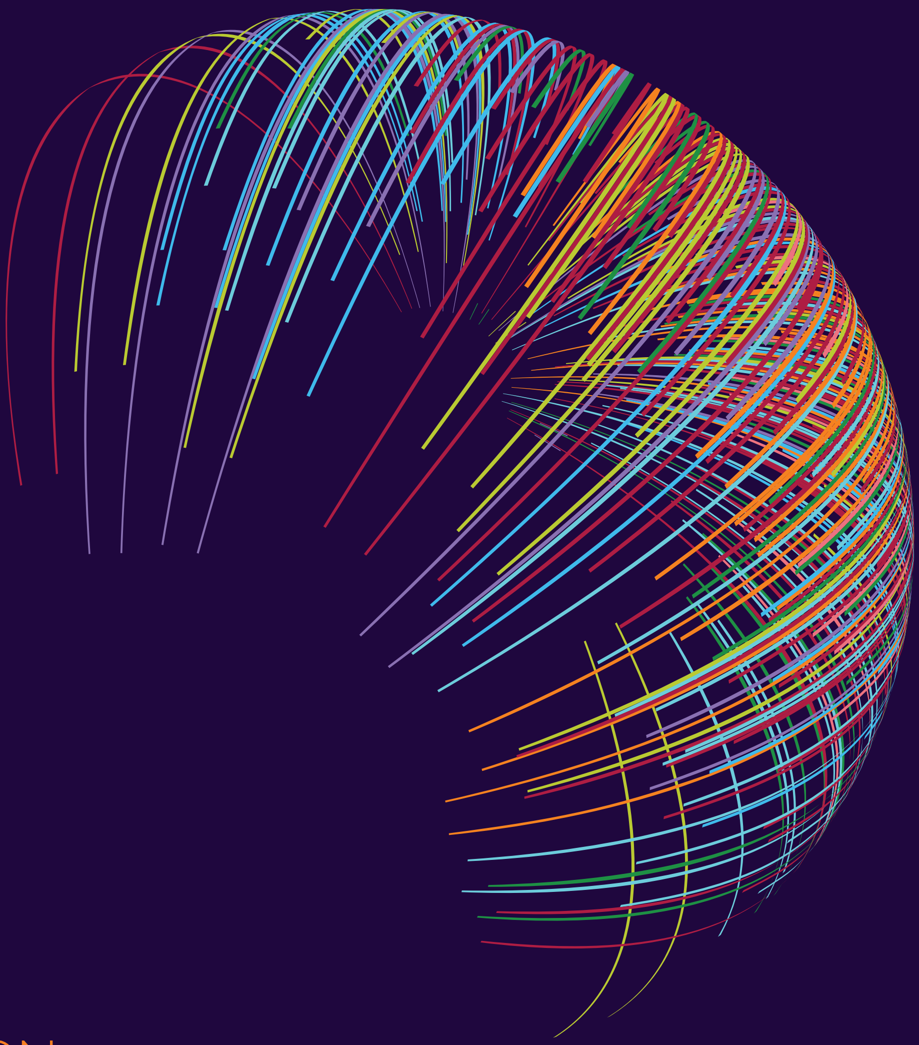
CENGAGE



EIGHTH EDITION

MULTIVARIATE DATA ANALYSIS

Joseph F. Hair Jr., William C. Black,
Barry J. Babin, Rolph E. Anderson

Multivariate Data Analysis

EIGHTH EDITION

Joseph F. Hair, Jr.

University of South Alabama

William C. Black

Louisiana State University

Barry J. Babin

Louisiana Tech University

Rolph E. Anderson

Drexel University

CENGAGE

Australia • Brazil • Mexico • South Africa • Singapore • United Kingdom • United States

CENGAGE

Multivariate Data Analysis,
Eighth Edition
Joseph F. Hair Jr, William C. Black,
Barry J. Babin, Rolph E. Anderson

Publisher: Annabel Ainscow

List Manager: Jenny Grene

Marketing Manager: Sophie Clarke

Content Project Manager: Melissa Beavis

Manufacturing Manager: Eyvett Davis

Typesetter: SPi Global

Text Design: SPi Global

Cover Designer: Simon Levy Associates

Cover Images: Istockphoto/liuzishan

For product information and technology assistance, contact us at **emea.info@cengage.com**

For permission to use material from this text or product and for permission queries, email **emea.permissions@cengage.com**

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library. This book was previously published by Pearson Education, Inc.

ISBN: 978-1-4737-5654-0

Cengage Learning, EMEA
Cheriton House, North Way
Andover, Hampshire, SP10 5BE
United Kingdom

Cengage Learning is a leading provider of customized learning solutions with employees residing in nearly 40 different countries and sales in more than 125 countries around the world. Find your local representative at: **www.cengage.co.uk.**

Cengage Learning products are represented in Canada by Nelson Education, Ltd.

For your course and learning solutions, visit **www.cengage.co.uk**

Purchase any of our products at your local college store or at our preferred online store **www.cengagebrain.com.**

Printed in China by RR Donnelley
Print Number: 01 Print Year: 2018

To my family, and particularly my wife Dale

—Joseph F. Hair, Jr., Mobile, Alabama

To Deb, Steve, Emily and especially Alden, my granddaughter,
for their love and support in this new stage of my career

