SOLUTIONS MANUAL

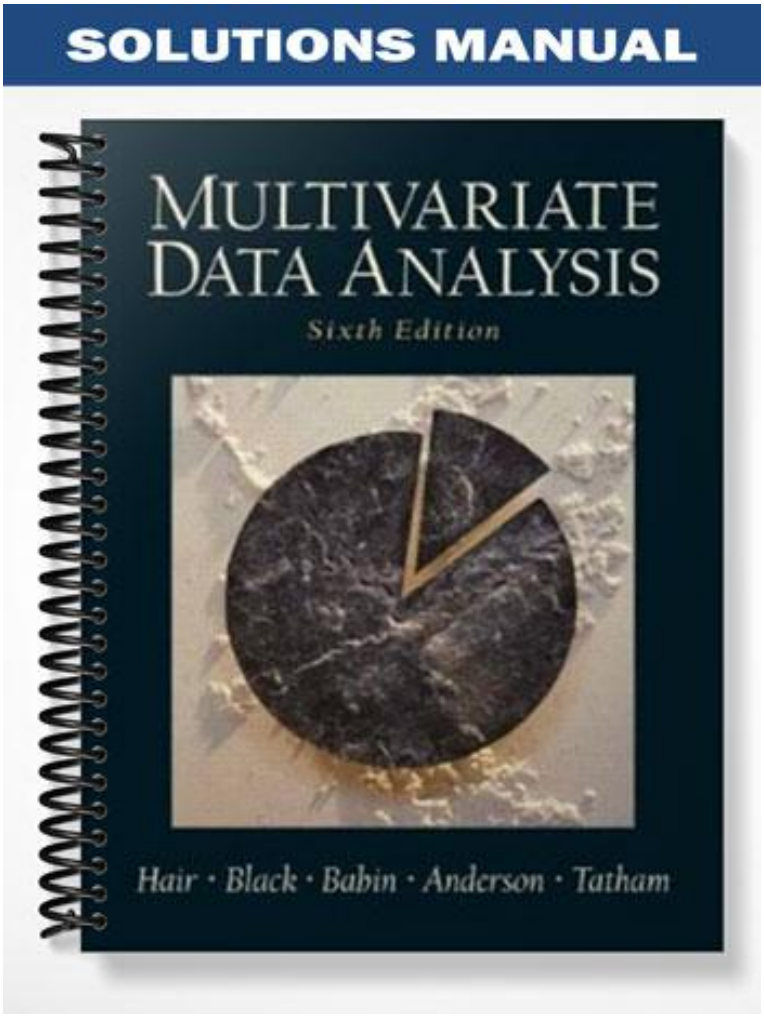

Table of Contents

INTRODUCTION

This manual has been designed to provide teachers using **Multivariate Data Analysis, 6th edition,** with supplementary teaching aids. The course suggestions made here are the result of years of experience teaching the basic content of this text in several universities. Obviously, the contents may be modified to suit the level of the students and the length of the term.

Multivariate data analysis is an interesting and challenging subject to teach. As an instructor, your objective is to direct your students' energies and interests so that they can learn the concepts and principles underlying the various techniques. You will also want to help your students learn to apply the techniques. Through years of teaching multivariate analysis, we have learned that the most effective approach to teaching the techniques is to provide the students with real-world data and have them manipulate the variables using several different programs and techniques. The text is designed to facilitate this approach, making available several data sets for analysis. Moreover, accompanying sample output and control cards are provided to supplement the analyses discussed in the text.

WHAT'S NEW AND WHAT'S CHANGED

The sixth edition has many substantial changes from prior editions that we feel will markedly improve the text for both faculty member and student. Three notable additions were made to the text:

- The most obvious change in the sixth edition is the new database—HBAT. The emphasis on improved measurement, particularly multi-item constructs, led us to develop HBAT. After substantial testing we believe it provides an expanded teaching tool with various techniques that are comparable to the HATCO database, which will still be available on the book's Web site.

- A second major addition is "Rules of Thumb" for the application and interpretation of the various techniques. The rules of thumb are highlighted throughout the chapters to facilitate their use. We are confident these guidelines will facilitate your utilization of the techniques.

- A third major change to the text is a substantial expansion in coverage of structural equations modeling. We now have three chapters on this increasingly important technique. Chapter 10 provides an overview of structural equation modeling, Chapter 11 focuses on confirmatory factor analysis, and Chapter 12 covers issues in estimating structural models. These three chapters provide a comprehensive introduction to this technique.

Each chapter has been revised to incorporate advances in technology, and several chapters have undergone more extensive change:

- Chapter 2 “Examining the Data” has an expanded section on missing data assessment, including a flowchart depicting a series of decisions that are involved in identifying and then accommodating missing data.

- Chapter 5, “Multiple Discriminant Analysis and Logistic Regression,” provides complete coverage of analysis of categorical dependent variables, including both discriminant analysis and logistic regression. An expanded discussion of logistic regression includes an illustrative example using the HBAT database.

- Chapter 7, “Conjoint Analysis,” has a revised examination of issues of research design that focuses on the development of the conjoint stimuli in a concise and straightforward manner.

An important development is the continuation of the Web site *“Great Ideas in Teaching Multivariate Statistics”* at www.mvstats.com, which can also be accessed as the Companion Web site at www.prenhall.com/hair. This Web site acts as a resource center for the textbook as well as everyone interested in multivariate analysis, providing links to resources for each technique as well as a forum for identifying new topics or statistical methods. In this way we can provide more timely feedback to researchers than if they were to wait for a new edition of the book. We also plan for the Web site to be a clearinghouse for materials on teaching multivariate statistics—providing exercises, datasets, and project ideas.

ORGANIZATION OF THE CHAPTERS IN THE TEXT

The text is designed to help make your teaching as enjoyable and as simple as possible. Each chapter begins with a "Chapter Preview" so that students will understand the major concepts they are expected to learn. To facilitate understanding of the chapter material and as a ready reference for clarification, definitions of key terms are presented at the front of each chapter. The text is designed for those individuals who want to obtain a conceptual understanding of multivariate methods—what they can do, when they should be used, and how the results should be interpreted.

Following this design, each chapter is structured in a step-by-step manner, including six steps. The end of each chapter includes an illustration of how to apply and interpret each technique. Basically, the approach is for the "data analyst," therefore, the math formulae and symbols are minimized. We believe it is the most practical, readable guide available to understanding and applying these otherwise complex statistical tools. At the end of each chapter, we review the Learning Objectives to provide the student with an overview of what has been covered in the chapter in relation to those major concepts or issues defined in the

Learning Objectives. Finally, a series of questions is provided to stimulate the student to evaluate what has been read and to translate this material into a workable knowledge base for use in future applications.

