify appropriate replacement cases
Hot and Cold Deck Imputation	Replaces missing data with actual values from the most similar case or best known value	Must define suitably similar cases or appropriate external values	Established replacement values are known, or Missing data process indicates variables upon which to base similarity
Imputation by Calculating Replacement Values			
Mean Substitution	Easily implemented Provides all cases with complete information	Reduces variance of the distribution Distorts distribution of the data Depresses observed correlations	Relatively low levels of missing data Relatively strong relationships among variables
Regression Imputation	Employs actual relationships among the variables Replacement values calculated based on an observation's own values on other variables Unique set of predictors can be used for each variable with missing data	Reinforces existing relationships and reduces generalizability Must have sufficient relationships among variables to generate valid predicted values Understates variance unless error term added to replacement value Replacement values may be "out of range"	Moderate to high levels of missing data Relationships sufficiently established so as to not impact generalizability Software availability
Model-Based Methods for MAR Missing Data Processes			
Maximum Likelihood/EM	Accommodates both MCAR and MAR "Single-step" imputation and model estimation Best statistical results (unbiased and efficient) Directly estimates interaction effects	Requires specialized statistical routines Limited in statistical methods available Harder to incorporate large number of auxiliary variables	MAR process well-defined Fewest decisions required by researcher
Multiple Imputation	Can be used with wide array of statistical techniques Imputed datasets allow for casewise diagnostics Accommodates both metric and nonmetric data Allows for large number of auxiliary variables	Requires more decisions and steps for imputation, then combination of results Results can vary slightly due to random sampling process in creating imputed datasets	Less well-defined MAR process Need for large number of auxiliary variables Use of several statistical models on same dataset

Imputation of Missing Data Based On Extent of Missing Data

Under 10%	Any of the imputation methods can be applied when missing data are this low, although the complete case method has been shown to be the least preferred
10% to 20%	The increased presence of missing data makes the all-available, hot deck case substitution, and regression methods most preferred for MCAR data, whereas model-based methods are necessary with MAR missing data processes
Over 20%	If it is deemed necessary to impute missing data when the level is over 20 percent, the preferred methods are: The regression method for MCAR situations Model-based methods when MAR missing data occur
Imputation Method By Type of Missing Data Process	
MCAR	Possible missing data process, but requires strict conditions not generally met Any imputation method can provide unbiased estimates if MCAR conditions met, but the model-based methods also provide protection against unidentified MAR relationships and provide appropriate estimates of standard errors
MAR	Most likely missing data process Only the model-based methods (maximum likelihood/EM and multiple imputation) can provide imputed data which results in unbiased estimates and correct standard errors

method is best in all situations. However, some general suggestions (see Rules of Thumb 2-2) can be made based on the extent of missing data.

Given the many imputation methods available, the researcher should also strongly consider following a multiple imputation strategy if the MCAR methods are used, whereby a combination of several methods is used. 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 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. The model-based methods, however, provide the best approaches to avoid biased estimates due to any underlying MAR missing data processes.

AN ILLUSTRATION OF MISSING DATA DIAGNOSIS WITH THE FOUR-STEP PROCESS

To illustrate the four-step process of diagnosing the patterns of missing data and the application of possible remedies, a new dataset is introduced (a complete listing of the observations and an electronic copy are available online). This dataset was collected during the pretest of a questionnaire used to collect the data described in Chapter 1. The pretest involved 70 individuals and collected responses on 14 variables (9 metric variables, V_1 to V_9, and 5 nonmetric variables, V_{10} to V_{14}). The variables in this pretest do not coincide directly with those in the HBAT dataset, so they will be referred to just by their variable designation (e.g., V_3).

In the course of pretesting, however, missing data occurred. The following sections detail the diagnosis of the missing data through the four-step process. All of the major software programs, including *R*, have missing data routines for performing both the descriptive analyses and the various imputation methods. The analyses described in these next sections were performed with the Missing Value Analysis module in IBM SPSS, but all of the analyses

can be replicated by data manipulation and conventional analysis. Examples are available in the online resources at the text's websites.

Step 1: Determine the Type of Missing Data All the missing data in this example are unknown and not ignorable because they are due to non-response by the respondent. As such, the researcher is forced to proceed in the examination of the missing data processes.

Step 2: Determine the Extent of Missing Data The objective in this step is to determine whether the extent of the missing data is sufficiently high enough to warrant a diagnosis of examining cases and variables for possible deletion or at a low enough level to proceed directly to ascertaining the randomness of the missing data process (step 3). Thus, the researcher is interested in the level of missing data on a case and variable basis, plus the overall extent of missing data across all cases.

Table 2.1 contains the descriptive statistics for the observations with valid values, including the percentage of cases with missing data on each variable. Viewing the metric variables (V_1 to V_9), we see that the lowest amount of missing data is six cases for V_6 (9% of the sample), ranging up to 30 percent missing (21 cases) for V_1. This frequency makes V_1 and V_3 possible candidates for deletion in an attempt to reduce the overall amount of missing data. All of the nonmetric variables (V_{10} to V_{14}) have low levels of missing data and are acceptable.

Table 2.1 Summary Statistics of Missing Data for Original Sample

				Missing Data	
Variable	Number of Cases	Mean	Standard Deviation	Number	Percent
V_1	49	4.0	93	21	30
V_2	57	1.9	93	13	19
V_3	53	8.1	1.41	17	24
V_4	63	5.2	1.17	7	10
V_5	61	2.9	78	9	13
V_6	64	2.6	72	6	9
V_7	61	6.8	1.68	9	13
V_8	61	46.0	9.36	9	13
V_9	63	4.8	.83	7	10
V_{10}	68	NA	NA	2	3
V_{11}	68	NA	NA	2	3
V_{12}	68	NA	NA	2	3
V_{13}	69	NA	NA	1	1
V_{14}	68	NA	NA	2	3

NA = Not applicable to nonmetric variables

Summary of Cases

Number of Missing Data per Case	Number of Cases	Percent of Sample
0	26	37
1	15	21
2	19	27
3	4	6
7	6	9
Total	70	100%

Moreover, the amount of missing data per case is also tabulated. Although 26 cases have no missing data, it is also apparent that six cases have 50 percent missing data, making them likely to be deleted because of an excessive number of missing values. Table 2.2 shows the missing data patterns for all the cases with missing data, and these six cases are listed at the bottom of the table. As we view the patterns of missing data, we see that all the missing data for the nonmetric variables occurs in these six cases, such that after their deletion there will be only valid data for these variables.

Table 2.2 Patterns of Missing Data by Case

			Missing Data Patterns													
Case	# Missing	% Missing	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}	V_{11}	V_{12}	V_{13}	V_{14}
205	1	7.1			S											
202	2	14.3	S		S											
250	2	14.3	S		S											
255	2	14.3	S		S											
269	2	14.3	S		S											
238	1	7.1	S													
240	1	7.1	S													
253	1	7.1	S													
256	1	7.1	S													
259	1	7.1	S													
260	1	7.1	S													
228	2	14.3	S			S										
246	1	7.1				S										
225	2	14.3			S	S										
267	2	14.3			S	S										
222	2	14.3			S		S									
241	2	14.3			S		S									
229	1	7.1					S									
216	2	14.3	S				S									
218	2	14.3	S				S									
232	2	14.3	S	S												
248	2	14.3	S	S												
237	1	7.1		S												
249	1	7.1		S												
220	1	7.1		S												
213	2	14.3		S	S											
257	2	14.3		S	S											
203	2	14.3		S					S							
231	1	7.1							S							
219	2	14.3							S	S						
244	1	7.1								S						
227	2	14.3		S						S						
224	3	21.4	S	S						S						
268	1	7.1									S					
235	2	14.3						S			S					
204	3	21.4	S		S						S					
207	3	21.4	S		S						S					
221	3	21.4	S		S				S							
245	7	50.0	S		S		S		S	S				S		S
233	7	50.0		S	S		S	S			S			S		S
261	7	50.0		S	S			S	S	S	S		S			
210	7	50.0				S	S	S	S	S	S	S				
263	7	50.0		S		S	S	S	S	S		S				
214	7	50.0	S			S		S	S	S			S		S	

Note: Only cases with missing data are shown.

S = missing data.

Even though it is obvious that deleting the six cases will improve the extent of missing data, the researcher must also consider the possibility of deleting a variable(s) if the missing data level is high. The two most likely variables for deletion are V_1 and V_3, with 30 percent and 24 percent missing data, respectively. Table 2.3 provides insight into the impact of deleting one or both by examining the patterns of missing data and assessing the extent that missing data will be decreased. For example, the first pattern (first row) shows no missing data for the 26 cases. The pattern of the second row shows missing data only on V_3 and indicates that only one case has this pattern. The far right column indicates the number of cases having complete information if this pattern is eliminated (i.e., these variables deleted or replacement values imputed). In the case of this first pattern, we see that the number of cases with

Table 2.3 Missing Data Patterns

	Missing Data Patterns													
Number of Cases	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}	V_{11}	V_{12}	V_{13}	V_{14}	Number of Complete Cases if Variables Missing in Pattern Are Not Used
26															26
1			X												27
4	X		X												37
6	X														32
1	X			X											34
1				X											27
2			X	X											30
2			X		X										30
1					X										27
2	X				X										35
2	X	X													37
3		X													29
2		X	X												32
1		X					X								31
1							X								27
1							X	X							29
1								X							27
1		X						X							31
1	X	X						X							40
1									X						27
1						X			X						28
2	X		X				X								40
1	X		X				X								39
1	X		X		X			X				X		X	47
1		X	X		X	X			X			X		X	38
1		X	X			X	X	X	X		X				40
1				X	X	X	X	X	X	X					34
1		X		X	X	X	X	X		X					37
	X			X		X	X	X			X		X		38

Notes: Represents the number of cases with each missing data pattern. For example, reading down the column for the first three values (26, 1, and 4), 26 cases are not missing data on any variable. Then, one case is missing data on V_3. Then, four cases are missing data on two variables (V_1 and V_3).

complete data would increase by one, to 27, by deleting V_3 because only one case was missing data on only V_3. If we look at the fourth row, we see that six cases are missing data on only V_1, so that if we delete V_1 32 cases will have complete data. Finally, row 3 denotes the pattern of missing data on both V_1 and V_3, and if we delete both variables the number of cases with complete data will increase to 37. Thus, deleting just V_3 adds one case with complete data, just deleting V_1 increases the total by six cases, and deleting both variables increases the cases with complete data by 11, to a total of 37.

For purposes of illustration, we will delete just V_1, leaving V_3 with a fairly high amount of missing data to demonstrate its impact in the imputation process. The result is a sample of 64 cases with now only eight metric variables. Table 2.4 contains the summary statistics on this reduced sample. The extent of missing data decreased markedly just by deleting six cases (less than 10% of the sample) and one variable. Now, one-half of the sample has complete data, only two variables have more than 10 percent missing data, and the nonmetric variables now have all complete data. Moreover, the largest number of missing values for any case is two, which indicates that imputation should not affect any case in a substantial manner.

Having deleted six cases and one variable, we move on to step 3 and diagnosing the randomness of the missing data patterns. This analysis will be limited to the metric variables because the nonmetric variables now have no missing data.

Step 3: Diagnosing the Randomness of the Missing Data Process 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. Hopefully the missing data will be judged MCAR, thus allowing a wider range of remedies in the imputation process. We will first employ a test of comparison between groups of missing and non-missing cases and then conduct an overall test for randomness.

Table 2.4 Summary Statistics for Reduced Sample (Six Cases and V_1 Deleted)

	Number of Cases	Mean	Standard Deviation	Missing Data	
				Number	Percent
V_2	54	1.9	.86	10	16
V_3	50	8.1	1.32	14	22
V_4	60	5.1	1.19	4	6
V_5	59	2.8	.75	5	8
V_6	63	2.6	.72	1	2
V_7	60	6.8	1.68	4	6
V_8	60	46.0	9.42	4	6
V_9	60	4.8	.82	4	6
V_{10}	64			0	0
V_{11}	64			0	0
V_{12}	64			0	0
V_{13}	64			0	0
V_{14}	64			0	0

Summary of Cases		
Number of Missing Data per Case	Number of Cases	Percent of Sample
0	32	50
1	18	28
2	14	22
Total	64	100

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 V_2 are placed in one group and those observations with valid responses for V_2 are placed in another group. Then, these two groups are compared to identify any differences on the remaining metric variables (V_3 through V_9). Once comparisons have been made on all of the variables, new groups are formed based on the missing data for the next variable (V_3) and the comparisons are performed again on the remaining variables. This process continues until each variable (V_2 through V_9; remember V_1 has been excluded) 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. The only noticeable pattern of significant *t* values occurs for V_2, for which three of the eight comparisons (V_4, V_5, and V_6) found significant differences between the two groups. Moreover, only one other instance (groups formed on V_4 and compared on V_2) showed a significant difference. This analysis indicates that although significant differences can be found due to the missing data on one variable (V_2), the effects are limited to only this variable, making it 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.

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. The MCAR missing data process is indicated by a *nonsignificant* statistical level (e.g., greater than .05), showing that the observed pattern *does not* differ from a random pattern. This test is performed in the Missing Value Analysis module of SPSS as well as several other software packages dealing with missing value analysis.

In this instance, Little's MCAR test has a significance level of .583, indicating a nonsignificant difference between the observed missing data pattern in the reduced sample and a random pattern. This result, coupled with the earlier analysis showing minimal differences in a nonrandom pattern, allow for the missing data process to be considered MCAR for all of the variables except for V_2 and perhaps V_4. As a result, the researcher may employ any of the remedies for missing data, because the extent of potential biases seems to be minimal in the patterns of missing data.

Step 4: Selecting an Imputation Method As discussed earlier, numerous imputation methods are available for both MAR and MCAR missing data processes. In this instance, the presence of both MCAR and MAR missing data processes allows researchers to apply all of the imputation methods and then compare their results. The other factor to consider is the extent of missing data. As the missing data level increases, methods such as the complete information method become less desirable due to restrictions on sample size, and the all-available method, regression, and model-based methods become more preferred.

The first option is to use only observations with complete data. The advantage of this approach in maintaining consistency in the correlation matrix is offset in this case, however, by its reduction of the sample to such a small size (32 cases) that it is not useful in further analyses. The next options are to still use only valid data through the all-available method or calculate replacement values through such methods as the mean substitution, the regression-based method (with or without the addition of residuals or error), or the model-building approaches (e.g., EM or multiple imputation without auxiliary variables and multiple imputation with auxiliary variables). All of these methods will be employed and then compared to assess the differences that arise between methods. They could also form the basis for a multiple imputation strategy where all the results are combined into a single overall result.