—William C. Black, Austin, TX

For Laurie, Amie, and James, and my mother Barbara

—Barry J. Babin, Choudrant, LA

To Rachel and Stuart for their unfaltering love and support

—Rolph E. Anderson, Philadelphia, PA

Brief contents

Contents

SECTION IV
Dependence Techniques – Non-metric Outcomes 469

7 Multiple Discriminant Analysis 471

Preface

In more than four decades since the first edition of *Multivariate Data Analysis*, the fields of multivariate statistics, and analytics in general, have evolved dramatically in several different directions for both academic and applied researchers. In the methodological domain, we have seen a continued evolution of the more "traditional" statistical methods such as multiple regression, ANOVA/MANOVA and exploratory factor analysis. These methods have been extended not only in their capabilities (e.g., additional options for variable selection, measures of variable importance, etc.), but also in their application (e.g., multi-level models and causal inference techniques). These "traditional" methods have been augmented by a new generation of techniques represented by structural equation modeling and partial least squares. These methods integrate many of the former methods (e.g., multiple regression and exploratory factor analysis) into new analytical techniques to perform confirmatory factor analysis and structural model estimation. But perhaps most exciting has been the integration of methods from the fields of data mining, machine learning and neural networks. These fields of study have remained separate for too long, representing different "cultures" in terms of approaches to data analysis. But as we discuss in Chapter 1 and throughout the text, these two fields provide complementary approaches, each of which has advantages and disadvantages. We hope that by acknowledging these complementarities we can in some small way increase the rate of integration between the two fields.

The development of these analytical methods has also been greatly facilitated by the tremendous increase in computing power available in so many formats and platforms. Today the processing power is essentially unlimited as larger and larger types of problems are being tackled. The availability of these techniques has also been expanded not only through the continued development of the traditional software packages such as SAS and it's counterpart JMP, IBM SPSS and STATA, as well as SmartPLS for PLS-SEM, but also the recent wide-spread use of free, open-source, software, typified by the R-project, which has been around as far back as 1992, with roots at Bell Labs previous to that time. Today researchers have at their disposal the widest range of software alternatives ever available.

But perhaps the most interesting and exciting development has occurred in the past decade with the emergence of "Big Data" and the acceptance of data-driven decisionmaking. Big Data has revolutionized the type and scope of analyses that are now being performed into topics and areas never before imagined. The widespread availability of both consumer-level, firm-level and event-level data has empowered researchers in both the academic and applied domains to address questions that only a few short years ago were not even conceptualized. An accompanying trend has been the acceptance of analytical approaches to decisionmaking at all levels. In some instances researchers had little choice since the speed and scope of the activities (e.g., many digital and ecommerce decisions) required an automated solution. But in other areas the widespread availability of heretofore unavailable data sources and unlimited processing capacity quickly made the analytical option the first choice.

The first seven editions of this text and this latest edition have all attempted to reflect these changes within the analytics landscape. As with our prior editions we still focus in the traditional statistical methods with an emphasis on design, estimation and interpretation. We continually strive to reduce our reliance on statistical notation and terminology and instead to identify the fundamental concepts which affect application of these techniques and then express them in simple terms—the result being an applications-oriented introduction to multivariate analysis for the non-statistician. Our commitment remains to provide a firm understanding of the statistical and managerial principles underlying multivariate analysis so as to develop a "comfort zone" not only for the statistical but also the practical issues involved.

New Features

But the emergence of "Big Data, " increased usage of data mining and machine learning techniques, and the acceptance of data-driven decision-making, has motivated us to try and provide a broader perspective on analytics than in the past. Our goal is to recognize these three emerging trends and address how they impact the analytical domain for both academicians and applied researchers. The eighth edition of Multivariate Data Analysis provides an updated perspective on data analysis of all types of data as well as introducing some new perspectives and techniques that are foundational in today's world of analytics:

- New chapter on partial least squares structural equation modeling (PLS-SEM), an emerging technique with equal applicability for researchers in the academic and organizational domains.
- Integration of the implications of Big Data into each of the chapters, providing some understanding of the role of multivariate data analysis in this new era of analytics.
- Extended discussions of emerging topics, including causal treatments/inference (i.e., causal analysis of non-experimental data as well as discussion of propensity score models) along with multi-level and panel data models (extending regression into new research areas and providing a framework for cross-sectional/time-series analysis).
- Updates in each chapter of technical improvements (e.g., multiple imputation for missing data treatments) as well as the merging of basic principles from the fields of data mining and its applications.
- In addition to the new PLS-SEM chapter, the chapters on SEM have greater emphasis on psychometrics and scale development, updated discussions on the use of reflective versus formative scaling, describe an alternative approach for handing interactions (orthogonal moderators), as well as in-depth discussion of higher-order CFA models, multi-group analyses, an introduction to Bayesian SEM, and an updated discussion of software availability. The multi-group discussion also includes an alternative to partial metric invariance when cross-group variance problems are small. Additionally, an added set of variables is included with HBATSEM as a way of illustrating diagnostic issues arising when the psychometric properties of a scale do not adhere to the rules of thumb for measuring a latent factor. The expanded data set is available as HBATSEM6CON.
- Online resources for researchers including continued coverage from past editions of all of the analyses from the latest versions of SAS, SPSS and SmartPLS (commands and outputs)

With the addition of the new chapter on PLS-SEM and other new content in each chapter, several chapters from prior editions were omitted from the current edition, but are still available on the website (www.mvstats.com). The chapters for conjoint analysis, multidimensional scaling and correspondence analysis have been formatted in "publication ready" formats and are available to all adopters as well as interested researchers on the website. Special thanks are due to Pei-ju Lucy Ting and Hsin-Ju Stephanie Tsai, both from University of Manchester, for their work on revising the chapter on canonical correlation analysis in our prior edition. They updated this chapter with an example using the HBAT database, added recently published material, and reorganized it to facilitate understanding. This is one of the chapters now available on our Web site for those who wish to learn more about this technique.

Each of these changes, and others not mentioned, will assist readers in gaining a more thorough understanding of both the statistical and applied issues underlying these techniques.