ORGANIZATION OF THE CHAPTER MATERIALS

This instructor's manual is designed to facilitate the preparation and conduct of classes, exams, and seminars. Materials included in the manual are organized in two sections for each chapter.

(1) Chapter summaries: to refresh the instructor's memory without the necessity of re-reading the entire chapter prior to class. Each chapter summary is organized around four major sections. The objective of these sections is to identify particular issues that may be useful in organizing class discussion. The four sections are:

a. **What** – an overview, or brief description, of the technique.

b. **Why** – a description of the basic objectives of the technique.

c. **When** – identification of the types of problems the technique may be used to address.

d. **How** – description of the assumptions applicable to the technique, the data requirements for its use, the major points which are essential to the successful implementation of the research plan and the key points contained in the computer output needed for a complete and accurate interpretation of the results.

(2) Answers to the end-of-chapter questions: suggested answers to the questions that can form the basis for further elaboration if desired.

Sample exam questions: while essay or short answer questions are probably preferable for examinations, many times multiple choice questions can be used as a method for assessing specific knowledge about the subject. To ease the burden of writing exam questions, multiple choice questions are provided for each chapter. All of the sample exam questions have been placed in a separate section at the end of the Instructor’s Manual.

THE COMPANION WEB SITE – MVSTATS.COM

The authors have established a Web site entitled "Great Ideas in Teaching Multivariate Statistics" with the objective of providing a clearinghouse for instructional materials and a forum for discussions about pedagogical issues. Accessed directly at www.mvstats.com or through the Companion Web site link at the Prentice-Hall Web site (www.prenhall.com/hair), the Web site will offer all of the supplementary materials for the text (datasets, control card files and output files) as well as links to additional datasets for use in class and Web-based materials for each technique. A complete set of datasets and related materials are available not only for the sixth edition, but the fifth edition (HATCO dataset) as well. We sponsor a mailing list MVSTAT-L that is open to the interested participants for asking questions related to either the teaching or application of multivariate statistics.

An important adjunct is a "faculty-only" section of the Web site where additional pedagogical materials will be made available to all adopters of the textbook. The permission-based section will allow for providing the text-related materials (e.g., PowerPoint slides and image files of all figures in the text) as well as acting as a forum for faculty interested in teaching multivariate statistics. We encourage any faculty member to contribute to the mailing list or even contribute class-related materials which we will disseminate among all those interested in the subject area. We envision this to be an evolutionary project, with its growth and focus guided primarily by its users and contributors. We hope that we can provide an easy and readily available forum for discussion and collaboration among interested faculty members.

The understanding and interpretation of multivariate analysis is enhanced immeasurably by the ability of students to actually analyze data (e.g., derive research questions and execute the analysis) and/or examine actual computer output to ascertain what is the form and content of the results. To avoid any unnecessary duplication of effort on the part of the instructor, the Web site contains a number of computer files related to the analyses conducted throughout the text:

Datasets

There are a number of datasets provided to students and faculty to perform all of the multivariate analyses described in the textbook. While some techniques require specialized datasets (e.g., multidimensional scaling, conjoint and structural equation modeling), many of the techniques can be performed using conventional survey data. To this end, a common dataset has been developed for use with many of the techniques to allow students to see the interrelationships among techniques as well as the techniques themselves. The HBAT dataset has three forms utilized throughout the text:

- HBAT – the primary database described in the text which has multiple metric and nonmetric variables allowing for use in most of the multivariate techniques.

- HBAT200 – an expanded dataset, comparable to HBAT except for 200 rather than 100 respondents, that allows for multiple independent variables in MANOVA while still maintaining adequate sample size in the individual cells.

- HBAT_MISSING – a reduced dataset with 70 respondents and missing data among the variables. It is utilized to illustrate the techniques for diagnosis and remedy of missing data described in Chapter 2.

In addition to these datasets, there are several others used with specific techniques, including conjoint analysis, multidimensional scaling and structural equation modeling. These datasets include:

- HBAT_CPLAN and HBAT_CONJOINT – the datafiles needed to perform the "full profile" conjoint analysis available in SPSS.

- HBAT_MDS and HBAT_CORRESP – the datafiles used in performing the multidimensional scaling and correspondence analyses described in the text.

- HBAT_SEM – the original data responses from 400 individuals which are the basis for the structural equation analyses of Chapters 10, 11 and 12. HBAT_COV is the covariance matrix derived from HBAT_SEM that is used as input to structural equation programs such as LISREL, EQS or AMOS.

Finally, two additional datasets are provided to allow students access to data other than the HBAT datafiles described in the textbook:

- HATCO – this dataset has been utilized in past versions of the textbook and provides a simplified set of variables amenable to all of the basic multivariate techniques.

- SALES – this dataset concerns sales training and is comprised of 80 respondents, representing a portion of data that was collected by academic researchers. Also, a copy of the sales training questionnaire is provided.

Given the widespread interchangeability of data formats among statistical programs, all of the datasets are provided in two formats. First is the .SAV format used in SPSS, which allows for documentation of variable descriptions, etc., in a standard format. Also, all of the datasets are contained in an EXCEL worksheet, amenable to input to any statistical program. Between these two formats the student or faculty member should be able to employ the data with any available statistical software program.

Control Card Files:

To assist the instructor in performing the analyses illustrated in the text, control card files containing program syntax commands are provided for the statistical software programs SPSS (Statistical Package for the Social Sciences, SPSS Inc.) and LISREL (Scientific Software, Inc.).

Computer Outputs

If computer access is not available for a particular technique, files of the actual outputs (from SPSS for Chapters 2 – 9 and LISREL for Chapters 10-12) for each analysis are also provided. This enables faculty and students with the complete computer outputs even without actually running the programs.

Acknowledgements

The authors wish to express their thanks to the following colleagues for their assistance in preparation of the current and previous versions of the Instructor's Manual and other course supplements: Rick Andrews, Scott Roach, Barbara Ross, Sujay Dutta, Bruce Alford, Neil Stern, Laura Williams, Jill Attaway, Randy Russ, Alan Bush, Sandeep Bhowmick and Ron Roullier.