We will start by examining how the values for individual cases are imputed across the various methods. Then we will examine the distributional characteristics (i.e., means and standard deviations) across methods to see how the aggregate results differ. Finally, we will compare empirical results, first with a set of correlations and then with regression coefficients, to assess how these types of results vary across the various methods. *We should note that this comparison is not to select the "best" imputation method, but instead to understand how each imputation method operates and that by choosing a particular method the researcher is impacting the imputed results.* We should also note that the multiple imputation method was performed both with V_{10} through V_{14} as auxiliary variables in the imputation process and also

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:	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9
V_2	*t* value	.	.7	−2.2	−4.2	−2.4	−1.2	−1.1	−1.2
	Significance	.	.528	**.044**	**.001**	**.034**	.260	.318	.233
	Number of cases (valid data)	54	42	50	49	53	51	52	50
	Number of cases (missing data)	0	8	10	10	10	9	8	10
	Mean of cases (valid data)	1.9	8.2	5.0	2.7	2.5	6.7	45.5	4.8
	Mean cases (missing data)	.	7.9	5.9	3.5	3.1	7.4	49.2	5.0
V_3	*t* value	1.4	.	1.1	2.0	.2	.0	1.9	.9
	Significance	.180	.	.286	.066	.818	.965	.073	.399
	Number of cases (valid data)	42	50	48	47	49	47	46	48
	Number of cases (missing data)	12	0	12	12	14	13	14	12
	Mean of cases (valid data)	2.0	8.1	5.2	2.9	2.6	6.8	47.0	4.8
	Mean cases (missing data)	1.6	.	4.8	2.4	2.6	6.8	42.5	4.6
V_4	*t* value	2.6	−.3	.	.2	1.4	1.5	.2	−2.4
	Significance	**.046**	.785	.	.888	.249	.197	.830	.064
	Number of cases (valid data)	50	48	60	55	59	56	56	56
	Number of cases (missing data)	4	2	0	4	4	4	4	4
	Mean of cases (valid data)	1.9	8.1	5.1	2.8	2.6	6.8	46.0	4.8
	Mean cases (missing data)	1.3	8.4	.	2.8	2.3	6.2	45.2	5.4
V_5	*t* value	−.3	.8	.4	.	−.9	−.4	.5	.6
	Significance	.749	.502	.734	.	.423	.696	.669	.605
	Number of cases (valid data)	49	47	55	59	58	55	55	55
	Number of cases (missing data)	5	3	5	0	5	5	5	5
	Mean of cases (valid data)	1.9	8.2	5.2	2.8	2.6	6.8	46.2	4.8
	Mean cases (missing data)	2.0	7.1	5.0	.	2.9	7.1	43.6	4.6
V_7	*t* value	.9	.2	−2.1	.9	−1.5	.	.5	.4
	Significance	.440	.864	.118	.441	.193	.	.658	.704
	Number of cases (valid data)	51	47	56	55	59	60	57	56
	Number of cases (missing data)	3	3	4	4	4	0	3	4
	Mean of cases (valid data)	1.9	8.1	5.1	2.9	2.6	6.8	46.1	4.8
	Mean cases (missing data)	1.5	8.0	6.2	2.5	2.9	.	42.7	4.7
V_8	*t* value	−1.4	2.2	−1.1	−.9	−1.8	1.7	.	1.6
	Significance	.384	.101	.326	.401	.149	.128	.	.155
	Number of cases (valid data)	52	46	56	55	59	57	60	56
	Number of cases (missing data)	2	4	4	4	4	3	0	4
	Mean of cases (valid data)	1.9	8.3	5.1	2.8	2.6	6.8	46.0	4.8
	Mean cases (missing data)	3.0	6.6	5.6	3.1	3.0	6.3	.	4.5
V_9	*t* value	.8	−2.1	2.5	2.7	1.3	.9	2.4	.
	Significance	.463	.235	.076	.056	.302	.409	.066	.
	Number of cases (valid data)	50	48	56	55	60	56	56	60
	Number of cases (missing data)	4	2	4	4	3	4	4	0
	Mean of cases (valid data)	1.9	8.1	5.2	2.9	2.6	6.8	46.4	4.8
	Mean cases (missing data)	1.6	9.2	3.9	2.1	2.2	6.3	39.5	.

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

without the auxiliary variables. The auxiliary variables, after deletion of the six cases earlier, did not have any missing data to be imputed. But they were included in the imputation method to demonstrate how auxiliary variables could be incorporated and to compare to the results obtained without using them as well. As a result the imputed values for the multiple imputation method with the auxiliary variables may differ somewhat from the other methods as it explicitly incorporates the additional information related to missingness available in these variables.

INDIVIDUAL IMPUTED VALUES We start by examining the actual imputed values for a set of selected cases. In this example, the focus is on V_2 and V_3 since these two variables have the largest extent of missing data. Moreover, V_2 was the variable that most likely had a MAR missing data process. As shown in Table 2.6, three cases were selected with missing data on V_3, three cases with missing data on V_2 and then the two cases that had missing data on both V_2 and V_3.

As we can see from the results, the mean substitution method imputed missing data for V_3 as a value of 8.1 and a value of 1.9 for missing data on V_2. As we will see when we examine these results graphically, some distinct patterns are formed. Also, while the various methods all used different approaches, there is a general consistency in the imputed values for the other methods for both V_2 and V_3. Finally, when we view any single case we see the variation across the five imputed data values in multiple imputation, which demonstrates how the method attempts to estimate a range of imputed values that would cover the range of possible values, versus just a single point estimate.

Figure 2.8 provides a graphical perspective on the imputation process by comparing the different imputed values for the entire data set for V_2 and V_3 across the various methods. Part A shows the cases with valid values for V_2 and

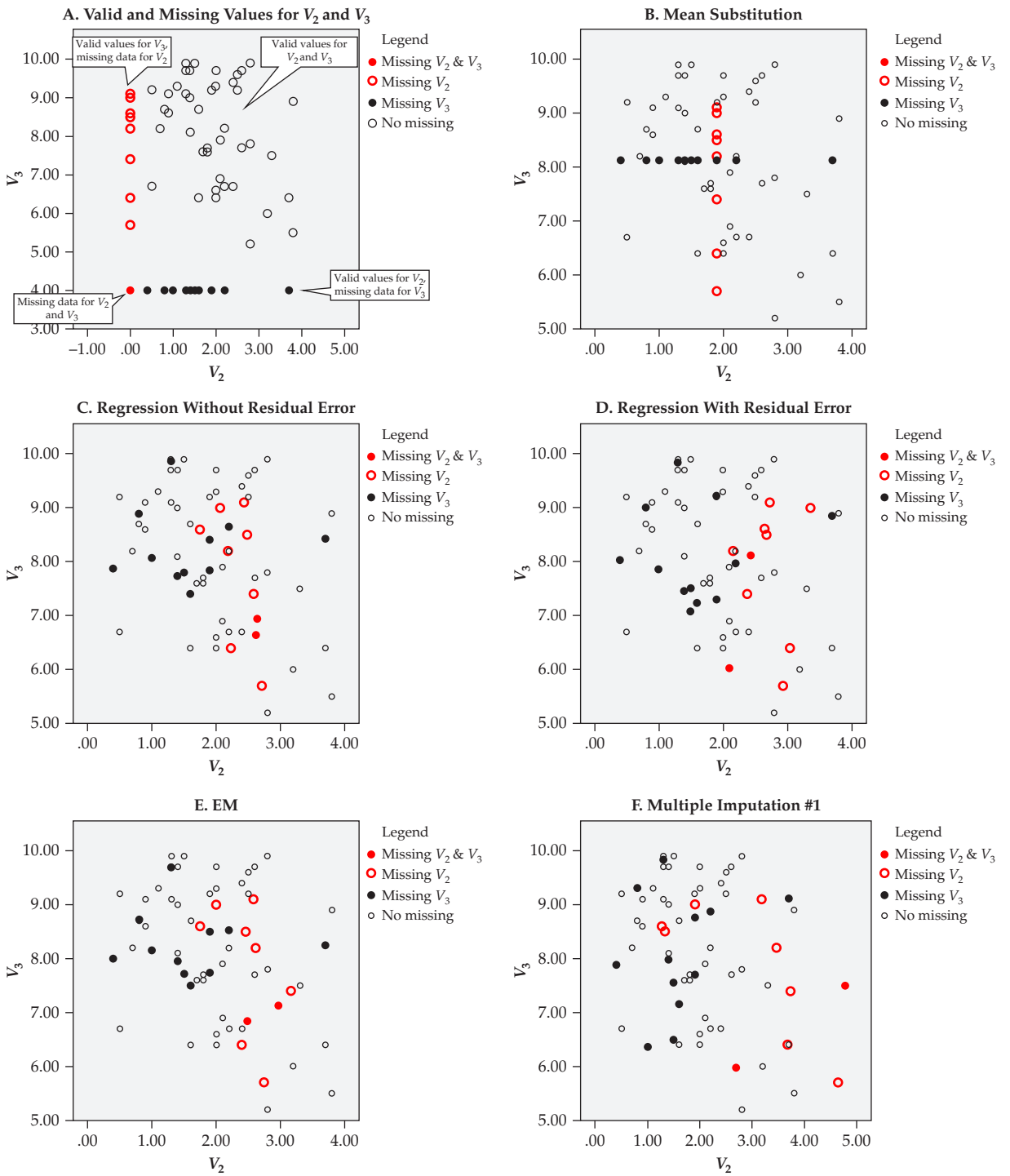
Table 2.6 Imputed Values for Selected Cases with Missing Data on V_2 and V_3

			Imputed Data Values								
ID	Variable	Values	Mean substitution	Regression w/o error	Regression with error	EM	Imputation 1	Imputation 2	Imputation 3	Imputation 4	Imputation 5
			Missing Data on V_3								
202	V_2	0.4									
	V_3	Missing	8.1	7.9	8.0	8.0	7.9	9.2	8.4	7.4	6.8
204	V_2	1.5									
	V_3	Missing	8.1	7.8	7.5	7.7	7.6	7.4	5.6	7.5	5.7
250	V_2	3.7									
	V_3	Missing	8.1	8.4	8.9	8.2	9.1	7.5	9.1	8.2	9.3
			Missing Data on V_2								
227	V_2	Missing	1.9	2.7	2.9	2.7	4.6	4.1	3.5	3.8	3.2
	V_3	5.7									
237	V_2	Missing	1.9	2.6	2.4	3.2	3.7	4.2	4.0	3.4	3.8
	V_3	7.4									
203	V_2	Missing	1.9	2.4	2.7	2.6	3.2	3.7	3.6	4.6	2.9
	V_3	9.1									
			Missing Data on V_2 and V_3								
213	V_2	Missing	1.9	2.6	2.1	2.5	2.7	3.6	4.9	3.4	4.4
	V_3	Missing	8.1	6.7	6.0	6.8	6.0	7.3	5.3	5.7	7.5
257	V_2	Missing	1.9	2.6	2.4	3.0	4.8	4.1	3.4	3.1	3.9
	V_3	Missing	8.1	7.0	8.1	7.1	7.5	6.7	5.8	5.8	8.4

Note: Multiple imputation values estimated using auxiliary values.

Figure 2.8

Comparing Imputed Values for Missing Data on V_2 and V_3



V_3, plus the cases missing data only on V_2 (arrayed on the left side of the chart indicating their valid values on V_3). In addition, the cases with data missing on V_3 are arrayed across the bottom indicating their valid value on V_2, and the cases at the bottom left that are missing on both V_2 and V_3. This illustrates the pattern of cases with valid values as well as the missing data being imputed for both V_2 and V_3.

The other parts of Figure 2.7 show the imputed values for the various imputation techniques. Perhaps most striking is the pattern seen with mean substitution, where the vertical pattern of cases at the value of 1.9 for V_2 and the horizontal pattern at 8.1 for V_3 indicate the single imputed value for missing values on those variables. Moreover, the two cases with missing data on both variables are at the intersection on those patterns. It becomes apparent how the mean substitution method decreases the correlations since there is no variation across the imputed values. This also illustrates how a large number of cases with missing data on both variables would decrease the correlation since all of those points would appear at one location —the mean values of V_2 and V_3—and therefore have no covariance at all to impact the correlation.

The patterns of points are generally consistent across the other imputation methods, although they differ somewhat from the multiple imputation method that used the auxiliary variables (V_{10} to V_{14}). While we do not know what the actual values are for the imputed values and thus cannot select the "best" method, we can see the need for consideration in the imputation method chosen as the results do vary across methods. This should emphasize the notion that there is not any "single best" imputation method and that using multiple imputation methods and noting any differences in the results is warranted. This will be demonstrated in a later section when we view some regression results from each imputation method.

DISTRIBUTIONAL CHARACTERISTICS The next form of comparison is to examine the distributional characteristics (i.e., mean and standard deviation) for each imputation method. Table 2.7 details the results of estimating means and

Table 2.7 Comparing the Estimates of the Means and Standard Deviations Across the Complete Case and Six Imputation Methods

	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9
Mean								
Complete Case (Listwise)	2.003	8.338	5.172	2.881	2.544	6.716	47.719	4.850
All Available (Pairwise)	1.896	8.130	5.147	2.839	2.602	6.790	45.967	4.798
Mean Substitution	1.896	8.130	5.147	2.839	2.602	6.790	45.967	4.798
Regression without error	1.971	8.107	5.139	2.835	2.585	6.844	45.679	4.776
Regression with error	2.014	8.094	5.160	2.833	2.596	6.780	45.578	4.799
EM	1.993	8.108	5.136	2.832	2.583	6.836	45.810	4.768
Multiple Imputation without auxiliary variables[a]	1.964	8.118	5.136	2.844	2.582	6.833	45.702	4.781
Multiple Imputation with auxiliary variables[a]	2.104	8.017	5.154	2.835	2.589	6.802	45.549	4.753
Imputation 1	2.079	8.078	5.177	2.827	2.591	6.827	45.409	4.769
Imputation 2	2.126	8.096	5.168	2.827	2.582	6.769	45.556	4.779
Imputation 3	2.111	7.981	5.137	2.837	2.597	6.782	45.389	4.725
Imputation 4	2.084	7.969	5.137	2.838	2.589	6.816	45.594	4.751
Imputation 5	2.120	7.960	5.149	2.848	2.586	6.815	45.796	4.740
Standard Deviation								
Complete Case (Listwise)	0.840	1.214	1.112	0.685	0.721	1.689	9.669	0.878
All Available (Pairwise)	0.859	1.319	1.188	0.754	0.719	1.675	9.420	0.819
Mean Substitution	0.788	1.164	1.149	0.724	0.713	1.621	9.116	0.793
Regression without error	0.815	1.221	1.151	0.744	0.726	1.641	9.224	0.804
Regression with error	0.848	1.251	1.155	0.747	0.715	1.697	9.327	0.804
EM	0.873	1.260	1.162	0.749	0.730	1.673	9.284	0.814
Multiple Imputation without auxiliary variables[a]	NC	NC	NC	NC	NC	NC	NC	NC
Multiple Imputation with auxiliary variables[a]	NC	NC	NC	NC	NC	NC	NC	NC
Imputation 1	1.014	1.283	1.171	0.769	0.718	1.658	9.404	0.825
Imputation 2	0.993	1.263	1.154	0.752	0.730	1.628	9.298	0.805
Imputation 3	1.001	1.337	1.176	0.759	0.715	1.692	9.506	0.856
Imputation 4	0.959	1.303	1.151	0.754	0.721	1.686	9.354	0.820
Imputation 5	0.974	1.346	1.160	0.749	0.724	1.641	9.276	0.850

[a] Combined results of the five imputations.