PEDAGOGY

Almost all statistics texts include formulas and require knowledge of calculus or matrix algebra. A very important question is "Can students comprehend what they are reading and apply it?" This book offers a wealth of pedagogical features, all aimed at answering this question positively. Presented here is a list of the major elements:

Learning Objectives. Each chapter begins with clear learning objectives that students can use to assess their expectations for the chapter in view of the nature and importance of the chapter material.

Key Terms and Concepts. These are bold-faced in the text and are listed at the beginning of the chapters to facilitate comprehension.

Chapter Summaries. These detailed summaries are organized by the learning objectives presented at the beginning of the chapter. This approach to organizing summaries helps students to remember the key facts, concepts, and issues. They also serve as an excellent study guide to prepare for in-class exercises or exams.

Questions for Review and Discussion. The review and discussion questions are carefully designed to enhance the self-learning process and to encourage application of the concepts learned in the chapter. There are six or seven questions in each chapter designed to provide students with opportunities to enhance their comprehension of the concepts.

Suggested Readings. A list of the most relevant additional readings is provided at the end of the chapter. These readings enable you to review many of the sources of the information summarized in the chapter as well as extend your knowledge to other more detailed information.

HBAT Database. The HBAT database is a continuing case of a business scenario embedded throughout the book for the purpose of illustrating the various statistical concepts and methods. The case is introduced in Chapter 1, and in each subsequent chapter it builds upon the previously learned concepts. A single research situation is used to illustrate various aspects of the process of applying each of the multivariate methods. The HBATSEM data is enhanced for the 8th edition.

Software commands/syntax. An expanded feature from prior editions are the software-specific resources to enable a researcher to replicate the analyses in the text in SAS, IBM SPSS, and SmartPLS for that chapter. In addition to these software commands, actual outputs for all of the analyses in the text are available for the student to examine and even compare to their results. The authors are also committed to continue development of these resources, hopefully extending theses supplements to include selected R code and outputs in the near future.

ONLINE RESOURCES

The book provides an extensive and rich ancillary package. The following is a brief description of each element in the package. These materials are available via Cengage Brain and the text's dedicated website (www.mvstats.com).

Instructor's Resources. PowerPoint slide presentations provide an easy transition for instructors teaching with the book the first time. For those who have used previous editions, there are many new support materials to build upon the notes and teaching enhancement materials available previously. A wealth of extra student projects and examples are available as additional classroom resources. All of these materials are available via Cengage Brain.

Website. Students can access the book's dedicated website (www.mvstats.com) for additional information about the various statistical methods and to evaluate their understanding of chapter material. Additional resources are offered for each chapter—look for prompts in the book that will guide you to the website for more useful information on various topics.

Data Sets. Data sets in SPSS format are available at the book's website (www.mvstats.com). The two primary data sets are the HBAT customer survey data set used in the first 8 chapters of the book, and the HBAT employee survey data set used in Chapters 9 to 13. These data sets can be used to illustrate the examples in the text as well as assign application exercises that go beyond the book examples.

Acknowledgments

We would like to acknowledge the comments and suggestions by Dr. Stephen Vaisey of Duke University, Dr. Helmut Schneider of Louisiana State University, and Dr. Marko Sarstedt, Otto-von-Guericke-University, Germany, on this edition. We would also like to acknowledge the assistance of the following individuals on prior editions of the text: Bruce Alford, Louisiana Tech University; David Andrus, Kansas State University; Jill Attaway, Illinois State University; David Booth, Kent State University; Jim Boles, University of North Carolina-Greensboro; Alvin C. Burns, Louisiana State University; Alan J. Bush, University of Memphis; Robert Bush, Louisiana State University at Alexandria; Rabikar Chatterjee, University of Michigan; Kerri Curtis, Golden Gate University; Chaim Ehrman, University of Illinois at Chicago; Joel Evans, Hofstra University; Thomas L. Gillpatrick, Portland State University; Andreas Herrman, University of St. Gallen; Dipak Jain, Northwestern University; Stavros Kalafatis, Kingston University; John Lastovicka, University of Kansas; Margaret Liebman, La Salle University; Arthur Money, Henley Management College; Peter McGoldrick, University of Manchester; Richard Netemeyer, University of Virginia; Ossi Pesamaa, Jonkoping University, Robert Peterson, University of Texas; Torsten Pieper, Kennesaw State University; Scott Roach, Northeast Louisiana University; Phillip Samouel, Kingston University; Marcus Schmidt, Copenhagen Business School; Muzaffar Shaikh, Florida Institute of Technology; Dan Sherrell, University of Memphis; Walter A. Smith, Tulsa University; Goren Svensson, University of Oslo; Ronald D. Taylor, Mississippi State University; Lucy Ting, University of Manchester; Arch Woodside, Boston College; and Jerry L. Wall, University of Louisiana-Monroe. Finally, we could not have written this or previous editions without the extensive feedback and input from our many graduate students who provided invaluable guidance on the content and direction of the book.

J.F.H.
W.C.B.
B.J.B.
R.E.A.

CENGAGE

Teaching & Learning Support Resources

Cengage's peer reviewed content for higher and further education courses is accompanied by a range of digital teaching and learning support resources. The resources are carefully tailored to the specific needs of the instructor, student and the course. Examples of the kind of resources provided include:

- A password protected area for instructors with, for example, a testbank, PowerPoint slides and an instructor's manual.
- An open-access area for students including, for example, useful weblinks and glossary terms.

Lecturers: to discover the dedicated lecturer digital support resources accompanying this textbook please register here for access: login.cengage.com.

Students: to discover the dedicated student digital support resources accompanying this textbook, please search for **Multivariate Data Analysis** on: cengagebrain.co.uk.

BE UNSTOPPABLE

Learn more at cengage.co.uk/education

1 Overview of Multivariate Methods

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

Explain what multivariate analysis is and when its application is appropriate.