CHAPTER ONE
INTRODUCTION TO MULTIVARIATE ANALYSIS

This presentation will approach the general idea of multivariate data analysis by attempting to answer the basic questions of "What?," "Why?," "When?," and "How?".

What is multivariate analysis?

1. Multivariate analysis, in this text, includes both multivariable techniques and multivariate techniques. The term "multivariate analysis" really stands for a family of techniques, and it is important that you realize this fact from the onset. The common characteristic is that multiple variables are used for the dependent and/or independent variables in the analysis. Each technique has its own special powers and unique instances of applicability. These will be revealed to you in detail in subsequent chapters in the text.

2. For now, it is only necessary that you realize in general terms the special powers represented by this family of analyses:

- ***Description*** – Some of the techniques have rather amazing abilities to describe phenomena. They can find patterns of relationships where the human eye and even univariate or bivariate statistics fail to do so.
- ***Explanation*** – Other techniques have special capabilities to explain or to help explain relationships. For instance, they can isolate the impact of one variable on another; show relative differences between the magnitude of impact for two or more variables; or even reveal how one set of variables impinges on another set.
- ***Prediction*** – Certainly every decision maker desires the ability to predict, and multivariate techniques afford that promise. Certain multivariate analyses are explicitly designed to yield predictive modes with specific levels of accuracy.
- ***Control*** – Finally, certain multivariate techniques can be used for control. Of course, the techniques themselves will not institute control, but their application can help a decision maker develop cause-and-effect relationships and models which will ultimately enhance their ability to control events to some degree.

3. There are primary and secondary uses of each of the multivariate techniques as well as cases in which they are not applicable. Given that you will be introduced to each one in the text chapters, perhaps it would be helpful to show the "fit" of each to the four abilities just enumerated. Figure 1 at the end of this section is a reference in this respect. In the Figure, "P" indicates a primary ability of the technique, while "S" represents a secondary ability. The "NA" means that the technique is not applicable.

4. While multivariate analyses are descended from univariate techniques, the extension to multiple variables or variates introduces a number of issues which must be understood before examining each technique in detail.

- ***The Role of the Variate*** – Multivariate techniques differ from univariate techniques in that they employ a variate.
    - Definition: a linear combination of variables with empirically determined weights.
    - Variables are specified by the researcher, and the weights are assigned by the technique.
    - Single value: When summed, the variate produces a single value that represents the entire set of variables. This single value, in addition to the specific contribution of each variable to the variate, is important in all multivariate techniques.
- ***Specification of measurement scales*** – Each multivariate technique assumes the use of a certain type of data. For this reason, the type of data held by the researcher is instrumental in the selection of the appropriate multivariate technique. The researcher must make sure that he or she has the appropriate type of data to employ the chosen multivariate technique.
    - Metric vs nonmetric: Data, in the form of measurement scales, can be metric or nonmetric.
        - Nonmetric data are categorical variables that describe differences in type or kind by indicating the presence or absence of a characteristic or property.
        - Metric data are continuous measures that reflect differences in amount or degree in a relative quantity or distance.

- ***Identification of measurement error***

  - Definition: the degree to which the observed values in the sample are not representative of the true values in the population.

  - Sources: several sources are possible, including data entry errors and the use of inappropriate gradations in a measurement scale.

  - All variables in a multivariate analysis have some degree of measurement error, which adds residual error to the observed variables.

  - Need for assessment of both the validity and the reliability of measures.

    - *Validity* is the degree to which a measure accurately represents what it is supposed to.
    - *Reliability* is the degree to which the observed variable measures the true value and is error-free.

- ***Statistical significance and statistical power***. Almost all of the multivariate techniques discussed in this text are based on ***statistical inference*** of a population's values from a randomly drawn sample of the population.

  - Specifying statistical error levels. Interpretation of statistical inferences requires the specification of acceptable levels of error.

    - *Type I error (alpha)* is the probability of rejecting the null hypothesis when it is actually true.
    - *Type II error (beta)* is the probability of failing to reject the null hypothesis when it is actually false. Alpha and beta are inversely related.

  - Power: the probability (1-beta) of correctly rejecting the null hypothesis when it should be rejected.

    - *Specifying type II error (beta)* also specifies the power of the statistical inference test.

    - Power is *determined by 3 factors*:
      - its inverse relationship with alpha;
      - the effect size; and
      - the sample size.

- When *planning* research, the researcher should:
  - estimate the expected effect size, and
  - then select the sample size and the alpha level to achieve the desired power level.

- Upon *completion* of the analyses, the actual power should be calculated in order to properly interpret the results.

Why is there multivariate data analysis?

The need for multivariate analysis comes from the simple fact of life that just about everything is in some way interrelated with other things. Inflation, for instance, is related to taxes, interest rates, the money supply, oil prices, the business cycle, foreign wars, and a good deal more. Buyers' reactions to an advertisement are related to the price of the item, competitors, warranty terms, previous experiences with the product, conversations with neighbors, credibility of the actor used in the commercial, and season of the year.

To managers or persons responsible for making decisions within an organization, each decision is affected by many "variables." These variables are identifiable entities which have some impact or relationship to the focal topic. In addition to the main impact of these variables, the variables interact with each other as well. In short, a manager is faced with a complete system of complicated relationships in any single decision-problem.

Humans, by their very nature, strive to make some sense out of their environment. They seek some order to the chaos of interrelationships. There are many ways of attempting to achieve order. You can use intuition, logical systems such as set theory, or simplification. You could even choose to ignore the impacts of some variables. But most intelligent managers prefer to rely on some framework which is more accepted.

From these two situations—the myriad of interrelated variables and the need for an accepted framework—springs the need for multivariate analysis techniques. They are the means of achieving parsimonious descriptions, explanations, and predictions of reality.

When do you use multivariate analysis?

Essentially, multivariate analysis is appropriate whenever you have two or more variables observed a number of times. In fact, it is most commonly the case that you have several variables observed a great many times, also known as a "data set." It is traditional to envision a data set as being comprised of rows and columns. The rows pertain to each observation, such as each person or each

completed questionnaire in a large survey. The columns, on the other hand, pertain to each variable, such as a response to a question or an observed characteristic for each person.

Data sets can be immense; a single study may have a sample size of 1,000 or more respondents, each of whom answers 100 questions. Hence the data set would be 1,000 by 100, or 100,000 cells of data. Obviously, the need for parsimony is very evident.