NC: Not computed by IBM SPSS.

standard deviations for the complete case approach and then seven imputation methods (mean substitution, all-available, regression imputation with and without stochastic error, EM, and multiple imputation with and without auxiliary variables). In comparing the means, we find a general consistency between the methods, with no noticeable patterns. For the standard deviations, however, we can see the variance reduction associated with the mean substitution method. Across all variables, it consistently provides the smallest standard deviation, attributable to the substitution on the constant value. The other methods again show a consistency in the results, except that multiple imputation with auxiliary variables for V_2, which has the greatest extent of missing data, is higher than the other methods. Again, this is indicative of the use of the auxiliary variables in the imputation process.

EMPIRICAL RESULTS Finally, Table 2.8 contains the correlations for four selected variables (V_2, V_3, V_4, and V_5) obtained using the valid and imputed values from the complete case and the six other imputation methods. These variables were selected since they include the variables with the greatest extent of missing data (V_2 and V_3) as well as two other variables with little missing data (V_4 and V_5). In most instances the correlations are similar, but several substantial

Table 2.8 Comparison of Correlations Across Imputation Methods for Selected Variables (V_2, V_3, V_4, and V_5)

		V_2	V_3	V_4	V_5
V_2	Complete Case (Listwise)	1.000			
	All Available (Pairwise)	1.000			
	Mean Substitution	1.000			
	Regression without error	1.000			
	Regression with error	1.000			
	EM	1.000			
	Multiple Imputation without auxiliary variables[a]	1.000			
	Multiple Imputation with auxiliary variables[a]	1.000			
V_3	Complete Case (Listwise)	−0.286	1.000		
	All Available (Pairwise)	−0.357	1.000		
	Mean Substitution	−0.289	1.000		
	Regression without error	−0.332	1.000		
	Regression with error	−0.270	1.000		
	EM	−0.343	1.000		
	Multiple Imputation without auxiliary variables[a]	−0.298	1.000		
	Multiple Imputation with auxiliary variables[a]	−0.292	1.000		
V_4	Complete Case (Listwise)	0.285	−0.075	1.000	
	All Available (Pairwise)	0.299	−0.065	1.000	
	Mean Substitution	0.245	−0.057	1.000	
	Regression without error	0.307	−0.094	1.000	
	Regression with error	0.305	−0.076	1.000	
	EM	0.317	−0.092	1.000	
	Multiple Imputation without auxiliary variables[a]	0.288	−0.102	1.000	
	Multiple Imputation with auxiliary variables[a]	0.351	−0.049	1.000	
V_5	Complete Case (Listwise)	0.285	0.248	0.259	1.000
	All Available (Pairwise)	0.440	0.047	0.432	1.000
	Mean Substitution	0.382	0.042	0.422	1.000
	Regression without error	0.481	0.064	0.413	1.000
	Regression with error	0.466	0.097	0.410	1.000
	EM	0.511	0.078	0.413	1.000
	Multiple Imputation without auxiliary variables[a]	0.455	0.036	0.400	1.000
	Multiple Imputation with auxiliary variables[a]	0.539	0.126	0.387	1.000

[a] Combined results of the five imputations.

differences arise. First is a consistency between the correlations obtained with the all-available, mean substitution, EM and multiple imputation without auxiliary variables. Consistent differences occur, however, between these values and the values from the complete case approach. Second, the notable differences are concentrated in the correlations with V_2 and V_3, the two variables with the greatest amount of missing data in the reduced sample (refer back to Table 2.6). These differences may indicate the impact of a missing data process, even though the overall randomness test showed no significant pattern. Finally, the multiple imputation method with auxiliary variables also has some noticeable differences between V_2 and V_4, V_5, and V_3 with V_5. Again, these differences from all of the other approaches relates to the use of the auxiliary variables. Although the researcher has no proof of greater validity for any of the approaches, 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.

As a final form of comparison, a multiple regression equation was estimated with V_9 as the dependent variable and V_2 through V_8 as independent variables. The purpose was to see if any differences occurred when a formal model was estimated. As before, the complete case method and then the various imputation method results were used in model estimation. Table 2.9 contains the regression coefficients and model R^2, while Figure 2.9 portrays the regression coefficients graphically to illustrate values that differ markedly. Outside of the estimates of the intercept, all of the estimated coefficients were quite similar except for (1) the complete case and all-available methods for V_3 and (2) the mean substitution and all-available methods for V_5. It is interesting to note that in both instances the methods that differ are those only using available data, or conversely, all of the imputation methods that involve some form of model are all very consistent. Even the multiple imputation with auxiliary variables, which has some differences on several variables and correlations, was similar to all of the other methods with models.

Table 2.9 Regression Results for Complete Case and Six Imputation Methods

	Intercept	V_2	V_3	V_4	V_5	V_6	V_7	V_8	R^2
Complete Case (Listwise)	0.140	−0.204	**0.483**	**0.311**	**0.492**	−0.015	**−0.153**	−0.018	0.809
Imputation Methods									
All Available (Pairwise)	−0.238	**−0.195**	**0.508**	**0.271**	**0.823**	−0.103	**−0.072**	**−0.037**	0.897
Mean Substitution	0.514	**−0.213**	**0.306**	**0.258**	**0.346**	−0.091	−0.072	0.013	0.715
Regression without Error	0.159	**−0.209**	**0.371**	**0.318**	**0.588**	−0.168	**−0.080**	−0.007	0.803
Regression with Error	0.456	**−0.195**	**0.351**	**0.299**	**0.509**	−0.126	**−0.085**	−0.004	0.758
EM	0.110	**−0.194**	**0.348**	**0.322**	**0.559**	−0.180	−0.072	−0.001	0.794
Multiple Imputation with Auxiliary Variables	0.419	−0.159	**0.304**	**0.351**	**0.537**	−0.208	−0.090	0.001	0.774

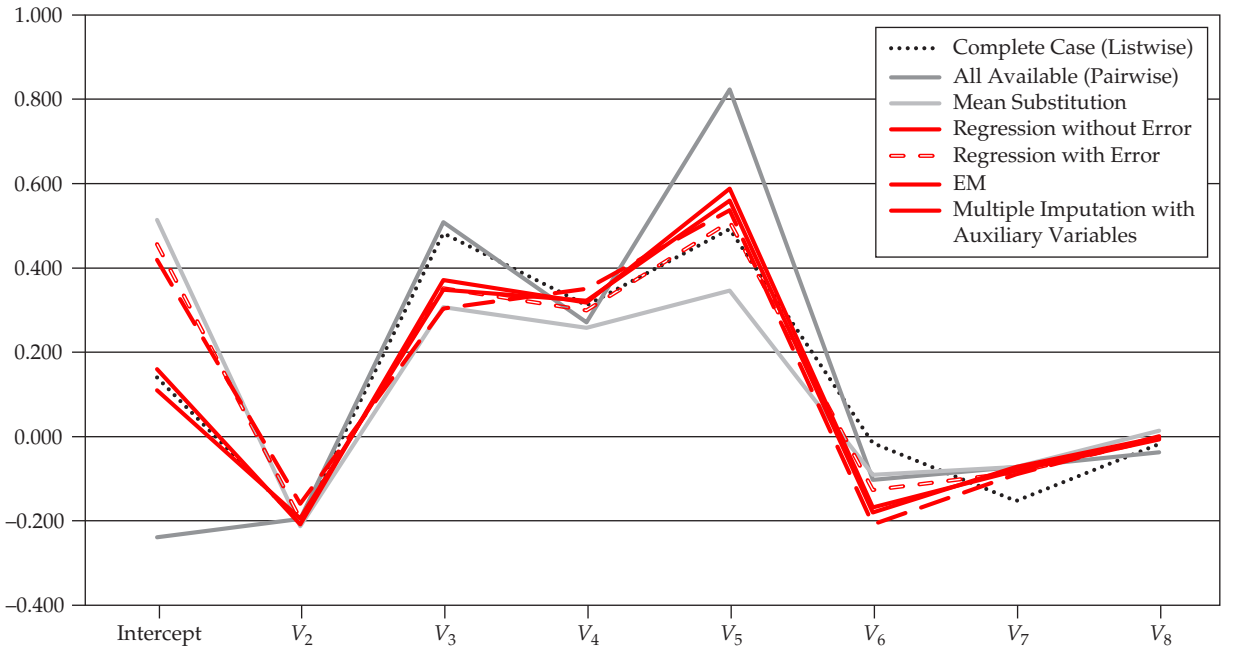
Note: Dependent measure: V_9.
Bolded items: significant at .05.

SUMMARY The task for the researcher is to 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 dataset with replacement values should be saved for further analysis. The model-based approaches, especially multiple imputation, provides a means to assess a wider range of imputed values and then make combined estimates of the results across these ranges of variables.

A Recap of the Missing Value Analysis Evaluation of the issues surrounding missing data in the dataset can be summarized in four conclusions:

1 **The missing data process is primarily MCAR.** All of the diagnostic techniques support the conclusion that no systematic missing data process exists for any of the variables except V_2 and possibly V_4, making the missing

Figure 2.9
Regression Parameter Estimates for Complete Case and Six Imputation Methods



Dependent Variable: V_9 Independent Variables: V_2, V_3, V_4, V_5, V_6, V_7, V_8

data process primarily MCAR (missing completely at random). Such a finding provides two advantages to the researcher. First, it should not involve any hidden impact on the results that need to be considered when interpreting the results except in one instance. Second, any of the imputation methods can be applied as remedies for the missing data with no marked biases expected. And the application of the model-based methods of EM and multiple imputation are appropriate for both MCAR and MAR Thus, the 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.** Even given the benefit of deleting cases and variables, the researcher is precluded from the simple solution of using the complete case method, because it results 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 any number of imputation techniques. In this situation, however, there are differences in the results among these methods. First, there are general consistencies among the all-available information, mean substitution, regression with and without error, EM and multiple imputation method without auxiliary variables methods. There are differences, however, with the complete case approach and the multiple imputation with auxiliary variable methods Even though 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 methods. The differences with the multiple imputation method with auxiliary variables can be attributed to the impact of the auxiliary variables, since multiple imputation without the auxiliary variables is consistent with the other methods. Interestingly, all of these approaches result is comparable regression coefficients except in a few instances for the complete case, all-available and mean substitution methods. Only in the instance of the complete case were the differences substantive enough to warrant concern.

4 **Multiple methods for replacing the missing data are available and appropriate.** As already mentioned, there are several methods that employ more of a model-based approach and provide comparable results. Moreover, the more complex model-based methods of EM and multiple imputation, while able to also accommodate MAR missing data processes, are readily available as needed. Researchers should consider employing several imputation methods and compare the results to ensure that no single method generates unique results. Moreover, multiple imputation provides a direct means of assessing a range of imputed values which are then combined for overall results, thus not making the imputation dependent on a single set of imputed values. Finally, EM and multiple imputation provide the methods needed to address MAR missing data processes that were not available in past years. Hopefully their use will spread to those instances in which MAR may bias imputed values from other methods and even be used in MCAR situations to provide additional assurance for the best imputation results possible.

In conclusion, the analytical tools and the diagnostic processes presented in the earlier section provide 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.

Outliers

Outliers, or **anomalies** in the parlance of data mining, are observations with a *unique combination of characteristics identifiable as distinctly different* from what is "normal." All of the analyses focused on outlier detection are based on establishing the norms of comparison so that individual observations can then be evaluated and outlier detection can be objective and routinized. This becomes critically important as the amount and diversity of the types of data being used increases. In this era of Big Data and continuous, real-time analytics, the ability to detect an outlier/anomaly must be formalized [29]. As a result, particularly with longitudinal data, the researcher must be constantly defining the "context" of the data to establish what is "normal." Consider a classic example where we are viewing an electrocardiogram—the series of spikes depicting heartbeats. Now if we focus on a very small timeframe, the spikes might seem like outliers and cause concern. But as we watch over time, we detect that they are part of a more general pattern and that they are the "normal" which we should expect. Here the researcher must be sure to correctly define the *context* upon which "normal" is defined.

TWO DIFFERENT CONTEXTS FOR DEFINING OUTLIERS

So how do we specify the context for defining outliers? It is perhaps easiest to distinguish contexts as pre-analysis and post-analysis, where "normal" is based on quite different criteria—comparison to the population versus comparison to the model expectations. We discuss these two contexts in the following sections.

Pre-analysis Context: A Member of a Population Here the focus is on each case as compared to the other observations under study. The examination involves the characteristics of the observations and how any particular observation(s) vary markedly from the other observations. Outliers are generally those observations that have extremely different values on one or a combination of variables. The objective at this stage is to ensure a representative sample of the population for analysis and identify observations that are *truly unique* in terms of their representativeness of the population. Thus, observations must be evaluated before any type of analysis has begun. This requires extensive domain knowledge of the research situation to ensure that observations are truly evaluated on what is unique in that context (e.g., our electrocardiogram example from before). Once the designation is made, the researcher must decide if the observation is retained as a member of a representative sample or designated as an outlier.

Post-analysis: Meeting Analysis Expectations The second perspective defines "normal" as the expectations (e.g., predicted values, group membership predictions, etc.) generated by the analysis of interest. Here the outlier designation occurs only after the analysis has been performed and we identify those observations for which the analysis did not perform well. The objective at this stage is model understanding and improvement. Outliers provide a basis for identifying what observations were not well predicted so that the model can be improved. For example, in regression we define outliers as those cases with large residuals, and residuals are defined as the difference between the observation's value of the dependent value and that predicted by the regression model. So while the observation's characteristics are inputs to the model, "normal" in this perspective is defined by the model predictions, not the overall characteristics of the observations. Moreover, an observation may be an outlier in one model application but may be considered "normal" in another model application.