Discuss the implications of Big Data, the emergence of algorithmic models and causal inference on multivariate analysis.

Discuss the nature of measurement scales and their relationship to multivariate techniques.

Understand the nature of measurement error and its impact on multivariate analysis.

Examine the researcher options for managing the variate and dependence models.

Understand the concept of statistical power and the options available to the researcher.

Determine which multivariate technique is appropriate for a specific research problem.

Define the specific techniques included in multivariate analysis.

Discuss the guidelines for application and interpretation of multivariate analyses.

Understand the six-step approach to multivariate model building.

Chapter Preview

Chapter 1 presents a simplified overview of multivariate analysis. It stresses that multivariate analysis methods will increasingly influence not only the analytical aspects of research but also the design and approach to data collection for decision making and problem solving. Although multivariate techniques share many characteristics with their univariate and bivariate counterparts, several key differences arise in the transition to a multivariate analysis. To illustrate this transition, Chapter 1 presents a classification of multivariate techniques. It then provides general guidelines for the application of these techniques as well as a structured approach to the formulation, estimation, and interpretation of multivariate results. The chapter concludes with a discussion of the databases utilized throughout the text to illustrate application of the techniques.

Key Terms

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

Algorithmic models See *data mining models*.

Alpha (α) See *Type I error*.

Beta (β) See *Type II error*.

Big Data The explosion in secondary data typified by increases in the volume, variety and velocity of the data being made available from a myriad set of sources (e.g., social media, customer-level data, sensor data, etc.).

Bivariate partial correlation Simple (two-variable) correlation between two sets of residuals (unexplained variances) that remain after the association of other independent variables is removed.

Bootstrapping An approach to validating a multivariate model by drawing a large number of subsamples and estimating models for each subsample. Estimates from all the subsamples are then combined, providing not only the "best" estimated coefficients (e.g., means of each estimated coefficient across all the subsample models), but their expected variability and thus their likelihood of differing from zero; that is, are the estimated coefficients statistically different from zero or not? This approach does not rely on statistical assumptions about the population to assess statistical significance, but instead makes its assessment based solely on the sample data.

Causal inference Methods that move beyond statistical inference to the stronger statement of "cause and effect" in non-experimental situations.

Composite measure Fundamental element of *multivariate measurement* by the combination of two or more *indicators*. See *summated scales*.

Cross-validation Method of validation where the original sample is divided into a number of smaller sub-samples (*validation samples*) and that the validation fit is the "average" fit across all of the sub-samples.

Data mining models Models based on algorithms (e.g., neural networks, decision trees, support vector machine) that are widely used in many *Big Data* applications. Their emphasis is on predictive accuracy rather than statistical inference and explanation as seen in *statistical/data models* such as multiple regression.

Data models See *statistical models*.

Dependence technique Classification of statistical techniques distinguished by having a variable or set of variables identified as the *dependent variable(s)* and the remaining variables as *independent*. The objective is prediction of the dependent variable(s) by the independent variable(s). An example is regression analysis.

Dependent variable Presumed effect of, or response to, a change in the *independent variable(s)*.

Dimensional reduction The reduction of *multicollinearity* among variables by forming *composite measures* of multicollinear variables through such methods as exploratory factor analysis.

Directed acyclic graph (DAG) Graphical portrayal of causal relationships used in causal inference analysis to identify all "threats" to causal inference. Similar in some ways to path diagrams used in structural equation modeling.

Dummy variable *Nonmetrically* measured variable transformed into a *metric* variable by assigning a 1 or a 0 to a subject, depending on whether it possesses a particular characteristic.

Effect size Estimate of the degree to which the phenomenon being studied (e.g., correlation or difference in means) exists in the population.

Estimation sample Portion of original sample used for model estimation in conjunction with *validation sample*

General linear model (GLM) Fundamental linear dependence model which can be used to estimate many model types (e.g., multiple regression, ANONA/MANOVA, discriminant analysis) with the assumption of a normally distributed dependent measure.

Generalized linear model (GLZ or GLIM) Similar in form to the *general linear model*, but able to accommodate non-normal *dependent* measures such as binary variables (logistic regression model). Uses maximum likelihood estimation rather than ordinary least squares.

Holdout sample See *validation sample*.

Independent variable Presumed cause of any change in the *dependent variable*.

Indicator Single variable used in conjunction with one or more other variables to form a *composite measure*.

Interdependence technique Classification of statistical techniques in which the variables are not divided into *dependent* and *independent* sets; rather, all variables are analyzed as a single set (e.g., exploratory factor analysis).

Measurement error Inaccuracies of measuring the "true" variable values due to the fallibility of the measurement instrument (i.e., inappropriate response scales), data entry errors, or respondent errors.

Metric data Also called quantitative data, interval data, or ratio data, these measurements identify or describe subjects (or objects) not only on the possession of an attribute but also by the amount or degree to which the subject may be characterized by the attribute. For example, a person's age and weight are metric data.

Multicollinearity Extent to which a variable can be explained by the other variables in the analysis. As multicollinearity increases, it complicates the interpretation of the *variate* because it is more difficult to ascertain the effect of any single variable, owing to their interrelationships.

Multivariate analysis Analysis of multiple variables in a single relationship or set of relationships.

Multivariate measurement Use of two or more variables as *indicators* of a single *composite measure*. For example, a personality test may provide the answers to a series of individual questions (indicators), which are then combined to form a single score (*summated scale*) representing the personality trait.

Nonmetric data Also called qualitative data, these are attributes, characteristics, or categorical properties that identify or describe a subject or object. They differ from *metric data* by indicating the presence of an attribute, but not the amount. Examples are occupation (physician, attorney, professor) or buyer status (buyer, non-buyer). Also called nominal data or ordinal data.