Multivariate techniques aid the researcher in obtaining parsimony by reducing the number of computations necessary to complete statistical tests. For example, when completing univariate tests, if an average were computed for each variable, 100 means would result, and if a correlation were computed between each variable and every other variable, there would be close to 5,000 separate values computed. In sharp contrast, multivariate analysis would require many less computations. For example, factor analysis might result in ten factors. Cluster analysis might yield five clusters. Multiple regression could identify six significant predictor variables. MANOVA might reveal 12 cases of significant differences. Multiple discriminant analysis perhaps would find seven significant variables. It should be evident that parsimony can be achieved by using multivariate techniques when analyzing most data sets.

How do you use multivariate data analysis techniques?

This text follows a structured approach to multivariate model building. The six steps defined in the following chapters should be viewed as guidelines for better understanding the process of applying each technique. The conceptual, as well as the empirical, issues for each technique are discussed. The following summary provides an overview of each of the six steps.

Step 1: Define the Research Problem, Objectives, and Multivariate Technique To Be Used.

Based on an underlying theoretical model, the researcher should define the research question(s). Depending on the type of question, the researcher should also identify the type of multivariate technique, dependence or interdependence, which is most appropriate.

Step 2: Develop the Analysis Plan

Each multivariate technique has an associated set of issues which are relevant to the design of an analysis plan. Sample sizes, scaling of variables, and estimation methods are common issues which must be evaluated by the researcher prior to data collection.

Step 3: Evaluate the Assumptions Underlying the Multivariate Technique

Once the data are collected, the researcher must evaluate the underlying assumptions of the chosen multivariate technique. All techniques have conceptual and statistical assumptions which must be met before the analysis can proceed.

Step 4: Estimate the Multivariate Model and Assess Overall Model Fit

Following a testing of the assumptions, a multivariate model is estimated with the objective of meeting specific characteristics. After the model is estimated, the overall model fit is compared to specified criteria. The model may be respecified for better fit if necessary.

Step 5: Interpret the Variate

Once an acceptable model is achieved, the nature of the multivariate relationship is investigated. Coefficients of the variables are examined. Interpretation may lead to model respecification. The objective is to identify findings which can be generalized to the population.

Step 6: Apply the Diagnostics to the Results

Finally, the researcher must assess whether the results are unduly affected by a single or a small set of observations and determine the degree of generalizability of the results by validation methods. These two actions provide the researcher with support for the research findings.

Some General Guidelines for Multivariate Analysis

Further, six guidelines which are applicable to all multivariate analyses are given below. As you review each chapter, you will find that these issues are identified several times across a number of multivariate techniques. For this reason, these guidelines serve as a "general philosophy" for completing multivariate analyses.

- ***Establish Practical Significance as Well as Statistical Significance***

  Not only must the researcher assess whether or not the results of a multivariate analysis are statistically significant, he or she must also determine whether or not the results have managerial, or practical, implications for action.

- ***Sample Size Affects All Results***

  The size of the sample will impact whether or not the results achieve statistical significance. Too little or too much power can be the result of sample sizes. Researchers should always assess the analysis results in light of the sample size.

- ***Know Your Data***

  Multivariate analyses require rigorous examination of data. Diagnostic measures are available to evaluate the nature of sets of multivariate variables.

- ***Strive for Model Parsimony***

  The researcher should evaluate those variables chosen for inclusion in the analysis. The objective is to create a parsimonious model which includes all relevant variables and excludes all irrelevant variables. Specification error (omission of relevant variables) and high multicollinearity (inclusion of irrelevant variables) can substantially impact analysis results.

- ***Look at Your Errors***

  Often, the first model estimation does not provide the best model fit. Thus, the researcher should analyze the prediction errors and determine potential changes to the model. Errors serve as diagnostics for achieving better model fit.

- ***Validate Your Results***

  The researcher should always validate the results. Validation procedures ensure the analyst that the results are not merely specific to the sample, but are representative and generalizable to the population.

Selecting a Multivariate Technique

Figure 1.2 in the text provides a general decision process with which to select the appropriate multivariate technique. In doing so, the analyst must address two issues: objective of the analysis and type of variables used. First, the analyst must evaluate the theoretical nature of the problem and determine if the objective is to assess a dependence relationship (predictive) or is an interdependence approach (structure seeking) needed. If the relationship is dependence based, the analyst must first determine the number of dependent variables. Second, the analyst must determine how the dependent variables will be measured. In selecting among the interdependence techniques, the analyst must define whether structure among variables, respondents or objects is desired.

FIGURE 1
MULTIVARIATE TECHNIQUES AND THEIR ABILITIES

Technique	Ability			
Control	Describe	Explain	Predict	
Multiple Regression	S	S	P	NA
Multiple Discriminant	S	S	P	NA
MANOVA	NA	S	S	P
Canonical Correlation	P	S	S	NA
Factor Analysis	P	S	NA	NA
Cluster Analysis	P	S	NA	NA
Multidimensional Scaling	S	P	S	NA
Conjoint Analysis	S	P	S	NA
Structural Equation Modeling	S	P	P	NA

Legend:

P: Primary ability
S: Secondary ability
NA: Not applicable

2222222 ANSWERS TO END-OF-CHAPTER QUESTIONS

(1) IN YOUR OWN WORDS, DEFINE MULTIVARIATE ANALYSIS.

Answer

a. The authors adopt a fairly inclusive description of the term. In so doing, they avoid becoming bogged down in the nuances of "multivariable" and "multivariate." The distinction between these terms is made as follows:

Multi-variable - usually referring to techniques used in the simultaneous analysis of more than two variables.

Multivariate - to be considered truly multivariate, all the variables must be random variables which are interrelated in such ways that their different effects cannot easily be studied separately. The multivariate character lies in the multiple combinations of variables, not solely in the number of variables or observations.

(2) NAME SEVERAL FACTORS THAT HAVE CONTRIBUTED TO THE INCREASED APPLICATION OF TECHNIQUES FOR MULTIVARIATE DATA ANALYSIS IN RECENT YEARS.

Answer

a. The ability to conceptualize data analysis has increased through the study of statistical methods;

b. Advances in computer technology which make it feasible to attempt to analyze large quantities of complex data;

c. The development of several fairly sophisticated "canned" computer programs for carrying out multivariate techniques; and

d. The research questions being asked are becoming more and more complex, and more sophisticated techniques for data analysis are needed.

(3) LIST AND DESCRIBE THE MULTIVARIATE DATA ANALYSIS TECHNIQUES DESCRIBED IN THIS CHAPTER. CITE EXAMPLES FOR WHICH EACH TECHNIQUE IS APPROPRIATE.