Summary The designation of an outlier occurs in two distinct stages of the analysis: pre-analysis and post-analysis. The criteria employed in defining "normal" vary in each stage as the objectives of outlier designation change. Our discussions in this chapter are focused on the pre-analysis stage, while the post-analysis designation of outliers will be addressed in the chapters for each statistical technique.

IMPACTS OF OUTLIERS

In assessing the impact of outliers, we must consider the practical and substantive considerations along with whether we should designate outliers as "good or bad."

Practical Impacts From a *practical* standpoint, outliers can have a marked effect on any type of empirical analysis. For example, assume that we sample 20 individuals to determine the average household income. In our sample we gather responses that range between $20,000 and $100,000, so that the average is $45,000. But assume that the 21st person has an income of $1 million. If we include this value in the analysis, the average income increases to more than $90,000. Obviously, the outlier is a valid case, but what is the better estimate of the average household income: $45,000 or $90,000? The researcher must assess whether the outlying value is retained or eliminated due to its undue influence on the results.

Substantive Impacts In *substantive* terms, the outlier must be viewed in light of how representative it is of the population. Again, using our example of household income, how representative of the more wealthy segment is the millionaire? If the researcher feels that it is a small, but viable segment in the population, then perhaps the value should be retained. If, however, this millionaire is the only one in the entire population and truly far above everyone else (i.e., unique) and represents an extreme value, then it may be deleted.

Are Outliers Good or Bad? 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. As such, they do not meet the framework of the analysis being conducted. While they may provide feedback for necessary adjustments to the analysis, they also provide the researcher a means of focusing the analysis and results on the intended population rather than being impacted by observations not even intended for inclusion. 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.

CLASSIFYING OUTLIERS

While the classification of outliers can take many forms, this discussion focuses on two perspectives: the impact on the analysis (i.e., the role they play for the researcher) and the basic nature/character of the outlier. While there is some overlap between the two perspectives, they each provide some different insights into how the researcher may wish to characterize outliers and then ultimately accommodate them in the analysis.

Types of Impacts on the Analysis A recent literature review [1] in the field of organizational science provided some interesting insights into the definition, identification and handling of outliers in this field. From 46 different academic sources, they identified 14 different outlier definitions, 39 identification techniques and 20 methods for handling outliers in the analysis. While it is beyond the scope of this chapter to review all these findings (we encourage the interested reader to read the article as it provides much greater detail on these issues), they did classify outliers into three types based on their contribution to the analysis:

- *Error outliers*. These are observations/cases that differ from the "normal" because of inaccuracies in data collection, etc. The remedy for this type of outlier is to correct the error or if not possible, remove the observation from the analysis.
- *Interesting outliers.* These observations are different and/or unique such that they may bring new insight into the analysis. The suggestion to study these observations underscores the need for domain knowledge of the context of the analysis to understand whether these observations add to existing knowledge of the context.
- *Influential outliers.* These observations are defined in terms of their impact on the analysis and are identified in the post-analysis stage. At this point, they had already been considered representative of the population in the pre-analysis stage, thus the researcher must either accommodate them in the analysis (perhaps through some robust methodology) or delete them from the analysis.

Reasons for Outlier Designation A second classification framework focuses on the basic nature/character of the observations and what makes them different from "normal." In understanding the source of their uniqueness, this approach focuses on the fundamental characteristics of observations and how they singly or perhaps in combination create that uniqueness for the observation. The four classes are:

- *Procedural error.* 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.
- *Extraordinary event.* The second class of outlier is the observation that occurs as the result of an extraordinary event, which accounts for the uniqueness of the observation. For example, assume we are tracking average daily rainfall, when we have a hurricane that lasts for several days and records extremely high rainfall levels. These rainfall levels are not comparable to anything else recorded in the normal weather patterns. If included, they will markedly change the pattern of the results. The researcher must decide whether the extraordinary event fits the objectives of the research. If so, the outlier should be retained in the analysis. If not, it should be deleted.
- *Extraordinary observations.* The third class of outlier comprises extraordinary observations for which the researcher has no explanation. In these instances, a unique and markedly different profile emerges. Although these outliers are the most likely to be omitted, they may be retained if the researcher feels they represent a valid element of the population. Perhaps they represent an emerging element, or an untapped element previously not identified. Here the researcher must use judgment in the retention/deletion decision.

- *Unique combinations.* The fourth and final class of outlier contains observations that fall within the ordinary range of values on each of the variables. These observations are not particularly high or low on 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.

Summary Hopefully our discussion to this point has illustrated the basic purpose of outliers—to provide "checkpoints" on first the sample to be analyzed and then on the analysis performed. The identification of outliers in relationship to the sample being analyzed requires that the researcher take a "big picture" perspective in making sure that the context being analyzed is understood and the observations of the sample are representative of this context. Once the analysis is complete, outliers provide feedback on what 'didn't work" versus all of the other effort to understand the analysis results and the conclusions that can be drawn. Many times it is this form of examination that provides the researcher with insights that can truly improve the analysis so that it addresses all of the observations within the sample.

DETECTING AND HANDLING OUTLIERS

The following sections detail the methods used in detecting outliers in univariate, bivariate, and multivariate situations. Once identified, they may be profiled to aid in placing them into one of the four classes just described. 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. As noted earlier, these methods are applicable to the preanalysis designation of outliers.

Methods of Detecting Outliers Outliers can be identified from a univariate, bivariate, or multivariate perspective based on the number of variables (characteristics) considered. The researcher should utilize as many of these perspectives as possible, looking for a consistent pattern across perspectives to identify outliers. The following discussion details the processes involved in each of the three perspectives.

UNIVARIATE DETECTION The univariate identification of outliers examines the distribution of observations for each variable in the analysis and selects as outliers those cases falling at the outer ranges (high or low) 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. An outlier designation then occurs when an observation falls well to the outer boundaries of the distribution of values, many times identified as cases with standardized values of ± 3, which makes them quite unique in terms of that characteristic.

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 will be seen as isolated points in the scatterplot. To assist in determining the expected range of observations in this two-dimensional portrayal, an ellipse representing a bivariate normal distribution's confidence interval (typically set at the 90% or 95% level) is superimposed over the scatterplot. This ellipse provides a graphical portrayal of the confidence limits and facilitates identification of the outliers. A variant of the scatterplot is termed the influence plot, with each point varying in size in relation to its influence on the relationship.

Each of these methods provides an assessment of the uniqueness of each observation in relationship to the other observation based on a specific pair of variables. A drawback of the bivariate method in general is the potentially

large number of scatterplots that arise as the number of variables increases. For three variables, it is only three graphs for all pairwise comparisons. But for five variables, it takes 10 graphs, and for 10 variables it takes 45 scatterplots! As a result, the researcher should limit the general use of bivariate methods to specific relationships between variables, such as the relationship of the dependent versus independent variables in regression. The researcher can then examine the set of scatterplots and identify any general pattern of one or more observations that would result in their designation as outliers.

MULTIVARIATE DETECTION Because most multivariate analyses involve more than two variables, the bivariate methods quickly become inadequate for several reasons. First, they require a large number of graphs, as discussed previously, when the number of variables reaches even moderate size. Second, they are limited to two dimensions (variables) at a time. Yet when more than two variables are considered, the researcher needs a means to objectively measure the *multidimensional* position of each observation relative to some common point. This issue is addressed by the Mahalanobis D^2 measure, a multivariate assessment of each observation across a set of variables. This method measures each observation's distance in multidimensional space from the mean center of all observations, providing a single value for each observation no matter how many variables are considered. Higher D^2 values represent observations farther removed from the general distribution of observations in this multidimensional space. This method, however, also has the drawback of only providing an overall assessment, such that it provides no insight as to which particular variables might lead to a high D^2 value.

For interpretation purposes, the Mahalanobis D^2 measure has statistical properties that allow for significance testing. The D^2 measure divided by the number of variables involved (D^2/df) is approximately distributed as a t value. Given the nature of the statistical tests, it is suggested that conservative levels of significance (e.g., .005 or .001) be used as the threshold value for designation as an outlier. Thus, observations having a D^2/df value exceeding 2.5 in small samples and 3 or 4 in large samples can be designated as possible outliers. Once identified as a potential outlier on the D^2 measure, an observation can be re-examined in terms of the univariate and bivariate methods discussed earlier to more fully understand the nature of its uniqueness.

The Mahalanobis D^2 measure is encountered across the entire spectrum of multivariate techniques, either as a component in the analysis or as a diagnostic (e.g., in operationalizing the concept of leverage in multiple regression). One feature of the Mahalanobis D^2 measure is that it many times is not readily available for general use, but instead is found within a specific technique (e.g., multiple regression). Researchers are encouraged to identify sources of calculating the Mahalanobis D^2 measure within their favored software package as it is a quite useful measure in many types of situations.

Impact of Dimensionality As researchers develop analyses with an increased number of variables, whether it be due to more constructs and the movement to multi-item scales or embracing Big Data, the analysis and identification of outliers becomes markedly more difficult. Many times the "**Curse of Dimensionality**" is ascribed to Big Data because as the number of variables considered increases dramatically, so do the attendant issues, and outlier detection is certainly included in this set. For purposes of outlier identification, at least three issues emerge:

1. **Distance measures become less useful.** One characteristic of high dimensionality (i.e., a large number of variables) is that in most samples the observations become widely dispersed and the distance measures generally used in multivariate detection become less distinctive among those observations. Research has demonstrated that as the number of dimensions increases the ability of these distance measures to identify the truly unique observations diminishes.
2. **Impact of Irrelevant Variables.** As the number of variables increases, the potential for irrelevant variables increases. As a result, the characterization of observations is "clouded" as measures that ultimately have no impact in the relationship are considered when identifying uniqueness. Researchers are cautioned to truly

understand what characteristics make an observation unique and if that uniqueness is due to factors of substantive interest to the research.

3 **Comparability of dimensions.** With an increase in the number of variables, and particularly different types of variables, the researcher must be aware of the potential impact of differing scales of measurement and variation across the variables may impact outlier detection. When facing this issue, standardization may be required so that comparability is maintained. An associated issue is outlier detection when nonmetric variables are included, which requires a different methodology [94].

Researchers will face these issues in even a small scale study to some extent, but the impact become magnified as the number of variables increases. Our HBAT example involves just a small number of variables, but imagine if you were employing hundreds or thousands of variables and had to first identify outliers in the sample. New techniques are being developed with the field of data mining that may provide researchers with more adaptable tools for this difficult task.

Outlier Designation With these univariate, bivariate, and multivariate diagnostic methods, the researcher has a complementary set of perspectives with which to examine observations as to their status as outliers. Each of these methods can provide a unique perspective on the observations and be used in a concerted manner to identify outliers (see Rules of Thumb 2-3).

When observations have been identified by the univariate, bivariate, and multivariate methods as possible outliers, the researcher must then select only observations that demonstrate real uniqueness in comparison with the remainder of the population across as many perspectives as possible. The researcher must 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 are identified, the researcher should generate profiles of each outlier observation and identify 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 (Chapter 7) or multiple regression (Chapter 5) to identify the differences between outliers and the other observations. If possible the

Outlier Detection

Univariate methods: Examine all metric variables to identify unique or extreme observations:

- For small samples (80 or fewer observations), outliers typically are defined as cases with standard scores of 2.5 or greater.
- For larger sample sizes, increase the threshold value of standard scores up to 4.
- If standard scores are not used, identify cases falling outside the ranges of 2.5 versus 4 standard deviations, depending on the sample size.

Bivariate methods: Focus their use on specific variable relationships, such as the independent versus dependent variables:

- Use scatterplots with confidence intervals at a specified alpha level.

Multivariate methods: Best suited for examining a complete variate, such as the independent variables in regression or the variables in exploratory factor analysis:

- Threshold levels for the D^2/df measure should be conservative (.005 or .001), resulting in values of 2.5 (small samples) versus 3 or 4 in larger samples.

researcher should assign the outlier to one of the four classes described earlier to assist in the retention or deletion decision to be made next. The researcher should continue this analysis until satisfied with understanding the aspects of the case that distinguish the outlier from the other observations.

Retention or Deletion of the Outlier After the outliers are identified, profiled and categorized, the researcher must decide on the retention or deletion of each one. Many philosophies among researchers offer guidance as to how to deal with outliers. Our belief is that they should be retained unless demonstrable proof indicates that they are truly aberrant and not representative of any observations in the population. If they do portray a representative element or 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. Techniques such as robust regression allow for the retention of outliers, but reduce their impact on the model results. Moreover, research has shown that employing more nonparametric tests can also reduce the influence of outliers [7].

AN ILLUSTRATIVE EXAMPLE OF ANALYZING OUTLIERS

As an example of outlier detection, the observations of the HBAT database introduced in Chapter 1 are examined for outliers. The variables considered in the analysis are the metric variables X_6 through X_{19}, with the context of our examination being a regression analysis, where X_{19} is the dependent variable and X_6 through X_{18} are the independent variables. The outlier analysis will include univariate, bivariate, and multivariate diagnoses. When candidates for outlier designation are found, they are examined, and a decision on retention or deletion is made.

Outlier Detection The first step is examination of all the variables from a univariate perspective. Bivariate methods will then be employed to examine the relationships between the dependent variable (X_{19}) and each of the independent variables. From each of these scatterplots, observations that fall outside the typical distribution can be identified and their impact on that relationship ascertained. Finally, a multivariate assessment will be made on all of the independent variables collectively. Comparison of observations across the three methods will hopefully provide the basis for the deletion/retention decision.

UNIVARIATE DETECTION The first step is to examine the observations on each of the variables individually. Table 2.10 contains the observations with standardized variable values exceeding ± 2.5 on each of the variables (X_6 to X_{19}). From this univariate perspective, only observations 7, 22, and 90 exceed the threshold on more than a single variable. Moreover, none of these observations had values so extreme as to affect any of the overall measures of the variables, such as the mean or standard deviation. We should note that the dependent variable had one outlying observation (22), which may affect the bivariate scatterplots because the dependent variable appears in each scatterplot. The three observations will be noted to see whether they appear in the subsequent bivariate and multivariate assessments.

BIVARIATE DETECTION For a bivariate perspective, 13 scatterplots are formed for each of the independent variables (X_6 through X_{18}) with the dependent variable (X_{19}). An ellipse representing the 95 percent confidence interval of a bivariate normal distribution is then superimposed on the scatterplot. Figure 2.10 contains examples of two such scatterplots involving X_6 and X_7. As we can see in the scatterplot for X_6 with X_{19}, the two outliers fall just outside the ellipse and do not have the most extreme values on either variable. This result is in contrast to the scatterplot of X_7 with X_{19}, where observation 22 is markedly different from the other observations and shows the highest values on both X_7 and X_{19}. The second part of Table 2.10 contains a compilation of the observations falling outside this ellipse for each variable. Because it is a 95 percent confidence interval, we would expect some observations normally to fall outside the ellipse. Only four observations (2, 22, 24, and 90) fall outside the ellipse more than two times.