Overfitting Estimation of model parameters that over-represent the characteristics of the sample at the expense of generalizability to the population at large.

Power Probability of correctly rejecting the null hypothesis when it is false; that is, correctly finding a hypothesized relationship when it exists. Determined as a function of (1) the statistical significance level set by the researcher for a *Type I error* (α), (2) the sample size used in the analysis, and (3) the *effect size* being examined.

Practical significance Means of assessing multivariate analysis results based on their substantive findings rather than their statistical significance. Whereas statistical significance determines whether the result is attributable to chance, practical significance assesses whether the result is useful (i.e., substantial enough to warrant action) in achieving the research objectives.

Reliability Extent to which a variable or set of variables is consistent in what it is intended to measure. If multiple measurements are taken, the reliable measures will all be consistent in their values. It differs from *validity* in that it relates not to what should be measured, but instead to how it is measured.

Specification error Omitting a key variable from the analysis, thus affecting the estimated effects of included variables.

Statistical models The form of analysis where a specific model is proposed (e.g., *dependent* and *independent* variables to be analyzed by the *general linear model*), the model is then estimated and a statistical inference is made as to its generalizability to the population through statistical tests. Operates in opposite fashion from *data mining models* which generally have little model specification and no statistical inference.

Summated scales Method of combining several variables that measure the same concept into a single variable in an attempt to increase the *reliability* of the measurement through *multivariate measurement*. In most instances, the separate variables are summed and then their total or average score is used in the analysis.

Treatment Independent variable the researcher manipulates to see the effect (if any) on the dependent variable(s), such as in an experiment (e.g., testing the appeal of color versus black-and-white advertisements).

Type I error Probability of incorrectly rejecting the null hypothesis—in most cases, it means saying a difference or correlation exists when it actually does not. Also termed *alpha* (α). Typical levels are five or one percent, termed the .05 or .01 level, respectively.

Type II error Probability of incorrectly failing to reject the null hypothesis—in simple terms, the chance of not finding a correlation or mean difference when it does exist. Also termed *beta* (β), it is inversely related to *Type I error*. The value of 1 minus the Type II error $(1 - \beta)$ is defined as *power*.

Univariate analysis of variance (ANOVA) Statistical technique used to determine, on the basis of one dependent measure, whether samples are from populations with equal means.

Validation sample Portion of the sample "held out" from estimation and then used for an independent assessment of model fit on data that was not used in estimation.

Validity Extent to which a measure or set of measures correctly represents the concept of study—the degree to which it is free from any systematic or nonrandom error. Validity is concerned with how well the concept is defined by the measure(s), whereas *reliability* relates to the consistency of the measure(s).

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

What Is Multivariate Analysis?

Today businesses must be more profitable, react quicker, and offer higher-quality products and services, and do it all with fewer people and at lower cost. An essential requirement in this process is effective knowledge creation and management. There is no lack of information, but there is a dearth of knowledge. As Tom Peters said in his book *Thriving on Chaos*, "We are drowning in information and starved for knowledge" [45].

The information available for decision making has exploded in recent years, and will continue to do so in the future, probably even faster. Until recently, much of that information just disappeared. It was either not collected or discarded. Today this information is being collected and stored in data warehouses, and it is available to be "mined" for improved decision-making. Some of that information can be analyzed and understood with simple statistics, but much of it requires more complex, multivariate statistical techniques to convert these data into knowledge.

A number of technological advances help us to apply multivariate techniques. Among the most important are the developments in computer hardware and software. The speed of computing equipment has doubled every 18 months while prices have tumbled. User-friendly software packages brought data analysis into the point-and-click era, and we can quickly analyze mountains of complex data with relative ease. Indeed, industry, government, and university-related research centers throughout the world are making widespread use of these techniques.

Throughout the text we use the generic term *researcher* when referring to a data analyst within either the practitioner or academic communities. We feel it inappropriate to make any distinction between these two areas, because research in both relies on theoretical and quantitative bases. Although the research objectives and the emphasis in interpretation may vary, a researcher within either area must address all of the issues, both conceptual and empirical, raised in the discussions of the statistical methods.

Three Converging Trends

The past decade has perhaps been the most complex, evolving and thus interesting with regards to analytics, whether within the academic domain or in the world of organizational decision-making. While there are many fields that can be identified as the "hot" topics or buzz terms of note (e.g., data scientists as the sexiest job of the 21st Century [16]), we feel that three topics merit discussion as they are emerging to radically transform what we think of as analytics in the near future. These topics are not focused just on the academic or organizational domains, as those worlds are converging as well. Instead they represent fundamental shifts in the inputs, processes/techniques and outputs of what we term analytics. We hope this discussion provides some broader context for you as an analyst in whatever domain you practice as the principles and objectives are similar anywhere you engage in analytics.

TOPIC 1: RISE OF BIG DATA

There is no factor impacting analytics that has been more publicized and talked about than "Big Data." And this is not just hyperbole, as there has been an explosion in data available today. The sources are varied: the world of social media and online behavior; the Internet of Things which has brought connectivity to almost every type of device; the almost incomprehensible amount of data in the sciences in such areas as genomics, neuroscience and astrophysics; the ability of storage devices to capture all this information and software (e.g., Hadoop and others) to manage that data; and finally the recognition by organizations of all types that knowing more about their customers through information can better inform decision-making. Everywhere you turn, the impact of data for improved decisions and knowledge is becoming increasingly important [14]. But what does this mean for the analyst—is it just more observations to be analyzed? We think not and hope to address in our brief discussion several important topics: What is Big Data? How does it impact organizational decisions and academic research? What impact does it have on analytics and the analyst? What are the problems Big Data presents for all analysts? And what can we expect moving forward?