Answer

a. DEPENDENCE TECHNIQUES - variables are divided into dependent and independent.

(1) Multiple Regression (MR) - the objective of MR is to predict changes in a single metric dependent variable in response to changes in several metric independent variables. A related technique is multiple correlation.

(2) Multiple Discriminant Analysis (MDA) - the objective of MDA is to predict group membership for a single nonmetric dependent variable using several metric independent variables.

(3) Multivariate Analysis of Variance (MANOVA) simultaneously analyzes the relationship of 2 or more metric dependent variables and several nonmetric independent variables. A related procedure is multivariate analysis of covariance (MANCOVA) which can be used to control factors other than the included independent variables.

(4) Canonical Correlation Analysis (CCA) simultaneously correlates several metric dependent variables and several metric independent variables. Note that this procedure can be considered an extension of MR, where there is only one metric dependent variable.

(5) Conjoint Analysis - used to transform nonmetric scale responses into metric form. It is concerned with the joint effect of two or more nonmetric independent variables on the ordering of a single dependent variable.

(6) Structural Equation Modeling - simultaneously analyzes several dependence relationships (e.g., several regression equations) while also having the ability to account for measurement error in the process of estimating coefficients for each independent variable.

b. INTERDEPENDENCE TECHNIQUES - all variables are analyzed simultaneously, with none being designated as either dependent or independent.

(1) Factor Analysis (FA) - used to analyze the interrelationships among a large number of variables and then explain these variables in terms of their common, underlying dimensions. The two major approaches are component analysis and common factor analysis.

(2) Cluster Analysis - used to classify a sample into several mutually exclusive groups based on similarities and differences among the sample components.

(3) Multidimensional Scaling (MDS) - a technique used to transform similarity scaling into distances in a multidimensional space.

(4) EXPLAIN WHY AND HOW THE VARIOUS MULTIVARIATE METHODS CAN BE VIEWED AS A FAMILY OF TECHNIQUES.

Answer

The multivariate techniques can be viewed as a "family" of techniques in that they are all based upon constructing composite linear relationships among variables or sets of variables. The family members complement one another by accommodating unique combinations of input and output requirements so that an exhaustive array of capabilities can be brought to bear on complex problems.

(5) WHY IS KNOWLEDGE OF MEASUREMENT SCALES IMPORTANT TO AN UNDERSTANDING OF MULTIVARIATE DATA ANALYSIS?

Answer

Knowledge and understanding of measurement scales is a must before the proper multivariate technique can be chosen. Inadequate understanding of the type of data to be used can cause the selection of an improper technique, which makes any results invalid. Measurement scales must be understood so that questionnaires can be properly designed and data adequately analyzed.

(6) WHAT ARE THE DIFFERENCES BETWEEN STATISTICAL AND PRACTICAL SIGNIFICANCE? IS ONE A PREREQUISITE FOR THE OTHER?

Answer

Statistical significance is a means of assessing whether the results are due to change. Practical significance assess whether the result is useful or substantial enough to warrant action. Statistical significance would be a prerequisite of practical significance.

(7) WHAT ARE THE IMPLICATIONS OF LOW STATISTICAL POWER? HOW CAN THE POWER BE IMPROVED IF IT IS DEEMED TOO LOW?

Answer

The implication of low power is that the researcher may fail to find significance when it actually exists. Power may be improved through decreasing the alpha level or increasing the sample size.

(8) DETAIL THE MODEL-BUILDING APPROACH TO MULTIVARIATE ANALYSIS, FOCUSING ON THE MAJOR ISSUES AT EACH STEP.

Answer

Stage One: Define the Research Problem, Objectives, and Multivariate Technique to Be Used The starting point for any analysis is to define the research problem and objectives in conceptual terms before specifying any variables or measures. This will lead to an understanding of the appropriate type of technique, dependence or interdependence, needed to achieve the desired objectives. Then based on the nature of the variables involved a specific technique may be chosen.

Stage Two: Develop the Analysis Plan A plan must be developed that addresses the particular needs of the chosen multivariate technique. These issues include: (1) sample size, (2) type of variables (metric vs. nonmetric, and (3) special characteristics of the technique.

Stage Three: Evaluate the Assumptions Underlying the Multivariate Technique All techniques have underlying assumptions, both conceptual and empirical, that impact their ability to represent multivariate assumptions. Techniques based on statistical inference must meet the assumptions of multivariate normality, linearity, independence of error terms, and equality of variances. Each technique must be considered individually for meeting these and other assumptions.

Stage Four: Estimate the Multivariate Model and Assess Overall Model Fit With assumptions met, a model is estimated considering the specific characteristics of the data. After the model is estimated, the overall model fit is evaluated to determine whether it achieves acceptable levels of statistical criteria, identifies proposed relationships, and achieves practical significance. At this stage the influence of outlier observations is also assessed.

Stage Five: Interpret the Variate With acceptable model fit, interpretation of the model reveals the nature of the multivariate relationship.

Stage Six: Validate the Multivariate Model The attempts to validate the model are directed toward demonstrating the generalizability of the results. Each technique has its own ways of validating the model.

CHAPTER TWO
EXAMINING YOUR DATA

Similar to the style of Chapter One, this presentation will address the basic questions of "Why?," "What?," "When?," and "How?" as applied to examining your data prior to the application of a multivariate technique.

Why examine data?

Data examination is needed for several reasons:

1. To ***gain a basic understanding of the data set***, including information about the relationships among the variables. The two approaches are:

- **Case-by-case evaluation** — although necessary in the examination of response bias, this approach is time consuming and does not enable the researcher to get the "big picture."

- **Compilation of cases** — this preferred method provides a more meaningful interpretation of the cases. Descriptive statistics, or data examination, provide the analyst with a means to present data descriptors in a manageable form.

Examining a compilation of cases reduces individual observations to easily interpretable summaries. In addition, variable associations, or relationships, can be calculated from the raw data and represented simply in reduced form.

2. To ensure that the ***statistical and theoretical underpinnings*** of the chosen multivariate technique are upheld.

Data examination enables the researcher to analyze the multivariate assumptions of normality, homoscedasticity, linearity, and independence of error terms. Each multivariate technique has underlying assumptions which will be highlighted in the following chapters.

3. To analyze the impacts of uncertainties inherent in data collection, including controllable and uncontrollable factors which may influence the data set.