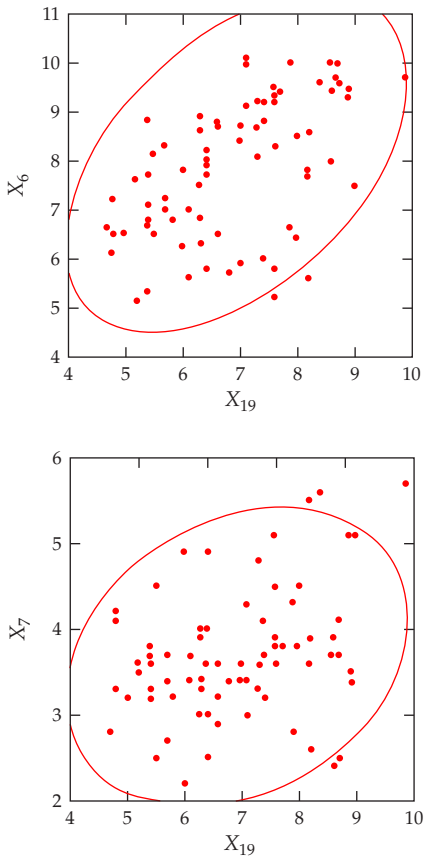


Figure 2.10
Selected Scatterplots for Bivariate Detection of Outliers: X_6 (Product Quality) and X_7 (E-Commerce Activities) with X_{19} (Customer Satisfaction)

Table 2.10 Univariate, Bivariate, and Multivariate Outlier Detection Results

UNIVARIATE OUTLIERS *Cases with Standardized Values Exceeding ±2.5*		*BIVARIATE OUTLIERS* *Cases Outside the 95% Confidence Interval Ellipse*		*MULTIVARIATE OUTLIERS* *Cases with a Value of D^2/df Greater than 2.5 (df = 13)*[a]		
		X_{19} with:		Case	D^2	D^2/df
X_6	No cases	X_6	44, **90**	98	40.0	3.08
X_7	13, 22, **90**	X_7	13, **22**, 24, 53, **90**	36	36.9	2.84
X_8	8, **7**	X_8	**22**, 87			
X_9	No cases	X_9	**2, 22**, 45, 52			
X_{10}	No cases	X_{10}	**22, 24**, 85			
X_{11}	**7**	X_{11}	**2**, 7, **22**, 45			
X_{12}	**90**	X_{12}	**22**, 44, **90**			
X_{13}	No cases	X_{13}	**22**, 57			
X_{14}	77	X_{14}	**22**, 77, 84			
X_{15}	6, 53	X_{15}	6, **22**, 53			
X_{16}	24	X_{16}	**22, 24**, 48, 62, 92			
X_{17}	No cases	X_{17}	**22**			
X_{18}	**7**, 84	X_{18}	**2**, 7, **22**, 84			
X_{19}	**22**					

[a]Mahalanobis D^2 value based on the 13 HBAT perceptions (X_6 to X_{18}).

Observation 22 falls outside in 12 of the 13 scatterplots, mostly because it is an outlier on the dependent variable. Of the remaining three observations, only observation 90 was noted in the univariate detection.

MULTIVARIATE DETECTION The final diagnostic method is to assess multivariate outliers with the Mahalanobis D^2 measure (see Table 2.10). This analysis evaluates the position of each observation compared with the center of all observations on a set of variables. In this case, all the metric independent variables were used. The calculation of the D^2/df value $(df = 13)$ allows for identification of outliers through an approximate test of statistical significance. Because the sample has only 100 observations, a threshold value of 2.5 will be used rather than the value of 3.5 or 4.0 used in large samples. With this threshold, two observations (98 and 36) 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 result 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 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. No observations are extreme on a sufficient number of variables 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. In the case of regression analysis, Chapter 5 will provide additional methods to assess the relative influence of each observation and provide more insight into the possible deletion of an observation as an outlier.

Testing the Assumptions of Multivariate Analysis

The final step in examining the data involves testing for the assumptions underlying the statistical bases for multivariate analysis. The earlier steps of missing data analysis and outlier detection attempted to clean the data to a format most suitable for multivariate analysis. Testing the data for compliance with the statistical assumptions underlying the multivariate techniques now deals with the foundation upon which the techniques make statistical inferences and results. Some techniques are less affected by violating certain assumptions, which is termed **robustness**, but in all cases meeting some of the assumptions will be critical to a successful analysis. Thus, it is necessary to understand the role played by each assumption for every multivariate technique.

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, particularly when the violations compound to become even more detrimental than if considered separately. Second, the complexity of the analyses and results may mask the indicators 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 for a univariate analysis, 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.

FOUR IMPORTANT STATISTICAL ASSUMPTIONS

Multivariate techniques and their univariate counterparts are all based on a fundamental set of assumptions representing the requirements of the underlying statistical theory. Although many assumptions or requirements come into play in one or more of the multivariate techniques we discuss in the text, four of them potentially affect every univariate and multivariate statistical 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, because 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 VERSUS 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 [36, 83], but specialized tests are available in the techniques most affected by departures from multivariate normality. In most cases assessing and achieving univariate normality for all variables is sufficient, and we will 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 always assess the normality for all metric variables included in the analysis.

ASSESSING THE IMPACT OF VIOLATING THE NORMALITY ASSUMPTION The severity of non-normality is based on two dimensions: the shape of the offending distribution and the sample size. As we will see in the following discussion, the researcher must not only judge the extent to which the variable's distribution is non-normal, but also the sample sizes involved. What might be considered unacceptable at small sample sizes will have a negligible effect at larger sample sizes.

Impacts Due to the Shape of the Distribution How can we describe the distribution if it differs from the normal distribution? The shape of any distribution can be described by two measures: kurtosis and skewness. **Kurtosis** refers to the "peakedness" or "flatness" of the distribution compared with the normal distribution. Distributions that are taller or more peaked than the normal distribution are termed *leptokurtic,* whereas a distribution that is flatter is termed *platykurtic.* Whereas kurtosis refers to the height of the distribution, **skewness** is used to describe the balance of the distribution; that is, is it unbalanced and shifted to one side (right or left) or is it centered and symmetrical with about the same shape on both sides? If a distribution is unbalanced, it is skewed. A positive skew denotes a distribution shifted to the left, whereas a negative skewness reflects a shift to the right.

Knowing how to describe the distribution is followed by the issue of how to determine the extent or amount to which it differs on these characteristics? Both skewness and kurtosis have empirical measures that are available in all statistical programs. In most programs, the skewness and kurtosis of a normal distribution are given values of zero. Then, values above or below zero denote departures from normality. For example, negative kurtosis values indicate a platykurtic (flatter) distribution, whereas positive values denote a leptokurtic (peaked) distribution. Likewise, positive skewness values indicate the distribution shifted to the left, and the negative values denote a rightward shift. To judge the "Are they large enough to worry about?" question for these values, the following discussion on

statistical tests shows how the kurtosis and skewness values can be transformed to reflect the statistical significance of the differences and provide guidelines as to their severity.

Impacts Due to Sample Size Even though it is important to understand how the distribution departs from normality in terms of shape and whether these values are large enough to warrant attention, the researcher must also consider the effects of sample size. As discussed in Chapter 1, sample size has the effect of increasing statistical power by reducing sampling error. It results in a similar effect here, in that larger sample sizes *reduce* the detrimental effects of non-normality. In small samples of 50 or fewer observations, and especially if the sample size is less than 30 or so, significant departures from normality can have a substantial impact on the results. For sample sizes of 200 or more, however, these same effects may be negligible. Moreover, when group comparisons are made, such as in ANOVA, the differing sample sizes between groups, if large enough, can even cancel out the detrimental effects. Thus, in most instances, as the sample sizes become large, the researcher can be less concerned about non-normal variables, except as they might lead to other assumption violations that do have an impact in other ways (e.g., see the following discussion on homoscedasticity).

TESTS OF NORMALITY Researchers have a number of different approaches to assess normality, but they primarily can be classified as either graphical or statistical. Graphical methods were developed to enable normality assessment without the need for complex computations. They provide the researcher with a more "in depth" perspective of the distributional characteristics than a single quantitative value, but they are also limited in making specific distinctions since graphical interpretations are less precise that statistical measures.

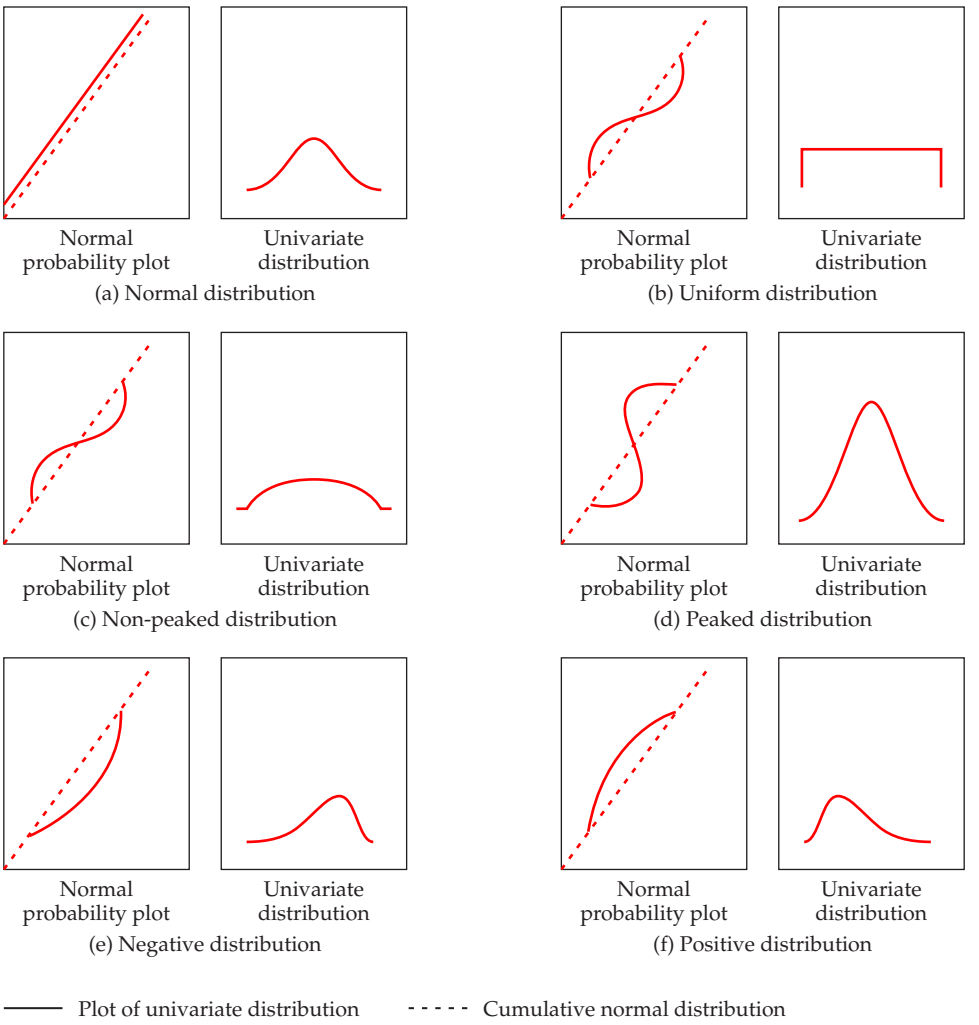
Graphical Analyses The simplest diagnostic test for normality is a visual check of the histogram that compares the observed data values with a distribution approximating the normal distribution (see Figure 2.1). Although appealing because of its simplicity, this method is problematic for smaller samples, 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 2.11 shows several departures from normality and their representation in the normal probability in terms of kurtosis and skewness. First, departures from the normal distribution in terms of kurtosis are easily seen in the normal probability plots. 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 2.11d), 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 non-peaked distribution has the opposite pattern (Figure 2.11c). Skewness is also easily seen, most often represented by a simple arc, either above or below the diagonal. A negative skewness (Figure 2.11e) is indicated by an arc below the diagonal, whereas an arc above the diagonal represents a positively skewed distribution (Figure 2.11f). An excellent source for interpreting normal probability plots, showing the various patterns and interpretations, is Daniel and Wood [26]. These specific patterns not only identify non-normality but also tell us the form of the original distribution and the appropriate remedy to apply.

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

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

Figure 2.11 Normal Probability Plots and Corresponding Univariate Distributions



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 either calculated z *value exceeds the specified critical value, then the distribution is non-normal in terms of that characteristic.* The critical value is from a z distribution, based on the significance level we desire. The most commonly used critical values are ± 2.58 (.01 significance level) and ± 1.96, which corresponds to a .05 error level. With these simple tests, the researcher can easily assess the degree to which the skewness and peakedness of the distribution vary from the normal distribution.

Specific statistical tests for normality are also available in all the statistical 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 NON-NORMALITY A number of data transformations available to accommodate non-normal distributions are discussed later in the chapter. This chapter confines the discussion to univariate normality tests and transformations. However, when we examine other multivariate methods, such as multivariate regression or multivariate analysis of variance, we discuss tests for multivariate normality as well. Moreover, many times when non-normality is indicated, it also contributes to other assumption violations; therefore, remedying normality first may assist in meeting other statistical assumptions as well. For those interested in multivariate normality, see references [36, 50, 88].

Homoscedasticity The next assumption is related primarily to dependence relationships between variables. **Homoscedasticity** 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.* 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 relatively equal at each value of the predictor variable. If this dispersion is unequal across values of the independent variable, the relationship is said to be **heteroscedastic**.

TYPE OF INDEPENDENT VARIABLE 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. The type of independent variable dictates how homoscedasticity is assessed.

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 dispersion of values for the dependent variable should be as large for small values of the independent values as it is for moderate and large values. In a scatterplot, it is seen as an elliptical distribution of points.

Nonmetric Independent Variables In these analyses (e.g., ANOVA and MANOVA) the focus now becomes the equality of the variance (single dependent variable) or the variance/covariance matrices (multiple dependent 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.