What is Big Data? While the definition of **Big Data** is still evolving, it is becoming more useful to define it in terms of its basic characteristics. Perhaps the most basic are the Vs of Big Data, first thought to encompass Volume, Variety and Velocity, but being expanded with the addition of Veracity, Variability and Value [20]. Let's look at each of these briefly for their impact on analytics.

VOLUME Perhaps no characteristic describes Big Data as well as Volume, since it is the sheer magnitude of information being collected that initiated the term. While quantifying the amount of data is always speculation, it is generally agreed upon that we are encountering amounts of data never before seen in history (i.e., actually equal to and perhaps more than everything gathered before by mankind), and a recent study estimates we will be gathering ten times the amount of information annually by 2025 [13]. So whatever the incomprehensible amount, the implications are huge!

Sample sizes will increase dramatically. For many types of studies the days of small scale studies of several hundred will be replaced by secondary data sources providing thousands if not millions of cases. Online experimentation will provide almost instant access to data, and longitudinal analyses will become much more common as data is collected over time. Every one of the techniques using statistical inferences will require new approaches for interpretation and impact when everything is statistically significant due to increased sample size. These and many other changes will impact not only the scale at which analysis is done, but also the fundamental approaches analysts take to address any research question.

VARIETY The variety of Big Data is in many ways the pathway to the increases in the volume of data described earlier. Whereas analysts used to rely on primary and to some extent secondary data gathered expressly for the purposes of research, today we have access to a myriad of sources (e.g., social media, Internet of Things, digital traces of behavior, etc.) that emerge from a digitization of social life [32] that are actual behavioral data. Rather than rely upon a respondent's reply to a question (e.g., website visits, social media posts and social graph information, search queries, products purchased), actual behaviors are available to provide more direct measures of each individual.

The increases in variety come with their own challenges. First, how can we incorporate hundreds or even thousands of variables into the analyses, especially if explanation and interpretability is required? As a result, techniques in data reduction (see Chapter 3) will play an increasingly important role. And when the variables are specified, variable selection techniques become critical (e.g., multiple regression in Chapter 5), as discussed in our later overview of managing the variate. And finally, all of this increased data will generally include a nonmetric quality, so how will our analyses adapt to the shift to these forms of measurement? What types of measures can we obtain from social media posts? Search queries? Past product purchases? The analyst now faces the task of "managing" the data so as to represent constructs beyond the actual measures themselves.

VELOCITY The third characteristic of velocity has its greatest impact in the implementation of analytics since decisions must be made in an automated fashion (e.g., online web auctions for ad placement taking only milliseconds, product recommendations available immediately and a host of other immediate benefits the customer now expects instantaneously). But researchers can leverage this velocity for their own benefit—one example being the ease of online data collection. As discussed in Chapter 6, we have seen a resurgence of experimentation and the interest in causal inference with the widespread availability of online surveys and the ease of administering online experiments in the digital domain.

VERACITY The characteristic of veracity is becoming of increased interest among Big Data analysts, since they must balance the increased variety of data sources versus data quality, among them the issues of missing data and measurement error (see Chapter 2). As secondary data become primary data sources for analysts in all areas, they must guard against a "blind trust" of the data and ensure the results and conclusions drawn from these varied data sources stand the scrutiny normally reserved for both academia and the applied domain. Just because we have new sources and an increased number of variables does not absolve the analyst from the same standards as applied to past research efforts.

VARIABILITY AND VALUE The variability seen in Big Data refers to the variation in the flow of the data, which may impact issues such as timeliness. The value of Big Data is a representation of the notion that abundance, not scarcity, is the driver of value in this new era. As such, the analyst must embrace these new sources of data and expand their perspectives on the types of information applicable to their research questions.

SUMMARY Even this quick overview of the characteristics of Big Data hopefully sensitizes the analyst to the range of issues arising from this new era of analytics. And it is important that analysts in all domains, academic and organizational, come to appreciate the opportunities presented by Big Data while also being cautious in the many pitfalls and implicit assumptions associated with the new sources of data. We discuss in the next sections some of the ways in which analytics are changing in these domains and then some explicit problems in Big Data use.

Impacts on Organizational Decisions and Academic Research The driving force behind the widespread adoption of Big Data analytics is the potential benefit in organizational decision-making. Whether the benefits are associated with for-profit or non-profit, governmental or private, the perceived potential is unquestioned. Perhaps this is best reflected in a statement from a recent article on the revolution of Big Data in management thought:

> Data-driven decisions are better decisions—it's as simple as that. Using Big Data enables managers to decide on the basis of evidence rather than intuition. For that reason it has the potential to revolutionize management [34, p. 5].

Examples abound extolling the benefits derived from the application of increased analytical focus with these newfound data sources. Indeed, management processes and expectations have been irrevocably changed with this desire for more "objectivity" through analytical perspectives. While there are obvious challenges in translating these outcomes to organizational value [27], and organizations must be willing to evolve and reorient their processes [46], there is little doubt that data-driven decision-making is here to stay. Thus, analysts will be more involved with providing critical inputs on decisions at all levels of the organization.

While academicians have generally been slower to adopt and integrate Big Data into their research, certain areas in the sciences (e.g., biology [33], neuroscience [31], biomedicine [3]) have already faced these challenges and are moving forward with research agendas predicated on enhanced analytics. Many other research areas, especially those that interface with organizations (e.g., the business disciplines) or those fields in the technical areas of computer science and informatics, are moving forward with the investigation of the applications of Big Data analytics in their fields. But these areas are also recognizing the increased potential Big Data provides in the types of research questions that can now be addressed as well as the expanded analytics "toolkit" that has become available [14, 22, 17, 51, 19]. As researchers become more aware of these opportunities in both data and techniques their utilization will naturally increase.