- **Controllable factors** — controlled by the researcher or analyst, such as the input of data. No matter how carefully the data is input, some errors will occur. For example, errors may result from incorrect coding or the misinterpretation of codes. Data examination provides the analyst an overview of the data, which will call attention to any impossible or improbable values which require further attention.

- **Uncontrollable factors** — characteristic of the respondent or the data collection instrument, may also be detected via data examination. For example, cases with a large number of missing values may be identified. In addition, outliers, or extreme cases, are designated in data examination techniques.

What is involved in examining data?

Data examination techniques vary from a simple visual examination of graphical representations to complex statistical analyses which address missing data problems and the assumptions underlying the multivariate technique. This chapter provides a detailed description of data examination in four phases:

- graphical representation analysis,

- evaluating missing data,

- identifying outliers, and

- assessing assumptions.

When do you examine your data?

Essentially, an analyst should examine every new data set and should re-examine any dataset being used for a new multivariate application. In fact, data examination is a necessary first step in any multivariate application. Not only does examination provide the analyst with a test of the underlying assumptions of the multivariate technique, but it also gives the analyst a better understanding of the nature of the data set.

Many techniques are available for examining data sets. Most statistical software packages offer techniques for the evaluation of data. Many packages refer to data examination as descriptive statistics. In addition to computer packages, data examination may also be computed by hand; however, the process is tedious and is not recommended given the computing power available.

How do you examine your data?

As outlined in the chapter, there are four phases of data examination:

- **graphical examination** of the variables in the analysis,
- evaluation of the possible causes and remedies for **missing dat**a in the variables in the analysis,
- identification of **outliers,** and
- assessment of the ability of the data to meet the **statistical assumptions** specific to the selected multivariate technique.

Phase 1: Graphical Examination of the Data

1. The nature of the variable can be evaluated by examining the shape of the distribution.

- ***Histogram*** — the most common form of graphical representation of the data. It displays the frequency of occurrences of the data values (X axis) with the data categories (Y axis). Histograms can be used to examine any type of variable.

- ***Stem and leaf diagram*** — similar to histograms, graphically displays the data distribution by frequencies and data categories, but also includes the actual data values. The stem is the root value to which each leaf is added to derive the actual data value.

2. Relationships between two or more variables may be examined by graphical plots.

- ***Scatterplot*** — the most common form of graphical display for examining the bivariate relationships among variables. The scatterplot is a graph of data points, where the horizontal axis is one variable and the vertical axis is another variable. The variable observations can be many values, including actual values, expected values, and residuals. The patterns of the data points represent the relationship between the two variables (i.e. linear, curvilinear, etc...).

- ***Scatterplot matrices*** — scatterplots computed for all combinations of variables. The diagonal of the matrix contains the histograms for each variable.

3. Testing for group differences requires examination of 1) how the values are distributed for each group, 2) if outliers are present in the groups, and 3) whether or not the groups are different from one another.

- ***Box plot*** — a pictorial representation of the data distribution of each group. Each group is represented by a box, with the upper and lower boundaries of the box marking the upper and lower quartiles of the data distribution.

    - Box length is the distance between the 25% percentile and the 75% percentile, such that the box contains the middle 50% of the values. The asterisk inside the box identifies the median.

    - Lines or whiskers extending from each box represent the distance to the smallest and the largest observations that are less than one quartile range from the box (also marked by an X).

    - Outliers (marked O) are observations which range between 1.0 and 1.5 quartiles away from the box.

    - Extreme values are marked E and represent those observations which are greater than 1.5 quartiles away from the end of the box.

4. When the analyst wishes to graphically examine more than two variables, one of three types of multivariate profiles is appropriate.

- ***Glyphs or Metroglyphs***: some form of circle with radii that correspond to a data value or a multivariate profile which portrays a bar-like profile for each observation.

- ***Mathematical transformation****:* transformation of the original data into a mathematical relationship which can be displayed graphically.

- ***Iconic representation****:* pictorially represents each variable as a component of a whole picture. This most common form is a face, with each variable representing a different feature.

Phase 2: Evaluating Missing Data

1. Missing data can produce hidden biases in the analysis results and can also adversely affect the sample size available for analysis.

- Without remedy, any observation with missing data on any of the variables will be excluded from the analysis.

- Exclusion of too many variables due to missing data can substantially affect the sample size. We know that sample size will impact the power of any statistical tests and affect whether or not the results achieve statistical significance.

2. The analyst must identify the missing data process (reasons underlying missing data) before he or she can select a remedy, or appropriate course of action.

3. A missing data process may be of two kinds: a systematic event external to the respondent (ignorable) or any action on the part of the respondent which leads to missing values (unidentifiable).

- ***Ignorable missing data*** — When the missing data process is known and is external to the respondent, and it can be accommodated in the research plan. Specific remedies are not needed since the allowances are inherent in the technique used.

    - Ignorable missing data operate at random; the observed values are a random sample of the total set of values, observed and missing.

    - Examples of ignorable missing data:

        - observations in a population which are not included in the sample, or

        - censored data (observations which are not complete because of their stage in the missing data process).

- ***Unidentifiable*** — When the missing data are due to an action of the respondent, they are often unidentifiable and cannot be accommodated in the research design. In this case, the researcher evaluates the pattern of the missing data and determines the potential for remedy.

4. Assessing the degree of randomness will identify one of two types: missing at random (MAR) and missing completely at random (MCAR).

- ***Missing at random (MAR)***: When the missing values of Y depend on X, but not on Y. This occurs when X biases the randomness of the observed Y values, such that the observed Y values do not represent a true random sample of all actual Y values in the population.

- ***Missing completely at random (MCAR)***: When the observed values of Y are truly a random sample of all Y values.

- ***Approaches for diagnosing the randomness of the missing data process***

    - Significance tests for a single variable: Form two groups, one group being those observations with missing data and another group being those observations with valid values, and test for significant differences between the two groups on any other variables of interest. If significant differences are found, a nonrandom missing data process is present, meaning that the missing data should be classified as MAR.

    - Dichotomized correlations for a pair of variables: For each of the two variables, replace each valid value with a value of one and each missing value with a value of zero, then compute correlations for the missing values of each variable. The correlations indicate the degree of association between the missing data on each variable pair. Low correlations denote randomness in the pair of variables. If all variable pairs have low correlations, the missing data can be classified as MCAR.

    - Overall test of randomness: Analyze the pattern of missing data on all variables and compare it to the pattern expected for a random missing data process. If no significant differences are found, the missing data can be classified as MCAR.