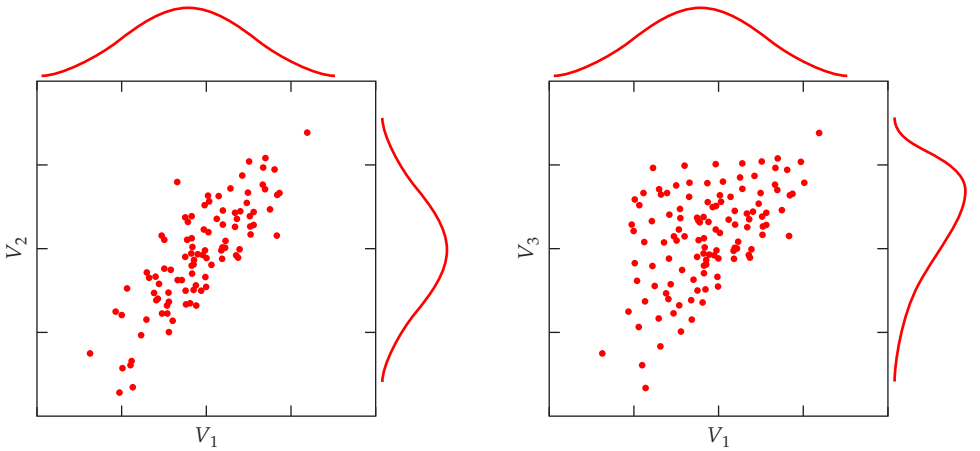
SOURCES OF HETEROSCEDASTICITY In each of these instances, the purpose is the same: to ensure that the variance used in explanation and prediction is distributed across the range of values, thus allowing for a "fair test" of the relationship across all values of the nonmetric variables. The two most common sources of heteroscedasticity are the following:

Variable Type Many types of variables have a natural tendency toward differences in dispersion. For example, as a variable increases in value (e.g., units ranging from near zero to millions) a naturally wider range of answers is possible for the larger values. Also, when percentages are used the natural tendency is for many values to be in the middle range, with few in the lower or higher values.

Skewed Distribution of One or Both Variables In Figure 2.12a, 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 2.12b 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.

The result of heteroscedasticity is to cause the predictions to be better at some levels of the independent variable than at others. This variability affects the standard errors and makes hypothesis tests either too stringent or too insensitive. The effect of heteroscedasticity is also often related to sample size, especially when examining the variance

Figure 2.12 Scatterplots of Homoscedastic and Heteroscedastic Relationships



(a) Homoscedasticity (b) Heteroscedasticity

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.

TESTS FOR HOMOSCEDASTICITY As we found for normality, there are a series of graphical and statistical tests for identifying situations impacted by heteroscedasticity. The researcher should employ both methods where the graphical methods provide a more in-depth understanding of the overall relationship involved and the statistical tests provide increased precision.

Graphical Tests of Equal Variance Dispersion The test of homoscedasticity for two metric variables is best examined graphically. Departures from an equal dispersion are shown by such shapes as cones (small dispersion at one side of the graph, large dispersion at the opposite side) or diamonds (a large number of points at the center of the distribution). The most common application of graphical tests occurs in multiple regression, based on the dispersion of the dependent variable across the values of either the metric independent variables. We will defer our discussion of graphical methods until we reach Chapter 5, which describes these procedures in much more detail.

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. Thus, heteroscedasticity would be portrayed by substantial differences in the length of the boxes and whiskers between groups representing the dispersion of observations in each group.

Statistical Tests for Homoscedasticity The statistical tests for equal variance dispersion assess the equality of variances within groups formed by nonmetric variables. The most common test, the Levene test, is 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 non-normality of one of the variables, and correction of the non-normality also remedies the unequal dispersion of variance. A later section discusses data transformations of the variables to "spread" the variance and make all values have a potentially equal effect in prediction.

We should also note that the issue of heteroscedasticity can be remedied directly in some statistical techniques without the need for transformation. For example, in multiple regression the standard errors can be corrected for heteroscedasticity to produce heteroscedasticity-consistent standard errors (HCSE) [90]. This method does not impact the coefficients, only the standard errors, thus leaving the coefficients to be interpreted in their original form. See Chapter 5 for more discussion of this feature in multiple regression.

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 omission 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 affect 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. Many scatterplot programs can show the straight line depicting the linear relationship, enabling the researcher to better identify any nonlinear characteristics. An alternative approach is to run a simple regression analysis (the specifics of this technique are covered in Chapter 5) 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. A third approach is to explicitly model a nonlinear relationship by the testing of alternative model specifications (also know as curve fitting) that reflect the nonlinear elements. A discussion of this approach and residual analysis is found in Chapter 5.

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 5.

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 One of the most common violations of the assumption that errors are uncorrelated is 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 combined effect leads to biased results because an unspecified cause is affecting the estimation of the relationship. The common example used is the collection of data within classes or other groups of respondents, where some form of group dynamic may impact each group differently. This situation is addressed in our discussion of multilevel models in Chapter 5.

Another common source of correlated errors is time series data. As we would expect, the data for any time period is highly related to the data at time periods both before and afterward. Thus, any predictions and any prediction errors will necessarily be correlated. This type of data led to the creation of specialized programs specifically for time series analysis and this pattern of correlated observations (see Chapter 5 for discussion of panel models).

Testing Statistical Assumptions

Normality can have serious effects in small samples (fewer than 50 cases), but the impact effectively diminishes when sample sizes reach 200 cases or more.

Most cases of heteroscedasticity are a result of non-normality in one or more variables; thus, remedying normality may not be needed due to sample size, but may be needed to equalize the variance.

Nonlinear relationships can be well defined, but seriously understated unless the data are transformed to a linear pattern or explicit model components are used to represent the nonlinear portion of the relationship.

Correlated errors arise from a process that must be treated much like missing data; that is, the researcher must first define the causes among variables either internal or external to the dataset; if they are not found and remedied, serious biases can occur in the results, many times unknown to the researcher.

To identify correlated errors, the researcher must first identify possible causes. Values for a variable should be grouped or ordered on the suspected variable and then examined for any patterns. In our earlier example of grouped data, once the potential cause is identified the researcher could see whether 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. For other types of data, such as time series data, we can see any trends or patterns when we order the data (e.g., by time period for time series data). This ordering variable (time in this case), if not included in the analysis in some manner, would cause the errors to be correlated and create substantial bias in the results.

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 belonged. 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. But beyond just including the variable, the intercorrelation among observations is best handled in some form of multi-level or panel model. Chapter 5 presents a discussion of these extensions of multiple regression which can accommodate grouped/clustered observations or time series data in a structured framework.

Overview of Testing for Statistical Assumptions The researcher is faced with what may seem to be an impossible task: satisfy all of these statistical assumptions or risk a biased and flawed analysis. We want to note that even though these statistical assumptions are important, the researcher must use judgment in how to interpret the tests for each assumption and when to apply remedies. Even analyses with small sample sizes can withstand small, but significant, departures from normality. What is more important for the researcher is to understand the implications of each assumption with regard to the technique of interest, striking a balance between the need to satisfy the assumptions versus the robustness of the technique and research context. The above guidelines in Rules of Thumb 2-4 attempt to portray the most pragmatic aspects of the assumptions and the reactions that can be taken by researchers.

Data Transformations

Data transformations provide the researcher a wide range of methods to achieve one of four basic outcomes: (1) enhancing statistical properties; (2) ease of interpretation; (3) representing specific relationship types; and (4) simplification. In each case, the original variable and its values are transformed in some manner to alter its characteristics so that it represents a different facet of the underlying information contained in the values. 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 [12]).

TRANSFORMATIONS RELATED TO STATISTICAL PROPERTIES

As discussed in the prior section, transformations play a pivotal role in ensuring that the variables in any statistical techniques meet the assumptions needed for statistical inference. For our purposes we will discuss two basic forms of transformations: normality/homoscedasticity and linearity.

Achieving Normality and Homoscedasticity Data transformations provide the principal means of correcting non-normality and heteroscedasticity. In both instances, patterns of the variables suggest specific transformations. For non-normal 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, squared, or cubed (X^2 or X^3) terms or even the inverse of the variable. Usually negatively skewed distributions are best transformed by employing a squared or cubed transformation, whereas the logarithm or square root typically works best on positive skewness. In many instances, the researcher may 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 a problem with the distribution of the variable(s). When examining the scatterplot, the most common pattern is the cone-shaped distribution. 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 $\left(X_{\text{new}} = 2 \arcsin\sqrt{x_{\text{old}}}\right)$ 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.

Achieving Linearity Numerous procedures are available for achieving linearity between two variables, but most simple nonlinear relationships can be placed in one of four categories (see Figure 2.13). In each quadrant, the potential transformations for both dependent and independent variables are shown. For example, if the relationship looks like Figure 2.13a, 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 5.

TRANSFORMATIONS RELATED TO INTERPRETATION

Transformations can also assist in improving the interpretation of a variable. The two most commonly used approaches in this category are standardization and centering. The objective in each case is to make the values of each observation relative to some specific value. This common reference value then allows for more direct comparison of the values across observations.

Standardization **Standardization** actually takes many forms, but the most commonly used is the z score or standard score. In this transformation each value is first differenced from the variable mean (i.e., value − variable mean) and then the difference is made relative to the variable's standard deviations (i.e., difference divided by the standard deviation). The result is a set of values with a mean of zero and a standard deviation of 1. A z score of zero means that

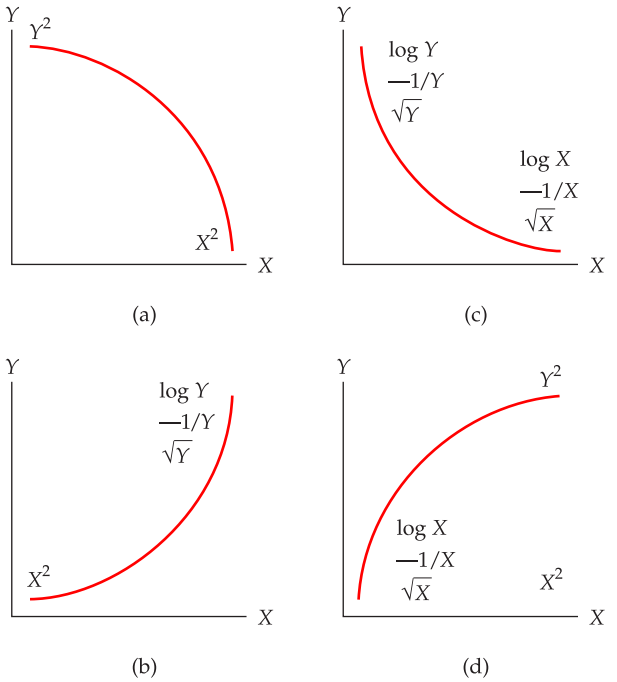


Figure 2.13
Selecting Transformations to Achieve Linearity
***Source*: F. Mosteller and J. W. Tukey, *Data Analysis and Regression*. Reading, MA: Addison-Wesley, 1977.**

the observation had a value exactly equal to the variable mean. Values above or below zero indicate the observation's difference from the variable mean in terms of standard deviations. So a value of .8 represents an observation whose original value was .8 standard deviations above the variable mean.

Standardization is widely used to directly compare observations on variables with different distributional characteristics (i.e., means and standard deviations). For example, a z score value of .8 has the same interpretation across variables on widely varying scales. As an example, standardization underlies the concept of Beta coefficients in multiple regression (see Chapter 5), and is also quite useful in cluster analysis to allow direct comparisons between disparate clustering variables.

Centering A variant of standardization is centering, either on a variable or person basis. **Centering** is typically associated with variable-based methods, the most common being subtracting the mean value from each observation's actual value. This is equivalent to standardization without the division by the standard deviation. The objective is to retain the original variation among the values, but make all variable relative to their mean. Centering is associated with improvements in interpretation of moderation and interaction effects in multiple regression, especially as a correction for multicollinearity [e.g., 72], but more recent research has cast doubt on the claims of these benefits even though it is still thought to improve interpretation of the moderation/interaction term [25, 55].

Another form of centering is **ipsatizing**, which is a person-centered form of centering. In this method, the value subtracted from each value of an observation is the observation's mean, thus making the values now different from the observation's mean value. This form of centering is limited to a set of variables with the same scale, but has been used to deal with response bias across a set of questions [23, 18]. While more limited in its applications, it represents a unique way to make responses relative to the observation rather than a population-derived value (e.g., variable mean). One other application of centering is in multi-level modeling to improve interpretation of effects at all levels [33].

TRANSFORMATIONS RELATED TO SPECIFIC RELATIONSHIP TYPES

Many times transformations are performed with the sole objective being an empirical outcome (e.g., meeting a statistical property) or to improve relationships with other variables (e.g., linearity or polynomials). But sometimes a transformation can lead to a specific concept that has both empirical consequences and conceptual meaning. This

is the case with the log transformation, which is well known for its ability to address nonlinear relationships. But beyond the empirical impact is a substantive meaning that translates into the concept of elasticity in its many forms. **Elasticity** in general is a ratio of the percentage change in two variables. It is widely used in a number of settings where some causal relationship exists between the two variables (e.g., elasticity of demand which details how demand changes relative to changes in price).

But elasticity is just one of a set of log-related transformations which extend past just empirical outcomes to represent specific concepts. Consider the relationship of dependent variable *Y* and independent variable *X*. We know that the coefficient in an untransformed regression equation would provide the change in *Y* (in units) for each unit change in *X*. But what about when we introduce log transformations of the dependent and independent variables, or both? Each of the combinations of transformations results in quite different interpretations [44].

- *Log-linear*: a log of the *Y* variable with an untransformed *X* variable provides an estimate of the percentage change in *Y* given a one unit change in *X*.
- *Linear-log*: a log of the *X* variable with an untransformed *Y* variable provides an estimate of the unit change in *Y* for a percentage change in *X*
- *Log-log*: a log of both *X* and *Y* provides the ratio of the percentage change of *Y* given a percentage change in *X*, the definition of elasticity.

The use of the log transformation extends past an empirical transformation to represent concepts with substantive implications. This example suggests that researchers always be ready to "look beyond" just the empirical outcomes of their transformations to the substantive meanings represented in these transformations.

TRANSFORMATIONS RELATED TO SIMPLIFICATION

Many times the examination phase of data analysis presents the researcher with a simple problem—How do I make the data simple? Faced with thousands or even millions of cases, each with their own unique values on certain variables (e.g., income), how does the researcher gain some perspective on the data? Researchers today have several approaches that can be characterized as either some form of binning or smoothing.