Impacts on Analytics and the Analyst To this point we have focused on the character of Big Data and its application within every field of inquiry. But what about the analytics process and the qualities required of the analyst? Is this a time for increasing specialization or diversification? What is the role of domain knowledge and the context within which the analysis takes place? These are only a few of the changes that are facing analysts today.

One area related to the Variety of Big Data is the wide range of analytic sub-domains. Emerging today are analytics focused on text processing, image analysis, and speech recognition, along with areas such as social media and network analysis [21]. All of these specialized domains can then provide inputs into the more generalized analysis (e.g., consumer sentiments from social media analytics) to extend the scope of the factors considered. We also see that analytics are being categorized based on the objective of the analysis, thus making the decision context even more important. Recent research [46] identified five types of analysis based on their objective: descriptive, inquisitive, predictive, prescriptive and pre-emptive. Much as the academicians differentiate their research, the applied domains are finding it useful to characterize their desired outcomes as well.

Finally, a number of skills are emerging that serve all analysts in this new era [20]. Visualization is becoming essential, not only in dealing with the vast numbers of observations, but the increasingly large number of dimensions/variables being included in the analysis. This also gives rise to methods to manage dimensionality, from concepts of sparsity (akin to parsimony) and even optimization to aid in the variable selection process. As noted earlier data management will be increasingly important and perhaps most important of all is the benefit of being multidisciplinary in interests, as fields of all types are pursuing research agendas that may assist in resolving research questions in other areas.

Problems in Big Data Use No discussion of Big Data analytics would be complete without addressing the rising concern with issues such as privacy, security, political disruption, invasive commercial strategies and social stratification [6,50]. All of these issues raise cautions in the use of Big Data analytics, for even those applications with the best of intentions may have unintended consequences. And the problems are not restricted to the outcomes of these efforts, but also reside in the implicit assumptions analysts may take for granted. Harford [28] identified four "articles of faith" that may cause serious problems: overrating accuracy if we ignore false positives; replacement of causation with correlation; the idea that sampling bias is eliminated when the issues still remain; and letting the "data speak" and ignoring the presence of spurious correlations [11]. These and many other issues require the analyst not only to define the problem correctly, select and apply the appropriate technique and then correctly interpret the results, but also to be cognizant of these more implicit concerns in every research situation.

Moving Forward Our brief discussion of Big Data and its impact on analytics and analysts, both academic and applied, was not to provide a definitive guide, but rather to expose analysts to the broader issues that create both benefits and consequences with the use of Big Data [18]. We believe awareness will encourage analysts to define not only their research questions more broadly, but also their scope of responsibility in understanding these types of issues and how they might impact their analyses. The impact of Big Data may be good or bad, but the distinction is heavily influenced by the analysts and the decisions made in their analysis.

TOPIC 2: STATISTICAL VERSUS DATA MINING MODELS

The era of Big Data has not just provided new and varied sources of data, but also placed new demands on the analytical techniques required to deal with these data sources. The result has been the recognition of two "cultures" of data analysis that are distinct and purposeful in their own way. Breiman [7] defined these two cultures as data models versus algorithmic models. As we will see in the following discussions, they represent two entirely different approaches towards analysis, with opposing assumptions as to the basic model formulations and the expected results. But these differences do not make one approach better than the other and we will discuss the strengths and weaknesses of each approach as a means of distinguishing their role in this new era of analytics.

While there are dramatic differences between these two approaches, they are both operating under the same conditions. A research problem can be simply defined as a set of predictor variables working some defined process to generate an outcome. But the analyst must create some representation of the process that provides two possible objectives: accurate prediction of the outcome and explanation/knowledge of how the process operates. As we will see, it is the how the process is defined and which of these two objectives takes precedence that distinguishes the two cultures.

Statistical or Data Models The concept of data models is one that closely aligns with our classical view of **statistical models** and analysis. Here the analyst typically defines some type of stochastic data model (e.g., a multiple or logistic regression model), such as the predictor variables and their functional form [7]. Thus, the **data model** is a researcher-specified model that is then estimated using the data available to assess model fit and ultimately its acceptability. The challenges for the researcher are to (a) correctly specify a model form that represents the process being examined, and (b) perform the analysis correctly to provide the explanation and prediction desired. Thus, the researcher must make some assumptions as to the nature of the process and specify a model form that best replicates its operations. Our basic models of statistical inference, such as multiple regression, MANOVA or discriminant analysis/logistic regression, are examples of this approach to analysis.

If the analytical results are successful (e.g., good model fit, acceptable model parameters) the researcher can then make inferences as to the underlying nature of the process being examined. From these inferences explanation of the process forms through repeated analyses and some body of knowledge is developed. But the researcher must also recognize that any conclusions are about the proposed model and not really about the underlying process. If the model is an incorrect representation of the process any inferences may be flawed. Science is replete with theories that were later disproved. This does not make the process flawed, but it does caution the researcher to always be aware that there may be many "alternative" models that can also represent the process and provide different conclusions.

Data Mining or Algorithmic Models An alternative field of modeling has developed outside of the field of statistics, generally known as data mining, where the focus is not on the specified model, but the technique of explanation. **Algorithmic models**, also known as data mining and even the contemporary terms of machine learning and artificial intelligence, take a different approach to understanding the process by shifting the focus from explanation of the process to prediction. The fundamental premise is that the process being studied is inherently so complex that specification of a precise model is impossible. Rather, the emphasis is on the algorithms, how they can represent any complex process and how well they ultimately predict the outcomes. Explanation of the process is secondary to the ability of the algorithm to complete its task of prediction. So image recognition models do not provide insight into how images might be distinguished, but rather at how well they differentiate among images. Examples of algorithmic models include neural networks, decision trees and even cluster analysis. In each instance the result is a prediction, not a generalization to the population (i.e., statistical inference).