5. Approaches are available for dealing with missing data that are selected based on the randomness of the missing data process.

- ***Use of only observations with complete data***. When conducting analysis, the researcher would include only those observations with complete data.

    - Default in many statistical programs.

    - Used only if the missing data are missing completely at random (MCAR); when used with data which are missing at random (MAR), the results are not generalizable to the population.

- ***Delete case(s) and/or variable(s).*** The researcher would delete the case(s) and/or variable(s) which exceed a specified level from the analysis.

    - Most effective for data which are not missing at random, but is an alternative which can be used if the data are MAR or MCAR.

- ***Imputation methods***. Imputation methods replace missing values with estimates based on the valid values of other variables and / or cases in the sample. Imputation methods should be used only if the data are MCAR.

    - Selecting values or observations to be used in the imputation process.

        - The complete case approach uses only data from observations that have no missing data.

        - The all-available approach uses all available valid observations to estimate missing data, maximizing pairwise information.

    - Five imputation methods are available:
        - Case substitution: observations with missing data are replaced by choosing another nonsampled observation.

        - Mean substitution: missing values for a single variable are replaced with the means value of that variable based on all responses.

        - Cold deck imputation: missing values are replaced with a constant value derived from external sources or previous research.

- Regression imputation: missing values are replaced with predicted estimates from a regression analysis. Estimated values are based on their relationship with other variables in the data set.

- Multiple imputation: a combination of several methods, two or more methods of imputation are used to derive a composite estimate for the missing value.

- ***Model-based procedures.*** Model-based procedures incorporate missing data into the analysis, either through a process specifically designed for missing data estimation or as an integral portion of the standard multivariate analysis.

Phase 3: Identification of Outliers

1. 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 that they may provide regarding the phenomenon under study.

- ***Beneficial outliers*** — when they are indicative of characteristics in the population that would not be discovered in the normal course of analysis.

- ***Problematic outliers*** — when they are not representative of the population and are counter to the objectives of the analysis.

2. Outliers can be classified into four categories.

- ***Outliers arising from a procedural error***. These outliers result from data entry errors or mistakes in coding. They should be identified and eliminated during data cleaning.

- ***Outliers resulting from an extraordinary event with an explanation*** These outliers can be explained. If found to be representative of the population, they should be kept in the data set.

- ***Outliers resulting from an extraordinary event with no explanation***. These outliers cannot be explained. Often, these observations are deleted from the data set.

- ***Ordinary values which become unique when combined with other variables***. While these values cannot be distinguished individually, they become very noticeable when combined with other values across other variables. In these cases, the observations should be retained unless specific evidence to the contrary is found.

3. Identification can be made from any of three perspectives: univariate, bivariate, or multivariate. If possible, multiple perspectives should be utilized to triangulate the identification of outliers.

- ***Univariate detection***: examine the distribution of observations for a variable and select as outliers those values which fall at the outer ranges of the distribution.
    - When standardized, data values which are greater than 2.5 may be potential outliers. (For large sample sizes, the value may increase to 3 or 4.)

- ***Bivariate detection:*** examine scatterplots of variable pairs and select as outliers those values which fall markedly outside the range of the other observations.
    - Ellipse representing a confidence interval may be drawn around the expected range of observations on the scatterplot. Those values falling outside the range are potential outliers.
    - Influence plot, where the point varies in size in proportion to its influence on the relationship and the largest points are potential outliers.

- ***Multivariate detection***: assess each observation across a set of variables and select as outliers those values which fall outside a specified range specific to the statistical test employed.
    - Mahalanobis D^2 is commonly used in multivariate analyses to identify outliers. It is a measure of the distance in multidimensional space of each observation from the mean center of the observations.
    - Conservative values (i.e. .001) for the statistical tests should be set for identification of potential outliers.

4. Only observations which are truly unique from the population should be designated as outliers. The researcher should be careful of identifying too many observations as outliers.

- ***Profiles*** on each outlier should be generated and the data should be examined for the variable(s) responsible for generating the outlier. Multivariate techniques may also be employed to trace the underlying causes of outliers. The researcher should be able to classify the outlier in one of the four categories discussed above.

- ***Unnecessary deletion*** of outliers will limit the generalizability of the analysis. Outliers should be deleted from the analysis only if they are proven to be not representative of the population.

Stage 4: Testing the Assumptions of Multivariate Analysis

1. Multivariate analyses require that the assumptions underlying the statistical techniques be tested twice: once for the individual variables and once for the multivariate model.

The following discussion relates only to assumptions underlying the individual variables. The assumptions for the variate for each technique will be discussed in the appropriate chapter.

2. Normality: Each variable in the analysis must be normally distributed.

- ***Most fundamental assumption*** in multivariate analyses.

- ***Sufficient non-normality invalidates statistical tests*** which use the F and t statistics.

- ***Visual checks***: The simplest way to evaluate the normality of a variable is to visually check a histogram or a normal probability plot.

  - Histogram -- the distribution should approximate a bell-shaped curve.

  - Normal probability plot -- the data points should closely follow the diagonal line.

- ***Statistical tests***: The two most common are the Shapiro-Wilks and Kolmogorov-Smirnov tests.

- ***Transformations***: When a distribution is found to be non-normal, data transformations should be computed.

- ***Skewness***: Skewness values exceeding ±2.58 are indicative of a non-normal distribution. Other statistical tests are available in specific statistical software programs.

3. Homoscedasticity: dependent variables should exhibit equal levels of variance across the range of predictor variables.

- ***Common sources***: Most problems with unequal variances stem from either the type of variables included in the model or from a skewed distribution.

- ***Impact***: Violation of this assumption will cause hypothesis tests to be either too conservative or too sensitive.

- ***Identification:*** graphical versus statistical.

    - Graphical plot of residuals will reveal violations of this assumption.

    - Statistical tests for equal variance dispersion relate to the variances within groups formed by nonmetric variables. The most common test is the **Levene test**, which is used to assess if the variances of a single metric variable are equal across any number of groups. When more than one variable is being tested, the **Box's M** test should be used.

- ***Remedies***: Heteroscedastic variables can be remedied through data transformations.

4. Linearity: variables should be linearly related.

- ***Identification***: Scatterplots of variable pairs are most commonly used to identify departures from linearity. Examination of the residuals in a simple regression analysis may also be used as a diagnostic method.

- ***Nonlinearity***: If a nonlinear relationship is detected, the most direct approach is to transform one or both of the variables. Other than a transformation, a new variable which represents the nonlinear relationship can be created.