Binning Distributional characteristics such as the mean and standard deviation tell us some basic characteristics, but don't describe the pattern of the distribution (e.g., bi-modal, relatively flat or peaked, etc.). We can use some additional measures (e.g., skewness and kurtosis) to provide empirical assessments, but many times a graphical examination is revealing. This is where the concept of binning comes into play. **Binning** is the categorization of values into "bins" or categories. We saw this earlier when we displayed the data in a frequency distribution and overlayed the normal distribution. The need for binning comes into play due to the **cardinality** of a variable – the number of unique data values for that variable across the observations. Nonmetric variables generally do not have issues with cardinality since they are already discrete values that represent a set of observations for each value (e.g., gender or occupational categories). Metric variables, however, in their attempt to provide more detailed information about each case, can result in potentially unique values for each observation (e.g., genomes in the study of DNA). Now while that may be quite useful in models, it creates problems in any form of data examination. How do we view the multitude of values and even relate them to just a single other variable?

Thus, as one might suspect, the process of binning can be quite complex—how many categories/bins, equal size in value or frequency, etc.? The emergence of Big Data has necessitated many routines in most software packages to automate the binning process to provide the researcher with an alternative means to examine data. The binning processes can either be performed to best describe the data distribution or it can be "optimized" to make the binned variable most predictive in some form of model. And the problem becomes even more difficult when trying to assess the relationship between two variables. How to avoid a scatterplot with just a large mass of points, making any form of relationship indistinguishable? Cross-tabulating binned variables and then color-coding the resulting bins creates a **heat map** that can distinguish magnitude and allow for relationships to emerge. So techniques are emerging that

enable researchers to handle these variables with high cardinality in ways that make them amenable to both data examination and analysis.

Researchers have also attempted to simplify variables for analysis by two other approaches: dichotomization and extreme groups. While they are quite distinct from the binning methods we just discussed, they still are based on the same principle: categorize observations into a much smaller number of discrete groups.

DICHOTOMIZATION The use of **dichotomization**—dividing cases into two classes based on being above or below a specified value—is generally discouraged because it is arbitrary and non-scientific, but still widely used. We have all heard from our basic statistics course not to move down in measurement level—don't transform a metric to a nonmetric variable—and lose the information in the metric variable. But we see it done all the time as a means of simplification (e.g., forming groups that are high vs low) or as a first step in assessing moderation (e.g., divide the observations into groups and then estimate separate models for each group). Yet in most research contexts dichotomization is not recommended. There are certain situations in which the independent variables are not correlated that dichotomization can be useful [48], but in most situations there is caution in their use because of unintended results and loss of statistical power [61, 76]. If researchers want dichotomization a much better approach is to apply cluster analysis to find the natural groupings.

EXTREME GROUPS A variation of dichotomization is the **extreme groups approach** where observations are formed into perhaps three groups (e.g., high, medium and low) and then the middle group is discarded for analytical purposes. The objective is to highlight the directionality of the relationship and provide for a more easily understood interpretation. While appealing in its basic objective, the approach has generally been seen as less useful for a number of reasons. It has a tendency to make nonlinear relationships hard to identify, reduces the statistical power of the analysis, in essence reducing analyses to bivariate relationships, and can lead to erroneous conclusions in certain situations [68]. Thus, researchers are cautioned in the use of the extreme groups approach unless the relationship is considered strong and the group-level comparison provides the most appropriate approach.

Smoothing A somewhat opposite approach to simplification is smoothing—fitting a relationship to the distribution that represents its shape. For many years a method used in time series analysis and also as a form of data reduction, smoothing is the underlying principle in tackling such problems as areal/geographic extent of a market area [47], the analysis of difference scores [30, 81] or the representation of a variable in a spatial setting (e.g., a surface of values across a map). In each of these instances an equation is used to describe the pattern of points that can be represented as a surface. An interesting application is **response surface** methodology, which is the application of polynomial regression to estimate a surface that best describes a multivariate response surface. Used in many multifactor experiments to provide a basis for optimization of the various factors [52], it also is used in fields as diverse as analytical chemistry [10] to process and product optimization [8]. Here the response surface is used to identify a functional form that can be analyzed by an optimization process to identify the optimal combination of input factors.

But in all of its forms, smoothing provides a mathematical formulation that can be used to describe the distribution of values, either as portrayal as a response surface or used in other analyses. As researchers face outcomes that have more complex patterns in the outcome values, some form of smoothing may be appropriate.

GENERAL GUIDELINES FOR TRANSFORMATIONS

Even our brief discussion above illustrates all of the potential transformations that researchers have at their disposal. Many times the type of transformation is dictated by quantitative or statistical demands, but in a wide number of situations the research can selectively employ some of these transformations to dramatically improve both the empirical results and the interpretability of those results.

One caveat for transformations:

> When explanation is important, beware of transformations!

Transforming Data

To judge the potential impact of a transformation, calculate the ratio of the variable's mean to its standard deviation:

- Noticeable effects should occur when the ratio is less than 4.
- When the transformation can be performed on either of two variables, select the variable with the smallest ratio.

Transformations should be applied to the independent variables except in the case of heteroscedasticity.

Heteroscedasticity can be remedied only by the 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.

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.

Use variables in their original (untransformed) format when profiling or interpreting results.

If the purpose of the analysis is only prediction, then obviously any type of transformation that improves the outcome is acceptable. But when explanation of the results is also desired, be cautious about the number of types of transformations employed. Some transformations, such as the log transformation described earlier, can have substantive meaning. But even just an inverse transformation, perhaps as a variance stabilizing measure, still impacts the interpretation as the direction of the relationship switches and the scale of the parameter is now not directly interpretable. We just caution researchers to strive for a balance in the use of transformations in order to achieve the best results possible from both empirical and interpretation perspectives. Apart from the technical issues of the type of transformation, several points to remember when performing data transformations are presented in Rules of Thumb 2-5.

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 dataset introduced in Chapter 1 will be examined. In the course of this 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 assumption 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 assessment of normality of the metric variables involves both empirical measures of a distribution's shape characteristics (skewness and kurtosis) and the normal probability plots. The empirical measures provide a guide as to the variables with significant deviations from normality, and the normal probability plots provide a visual portrayal of the shape of the distribution. The two portrayals complement each other when selecting the appropriate transformations.

Table 2.11 and Figure 2.14 contain the empirical measures and normal probability plots for the metric variables in our data set. Our first review concerns the empirical measures reflecting the shape of the distribution (skewness and kurtosis) as well as a statistical test for normality (the modified Kolmogorov–Smirnov test). Of the 17 metric variables, only 6 (X_6, X_7, X_{12}, X_{13}, X_{16}, and X_{17}) show any deviation from normality in the overall

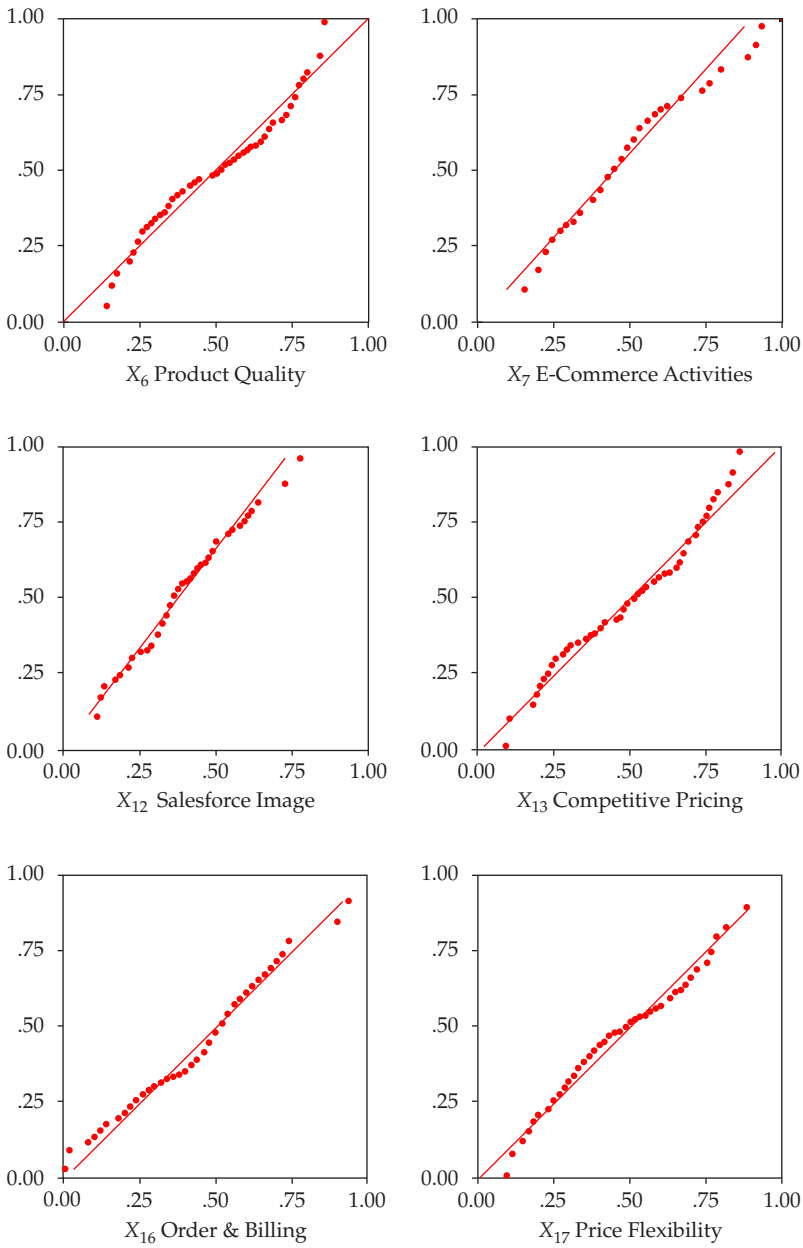
Table 2.11 Distributional Characteristics, Testing for Normality, and Possible Remedies

	SHAPE DESCRIPTORS								
	Skewness		*Kurtosis*		*Tests of Normality*		*Applicable Remedies*		
Variable	Statistic	*z* value	Statistic	*z* value	Statistic	Significance	Description of the Distribution	Transformation	Significance After Remedy
Firm Characteristics									
X_6	−.245	−1.01	−1.132	**−2.37**	.109	**.005**	Almost uniform distribution	Squared term	.015
X_7	.660	**2.74**	.735	1.54	.122	**.001**	Peaked with positive skew	Logarithm	.037
X_8	−.203	−.84	−.548	−1.15	.060	.200[a]	Normal distribution		
X_9	−.136	−.56	−.586	−1.23	.051	.200[a]	Normal distribution		
X_{10}	.044	.18	−.888	−1.86	.065	.200[a]	Normal distribution		
X_{11}	−.092	−.38	−.522	−1.09	.060	.200[a]	Normal distribution		
X_{12}	.377	1.56	.410	.86	.111	**.004**	Slight positive skew and peakedness		
X_{13}	−.240	−1.00	−.903	−1.89	.106	**.007**	Peaked	Cubed term	-
X_{14}	.008	.03	−.445	−.93	.064	.200[a]	Normal distribution		
X_{15}	.299	1.24	.016	.03	.074	.200[a]	Normal distribution		
X_{16}	−.334	−1.39	.244	.51	.129	**.000**	Negative skewness	Squared term	.066
X_{17}	.323	1.34	−.816	−1.71	.101	**.013**	Peaked, positive skewness	Inverse	.187
X_{18}	−.463	−1.92	.218	.46	.084	.082	Normal distribution		
Performance Measures									
X_{19}	.078	.32	−.791	−1.65	.078	.137	Normal distribution		
X_{20}	.044	.18	−.089	−.19	.077	.147	Normal distribution		
X_{21}	−.093	−.39	−.090	−.19	.073	.200[a]	Normal distribution		
X_{22}	−.132	−.55	−.684	−1.43	.075	.180	Normal distribution		

[a]Lower bound of true significance.

Note: 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.

Figure 2.14 Normal Probability Plots (NPP) of Non-normal Metric Variables (X_6, X_7, X_{12}, X_{13}, X_{16}, and X_{17})



normality tests. When viewing the shape characteristics, significant deviations were found for skewness (X_7) and kurtosis (X_6). One should note that only two variables were found with shape characteristics significantly different from the normal curve, while six variables were identified with the overall tests. The overall test provides no insight as to the transformations that might be best, whereas the shape characteristics provide guidelines for possible transformations. The researcher can also use the normal probability plots to identify the shape of the distribution. Figure 2.14 contains the normal probability plots for the six variables found to have the non-normal distributions. By combining information, from the empirical and graphical methods, the researcher can characterize the non-normal distribution in anticipation of selecting a transformation (see Table 2.11 for a description of each non-normal distribution).

Table 2.11 also suggests the appropriate remedy for each of the variables. Two variables (X_6 and X_{16}) were transformed by taking the square root. X_7 was transformed by logarithm, whereas X_{17} was squared and X_{13} was cubed. Only X_{12} could not be transformed to improve on its distributional characteristics. For the other five variables, their tests of normality were now either nonsignificant (X_{16} and X_{17}) or markedly improved to more acceptable levels (X_6, X_7, and X_{13}). Figure 2.15 demonstrates the effect of the transformation on X_{17} in achieving normality. The transformed X_{17} appears markedly more normal in the 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 remaining variable (X_{12}), none of the transformations could improve the normality. This variable will have to be used in its original form. In situations where the normality of the variables is critical, the transformed variables can be used with the assurance that they meet the assumptions of normality. But the departures from normality are 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 dataset. These analyses are appropriate 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. Among the performance factors, only X_4 (Region) has notable problems with heteroscedasticity. For the 13 firm characteristic variables, only X_6 and X_{17} show patterns of heteroscedasticity on more than one of the nonmetric variables. Moreover, in no instance do any of the nonmetric variables have more than two problematic metric variables. The actual 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. The relative lack of either numerous problems or any consistent patterns across one of the nonmetric variables suggests that heteroscedasticity problems will be minimal. If the assumption violations are found, variable transformations are available to help remedy the variance dispersion.

The ability for transformations to address the problem of heteroscedasticity for X_{17}, if desired, is also shown in Figure 2.15. Before a logarithmic transformation was applied, heteroscedastic conditions were found on three of the five nonmetric variables. The transformation not only corrected the nonnormality problem, but also eliminated the problems with heteroscedasticity. It should be noted, however, that several transformations "fixed" the normality problem, but only the logarithmic transformation also addressed the heteroscedasticity, which demonstrates the relationship between normality and heteroscedasticity and the role of transformations in addressing each issue.