Why Distinguish Between Models? As Brieman [7] describes, these two models truly represent different "cultures" of model building, coming from different research disciplines, operating on fundamentally different assumptions about how the models should operate and what are the most important objectives. While each will continue to evolve separately, recognition of their different strengths and weaknesses can enable researchers from both "cultures" to appreciate the differences and better understand the situations and research questions most suited to each approach. This becomes critically important as the era of Big Data analytics demands more insights from analysts into an ever increasing array of issues and contexts.

Figure 1.1 provides some differences between these two approaches on a number of fundamental characteristics of the research process. As we see, the statistical or data model approach approximates what many consider the scientific method, postulating a model based upon theory and then executing a research design to rigorously test that model and ultimately the underlying theory. The data mining or algorithmic models approach the problem differently, and bring little in the way of conceived speculation on how the process should be described. They instead focus on the best algorithms what can replicate the process and achieve predictive accuracy. Moreover, there is little theory testing and thus little explanation of how the process of interest actually operates, just that is can be mimicked by the algorithm as represented by its predictive accuracy.

Figure 1.1
Comparing Between Statistical/Data Models and Data Mining/Algorithmic Models

Characteristic	Statistical/Data Models	Data Mining/Algorithmic Models
Research Objective	Primarily Explanation	Prediction
Research Paradigm	Theory-based (deductive)	Heuristic-based (inductive)
Nature of Problem	Structured	Unstructured
Nature of Model Development	Confirmatory	Exploratory
Type of Data Analyzed	Well defined, collected for purpose of the research	Undefined, generally analysis used data available
Scope of the Analysis	Small to large datasets (number of variables and/or observations)	Very large datasets (number of variables and/or observations)

Our purpose in this discussion is not to "pick a winner" or build a case for one method over another, but instead to expose researchers in both "cultures" to the assumptions underlying their two approaches and how they might complement one another in many situations. There are obviously situations in which the process of interest (e.g., autonomous vehicles) is so complex that it would seem impossible to completely specify the underlying models (e.g., rules of the road) and the algorithmic models like machine learning seem most applicable. But how do analysts and managers deal with situations that require managerial action? How do we best engage in product design, select promotional appeals or increase customer satisfaction? In these situations the researcher is looking more for explanation than just prediction, and an understanding of the process is required before it can be managed. In an academic setting, the strong theory-based approach favors the data model method, but there may be situations in which the algorithmic model provides insight not otherwise possible. One intriguing possibility is the "automation of discovery" where patterns of correlations can be analyzed and causal structure identified [38, 24]. We encourage analysts from both cultures to try and understand the basic differences and ultimately the benefits that can be achieved through these two different, but increasingly complementary, approaches to analytics.

TOPIC 3: CAUSAL INFERENCE

Our last topic spans all of the issues we have discussed to this point—the impact of Big Data and the distinction between data models and algorithmic models. **Causal inference** is the movement beyond statistical inference to the stronger statement of "cause and effect" in non-experimental situations. While causal statements have been primarily conceived as the domain of randomized controlled experiments, recent developments have provided researchers with (a) the theoretical frameworks for understanding the requirements for causal inferences in non-experimental settings, and (b) some techniques applicable to data not gathered in an experimental setting that still allow some causal inferences to be drawn [36].

The move from statistical inference to causal analysis is not a single technique, but instead a paradigm shift incorporating an emphasis on the assumptions that are the foundation of all causal inferences and a framework for formulating and then specifying these assumptions [43]. The result is a general theory of causation based on the Structural Causal Model (SCM) first proposed by Pearl [42]. A key component of the SCM is the **directed acyclic graph (DAG)**, which is a graphical representation of the causal relationships impacting the relationship of interest. Similar in some regards to the path models used in structural equation modeling (see Chapter 9), it differs in that its sole focus is on causal pathways and the ability of the researcher to "control for" all of the confounds that may impact the relationship of interest [30]. In Chapter 6 we introduce the notion of confounds and other "threats" to causal inference as well as one of the most popular methods for causal inference—propensity score models.

In many ways the increase in the use of causal inference is the result of the emergence of Big Data (i.e., the wealth of non-experimental data) and the desire for a more rigorous framework for analysis than statistical inference. Causal inference has become widespread in disciplines ranging from accounting [25] to the health sciences [47] and can be employed not only with basic statistical models, but more complex effects such as mediation [52] and structural equation modeling [10, 5]. Moreover, the techniques from many disciplines are being combined to generate an entirely new analytical framework for non-experimental data [37, 26, 44]. In the not too distant future we believe that all analysts will employ these causal inference methods to increase the rigor of their analysis and help overcome the doubts raised by many concerning the pitfalls of Big Data analytics [18].

SUMMARY

While some might question the relevance of these topics in a text oriented towards multivariate techniques, we hope that exposure to the ideas and issues raised in these areas will better inform the researcher in their activities. As evidenced in these discussions, the analyst of today is much more than just a technician trained to select and then apply the appropriate statistical technique. Yes, there are unique technical challenges facing today's analyst with the many issues in Big Data and the competing approaches of data models versus algorithmic models, always striving to make causal inferences if possible. But these are more than technical problems, and analysts in both the academic and applied fields must develop their own approach for assessing these issues in each research situation they face. Only then can they design a research plan that best meets the overall needs of the research question.

Multivariate Analysis in Statistical Terms

Multivariate analysis techniques are popular because they enable organizations to create knowledge and thereby improve their decision-making. **Multivariate analysis** refers to all statistical techniques that simultaneously analyze multiple measurements on individuals or objects under investigation. Thus, any simultaneous analysis of more than two variables can be loosely considered multivariate analysis.