5. Prediction errors should not be correlated.

- ***Patterns in the error terms*** reflect an underlying systematic bias in the relationship.

- ***Residual plots*** should not contain any recognizable pattern.

- ***Violations of this assumption*** often result from problems in the data collection process.

6. Data transformations enable the researcher to modify variables to correct violations of the assumptions of normality, homoscedasticity, and linearity and to improve the relationships between variables.

- ***Basis***: Transformations can be based on **theoretical or empirical** reasons.

- ***Distribution shape***: The shape of the distribution provides the basis for selecting the appropriate transformation.

  - Flat distribution is the most common transformation is the inverse.

  - Positively skewed distributions transformed by taking logarithms

  - Negatively skewed distributions transformed by taking the square root.

  - Cone-shaped distribution which opens to the right should be transformed using an inverse. A cone shaped distribution which opens to the left should be transformed by taking the square root.

  - Nonlinear transformations can take many forms, including squaring the variable and adding additional variables termed polynomials.

- ***General guidelines*** for performing data transformations.

  - Ratio of a variable's mean divided by its standard deviation should be less than 4.0.

  - Select the variable with the smallest ratio from item 1.

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

- Heteroscedasticity can only be remedied by transformation of the dependent variable in a dependence relationship. If a heteroscedasticity relationship is also nonlinear, the dependent and perhaps the independent variables must be transformed.

- Transformations may change the interpretation of the variables.

Incorporating Nonmetric Data with Dummy Variables

When faced with nonmetric variables in the data the researcher may wish to represent these categorical variables as metric through the use of dummy variables. Any nonmetric variable with *k* groups may be represented as *k* - 1 dummy variables. There are two general methods of accomplishing this task.

- ***Indicator coding*** assigns a value of 1 to one group, for instance females, and zero to the comparison group (males).

- ***Effects coding*** assign a value of -1 to the comparison group while still using 1 to designate the other group.

ANSWERS TO END-OF-CHAPTER QUESTIONS

(1) LIST POTENTIAL UNDERLYING CAUSES OF OUTLIERS. BE SURE TO INCLUDE ATTRIBUTIONS TO BOTH THE RESPONDENT AND THE RESEARCHER.

Answer

a. Respondent:
   1) Misunderstanding of the question
   2) Response bias, such as yea-saying
   3) Extraordinary experience

b. Researcher:
   1) Data entry errors
   2) Data coding mistakes

c. An extraordinary observation with no explanation.

d. An ordinary value which is unique when combined with other variables.

(2) DISCUSS WHY OUTLIERS MIGHT BE CLASSIFIED AS BENEFICIAL AND AS PROBLEMATIC.

Answer

a. Beneficial outliers are indicative of some characteristic of the population which would not have been otherwise known. For example, if only one respondent from a lower income group is included in the sample and that respondent expresses an attitude atypical to the remainder of the sample, this respondent would be considered beneficial.

b. Problematic outliers are not indicative of the population and distort multivariate analyses. Problematic outliers may be the result of data input errors, a respondent's misunderstanding of the question, or response bias. These extreme responses must be evaluated as to the type of influence exerted and dealt with accordingly.

(3) DISTINGUISH BETWEEN DATA WHICH ARE MISSING AT RANDOM (MAR) AND MISSING COMPLETELY AT RANDOM (MCAR). EXPLAIN HOW EACH TYPE WILL IMPACT THE ANALYSIS OF MISSING DATA.

Answer

a. Missing at Random (MAR): If the missing values of Y depend on X, but not on Y, the missing data are at random. This occurs when X biases the randomness of the observed Y values, such that the observed Y values do not represent a true random sample of all actual Y values in the population.

b. Missing Completely at Random (MCAR): When the observed values of Y are truly a random sample of all Y values.

c. When the missing data are missing at random (MAR), the analyst should only use a modeling-based approach which accounts for the underlying processes of the missing data. When the missing data are missing completely at random (MCAR), the analyst may use any of the suggested approaches for dealing with missing data, such as using only observations with complete data, deleting case(s) or variable(s), or employing an imputation method.

(4) DESCRIBE THE CONDITIONS UNDER WHICH A RESEARCHER WOULD DELETE A CASE WITH MISSING DATA VERSUS THE CONDITIONS UNDER WHICH A RESEARCHER WOULD USE AN IMPUTATION METHOD.

Answer

The researcher must first evaluate the randomness of the missing data process. If the data are missing at random, deleting a case is the only acceptable alternative of the two. Data which are missing at random cannot employ an imputation method, as it would introduce bias into the results. Only cases with data which are missing completely at random would utilize an imputation method.

If the data are missing completely at random, the choice of case deletion versus imputation method should be based on theoretical and empirical considerations.

If the sample size is sufficiently large, the analyst may wish to consider deletion of cases with a great degree of missing data. Cases with missing data are good candidates for deletion if they represent a small subset of the sample and if their absence does not otherwise distort the data set.

For instance, cases with missing dependent variable values are often deleted.

If the sample size is small, the analyst may wish to use an imputation method to fill in missing data. The analyst should, however, consider the amount of missing data when selecting this option. The degree of missing data will influence the researcher's choice of information used in the imputation (i.e. complete case vs. all-available approaches) and the researcher's choice of imputation method (i.e. case substitution, mean substitution, cold deck imputation, regression imputation, or multiple imputation).

(5) EVALUATE THE FOLLOWING STATEMENT, "IN ORDER TO RUN MOST MULTIVARIATE ANALYSES, IT IS NOT NECESSARY TO MEET ALL OF THE ASSUMPTIONS OF NORMALITY, LINEARITY, HOMOSCEDASTICITY, AND INDEPENDENCE."

Answer

As will be shown in each of the following chapter outlines, each multivariate technique has a set of underlying assumptions which must be met. The degree to which a violation of any of the four above assumptions will distort data analyses is dependent on the specific multivariate technique. For example, multiple regression analysis is sensitive to violations of all four of the assumptions, whereas multiple discriminant analysis is primarily sensitive to violations of multivariate normality.

(6) DISCUSS THE FOLLOWING STATEMENT, "MULTIVARIATE ANALYSES CAN BE RUN ON ANY DATA SET, AS LONG AS THE SAMPLE SIZE IS ADEQUATE."

Answer

False. Although sample size is an important consideration in multivariate analyses, it is not the only consideration. Analysts must also consider the degree of missing data present in the data set and examine the variables for violations of the assumptions of the intended techniques.