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 5 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 linearity relates to the patterns of association between each pair of variables and the ability of the correlation coefficient to adequately 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

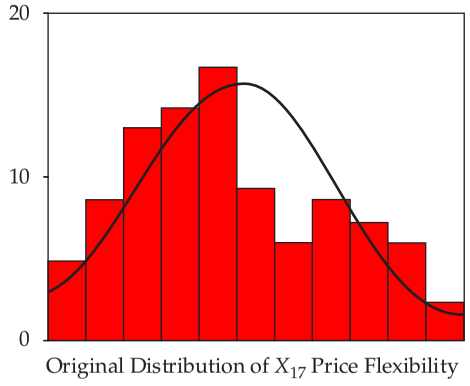
Table 2.12 Testing for Homoscedasticity

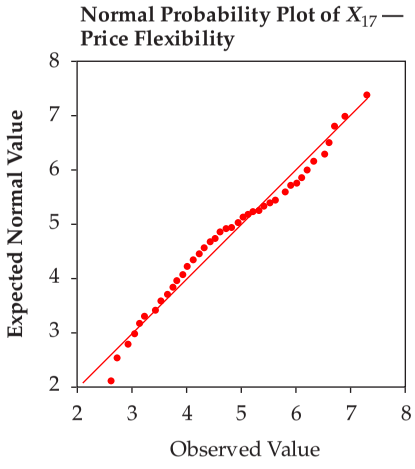
	NONMETRIC/CATEGORICAL VARIABLE									
	X_1 *Customer Type*		X_2 *Industry Type*		X_3 *Firm Size*		X_4 *Region*		X_5 *Distribution System*	
Metric Variable	Levene Statistic	Sig.	Levene Statistic	Sig.	Levene Statistic	Sig.	Levene Statistic	Sig.	Levene Statistic	Sig.
Firm Characteristics										
X_6	**17.47**	**.00**	.01	.94	.02	.89	**17.86**	**.00**	.48	.49
X_7	.58	.56	.09	.76	.09	.76	.05	.83	2.87	.09
X_8	.37	.69	.48	.49	1.40	.24	.72	.40	.11	.74
X_9	.43	.65	.02	.88	.17	.68	.58	.45	1.20	.28
X_{10}	.74	.48	.00	.99	.74	.39	1.19	.28	.69	.41
X_{11}	.05	.95	.15	.70	.09	.76	3.44	.07	1.72	.19
X_{12}	2.46	.09	.36	.55	.06	.80	1.55	.22	1.55	.22
X_{13}	.84	.43	**4.43**	**.04**	1.71	.19	.24	.63	2.09	.15
X_{14}	2.39	.10	2.53	.11	**4.55**	**.04**	.25	.62	.16	.69
X_{15}	1.13	.33	.47	.49	1.05	.31	.01	.94	.59	.45
X_{16}	1.65	.20	.83	.37	.31	.58	2.49	.12	**4.60**	**.03**
X_{17}	**5.56**	**.01**	2.84	.10	**4.19**	**.04**	**16.21**	**.00**	.62	.43
X_{18}	.87	.43	.30	.59	.18	.67	2.25	.14	**4.27**	**.04**
Performance Measures										
X_{19}	**3.40**	**.04**	.00	.96	.73	.39	**8.57**	**.00**	.18	.67
X_{20}	1.64	.20	.03	.86	.03	.86	**7.07**	**.01**	.46	.50
X_{21}	1.05	.35	.11	.74	.68	.41	**11.54**	**.00**	2.67	.10
X_{22}	.15	.86	.30	.59	.74	.39	.00	.99	1.47	.23

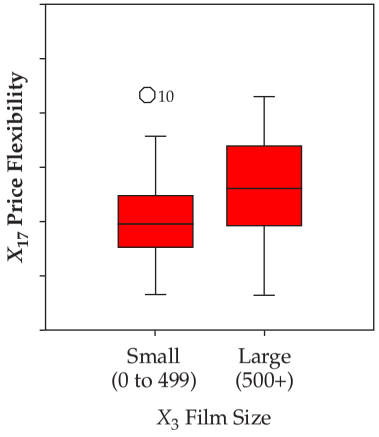
Notes: Values represent the Levene statistic value and the statistical significance in assessing the variance dispersion of each metric variable across the levels of the nonmetric/categorical variables. Values in bold are statistically significant at the .05 level or less.

Figure 2.15 Transformation of X_{17} (Price Flexibility) to Achieve Normality and Homoscedasticity

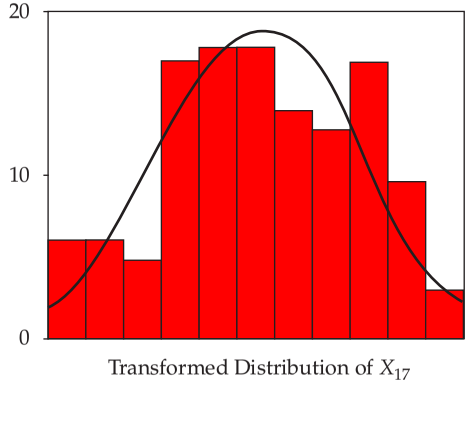
ORIGINAL VARIABLE

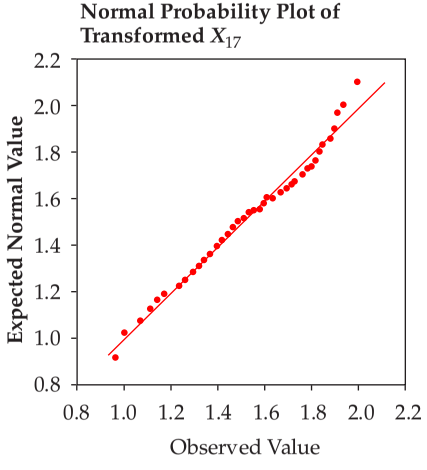


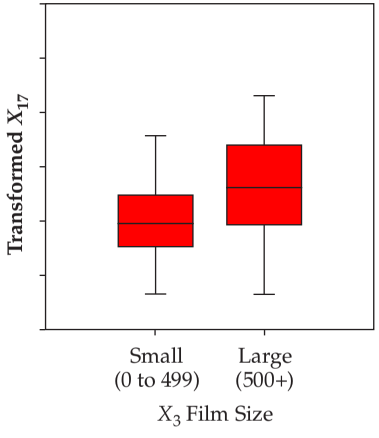




TRANSFORMED VARIABLE







(Continued)

Figure 2.15 Continued

Distribution Characteristics Before and After Transformation						
	SHAPE DESCRIPTORS				*Test of Normality*	
	Skewness		*Kurtosis*			
Variable Form	**Statistic**	***z* value[a]**	**Statistic**	***z* value[a]**	**Statistic**	**Significance**
Original X_{17}	.323	1.34	−.816	−1.71	.101	.013
Transformed X_{17}^{b}	−.121	.50	−.803	−1.68	.080	.117

[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.

[b] Logarithmic transformation.

	Levene Test Statistic				
Variable Form	**X_1 Customer Type**	**X_2 Industry Type**	**X_3 Firm Size**	**X_4 Region**	**X_5 Distribution System**
Original X_{17}	5.56**	2.84	4.19*	16.21**	.62
X_{17}	2.76	2.23	1.20	3.11**	.01

* Significant at .05 significance level.

** Significant at .01 significance level.

relationships are present. The reader can also refer to Figure 2.2, the scatterplot matrix containing the scatterplot for selected metric variables in the dataset. Examination of the scatterplots does not reveal any apparent nonlinear relationships. Review of the scatterplots not shown in Figure 2.2 also did not reveal any apparent nonlinear patterns. 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 (see Chapter 1 for a more detailed discussion of selecting multivariate techniques). Some of the techniques, such as discriminant analysis or multivariate analysis of variance, specifically require nonmetric data as dependent or independent variables. 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 assumed metric measurement for variables. 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.

CONCEPT OF DUMMY VARIABLES

The researcher has available a method for using dichotomous variables, known as **dummy variables**, which act as replacement variables for the nonmetric variable. *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 Figure 2.16). To represent the nonmetric variable gender, we would create two new dummy variables (X_1 and X_2), as shown in Figure 2.16. X_1 would represent those individuals who are

Figure 2.16 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 & $\leq$\$25,000	$X_4 = 1$, else $X_4 = 0$
		if $>$ \$25,000	$X_5 = 1$, else $X_5 = 0$

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, 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 Figure 2.16, we would first define three dummy variables (X_3, X_4, and X_5). 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.

DUMMY VARIABLE CODING

In constructing dummy variables, two approaches can be used to represent the categories, and more importantly, the category that is omitted, known as the **reference category** or **comparison group**.

Indicator Coding The first approach, known as **indicator coding**, uses three ways to represent the household income levels with two dummy variables, as shown in Figure 2.17. *An important consideration is the reference category or comparison group, 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 dependent variable.* The deviations represent the differences between the dependent variable mean score for each group of respondents (represented by a separate dummy variable) and the comparison group. This form is most appropriate in 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.

Figure 2.17 Alternative Dummy Variable Coding Patterns for a Three-Category Nonmetric Variable

	Pattern 1		Pattern 2		Pattern 3	
Household Income Level	X_1	X_2	X_1	X_2	X_1	X_2
If $<$ \$15,000	1	0	1	0	0	0
If $<$ \$15,000 and $\leq$\$25,000	0	1	0	0	1	0
If $>$ \$25,000	0	0	0	1	0	1

Effects Coding 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.

USING DUMMY VARIABLES

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 correlation patterns, such as exploratory 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.

Researchers should examine and explore the nature of the data and the relationships among variables before the application of any of the multivariate techniques. This chapter helps the researcher to do the following:

Select the appropriate graphical method to examine the characteristics of the data or relationships of interest. Use of multivariate techniques places an increased burden on the researcher to understand, evaluate, and interpret the more complex results. It requires a thorough understanding of the basic characteristics of the underlying data and relationships. The first task in data examination is to determine the character of the data. A simple, yet powerful, approach is through graphical displays, which can portray the univariate, bivariate, and even multivariate qualities of the data in a visual format for ease of display and analysis. The starting point for understanding the nature of a single variable is to characterize the shape of its distribution, which is accomplished with a histogram. The most popular method for examining bivariate relationships is the scatterplot, a graph of data points based on two variables. Researchers also should examine multivariate profiles. Three types of graphs are used. The first graph type is a direct portrayal of the data values, either by glyphs that display data in circles or multivariate profiles that provide a barlike profile for each observation. A second type of multivariate display involves a transformation of the original data into a mathematical relationship, which can then be portrayed graphically. The most common technique of this type is the Fourier transformation. The third graphical approach is iconic representativeness, the most popular being the Chernoff face.

Assess the type and potential impact of missing data. Although some missing data can be ignored, missing data is still one of the most troublesome issues in most research settings. At its best, it is a nuisance that must be remedied to allow for as much of the sample to be analyzed as possible. In more problematic situations, however, it can cause serious biases in the results if not correctly identified and accommodated in the analysis. The four-step process for identifying missing data and applying remedies is as follows:

Determine the type of missing data and whether or not it can be ignored.
Determine the extent of missing data and decide whether respondents or variables should be deleted.
Diagnose the randomness of the missing data.
Select the imputation method for estimating missing data.

Understand the different types of missing data processes. A missing data process is the underlying cause for missing data, whether it be something involving the data collection process (poorly worded questions, etc.) or the individual (reluctance or inability to answer, etc.). When missing data are not ignorable, the missing data process can be classified into one of two types. The first is MCAR, which denotes that the effects of the missing data process are randomly distributed in the results and can be remedied without incurring bias. The second is MAR, which denotes that the underlying process results in a bias (e.g., lower response by a certain type of consumer) and any remedy must be sure to not only "fix" the missing data, but not incur bias in the process.

Explain the advantages and disadvantages of the approaches available for dealing with missing data. The remedies for missing data can follow one of two approaches: using only valid data or calculating replacement data for the missing data. Even though using only valid data seems a reasonable approach, the researcher must remember that doing so assures the full effect of any biases due to nonrandom (MAR) data processes. Therefore, such approaches can be used only when random (MCAR) data processes are present, and then only if the sample is not too depleted for the analysis in question (remember, missing data excludes a case from use in the analysis). The calculation of replacement values attempts to impute a value for each missing value, based on criteria ranging from the sample's overall mean score for that variable to specific characteristics of the case used in a predictive relationship. Again, the researcher must first consider whether the effects are MCAR or MAR, and then select a remedy balancing the specificity of the remedy versus the extent of the missing data and its effect on generalizability.

Identify univariate, bivariate, and multivariate outliers. Outliers are observations with a unique combination of characteristics indicating they are distinctly different from the other observations. These differences can be on a single variable (univariate outlier), a relationship between two variables (bivariate outlier), or across an entire set of

variables (multivariate outlier). Although the causes for outliers are varied, the primary issue to be resolved is their representativeness and whether the observation or variable should be deleted or included in the sample to be analyzed.

Test your data for the assumptions underlying most multivariate techniques. Because our analyses involve the use of a sample and not the population, we must be concerned with meeting the assumptions of the statistical inference process that is the foundation for all multivariate statistical techniques. The most important assumptions include normality, homoscedasticity, linearity, and absence of correlated errors. A wide range of tests, from graphical portrayals to empirical measures, is available to determine whether assumptions are met. Researchers are faced with what may seem to be an impossible task: satisfy all of these statistical assumptions or risk a biased and flawed analysis. These statistical assumptions are important, but judgment must be used in how to interpret the tests for each assumption and when to apply remedies. Even analyses with small sample sizes can withstand small, but significant, departures from normality. What is more important for the researcher is to understand the implications of each assumption with regard to the technique of interest, striking a balance between the need to satisfy the assumptions versus the robustness of the technique and research context.

Determine the best method of data transformation given a specific problem. When the statistical assumptions are not met, it is not necessarily a "fatal" problem that prevents further analysis. Instead, the researcher may be able to apply any number of transformations to the data in question that will solve the problem and enable the assumptions to be met. 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. Most of the transformations involve modifying one or more variables (e.g., compute the square root, logarithm, or inverse) and then using the transformed value in the analysis. It should be noted that the underlying data are still intact, just their distributional character is changed so as to meet the necessary statistical assumptions.

Understand how to incorporate nonmetric variables as metric variables. An important consideration 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. In many instances, the multivariate methods require that metric variables be used. Yet nonmetric variables are often of considerable interest to the researcher in a particular analysis. A method is available to represent a nonmetric variable with a set of dichotomous variables, known as dummy variables, so that it may be included in many of the analyses requiring only metric variables. A dummy variable is a dichotomous variable that has been converted to a metric distribution and represents one category of a nonmetric independent variable.

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.

Explain how graphical methods can complement the empirical measures when examining data.

List potential underlying causes of outliers. Be sure to include attributions to both the respondent and the researcher.

Discuss why outliers might be classified as beneficial and as problematic.

Distinguish between data that are missing at random (MAR) and missing completely at random (MCAR). Explain how each type affects the analysis of missing data.

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.

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.

Discuss the following statement: Multivariate analyses can be run on any data set, as long as the sample size is adequate.